Numerik I

Vorlesung
von

Eugen Schäfer

L M U München

S S 2006

Inhaltsverzeichnis

Letzter Stand der **Korrekturen** : 18. August 2006

Abbildungsverzeichnis

Literatur zur Numerik I

Im folgenden eine kleine Auswahl unter vielen vorlesungsbegleitenden deutschsprachigen Büchern – die an Stofffülle meist deutlich über den Inhalt der Vorlesung hinausgehen. Natürlich gibt es auch viele englisch– und französischsprachige Bücher zu dieser Thematik:

P. Deuflhard, A. Hohmann [2002]: Numerische Mathematik I. Eine algorithmisch orientierte Einführung. (3. überarb. und erw. Auflage), de Gruyter-Verlag Berlin u.a., ISBN 3-11-017182-1, EUR 24,95.

G. Hämmerlin, K.-H. Hoffmann [1994]: Numerische Mathematik. (4. Auflage) Springer-Verlag Berlin u.a., ISBN 3-540-58033-6, EUR 23,32 *(zur Zeit vergriffen)*.

R. Schaback, H. Wendland [2005]: Numerische Mathematik. (5. Auflage) Springer-Verlag Berlin u.a., ISBN 3-540-21394-5, EUR 24,95.

J. Werner [1992]: Numerische Mathematik 1 / 2. Vieweg-Verlag Braunschweig, ISBN 3-528-07232-6, EUR 24,90 / ISBN 3-528-07233-4, EUR 24,90.
zwar zweibändig, dafür jedoch mit ausführlicher Behandlung von Optimierungsverfahren und nichtlinearen Gleichungssystemen

Numerik II mit abdeckend:

P. Deuflhard, F. Bornemann [2002]: Numerische Mathematik II. Gewöhnliche Differentialgleichungen. (2. vollst. überarb. und erw. Auflage), de Gruyter-Verlag Berlin u.a., ISBN 3-11-017181-3, EUR 29,95.

M. Hanke–Bourgeois [2002]: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens. Teubner-Verlag Stuttgart u.a., ISBN 3-519-00356-2, 836 S., EUR 64,90.
auch Numerik partieller Differentialgleichungen

A. Quarteroni, R. Sacco, F. Saleri [2002]: Numerische Mathematik 1. Springer-Verlag Berlin u.a., ISBN 3-540-67878-6, EUR 23,32.
Numerische Mathematik 2. Springer-Verlag Berlin u.a., ISBN 3-540-43616-2, EUR 27,99.

J. Stoer [1999]: Numerische Mathematik 1. (9. Auflage 2005) Springer-Verlag Berlin u.a., ISBN 3-540-21395-3, EUR 24,95.

J. Stoer, R. Bulirsch [2000]: Numerische Mathematik II. (4. Auflage) Springer-Verlag Berlin u.a., ISBN 3-540-67644-9, EUR 23,32.

Stand der Angaben: Mai 2006

Kapitel 0

Einführung

Die numerische Mathematik hat die zahlenmäßige Lösung von Problemen zum Ziel. Die Berechnungsmethoden dafür nennt man Algorithmen. Die meisten Probleme können nur näherungsweise gelöst werden. Dann sind wesentliche Aufgaben der numerischen Mathematik

Bewertung der gelieferten Resultate:
– Genauigkeit einer bestimmten Näherung
– 'schließliche' Konvergenz einer Folge von Näherungen

Bewertung einer Methode im Vergleich zu anderen:
– Durchführbarkeit
– Aufwand, sog. arithmetische Komplexität
– *Genauigkeit* bzw. asymptotische *Konvergenzgeschwindigkeit*
– *Stabiblität*, i.e. nicht zu große Änderung der Resultate bei Störungen in den Eingabedaten z.B. durch Messfehler in den Daten, oder durch Verwendung von nur endlich vielen 'Gleitpunktzahlen' im Computer.

Gleitpunktzahlen und Gleitpunktarithmetik

Bekannt ist aus der Analysis die Darstellung reeller Zahlen als Dezimal– oder Dualentwicklung, allgemein zur *Basis* $B \in \mathbb{N}$, $B \geq 2$,

$$0 \neq x = \underbrace{\sigma}_{=sign(x)\in\{-1,+1\}} \quad B^N \underbrace{\sum_{\nu=1}^{\infty} x_{-\nu} B^{-\nu}}_{\in]0,1[} \ ,$$

mit dem *Exponenten* $N \in \mathbb{Z}$ und den *Koeffizienten* $x_{-\nu} \in \{0, 1, ..., B-1\}$. Diese Darstellung ist eindeutig unter den *Normalisierungsbedingungen*

$$x_1 \neq 0 \quad \text{und} \quad \textit{nicht fast alle } x_\nu = B - 1 \ .$$

(0.1) Gleitpunkt-Maschinenzahlen zur Basis B

$$0 \neq x = \underbrace{\sigma}_{=sign(x)\in\{-1,+1\}} \quad B^N \quad \underbrace{\sum_{\nu=1}^{t} x_{-\nu} B^{-\nu}}_{\textit{sog. Mantisse von } x} \quad ,$$

mit dem *Exponenten* $N \in \mathbb{Z}$ aus dem *Exponentenbereich* $N_- \leq N \leq N_+$, und der sog. *Mantissenlänge* $t \in \mathbb{N}$. Wieder normalisieren wir duch die Bedingung

$$x_1 \neq 0 \ .$$

□

Für bestimmte Anwendungen verwendet man Festpunktzahlen, d.h. N ist fixiert. Der Vorteil ist eine exakte Arithmetik, was bei Indexrechnung, zahlentheoretischen Algorithmen, und z.B. im Bankenbereich wünschenswert ist.
Für wissenschaftlich–technische Anwendungen ist diese Art nicht so geeignet, da u.U. Zahlen sehr unterschiedlicher, nicht im vorhinein bekannter Größenordnung auftreten.

(0.2) Runden
Für $x = \sigma B^N \sum_{\nu=1}^{\infty} x_{-\nu} B^{-\nu} \neq 0$ ist

$$rd\big(\sigma B^N \sum_{\nu=1}^{\infty} x_{-\nu} B^{-\nu}\big) \ := \ \begin{cases} \sigma B^N \sum\limits_{\nu=1}^{t} x_{-\nu} B^{-\nu} & , \quad \text{falls } x_{-(t+1)} < \frac{1}{2}B \ , \\ \sigma B^N \big(\sum\limits_{\nu=1}^{t} x_{-\nu} B^{-\nu} + B^{-t} \big) & , \quad \text{falls } \frac{1}{2}B \leq x_{-(t+1)} \leq B-1 \ , \end{cases}$$

und $rd(x) = 0$ für $x = 0$.

□

Die *Genauigkeit dieser Zahldarstellung* beschreibt

(0.3) Bemerkung
Es gilt und $rd(x) \neq 0$ gilt

$$\frac{|rd(x) - x|}{|x|} \ \leq \ \frac{B}{2} B^{-t} \ , \ x \neq 0 \quad (und) \quad \frac{|rd(x) - x|}{|rd(x)|} \ \leq \ \frac{B}{2} B^{-t} \ , \ rd(x) \neq 0 \ .$$

Beweis:

$rd(x)$ und x unterscheiden sich in der (t+1)–ten Nachkommastelle höchstens um $\frac{B}{2} - 1$. Damit folgt

$$|rd(x) - x| \leq B^N \big[\big(\frac{B}{2} - 1\big) B^{-(t+1)} + \sum_{\nu = t+2}^{\infty} x_{-\nu} B^{-\nu} \big]$$

$$\leq B^N \big[\big(\frac{B}{2} - 1\big) B^{-(t+1)} + B^{-(t+2)} \sum_{\nu = 0}^{\infty} (B-1) B^{-\nu} \big].$$

Wegen $B^{-(t+2)} \sum_{\nu=0}^{\infty} (B-1) B^{-\nu} = B^{-(t+2)} (B-1) \frac{1}{1 - 1/B} = B^{-(t+1)}$ und $|x|, |rd(x)| \geq B^N x_{-1} B^{-1} \geq B^N \cdot 1 \cdot B^{-1} = B^N - 1$ folgt die Behauptung.

□

Bereits in der Zahldarstellung etwa von $\sqrt{2} \notin \mathbb{Q}$ macht man also Fehler. Welche Fehler entstehen bei den (angenäherten) arithmetischen Operationen, den sog. Gleitpunktoperationen? Bei guten Rechnern sind diese zumindest kommutativ, jedoch nicht assoziativ und nicht distributiv. Ein Modell – das für viele Rechner gilt, und das ich immer voraussetzen werde – ist

(0.4) Modell der Gleitpunktarithmetik

(a) Aus zwei Operanden einer arithmetischen Grundoperation $\cdot \in \{+, -, \times, /\}$ wird ein Zwischenresultat mit doppelter Mantissenlänge $2t$ gebildet.

(b) Anschließend wird dieses Ergebnis auf t Stellen gerundet.

(c) Gegebenenfalls wird dieses gerundete Ergebnis normalisiert.

Die Notation für die Vorgehensweise (a) – (c) ist $\quad \odot :\equiv gl(x \cdot y)$ für $\cdot \in \{+, -, \times, /\}$.

(d) Treten keine Exponentenbereichsüberschreitungen oder –unterschreitungen auf, so gilt für jede der Operationen $\odot \in \{\oplus, \ominus, \otimes, \oslash\}$

$$x \odot y = (x \cdot y)(1 + \varepsilon) = \frac{(x \cdot y)}{1 + \eta} \quad \text{mit} \quad |\varepsilon|, |\eta| \leq \frac{B}{2} B^{-t} \ .$$

Beweis:

Man zeigt dazu: $x \odot y = rd\ (x \cdot y)$. Mit (0.3) folgt die Behauptung.

□

Wie pflanzen sich Fehler (Rundungs–, Mess–, Rechenfehler) fort bei arithmetischen Gleitpunktoperationen?

Besonders kritisch ist die Subtraktion – genauer die Addition zweier Zahlen etwa gleich großen Betrags und unterschiedlichen Vorzeichens.

(0.5) Beispiel

$$B = 10, t = 3, x = 0.9995, y = -0.9984 \implies$$

$$x \oplus y \equiv gl(rd(x) + rd(y)) = 0.2 \cdot 10^{-2} \ , \ x + y = 0.11 \cdot 10^{-2} \implies$$

$$\frac{(x \oplus y) - (x + y)}{x + y} = \frac{0.09 \cdot 10^{-2}}{0.11 \cdot 10^{-2}} \ = \ 0.\overline{81} \ \doteq \ 0.82 \ ,$$

wohingegen $\frac{rd(x) - x}{x} \doteq 0.0005$, und $\frac{rd(y) - y}{y} \doteq 0.0004$; d.h. es tritt hier eine 'Verstärkung' der relativen Fehler um einen Faktor $1.64 \cdot 10^3$ auf.

□

Wirklich beheben kann man diesen Effekt nur durch Festpunktarithmetik. Diesen Effekt der sog. *Auslöschung führender Stellen* bei der Subtraktion kann man manchmal mildern/verhindern durch mathematisch äquivalente Umformung in stabil auswertbare Formeln bzw. Näherungen:

(0.6) Beispiele (für stabil auswertbare Formeln bzw. Näherungen)
(a) Nullstellen quadratischer Gleichungen $ax^2 + bx + c = 0$.

$$x_1 = -\frac{1}{2a}\big(+ b + sign(b)\sqrt{b^2 - 4ac}\big) \quad \text{'genau' für } |4ac| << b^2 \ ;$$

$$x_2 = -\frac{1}{2a}\big(+ b - sign(b)\sqrt{b^2 - 4ac}\big) \quad \text{'ungenau' für } |4ac| << b^2 \ ;$$

Berechnung von x_2 *ohne Auslöschung* durch Auswertung der Vietaschen Formel $ax_1x_2 = c$.

(b) $\log(x - \sqrt{x^2 - 1})$ 'ungenau' für $x >> 1$ wie in (a), gemildert durch log. Durch Erweitern mit $x + \sqrt{x^2 - 1}$ ergibt sich der Ausdruck $-\log(x + \sqrt{x^2 - 1})$, der 'genau' auswertbar ist für $x >> 1$.

(c) $\sinh(x) = \frac{e^x - e^{-x}}{2}$ ist 'ungenau' auswertbar ist für $|x| << 1$. Taylorentwicklung liefert für $|x| << 1$ die sehr 'genau auswertbare' Näherung $x + \frac{x^3}{3!} + \frac{x^5}{5!}$.

□

Normen

Zur quantitativen Fassung der Größe von Störungen, Fehlern, und verfälschten Resultaten bei Vektoren und Abbildungen vom $\mathbb{R}^n$ in den $\mathbb{R}^m$ verwendet man Normen. Ich verwende hier nur normierte Vektorräume über $\mathbb{R}$ oder $\mathbb{C}$ Falls beide Körper erlaubt sind, schreibe ich $\mathbb{K}$.

Definition
Sei V Vektorraum über $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}, \mathbb{K} = \mathbb{C}$). Dann heißt eine Abbildung $\| \ \| : V \longrightarrow \mathbb{R}$, $V \ni x \longmapsto \|x\| \in \mathbb{R}$, *Norm* auf V , falls gilt

(a) $\|x\| = 0 \Longleftrightarrow x = 0 \, ; \, x \in V$.

(b) $\|cx\| = |c|\|x\| \, ; \, x \in V \, , c \in \mathbb{K}$.

(c) $\|x + y\| \leq \|x\| + \|y\| \, ; \quad x, \, y \in V$.

$(V, \| \ \|)$ heißt *normierter (Vektor-)Raum*, oder kurz *'V normierter Raum'*.

□

(0.7) l_p-Normen $(1 \leq p \leq \infty)$
Für $v = [v_i]_{1 \leq i \leq n} \in \mathbb{K}^n$ definiert man

$$\|v\|_\infty := \max_{1 \leq i \leq n} |v_i| \, , \qquad l_\infty\text{– oder } Maximum\text{–}Norm;$$

$$\|v\|_p := \Big(\sum_{i=1}^{n} |v_i|^p \Big)^{1/p} \, , \qquad l_p - \text{Norm}, 1 \leq p < \infty \ .$$

Für $p = 2$ erhält man die

$$\|v\|_2 = \Big(\sum_{i=1}^{n} |v_i| \Big)^{1/2} \, , \qquad Euklidische \ Norm \ ,$$

für $p = 1$ die

$$\|v\|_1 = \sum_{i=1}^{n} |v_i| \, , \qquad Betragssummen - Norm \ .$$

□

Das Aussehen der verschiedenen Einheitskugeln erkennt man in Abbildung 1. Offensichtlich sind diese Normen nicht differenzierbar für $p = 1, \infty$ – in allen anderen Fällen sind die l_p–Normen differenzierbar. Die Äquivalenz dieser Normen ist geometrisch offensichtlich (vgl. Abbildung 1)

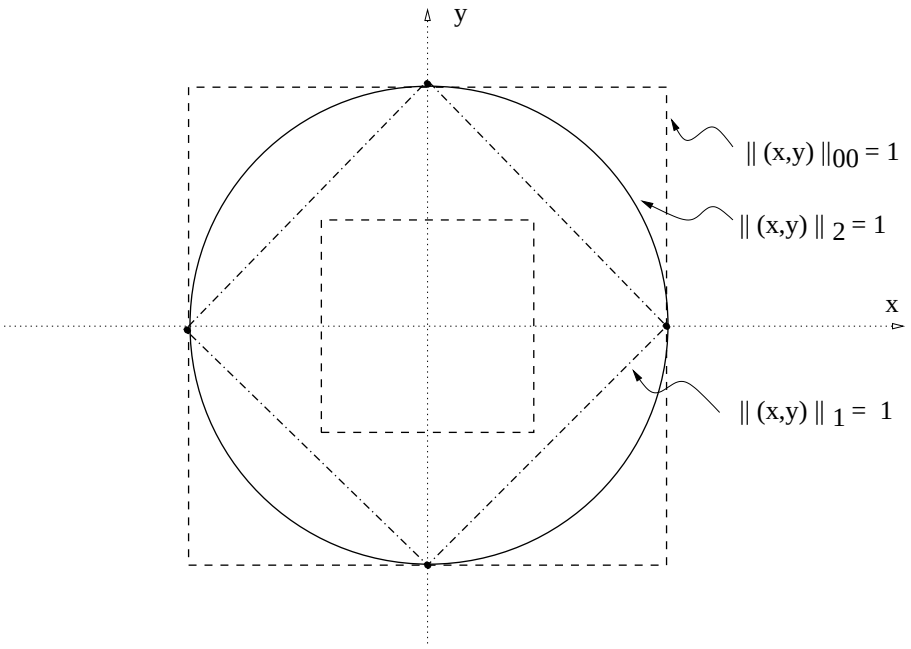

Abbildung 1: Einheitskreislinie für $p = 1, 2, \infty$

– analytisch bedeutet dies für zwei Normen $\|.\|'$ und $\|.\|''$ auf V, daß es $\underline{c} > 0$ und $\overline{c} > 0$ gibt mit

$$\underline{c}\|v\|'' \leq \|v\|' \leq \overline{c}\|v\|'' \ , \ v \in V \ .$$

Für l_p–Normen kann man die Konstanten $\underline{c}$ und $\overline{c}$ explizit angeben:

(0.8) Bemerkung (Beweis als Übung)
Einen Vergleich beliebiger l_p–Normen für $v \in \mathbb{K}^n$ ermöglichen die Aussagen

(a) $[1, +\infty[\ni p \longrightarrow \|v\|_p$ ist monoton fallend mit $\lim_{p\to\infty} \|v\|_p = \|v\|_\infty$ (was auch die Bezeichnung $\|.\|_\infty$ für die Maximumnorm sinnvoll macht).

(b) $[1, +\infty[\ni p \longrightarrow n^{-1/p}\|v\|_p$ ist monoton wachsend mit $\lim_{p\to\infty} n^{-1/p}\|v\|_p = \|v\|_\infty$.

□

Welche der l_p–Normen passend ist, hängt von der jeweiligen Situation ab (siehe auch Übung):

l_∞–Norm: gleichmäßig in allen Komponenten gültige Abschätzung bzw. Fehlergröße,

l_2–Norm: geeignet bei normal verteilten Meßwertfehlern,

l_1–Norm: geeignet bei vereinzelten starken Ausreißern.

Matrixnormen

$A \in \mathbb{K}^{m\times n}$, $A : \big(\mathbb{K}^n, \|.\|'\big) \longrightarrow \big(\mathbb{K}^m, \|.\|''\big)$, $Ax \in \mathbb{K}^m$ für $x \in \mathbb{K}^n$. Dann gilt für $x \neq y$

$$\|Ax - Ay\|'' = \|A(x-y)\|'' = \|A\Big(\frac{x-y}{\|x-y\|'}\Big)\|''\|x-y\|' \leq \big(\max_{\|z\|'=1} \|Az\|''\big)\|x-y\|' \ .$$

(0.9) Induzierte Matrixnormen
Sei $A \in \mathbb{K}^{m\times n}$. Dann ist

(a) zu $A : \left(\mathbb{K}^n, \|.\|'\right) \longrightarrow \left(\mathbb{K}^m, \|.\|''\right)$ die *induzierte Matrixnorm* gegeben durch

$$\|A\|_{ind} := \max_{x\in\mathbb{K}^n \,:\, \|x\|'=1} \|Ax\|'' = \max_{x\in\mathbb{K}^n \,:\, x\neq 0} \frac{\|Ax\|''}{\|x\|'} \ .$$

(b) $\|A\|_p := \max_{\|x\|_p=1} \|Ax\|_p$ die l_p–Matrixnorm.

(c) $\|A\|_F := \Big(\sum\limits_{1\le i\le m \ , \ 1\le j\le n} |A_{ij}|^2 \Big)^{1/2}$ die sog. *Frobenius*–Norm.

□

Aus der Definition folgen offensichtlich die sog. *Submultiplikivität* induzierter Matrixnormen

$$\|AB\|_{ind} \le \|A\|_{ind}\|B\|_{ind} \ ,$$

und die *Abschätzung* (vgl. auch (0.12) (c))

(0.10) $$\max_{\lambda \ Eigenwert \ von \ A} |\lambda| \ \le \ \|A\|_{ind} \ .$$

Theoretisch wichtig ist die $\|A\|_2$–Norm; praktisch verwendbar nur die $\|A\|_1$– und die $\|A\|_\infty$–Norm. Denn es gilt mit folgenden

(0.11) Bezeichnungen
Zu $A \in \mathbb{K}^{m\times n}$ ist, wenn nichts anderes vereinbart wird, A_{ij} das (i,j)–te Element;
$A_{i\,.} := [A_{i1}...A_{in}] = {e_i}^t \cdot A$ der i–te Zeilenvektor;

$$A_{.\,j} := \begin{bmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{bmatrix} = A \cdot e_j \text{ der } j\text{–te Spaltenvektor.}$$

□

(0.12) p–Normen von $A \in \mathbb{R}^{m\times n}$

(a) $\|A\|_1 = \max\limits_{1\le j\le n} \sum\limits_{i=1}^{m} |A_{ij}| \equiv \max\limits_{1\le j\le n} \|A_{.j}\|_1$ sog. *Spaltensummennorm*;

(b) $\|A\|_\infty = \max\limits_{1\le i\le m} \sum\limits_{j=1}^{n} |A_{ij}| \equiv \max\limits_{1\le i\le m} \|A_{i.}\|_1$ sog. *Zeilensummennorm*;

(c) $\|A\|_2 = \max\limits_{\mu \ Eigenwert \ von \ A^tA} \sqrt{\mu}$ der größte sog. *Singulärwert* von A.

Beweis:

(a) $\|Ax\|_1 = \sum\limits_{i=1}^{m} |(Ax)_i| = \sum\limits_{i=1}^{m} |\sum\limits_{j=1}^{n} A_{ij}x_j| \le \sum\limits_{j=1}^{n}\sum\limits_{i=1}^{m} |A_{ij}||x_j| \le \Big(\sum\limits_{j=1}^{n} |x_j|\Big) \max\limits_{1\le j\le n} \sum\limits_{i=1}^{m} |A_{ij}| = \|x\|_1 \max\limits_{1\le j\le n} \|A_{.j}\|_1$;

damit gilt $\|A\|_1 \le \max\limits_{1\le j\le n} \|A_{.j}\|_1$. Für geeignet konstruiertes x gilt $' ='$.

(b) zeigt man ähnlich.

(c) Es gilt $\|A\|_2{}^2 = \max\limits_{\|x\|_2=1} \|Ax\|_2{}^2 = \max\limits_{\|x\|_2=1} (Ax)^t Ax = \max\limits_{\|x\|_2=1} x^t A^t Ax$. Für die symmetrische Matrix A^tA ist dieser Ausdruck aber gerade der größte Eigenwert.

□

Zum Abschluß ohne Beweis (vgl. Übungen) noch zwei nützliche 'technische' Aussagen über l_p–Normen, die man mit Hilfe der Hölderschen Ungleichung zeigen kann.

(0.13) Satz

(a) $\|x\|_p = \max\limits_{\|y\|_{p'}=1} |y^h x|$, wobei $1/p + 1/p' = 1$ (p und p' sind sog. konjugierte Exponenten).

(b) $\|A\|_p = \|A^h\|_{p'}$ für $A \in \mathbb{K}^{m\times n}$, wobei $1/p + 1/p' = 1$.

□

Damit kann man z.B. Aussage (b) von Satz (0.12) direkt aus $(0.12)(a)$ folgern.

Kapitel 1

Direkte Lösung linearer Gleichungssysteme

Aufgabe ist:

$$Ax = b\,;\ \text{gegeben } A \in \mathbb{R}^{n\times n}\,, b \in \mathbb{R}^n\ ;\ \text{gesucht } x \in \mathbb{R}^n\ .$$

Ich erinnere an *MIB - Gaußsches Eliminationsverfahren:*
— löse eine der Gleichungen, etwa die i-te, nach x_1 auf; setze den erhaltenen Ausdruck in die restlichen $(n-1)$ Gleichungen ein, die damit $(n-1)$ Unbekannte haben.

— nach $(n-2)$ weiteren solchen Schritte erhält man eine Gleichung in der Unbekannten x_n.

— daraus berechnet man x_n, und anschließend durch Rücksubstitution $x_{n-1}, x_{n-2}, \ldots, x_1$ in dieser Reihenfolge.

Die *Analyse der berechneten Ergebnisse* dieses bekannten Verfahrens bei Durchführung auf einem Rechner wird ermöglicht bzw. erleichtert durch

— *Formulierung* obiger Schritte *als Matrix-Multiplikationen*

— Untersuchung der Auswirkungen von Störungen mittels *Verwendung von Vektor/Matrix-Normen.*

Störungen sind gegeben z.B. durch *Gleitpunktarithmetik* anstelle reeller Arithmetik.

Diese Vorgehensweise werde ich zuerst durchführen. Anschließend werde ich eine besonders günstige Situation für die LR–Zerlegung behandeln , d.h. $A = LR$ mit einer unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R. Abschließend bespreche ich ein anderes direktes Verfahren, die *QR-Zerlegung.* Diese ist stabiler als das Gaußsche Eliminationsverfahren, und eignet sich daher besonders für kritischere Situationen und Aufgabenstellungen.

Daß man zur stabileren Durchführung der LR-Zerlegung unbedingt mindestens eine sog. *Spaltenpivot-Strategie* (oder sogar die aufwändigere *Totalpivot-Strategie*) durchführen sollte, möchte ich veranschaulichen am folgenden

Beispiel

Zu $\varepsilon \in \mathbb{R}$ sei $A_\varepsilon = \begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}$. A_ε ist nichtsingulär für $\varepsilon \neq 1$. Für $b_\varepsilon = \begin{bmatrix} 1+\varepsilon \\ 2 \end{bmatrix}$ hat $A_\varepsilon x = b_\varepsilon$ die Lösung $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Für $\varepsilon \neq 0$ ist die LR-Zerlegung durchführbar, und man erhält in reeller Arithmetik

$$L\ R\ =\ \begin{bmatrix} 1 & 0 \\ \varepsilon^{-1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon & 1 \\ 0 & 1-\varepsilon^{-1} \end{bmatrix}.$$

In Gleitpunktarithmetik mit $MATLAB$ durchgeführt ($macheps \doteq 10^{-16}$), erhält man für $eps = 10^{-12}$ die Näherung $\tilde{x}_\varepsilon = \begin{bmatrix} 0.99987 \\ 1.0 \end{bmatrix}$ für die Lösung $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, also nur etwa vier gültige Ziffern in der 1. Komponente.

1.1 Gauß-Elimination

Durch die Gaußelimination werden sukzessive Nullrestspalten unterhalb der Diagonale erzeugt. Ich veranschauliche dies schematisch

$$\begin{bmatrix} \times & \times & \times & \times \\ \otimes & \times & \times & \times \\ \otimes & \times & \times & \times \\ \otimes & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \otimes & \times & \times \\ 0 & \otimes & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \otimes & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \end{bmatrix}$$

Bewirkt wir dies durch Linksmultiplikation mit Eliminationsmatrizen (vgl. (1.1.5')).

(1.1.1) Gaußelimination *(mit Spaltenpivotsuche)*
Gegeben: $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$;

Gesucht: $x \in \mathbb{R}^n$: $A\,x = b$.

$$\widetilde{A}^{(1)} := \left[\begin{array}{c|c} A & b \end{array}\right]$$

für $k := 1, 2, ..., n-1$ *führe durch*

(0) *gegeben* $\widetilde{A}^{(k)}$ *mit* $\widetilde{A}^{(k)}_{ij} = 0,\ j < i \le n,\ 1 \le j < k$;

(1) *bestimme (zur algorithmischen Festlegung z.B. 'kleinstes')* $i_k \ge k$ *mit*

$$|\widetilde{A}^{(k)}_{i_k k}| = \max_{k\le j\le n} |\widetilde{A}^{(k)}_{jk}| .$$

Falls *(2)* $|\widetilde{A}^{(k)}_{i_k k}| > 0$, *dann*

(3)
$$\widetilde{\widetilde{A}}^{(k)}_{j\,\cdot} := \begin{cases} \widetilde{A}^{(k)}_{i_k\,\cdot} & , \quad j = k \qquad \textit{Zeilenver-} \\ \widetilde{A}^{(k)}_{k\,\cdot} & , \quad j = i_k \qquad \textit{tauschung} \\ \widetilde{A}^{(k)}_{j\,\cdot} & , \quad \textit{sonst} \end{cases}$$

(4)
$$\widetilde{A}^{(k+1)}_{j\,\cdot} := \begin{cases} \widetilde{\widetilde{A}}^{(k)}_{j\,\cdot} & , \quad 1 \le j \le k \\ \widetilde{\widetilde{A}}^{(k)}_{j\,\cdot} - \frac{\widetilde{\widetilde{A}}^{(k)}_{jk}}{\widetilde{\widetilde{A}}^{(k)}_{kk}} \widetilde{\widetilde{A}}^{(k)}_{k\,\cdot} & , \quad k < j \le n \end{cases} \qquad \begin{array}{l} \textit{Elimination} \\ \textit{der k-ten} \\ \textit{Restspalte} \end{array}$$

sonst

A *singulär* $\rightsquigarrow$ *Abbruch*

□

(1.1.2) Satz
Ist A nichtsingulär, so ist der Algorithmus (1.1.1) durchführbar. Mit $\widetilde{A}^{(n)} =: \left[\begin{array}{c|c} A^{(n)} & b^{(n)} \end{array}\right]$ ist $Ax = b$ äquivalent zu $A^{(n)}x = b^{(n)}$. $A^{(n)}$ ist nichtsinguläre obere Dreiecksmatrix.

Beweis:

Durchführbarkeit : $\iff \widetilde{A}^{(n)}_{i_k,k} \neq 0$ in (1.1.1)*(2)*
$\iff det(A^{(k)}) \neq 0$ mit $\left[\begin{array}{c|c} A^{(k)} & b^{(k)} \end{array}\right] = \widetilde{A}^{(k)}$ (elementare Zeilenoperationen ändern Determinante nicht)
$\iff det(A) \neq 0$; dies gilt aber nach Voraussetzung.

$\diamond$

Äquivalenz ist klar.

$\diamond$

$A^{(n)}$ *nichtsinguläre obere Dreiecksmatrix* :
Nach (1.1.1)*(4)* hat $\widetilde{A}^{(k+1)}$ wieder die Eigenschaft (1.1.1)*(0)* für $k+1$ anstelle von k
$\implies$ $\widetilde{A}^{(n)}$ ist obere Dreiecksmatrix $\implies A^{(n)}$ obere Dreiecksmatrix, und
$$det(A^{(n)}) = A^{(n)}_{11} \cdot .... \cdot A^{(n)}_{nn} \neq 0$$

$\square$

(1.1.3) Bemerkung
(a) Schritt *(1)* von (1.1.1) ist sog. *Spaltenpivot-Strategie*, da man in jeder Restspalte ein betragsgrößtes Element zur Elimination nimmt.

(b) Bei DV-Realisierung ersetzt man die Abfrage in Schritt *(2)* von (1.1.1) durch

$$max_{1 \leq i \leq n} |\widetilde{A}^{(k)}_{i,k}| \; > \; tol \quad \text{mit } tol > 0.$$

Üblich ist $tol := macheps$, wobei $macheps$ die Maschinengenauigkeit bei Addition gegenüber der 1 ist. Gilt

$$\max_{k \leq j \leq n} |\widetilde{A}^{(k)}_{jk}| \; \leq tol \; ,$$

so könnte man A als *numerisch singulär* bezeichnen.

$\square$

(1.1.4) Bemerkung (arithmetischer Aufwand)
Für $A \in \mathbb{R}^{n \times n}$ benötigt man für Algorithmus (1.1.1)

$\frac{n^3}{3} + \ldots$ Additionen/Subtraktionen, $\frac{n^3}{3} + \ldots$ Multiplikationen/Divisionen,

Dagegen benötigt man zur *Lösung eines Dreiecksgleichungssystems* $Lx = b$ oder $Rx = y$ mit einer unteren Dreiecksmatrix L oder einer oberen Dreiecksmatrix R

$\frac{n^2}{2} + \ldots$ Additionen/Subtraktionen, $\frac{n^2}{2} + \ldots$ Multiplikationen/Divisionen.

Dabei stehen die Punkte $'\ldots'$ für ein Polynom in n von niedrigerem Grad als der angegebene Summand, der sog. *Hauptterm*.

Beweis:

In Algorithmus (1.1.1) sind für jedes $k \in \{1, ..., n-1\}$

in Schritt *(1)*: $n-k$ Vergleiche

in Schritt *(4)*: $n-k$ Divisionen für $\frac{\widetilde{A}_{ik}^{(k)}}{\widetilde{A}_{kk}^{(k)}}$, $i = k+1, ..., n$; und $n-k$ Multiplikationen und $n-k$ Subtraktionen für $\widetilde{A}_{i\cdot}^{k+1}$, $i = k+1, ..., n$; insgesamt also:

Vergleiche: $\sum_{k=1}^{n-1}(n-k) = (n-1)n/2$

Mult./Div.: $\sum_{k=1}^{n-1}\left((n-k)_{Div} + \sum_{i=k+1}^{n}(n-k)_{Mult}\right)$

$= \sum_{k=1}^{n-1}(n-k)(n-k)_{Mult/Div} +$ Polynom vom Grad kleiner 3 in n

$= \sum_{l=1}^{n-1} l^2 + \ldots$ Mult./Div. $= \frac{(n-1)n(2n-1)}{6} + \ldots$ Mult./Div. $= \frac{2n^3}{6} + \ldots$ Mult./Div.

Dabei stehen die $'\ldots'$ für (irgend)ein Polynom höchstens 2. Grades in n.
Die Anzahl der Additionen/Subtraktionen erhält man analog.

◇

Der Aufwand zur *Lösung von* $Rx = y$ mit einer oberen Dreiecksmatrix R als Übung.

□

Algorithmus (1.1.1) hat also eine höhere Komplexität als das Lösen von Dreiecksgleichungssystemen:

$$\text{Anzahl der } \textit{flops} \text{ (1 Add/Sub + 1 Mult/Div)} = \begin{cases} \frac{n^3}{3} + \ldots & \text{für Algorithmus (1.1.1)} \\ \frac{n^2}{2} + \ldots & \text{für Dreiecksgleichungssysteme} \end{cases}$$

Eine bequeme Sprechweise für diesen Hauptanteil der arithmetischen Operatinen liefern die *Landau*schen $\mathcal{O}$– und o–Symbole:

$$\kappa(n) = \mathcal{O}(n^\alpha)\ ,\ n \to \infty \quad \Longleftrightarrow \quad \exists\, c > 0\ \forall\ n \in \mathbb{N} \quad |\kappa(n)| \ \leq\ c|n^\alpha|\ .$$

Für Algorithmus (1.1.1) ist also die Anzahl $\kappa(n)$ der flops, die sog. arithmetische Komplexität,

$$\kappa(n) = \mathcal{O}(n^3)\ ,\ n \to \infty.$$

Will man die Größe der Konstante bei n^3 herausstreichen, kann man auch schreiben

$$\kappa(n) = \frac{n^3}{3} \ +\ \mathcal{O}(n^2)\ ,\ n \to \infty.$$

Will man mit ein- und derselben Matrix A und mehreren rechten Seiten b das Gleichungssystem $Ax = b$ lösen, so muß man die Hauptarbeit, i.e. den Algorithmus (1.1.1) ohne rechte Seite b, *nur einmal* machen. Um dies einzusehen, stelle ich Algorithmus (1.1.1) als eine Folge von Matrix-Multiplikationen dar, die eine *LR-Faktorisierung* der Matrix A liefern.

Darstellung der Gaußelimination als LR – Zerlegung

Für die auftretenden Matrizen folgende

(1.1.5)' Bezeichnungen

(a) Sei σ Permutation der Zahlen $\{1, 2, \ldots, n\}$. Dann heißt $P_\sigma = [e_{\sigma(1)} \vdots e_{\sigma(2)} \vdots \ldots \vdots e_{\sigma(n)}] \in \mathbb{R}^{n\times n}$ *Permutationsmatrix*. Dabei ist e_j der j–te Einheits(spalten)vektor.

(b)

$$L = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & -L_{k+1,k} & 1 & \\ & & & \vdots & & \ddots \\ & & & -L_{n,k} & & 1 \end{bmatrix} = E - l \cdot e_k{}^t \quad \text{mit} \quad l = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ L_{k+1\ k} \\ \vdots \\ L_{n\ k} \end{bmatrix}$$

heißt *elementare Eliminationsmatrix* (spezielle untere Dreiecksmatrix mit *Einsen-Diagonale*).

(c) Mit den Bezeichnungen von Algorithmus (1.1.1) sei

$$P^{(k)} := P_{\sigma_k}{}^t = P_{\sigma_k} \quad \text{mit} \quad \sigma_k(j) := \begin{cases} i_k & \text{für} \quad j = k \\ k & \text{für} \quad j = i_k \\ j & \text{sonst} \end{cases}$$

$$L^{(k)} := E - l^{(k)}\ e_k{}^t\ ,\ l^{(k)} := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \tilde{\tilde{A}}^{(k)}_{k+1\ k} / \tilde{\tilde{A}}^{(k)}_{k\ k} \\ \vdots \\ \tilde{\tilde{A}}^{(k)}_{n\ k} / \tilde{\tilde{A}}^{(k)}_{k\ k} \end{bmatrix}$$

$$\tilde{l}^{(k)} := \begin{cases} l^{(k)} & , \quad k = n-1 \\ P^{(n-1)} P^{(n-2)} \cdots P^{(k+1)} l^{(k)} & , \quad k < n-1 \end{cases}$$

$$P := P^{(n-1)} P^{(n-2)} \cdots P^{(2)} P^{(1)}\ ;$$

$$L := (E + \tilde{l}^{(1)}\ e_1{}^t) \cdot (E + \tilde{l}^{(2)}\ e_2{}^t) \cdots (E + \tilde{l}^{(n-1)}\ e_{n-1}{}^t)\ .$$

□

Dann gilt

(1.1.5) Satz

Sei $A \in \mathbb{R}^{n\times n}$ nicht singulär. Dann liefert Algorithmus (1.1.1) die Zerlegung

$$PA = LR\ ,$$

wobei mit obigen Bezeichnungen L untere Dreiecksmatrix mit Einsen-Diagonale, R obere Dreiecksmatrix und P eine Permutationsmatrix ist.

Beweis:

$L^{(k+1)}P^{(k+1)}A^{(k+1)} = L^{(k+1)}P^{(k+1)}\left(L^{(k)}P^{(k)}A^{(k)}\right) = L^{(k+1)}\widetilde{L}^{(k)}P^{(k+1)}P^{(k)}A^{(k)}$,
denn

$P^{(k+1)}(E - P^{(k+1)}l^{(k)}{e_k}^t) = \widetilde{L}^{(k)}P^{(k+1)}$ und ganz analog

$P^{(j)}(E - l^{(k)}{e_k}^t) = (E - P^{(j)}l^{(k)}{e_k}^t)P^{(j)}$ für $j > k$.

Man rechnet nach, daß $(E - l^{(k)}{e_k}^t)^{-1} = (E + l^{(k)}{e_k}^t)$ gilt. Daher folgt $PA = LR$.

□

(1.1.6) Lösung von $Ax = b$

Sei $PA = LR$ wie in (1.1.5). Dann gilt $Ax = b$ genau dann, wenn $Ly = Pb$, $Rx = y$ gilt. Diese Dreiecksgleichungssysteme sind durch Rekursion in der Reihenfolge

$$y_1, y_2, ..., y_n \quad \text{(Vorwärtseinsetzen)} \quad \text{und} \quad x_n, x_{n-1}, ..., x_1 \quad \text{(Rückwärtseinsetzen)}$$

zu lösen.

□

(1.1.7) Beispiel (*ohne* vs. *mit* Pivotstrategie)

vgl. Übungsaufgabe 3.1(c): LR–Zerlegung zur Lösung von $\begin{bmatrix} 0.001240 & 1.100 \\ 1.200 & 1.000 \end{bmatrix} x = \begin{bmatrix} 1.000 \\ 2.000 \end{bmatrix}$

liefert in 4–stelliger Gleitpunktarithmetik *ohne Pivotstrategie* die Näherung $\widetilde{x} = \begin{bmatrix} 0.4839 \\ 0.9085 \end{bmatrix}$,

mit Pivotstrategie die Näherung $\widetilde{x} = \begin{bmatrix} 0.9100 \\ 0.9080 \end{bmatrix}$ für die auf 6 Stellen gerundete wahre Lösung

$x = \begin{bmatrix} 0.909946 \\ 0.908065 \end{bmatrix}$.

□

Analyse der in Gleitpunktarithmetik berechneten Ergebnisse– Rückwärtsfehleranalyse

Algorithmus (1.1.1) werde in Gleitpunktarithmetik durchgeführt, wobei (0.4)(d) mit $\varepsilon_0 :\equiv$ *macheps* für $\cdot \in \{+, -, \times, /\}$ gelte:

$$(0.4) \quad x \odot y = (x \cdot y)(1 + \varepsilon) \quad \text{und} \quad x \odot y = (x \cdot y)/(1 + \eta) \text{ mit } |\varepsilon| \leq \varepsilon_0 \ , \ |\eta| \leq \varepsilon_0,$$

Durch Verfolgung der aufgelaufenen Rundungsfehler kann man unter diesen Voraussetzungen zeigen

(1.1.8) Satz
Es gelte (0.4)(d). Sei $A = LR$, d.h. A besitze auch ohne Zeilenvertauschungen eine LR–Zerlegung. Algorithmus (1.1.1) in Gleitpunktarithmetik laufe *ohne* Zeilenvertauschungen ab, und liefere $\widetilde{L}$ und $\widetilde{R}$. Die Lösung von $\widetilde{L}y = b$ in Gleitpunktarithmetik liefere $\widetilde{y}$, die Lösung von $\widetilde{R}x = \widetilde{y}$ in Gleitpunktarithmetik liefere $\widetilde{x}$. Dann gilt für n mit $n\varepsilon_0 < 1$, $\varepsilon_0 \equiv macheps$, für $\varepsilon_0{}' := (n\varepsilon_0)/(1 - n\varepsilon_0)$ (beachte $0 < \varepsilon_0{}' < 1$):

(a) $|(A - \widetilde{L}\widetilde{R})_{i,j}| < \varepsilon_0{}' \sum_{k=1}^{\min\{i,j\}} |\widetilde{L}_{ik}||\widetilde{R}_{kj}|$, $1 \le i, j \le n$;

(b) $(\widetilde{L} + \Delta L)\widetilde{y} = b + \Delta b$; $(\widetilde{R} + \Delta R)\widetilde{x} = \widetilde{y}$ mit
$|(\Delta L)_{ij}| < \varepsilon_0{}'|\widetilde{L}_{ij}|$; $|(\Delta R)_{ij}| < \varepsilon_0{}'|\widetilde{R}_{ij}|$; $|(\Delta b)_i| < \varepsilon_0|b_i|$; $1 \le i, j \le n$.

(1.1.9) Bemerkung
(a) In den Abschätzungen (1.1.8) tauchen auf der rechten Seite die erst *nach* Durchführung des Algorithmus bekannten Größen $\sum_{k=1}^{\min\{i,j\}} |\widetilde{L}_{ik}||\widetilde{R}_{kj}|$ auf. Deshalb nennt man sie *a–posteriori–Abschätzungen.*

(b) In *Stoer [1999]* werden analoge *a–priori–Abschätzungen* gegeben mit den im vorhinein gegebenen Daten $\sum_{k=1}^{\min\{i,j\}} |L_{ik}||R_{kj}|$ an Stelle von $\sum_{k=1}^{\min\{i,j\}} |\widetilde{L}_{ik}||\widetilde{R}_{kj}|$.

□

Zum Nachweis von (1.1.8) beachte man, daß die Verstärkungsfaktoren für die Rundungsfehler von der Form

$$\prod_{j=1}^{k}(1+\varepsilon_j)^{\sigma_j} \quad \text{mit} \quad \sigma_j \in \{-1,+1\} \ , \ |\varepsilon_j| \le \varepsilon_0 \ , \ 1 \le j \le k \ ; \ 1 \le k \le n \ ,$$

sind. Als Übung zeige man mit der Bernoullischen Ungleichung

$$\prod_{j=1}^{k}(1+\varepsilon_j)^{\sigma_j} \ = \ 1 + \varepsilon' \text{ mit } |\varepsilon'| \le \varepsilon_0{}' \ , \ 1 \le j \le k \ ; \ 1 \le k \le n \ .$$

Ich möchte nochmals betonen: In (1.1.8) sind die Matrixmultiplikationen in reeller Arithmetik gemeint, daher der Name

Rückwärtsfehleranalyse: $\widetilde{x}$ (und $\widetilde{y}$) ist exakte Lösung eines benachbarten, für $\varepsilon_0{}' << 1$ nur wenig gestörten Problems: Setzt man per definitionem

$$\Delta A \ := \ (\widetilde{L} + \Delta L)(\widetilde{R} + \Delta R) - A \ ,$$

(Matrixoperationen in reeller (exakter) Arithmetik), dann gilt

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

im Vergleich zu

$$Ax = b \ .$$

Die Größe der Störmatrix ΔA bzw. des Störvektors Δb kann man nach Satz (1.1.8) abschätzen. Sie ist also klein für $n\varepsilon_0{}' << 1$.

(1.1.10) Folgerung
Unter den Voraussetzungen und mit den Bezeichnungen von Satz (1.1.8) gilt mit
$\Delta A := (\widetilde{L} + \Delta L)(\widetilde{R} + \Delta R) - A$

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

wobei

(a) $|\Delta A_{ij}| < (3\varepsilon_0{}' + \varepsilon_0{}'^2) \sum_{k=1}^{\min(i,j)} |\widetilde{L}_{ik}\widetilde{R}_{kj}| \ .$

(b) $\|\Delta A\|_\infty < n(3\varepsilon_0{}' + \varepsilon_0{}'^2) \max_{1\le i,j\le n} \sum_{k=1}^{\min(i,j)} |\widetilde{L}_{ik}\widetilde{R}_{kj}| \ .$

(c) $\|\Delta b\|_\infty < \varepsilon_0 \|b\|_\infty \ .$

Beweis:

$\Delta A = (\widetilde{L} \cdot \widetilde{R} - A) + (\Delta L \cdot \widetilde{R}) + (\widetilde{L} \cdot \Delta R) + (\Delta L \cdot \Delta R).$

$(\widetilde{L} \cdot \widetilde{R}) - A)_{ij}$ nach (1.1.8)(a) abschätzbar mit Faktor $\varepsilon_0{}'$,
$(\widetilde{L} \cdot \Delta R)_{ij}$ nach (1.1.8)(b) abschätzbar mit Faktor $\varepsilon_0{}'$,
$(\widetilde{L} \cdot \Delta R)_{ij}$ nach (1.1.8)(b) abschätzbar mit Faktor $\varepsilon_0{}'$,
$(\Delta L \cdot \Delta R)_{ij}$ nach (1.1.8)(b) abschätzbar mit Faktor $\varepsilon_0{}'^2$.
Die Abschätzung für $\|\Delta b\|_\infty$ ist nach (1.1.8)(b) klar.

□

Wie wirken sich solche Fehler in den Matrixelementen oder der rechten Seite auf die Lösung aus? Dies ist gerade die *Kondition* des Problems, ein lineares Gleichungssystem zu lösen:

(1.1.11) Definition (*Kondition* einer Matrix)
Sei $(\mathbb{K}^n, \|.\|)$ und die nichtsinguläre Matrix $A \in \mathbb{K}^{n\times n}$ gegeben. Dann heißt

$$cond_{ind}(A) := \|A\|_{ind}\|A^{-1}\|_{ind}$$

die *Kondition* der Matrix $A \in \mathbb{K}^{n\times n}$ (bezüglich der induzierten Matrixnorm).
Bei Verwendung der l_p–Norm schreiben wir

$$cond_p(A) := \|A\|_p\|A^{-1}\|_p$$

□

Natürlich ist die Größe $cond(A)$ normabhängig. Die Abschätzungen zwischen den verschiedenen p–Vektornormen ermöglichen aber Abschätzungen zwischen den entsprechenden induzierten p–Matrixnormen.
Aus (0.10) erhält man eine untere Abschätzung für die Kondition

$$\Big(\max_{\lambda \ Eigenwert \ von \ A} |\lambda|\Big) \ / \ \Big(\min_{\lambda \ Eigenwert \ von \ A} |\lambda|\Big) \ \le \ cond(A) \ .$$

Die *Abhängigkeit der Lösung* eines linearen Gleichungssystems *von den Daten* beschreibt

(1.1.12) Satz
Betrachte $(\mathbb{K}^n, \|.\|)$ und die induzierte Matrixnorm. Sei $A \in \mathbb{K}^{n\times n}$ nichtsingulär, $\Delta A \in \mathbb{K}^{n\times n}$ mit $\|A^{-1}\|\|\Delta A\| < 1$, $\Delta b \in \mathbb{K}^n$, und $\mathbb{K}^n \ni b \neq 0$. Dann gilt für

$$Ax = b \quad , \quad (A + \Delta A)(x + \Delta x) = b + \Delta b$$

die Abschätzung

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

Beweis:

Aus den Gleichungen für $x + \Delta x$ und x erhält man durch Ausmultiplizieren und Einsetzen

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

$\Longrightarrow$

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

nach dem folgenden Satz (1.1.13). Beachtet man

$$\|b\| = \|Ax\| \leq \|A\|\|x\| \quad \text{d.h.} \quad \frac{1}{\|A\|\|x\|} \leq \frac{1}{\|b\|} \text{ , falls } b \neq 0 \; ,$$

so folgt die Behauptung.

□

$$\frac{cond(A)}{1 - cond(A)\frac{\|\Delta A\|}{\|A\|}} \doteq cond(A) \quad \text{, falls } cond(A)\frac{\|\Delta A\|}{\|A\|} << 1 \; ,$$

z.B. für $\Delta A = 0$. In diesem Fall erhält man die Aussage (1.1.12) sofort aus den Abschätzungen

$$\|\Delta x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\|\|\Delta b\| \; , \; \|b\| = \|Ax\| \leq \|A\|\|x\| \; .$$

(1.1.13) Satz
Sei $A \in \mathbb{K}^{n\times n}$ invertierbar und $\Delta A \in \mathbb{K}^{n\times n}$ mit $\|A^{-1}\|\|\Delta A\| < 1$. Dann ist $A + \Delta A$ invertierbar und es gilt

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

Beweis:

Wegen

$$\|\Delta A y\| \le \|\Delta A\|\,\|y\| \quad \text{bzw.} \quad -\|\Delta A\ y\| \ge -\|\Delta A\|\,\|y\|$$

und

$$\|y\| = \|A^{-1}\ A\ y\| \le \|A^{-1}\|\,\|A\ y\| \quad \text{bzw.} \quad \|A\ y\| \ge \frac{1}{\|A^{-1}\|}\|y\|$$

folgt

$$\|(A+\Delta A)y\| \ge \|Ay\| - \|\Delta A\ y\| \ge \left(\frac{1}{\|A^{-1}\|} - \|\Delta A\|\right)\|y\| \ = \underbrace{\left(\frac{1-\|A^{-1}\|\|\Delta A\|}{\|A^{-1}\|}\right)}_{>\,0 \text{ nach Vor.}}\|y\|\ ,$$

insgesamt also

$$\frac{1-\|A^{-1}\|\,\|\Delta A\|}{\|A^{-1}\|}\ \|y\| \ \le\ \|(A+\Delta A)y\|\ .$$

Daraus folgt aber sowohl, daß $(A+\Delta A)^{-1}$ existiert (denn Kern $N(A+\Delta A) = \{0\}$), als auch für $y := (A+\Delta A)^{-1}z$, daß gilt

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

□

(1.1.14) Beispiel (für Satz (1.1.12))

$$A = \begin{bmatrix} 2 & 2 \\ 1+\alpha & 2 \end{bmatrix},\ 0 \le \alpha < 1\,; \quad A^{-1} = \frac{1}{2(1-\alpha)}\begin{bmatrix} 2 & -2 \\ -(1+\alpha) & 2 \end{bmatrix}$$

$$b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \quad \tilde{b} = \begin{bmatrix} 1+\delta \\ 1-\delta \end{bmatrix} = b + \begin{bmatrix} +\delta \\ -\delta \end{bmatrix}, \quad \delta > 0\,;\ x : Ax = b \implies x = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}$$

$$\widetilde{x} : A\widetilde{x} = \tilde{b} \implies \widetilde{x} = A^{-1}\tilde{b} = x + A^{-1}\begin{bmatrix} +\delta \\ -\delta \end{bmatrix} = x + \frac{1}{2(1-\alpha)}\begin{bmatrix} 4\delta \\ -(3+\alpha)\delta \end{bmatrix};$$

Für $p = \infty$ erhält man für die in Satz (1.1.12) vorkommenden Größen

$$\|b-\tilde{b}\|_\infty = \delta\,;\ \|b\|_\infty = 1\,;\ \|A\|_\infty = max\{|2|+|2|, |1+\alpha|+|2|\} = 4\,;$$

$$\|A^{-1}\|_\infty = \frac{1}{2(1-\alpha)}\max\{|2|+|-2|, |-(1+\alpha)|+|2|\} = \frac{2}{(1-\alpha)}\,;$$

$$\frac{\|x-\widetilde{x}\|_\infty}{\|x\|_\infty} = [\frac{1}{2(1-\alpha)}\,(\max\{4, 3+\alpha\})\,\delta]/(\frac{1}{2}) = \frac{4}{(1-\alpha)}\,\delta\,;$$

Andererseits gilt

$$cond_\infty(A)\frac{\|\Delta b\|_\infty}{\|b\|_\infty} = 4\cdot\frac{2}{(1-\alpha)}\,\delta\,,$$

d.h. der tatsächliche Fehler wird durch die Abschätzung in Satz (1.1.12) (nur) um den Faktor 2 überschätzt, gleichmäßig für $0 \le \alpha < 1$ (beachte $\|x-\widetilde{x}\|_\infty/\|x\|_\infty \longmapsto +\infty$ für $\alpha \longmapsto 1$) .

□

Wendet man Satz (1.1.12) an mit den in Folgerung (1.1.10) gegebenen Abschätzungen für die Größe der Störung, so erhält man *gute* Schranken für den relativen Fehler $\|\Delta x\|/\|x\|$, wenn

$$cond(A) \quad \text{und} \quad \sum_k |\widetilde{L}_{ik}||\widetilde{R}_{kj}| \quad \text{nicht zu groß ist.}$$

Pivotstrategie verhindert i.Allgem. ein zu starkes Anwachsen des zweiten Ausdrucks. Die Eigenschaft eines gegebenen linearen Gleichungssystems,

'gut' oder *'schlecht' konditioniert*

zu sein, ist natürlich abhängig von der zur Verfügung stehenden *Rechengenauigkeit*:

$$\textit{Regel}: \quad cond(A) \leq \frac{1}{\sqrt{macheps}} \quad \longleftrightarrow \quad A \text{ gut/nicht zu schlecht konditioniert}$$

$$cond(A) >> \frac{1}{\sqrt{macheps}} \quad \longleftrightarrow \quad A \text{ schlecht/kraß schlecht konditioniert}$$

□

Ist nach dieser Regel das Problem schlecht gestellt, sind zwei Reaktionen möglich:

– Erhöhung der Rechengenauigkeit, d.h Verkleinerung von *macheps* und damit Vergrößerung von $\frac{1}{\sqrt{macheps}}$.

– Transformation in ein äquivalentes Problem mit kleinerer Kondition, sog.

Vorkonditionierung
Finde $C, X \in \mathbb{R}^{n\times n}$ nichtsingulär, so daß für $A' := CAX$ gilt

(a) $cond(A') < cond(A)$ (am liebsten $cond(A') << cond(A)$).

(b) $A'x' = b'$, $b' = Cb$, $x = Xx'$; ist nicht viel aufwändiger aufstellbar und lösbar als $Ax = b$.

□

Nach praktischer Erfahrung günstig sind Gleichungssysteme mit *äquilibrierter Matrix* A', d.h. $\sum_j |A'_{ij}| = \sum_i |A'_{ij}| = const, 1 \leq i, j \leq n$.
Leicht zu erreichen ist folgender 'Ersatz': $\sum_i |A_{ij}| = const, 1 \leq i \leq n$; d.h.

$$A'_{i.} := \frac{1}{\|A_{i.}\|_1} A_{i.}, 1 \leq i \leq n.$$

Dies entspricht in obigem Vorkonditionierungsschema gerade dem Fall

$X = E$, $C = diag(1/(\sum_j |A_{1j}|), ..., 1/(\sum_j |A_{nj}|))$.

Diese Vorkonditionierung wird in jedem Softwarepaket automatisch durchgeführt.

Symmetrische positiv definite Systeme – Cholesky-Verfahren

Im Fall symmetrischer positiv definiter Matrizen existiert eine LR–Zerlegung ohne Zeilenpermutationen. Dann ist also die Voraussetzung in Satz (1.1.8) erfüllt. Außerdem ist dann die Gaußelimination stabil (vgl. (1.1.9)(b)).

(1.1.15) Definition
Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ heißt *positiv definit*, wenn $x^t Ax > 0$ gilt für $x \neq 0$.

□

Zur Vorbereitung der Faktorisierungen $A = LL^t$ resp. $A = L_1 D {L_1}^t$ ein

(1.1.16) Hilfssatz
Sei $X \in \mathbb{R}^{n\times n}$ nichtsingulär. Dann ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit (resp. semidefinit) genau dann, wenn XAX^t diese Eigenschaften hat. Mit A hat dann auch A^{-1} diese beiden Eigenschaften.

Beweis:

$$A \text{ symmetrisch } \Longrightarrow (XAX^t)^t = (X^t)^t \underbrace{A^t}_{=A} X^t = XAX^t \text{ , d.h. } XAX^t \text{ symmetrisch .}$$

Mit X ist auch X^t nichtsingulär. Damit gilt

$$A\text{positiv definit } \Longrightarrow y^t(XAX^t)y = (X^t y)^t A(X^t y) > 0 \text{ für } y \neq 0, \text{ denn } X^t y \neq 0.$$

Die Umkehrung ist klar, da auch X^{-1} nichtsingulär ist.

□

(1.1.17) Satz
Zu jeder symmetrischen positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gibt es eine eindeutige Faktorisierung

$$A = L_1 D (L_1)^t ,$$

wobei L_1 untere Dreiecksmatrix mit Einsen-Diagonale und $D = diag(A_{11}, A_{22}, ...A_{nn})$ eine Diagonalmatrix mit positiven Diagonalelementen ist.

Beweis:

$n = 1$: klar, denn $A_{11} = {e_1}^t A e_1 > 0$ wegen 'pos.def.'

$n - 1 \to n$: Wegen $A_{11} > 0$ ist der erste Gauß-Eliminationsschritt ohne Zeilenvertauschungen möglich
$\Longrightarrow$

$$L^{(1)}A = \left[\begin{array}{c|c} A_{11} & *\cdots* \\ \hline 0 & \\ \vdots & B^{(0)} \\ 0 & \end{array}\right] \quad , L^{(1)} = (E - l \cdot {e_1}^t) \quad ; \quad l = \left[\begin{array}{c} 0 \\ \hline l' \end{array}\right] \quad ;$$

Nach Hilfssatz (1.1.16) ist $L^{(1)}\ A\ {L^{(1)}}^t$ symmetrisch und positiv definit, also gilt

$$\left(L^{(1)}\ A\right)\ {L^{(1)}}^t \ = \ \left[\begin{array}{c|c} A_{11} & *\cdots* \\ \hline 0 & B^{(0)} \end{array}\right] \ {L^{(1)}}^t \ = \ \left[\begin{array}{c|c} A_{11} & 0^t \\ \hline 0 & B^{(1)} \end{array}\right] \text{ auf Grund der Symmetrie,}$$

wobei $B^{(1)}$ mit $L^{(1)}\ A\ {L^{(1)}}^t$ wieder symmetrisch und positiv definit ist. Nach Induktionsvoraussetzung gilt daher

$$B^{(1)} \ = \ M^{(1)}\ D^{(1)}\ {M^{(1)}}^t ,$$

wobei $M^{(1)}$ untere Dreiecksmatrix mit Einsen-Diagonale und $D^{(1)}$ eine Diagonalmatrix ist. Damit folgt

$$\left[\begin{array}{c|c} A_{11} & 0^t \\ \hline 0 & B^{(1)} \end{array}\right] = \left[\begin{array}{c|c} A_{11} & 0^t \\ \hline 0 & M^{(1)}\ D^{(1)}\ M^{(1)^t} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0^t \\ \hline 0 & M^{(1)} \end{array}\right] \left[\begin{array}{c|c} A_{11} & 0^t \\ \hline 0 & D^{(1)} \end{array}\right] \left[\begin{array}{c|c} 1 & 0^t \\ \hline 0 & M^{(1)} \end{array}\right]^t$$

Mit

$$L_1 := (L^{(1)})^{-1} \left[\begin{array}{c|c} 1 & 0^t \\ \hline 0 & M^{(1)} \end{array}\right] = \left[\begin{array}{c|c} 1 & 0^t \\ \hline l' & M^{(1)} \end{array}\right] \quad , \quad D := \left[\begin{array}{c|c} A_{11} & \\ \hline & D^{(1)} \end{array}\right]$$

folgt die Identität

$$A = L_1 D (L_1)^t \ .$$

Da L_1 untere Dreiecksmatrix mit Einsen-Diagonale ist, folgt

$$A_{jj} \ = \ e_j{}^t \ (L_1 D L_1{}^t) \ e_j \ = \ (L_1)_{j.} \ D \ (L_1{}^t)_{.j} \ = \ D_{jj} \ .$$

□

(1.1.18) Folgerung
Zu jeder symmetrischen positiv definiten Matrix $A \in \mathbb{R}^{n \times n}$ gibt es eine eindeutige Faktorisierung

$A = LL^t$,

wobei L untere Dreiecksmatrix mit positiven Diagonalelementen ist.

Beweis:

Nach (1.1.17) gilt $A = L_1 D L_1{}^t$. Wegen $D = \sqrt{D}\sqrt{D}$ mit $\sqrt{D} := diag(\sqrt{D_{11}}, ..., \sqrt{D_{nn}})$ folgt mit

$$L := L_1 \ \sqrt{D}$$

die *Existenz*.

Die *Eindeutigkeit* ergibt sich hier und in Satz (1.1.17) durch rekursive Auflösung der definierenden Gleichungen:

für $i = 1, ..., n$
  für $j = 1, ..., i$ löse

$$L_{i.} \cdot \ (L_{j.})^t \ = \ A_{ij} \ .$$

Beachte dazu $L_{i.} \cdot \ (L^t)_{.j} = L_{i.} \cdot \ (L_{j.})^t$.

□

Die *rekursive Auflösung* – von oben nach unten – der L definierenden Gleichungen ist der *Cholesky-Algorithmus:*

$$L_{i\cdot} \cdot (L_{j\cdot})^t = A_{ij} \ , \ 1 \le j \le i \ , \ 1 \le i \le n \ .$$

Dieser Algorithmus erkennt auch die (zumindest in Gleitpunktarithmetik unter Umständen vorhandene) Nichtdefinitheit. Seine guten Stabilitätseigenschaften kann man nach Folgerung (1.1.9) analysieren.

Besonders einfach kann man die im Anschluß an Satz (1.1.8) zitierte a–priori–Rückwärtsanalyse (1.1.9)'(b) anwenden, denn mit der Cauchy–Schwarzschen Ungleichung folgt aus $A = LL^t$

$$\max_{1 \le i,j \le n} | A_{ij} | = \max_{1 \le i \le n} A_{ii} \ .$$

1.2 QR-Zerlegung

Um zusätzliche Konditionsverschlechterungen zu vermeiden, verwendet man nicht die *LR–Zerlegung*, sondern eine *QR–Zerlegung*:

(1.2.1) QR–Zerlegung
Gegeben $A \in \mathbb{R}^{m \times n}$, $m \ge n$. Dann heißt die Faktorisierung

$A = Q\,R$, $Q \in \mathbb{R}^{m \times m}$ orthogonal , $R \in \mathbb{R}^{m \times n}$ obere Dreiecksmatrix ;

eine *QR–Zerlegung* von A.

□

Eine Möglichkeit zur Erzeugung einer QR-Zerlegung bietet

(1.2.2) Householder–Spiegelung
Sei $w \in \mathbb{R}^{n \times 1}$, $\|w\|_2 = 1$. Dann heißt

$H = E - 2ww^t$,

Householder–Spiegelung. Für $u \in span(w)$, $u \neq 0$, gilt

$H = E - \pi^{-1} uu^t$ mit $\pi = \frac{u^t u}{2}$.

□

(1.2.3) Bemerkung
Die Householder–Spiegelung ist symmetrisch und orthogonal.

Beweis:

$$H = E - 2w\ w^t \ , \ w^t\ w = 1 \Longrightarrow H^t = E^t - 2(w^t)^t\ w^t = H;$$

$$H^t\ H = (E - 2w\ w^t)(E - 2w\ w^t) = E - 2w\ w^t - 2w\ w^t + 4w \underbrace{w^t\ w}_{=1} w^t = E \ .$$

□

(1.2.4) Satz
Sei $x \in \mathbb{R}^n$, $x \notin span(e_1)$. Setzt man $\sigma \in \{+\|x\|_2, -\|x\|_2\}$, $u := x + \sigma e_1$, $\pi := (u^t\ u)/2$, so ist

$$H \ = \ E \ - \ \pi^{-1}\ u\ u^t$$

eine Householderspiegelung mit

$$Hx \ = \ -\sigma e_1 \ .$$

Beweis:

Wegen $x \notin span(e_1)$ ist $u \neq 0$. Mit $w = u/\|u\|_2$ folgt

$(E - 2ww^t)x \ = \ x - u\big(2u^t x/u^t u\big) \overset{(*)}{=} x - u \ = \ -\sigma e_1$; $(*)$ gilt wegen

$u^t u \ = \ (x + \sigma e_1)^t (x + \sigma e_1) \ = \ x^t x + \sigma {e_1}^t x + \underbrace{\sigma x^t e_1}_{= \sigma {e_1}^t x} + \underbrace{\sigma^2 {e_1}^t e_1}_{= x^t x} \ = \ 2(x + \sigma e_1)^t x \ = \ 2u^t x$.

□

Für eine tatsächliche Programmierung geeignet ist folgende Version, aus der man auch die arithmetische Komplexität ablesen kann:

(1.2.4') Elementare Householder–Spiegelung $\boldsymbol{H_u \equiv (E - \pi^{-1} u u^t)}$
Gegeben: $x \in \mathbb{R}^m$, $x \neq 0$, $x \notin span(e_1)$.

Gesucht: $u \in \mathbb{R}^m$, $\sigma \in \mathbb{R}$ *mit* $(E - \pi^{-1} u u^t)x \ = -\sigma e_1$ *mit* $\pi = \frac{1}{2} u^t u$.

Bestimmung von $\boldsymbol{u}$, $\boldsymbol{\pi}$ und $\boldsymbol{\sigma}$ durch:

(1) $\eta := \max_{1 \leq i \leq m} |x_i|$; ('*Skalierung*' *zur Vermeidung von*

(2) $\widetilde{x}_i := x_i/\eta$, $i = 1, ..., m$; *Exponentenunter/überlauf in* (3))

(3) $\nu := sign(\widetilde{x}_1) \cdot \sqrt{\widetilde{x}_1^{\,2} + ... + \widetilde{x}_m^{\,2}}$; ($sign(\widetilde{x}_1) := 1$, $falls$ $\widetilde{x}_1 = 0$)

(4) $u_1 := \widetilde{x}_1 + \nu$; $(u = \widetilde{x} + sign(\widetilde{x}_1)\|\widetilde{x}\|_2 e_1$, $H = (E - (\frac{u^t u}{2})^{-1} u u^t)$

(5) $\pi := \nu \cdot |u_1|$; (*denn* $\frac{1}{2} u^t u = u^t \widetilde{x}$; $u^t \widetilde{x} = \widetilde{x}^t \widetilde{x} + sign(\widetilde{x}_1)\|\widetilde{x}\|_2 \widetilde{x}_1$)

(6) $\sigma := \eta \cdot \nu$;

□

Bemerkung
Die Wahl $\sigma := sign(x_1)\|x\|_2$ für $x = [x_i : 1 \leq i \leq n]$ ergibt *keine Auslöschung* bei der Berechnung von u und π^{-1}.

□

Veranschaulichung der beiden Vorzeichenwahl–Möglichkeiten in den Abbildungen 1.1 und 1.2.

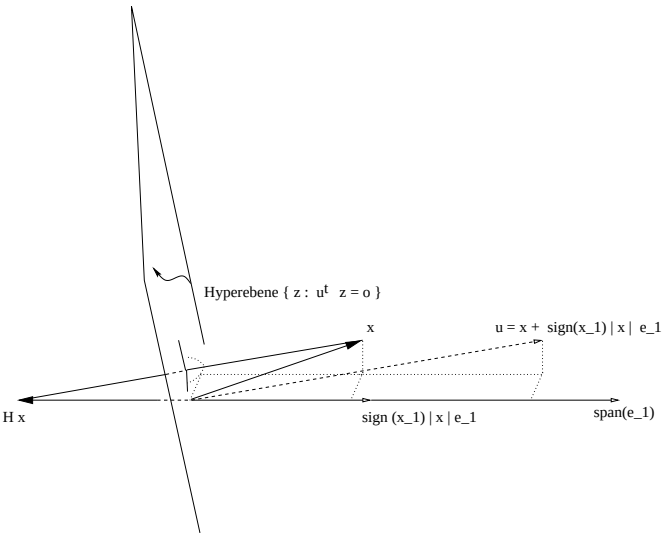

Abbildung 1.1: Stabile Householder-Spiegelung

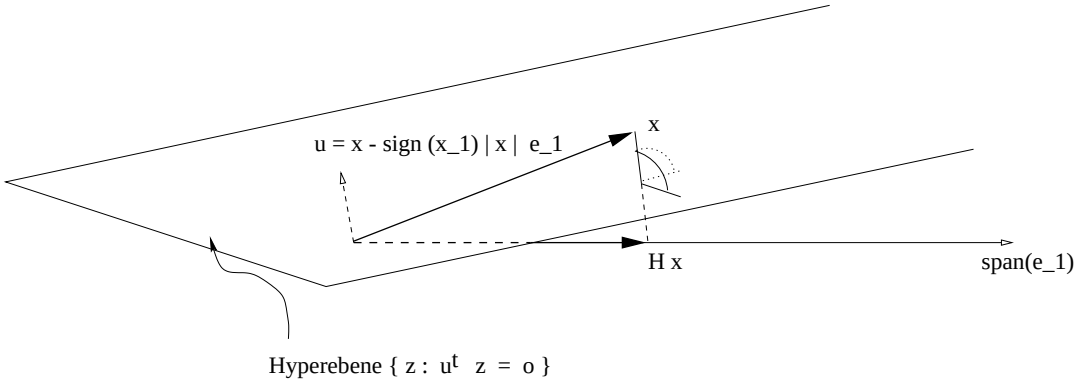

Abbildung 1.2: Instabile Householder-Spiegelung

(1.2.5) Satz (QR–Zerlegung)
Sei $A \in \mathbb{R}^{m\times n}$, $m \geq n$. Dann existiert $Q \in \mathbb{R}^{m\times m}$ orthogonal, und eine obere Dreiecksmatrix $R \in \mathbb{R}^{m\times n}$ mit

$$A = Q\,R\,.$$

Beweis:

Wir führen den Beweis konstruktiv über den folgenden

(1.2.5') Algorithmus zur QR–Zerlegung (*mittels Householder–Spiegelungen*)
Gegeben: $A \in \mathbb{R}^{m\times n}$, $m \geq n$.

Gesucht: $Q \in \mathbb{R}^{m\times m}$ *orthogonal derart, daß*
$Q^t\,A = R$ obere Dreiecksmatrix ist.

$A^{(1)} := A$; $\widetilde{n} := min\{m-1, n\}$;

für $k := 1, 2, ..., \widetilde{n}$ führe durch

(0) gegeben $A^{(k)}$ mit $A^{(k)}_{ij} = 0,\ j < i \leq m,\ 1 \leq j < k$;

Falls (1) $\max_{k<i\leq m} |A^{(k)}_{ik}| > 0$, *dann*

(2) berechne zu $x := [A^{(k)}_{ik} : k \leq i \leq m] \in \mathbb{R}^{(m+1-k)}$ eine elementare Householder–Spiegelung $H \in \mathbb{R}^{(m+1-k)\times(m+1-k)}$ mit $Hx = -\sigma\,[\delta_{ik}]_{k\leq i\leq m} \in \mathbb{R}^{m+1-k}$.
Mit der Einheitsmatrix $E \in \mathbb{R}^{(k-1)\times(k-1)}$ setze

$$H^{(k)} := \left[\begin{array}{c|c} E & 0 \\ \hline 0 & H \end{array}\right] \in \mathbb{R}^{m\times m} \textit{ orthogonal}$$

(3) $A^{(k+1)} := H^{(k)} A^{(k)}$;

sonst

(4) $A^{(k+1)} := A^{(k)}$; $H^{(k)} := E \in \mathbb{R}^{m\times m}$;

◇

Ergebnis: $R := A^{(\tilde{n}+1)} \in \mathbb{R}^{m\times n}$ ist obere Dreiecksmatrix, $R = H^{(\tilde{n})} \cdot \ldots \cdot H^{(1)} A$ bzw.

$A = Q\,R$ *mit* (*denn* $H^{(k)^t} = H^{(k)}$)

$Q := H^{(1)} \cdot \ldots \cdot H^{(\tilde{n})} \in \mathbb{R}^{m\times m}$ orthogonal.

□

(1.2.6) Bemerkung (arithmetischer Aufwand für (1.2.5'))
Für $A \in \mathbb{R}^{m\times n}$ benötigt man für obige *QR–Zerlegung* (1.2.5′)

$n^2(m - \frac{n}{3}) + \ldots$ *flops* (1 *flop* = 1 Addition/Subtraktion und 1 Multiplikation/Division),

im Fall $n = m$ also $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ *flops* für die *QR–Zerlegung*, und zur *Lösung von* $Qy = b$ und *des Dreiecksgleichungssystems* $Rx = y$

$\frac{3}{2}n^2 + \mathcal{O}(n)$ *flops*.

Dabei stehen die Punkte $'\ldots'$ für ein Polynom in n und m von niedrigerem Gesamtgrad als der angegebene Hauptterm.

□

Beweis als Übung.

Algorithmus (1.2.5') hat also für $m = n$ eine doppelt so hohe Komplexität wie die Gauß-Elimination.

$$\text{arithmetischer Aufwand bei Gauß-Elimination} = \begin{cases} \frac{n^3}{3} + \mathcal{O}(n^2) & \text{für Gauß-Elimination} \\ \frac{n^2}{2} + \mathcal{O}(n) & \text{für Dreiecksgleichungssystem} \end{cases}$$

Die guten Stabilitätseigenschaften der QR–Zerlegung erkennt man auch daran, daß keine Konditionsverschlechterung beim Lösen der zerlegten Gleichungssysteme auftritt (Beweis der folgenden Bemerkung als Übung).

(1.2.7) Bemerkung
Sei $A = QR \in \mathbb{R}^{n\times n}$ nichtsingulär und $b \in \mathbb{R}^n$. Die Lösung von $Ax = b$ erhält man durch Lösen von $Rx = Q^t b$. Dabei gilt

$$cond_2(R) \;=\; cond_2(A) \;.$$

□

Eine Anwendung der QR–Zerlegung ist die 'Lösung' überbestimmter Gleichungssysteme

$$Ax \doteq b \;,\; A \in \mathbb{R}^{m\times n} \;,\; b \in \mathbb{R}^m \;\text{ mit }\; m > n \;.$$

Als Verallgemeinerung des Lösungsbegriffs suchen wir ein $x^+ \in \mathbb{R}^n$ mit minimalem Defekt in der 2-Norm, eine sogenannte 'Least–Squares–Lösung':

$$x^+ \in \mathbb{R}^n \;\text{ mit }\; \|Ax^+ - b\|_2 \;=\; \min_{y\,\in\,\mathbb{R}^n} \|Ay^+ - b\|_2 \;.$$

(1.2.8) Satz
Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ mit $m > n$. Hat A vollen Rang, $rang(A) = n$, so gilt für $A = QR$ mit den Blockzerlegungen

$$R \;=\; \left[\begin{array}{c} \widetilde{R} \\ \hline 0 \end{array}\right] \;,\; Q^t b \;=\; \left[\begin{array}{c} \widetilde{b_1} \\ \hline \widetilde{b_2} \end{array}\right] \;,$$

daß die obere Dreiecksmatrix $\widetilde{R} \in \mathbb{R}^{n\times n}$ nichtsingulär ist und $x^+ \;=\; \widetilde{R}^{-1}\widetilde{b_1}$ ist, d.h. x^+ löst

$$\widetilde{R}x^+ \;=\; \widetilde{b_1} \;.$$

Beweis:

$rang(R) = rang(QR) = n$ nach Voraussetzung über A. Damit ist $\widetilde{R}$ nichtsingulär und $x^+ \;=\; \widetilde{R}^{-1}\widetilde{b_1}$ wohldefiniert. Für $y \in \mathbb{R}^n$ gilt

$$Ry \;=\; \left[\begin{array}{c} \widetilde{R}y \\ \hline 0 \end{array}\right] \;,$$

und damit

$$\|Ay - b\|_2{}^2 \;=\; \|Q^t\big(Ay - b\big)\|_2{}^2 \;=\; \|Q^t(QR)y - Q^t b\|_2{}^2 \;=\; \|Ry - Q^t b\|_2{}^2$$

$$=\; \|\widetilde{R}y - \widetilde{b_1}\|_2{}^2 \;+\; \|0 - \widetilde{b_2}\|_2{}^2 \;.$$

Da $\|\widetilde{R}y - \widetilde{b_1}\|_2 > 0$ ist für $y \neq x^+$ (und $\|\widetilde{R}x^+ - \widetilde{b_1}\|_2 = 0$ ist), folgt die Behauptung.

□

Kapitel 2

Interpolation

2.1 Allgemeine Betrachtungen zur Interpolation

In diesem und dem nächsten Kapitel behandle ich die näherungsweise Darstellung von Funktionen. Die erste Möglichkeit ist die Interpolation , die zweite Möglichkeit die sog. beste Approximation . Die Interpolation ist die '*exakte Wiedergabe in endlich vielen Punkten*'. Zur Wiedergabe verwendet man Funktionen einfacher Bauart, die günstig weiterverarbeitet werden können z.B. bezüglich der *Auswertung* in anderen Argumentpunkten, bezüglich der *Integration* usw. Die Bauart wird gegeben durch

$$V \subset \mathbb{R}^J := \{f : J \longrightarrow \mathbb{R} \text{ Abbildung } \}.$$

I.a. wird

$$V \subset C(J) \equiv C^0(J) := \{f : J \longrightarrow \mathbb{R} \mid f \text{ stetig } \},$$

sein. Analog zu $C^0(J)$ sei

$$C^k(J) := \{f : J \longrightarrow \mathbb{R} \mid f \; k\text{-mal stetig differenzierbar } \} \; , \; k \in \mathbb{N}.$$

(2.1.1) Definition
Gegeben sei $J \subset \mathbb{R}^d, V \subset \mathbb{R}^J$ und eine *endliche Punktmenge* $I = \{x_1, ..., x_n\} \subset J$ von n paarweise verschiedenen Punkten $\{x_1, ..., x_n\}$.

(a) Zu n Werten $y_1, ..., y_n \in \mathbb{R}$ heißt $v \in V$ *Interpolierende* aus V für die Wertepaare $(x_i, y_i), i = 1, ..., n$, wenn

$$v(x_i) = y_i \; , \; 1 \leq i \leq n.$$

(b) Zu $f \in C(J)$ heißt $v \in V$ *Interpolierende* aus V für f auf I, wenn

$$v(t) = f(t) \; , \; t \in I.$$

□

(2.1.2) Bemerkungen (und Bezeichnungen)
(a) Ist $V \subset C(J)$ Teilraum, $dim V < \infty$, dann spricht man von *linearer Interpolation.* *Das Beispiel* ist die Polynominterpolation

$$V = \mathbb{P}_n|_J = \{\ p|_J : J \to \mathbb{R} \mid p \text{ Polynom }, \ grad(p) < n\}\ .$$

$\mathbb{P}_n$ sind die Polynome der sog. *Ordnung* n , d.h. vom Höchstgrad $n-1$. I.a. unterdrücke ich die ausdrückliche Angabe des Definitionsbereiches J und schreibe kurz $\mathbb{P}_n$ anstelle von $\mathbb{P}_n|_J$.

(b) Ist V kein Teilraum, so spricht man von *nichtlinearer Interpolation.* Beispiel ist die *'rationale Interpolation'* durch

$$V = \{p/q : p \in \mathbb{P}_n, q \in \mathbb{P}_m : q(t) \neq 0, t \in J.\}$$

(c) $J \subset \mathbb{R}^d, d \geq 2$, ist sog. *'mehrdimensionale' Interpolation* .

□

(2.1.3) Satz (Existenz und Eindeutigkeit bei linearer Interpolation)
Sei $V = span(\{u_1, ..., u_n\}) \subset \mathbb{R}^J$. Dann gilt

(a) Zu den n paarweise verschiedenen Punkten $x_1, x_2, \ldots, x_n \in J$ existiert zu beliebigen $y_1, y_2, \ldots, y_n \in \mathbb{R}$ genau eine Interpolierende aus V auf $I := \{x_1, x_2, \ldots, x_n\}$ dann und nur dann, wenn

$$(*) \quad det\big([u_j(x_i)]_{1\leq i,j\leq n}\big) \neq 0\ .$$

(b) Sei $dim\ V = n$. Zu je n paarweise verschiedenen Punkten $x_1, x_2, \ldots, x_n \in J$ ist die Interpolationsaufgabe auf $I = \{x_1, x_2, \ldots, x_n\}$ bzgl. V für beliebige $y_1, y_2, \ldots, y_n \in \mathbb{R}$ eindeutig lösbar genau dann, wenn jedes $v \in V, v \neq 0$, höchstens $n-1$ Nullstellen in J hat.

Beweis:

(a) $v \in V$ interpoliert (x_i, y_i) , $i = 1, .., n$; für beliebige $y_1, y_2, \ldots, y_n$ $\Longleftrightarrow$

$$v = \sum_{j=1}^{n} a_j u_j\ ,\ \sum_{j=1}^{n} a_j u_j(x_i) = y_i\ ,\ 1 \leq i \leq n\ ; \quad \text{für beliebige } y_1, y_2, \ldots, y_n \quad \Longleftrightarrow$$

für die Matrix dieses linearen Gleichungssystems gilt

$$det\big([u_j(x_i)]_{1\leq i,j\leq n}\big) \neq 0$$

◇

(b) Es existiert $v = \sum_{j=1}^{n} a_j u_j \neq 0$, $v(x_i) = 0$, $1 \leq i \leq n$

$\Longleftrightarrow$ (da $\{u_1, \ldots, u_n\}$ linear unabhängig ist)

das homogene lineare Gleichungssystem $[u_j(x_i)]_{1\leq i,j\leq n}\ [y_j]_{1\leq j\leq n} = 0$ hat eine nichttriviale Lösung $[a_j]_{1\leq j\leq n} \neq 0$

$\Longleftrightarrow$ $det\big([u_j(x_i)]_{1\leq i,j\leq n}\big) = 0$.

□

(2.1.4) Haarsche Teilräume
n–dimensionale Teilräume $V = span(\{u_1, ..., u_n\})$ heißen *Haarsche Teilräume*, wenn Eigenschaft (2.1.3)$(*)$ gilt für je n paarweise verschiedene Punkte $x_1, x_2, \ldots, x_n \in J$.

□

(2.1.5) Beispiele (für Haarsche Teilräume)
(a) $V = \mathbb{P}_n$, $J \subset \mathbb{R}$ beliebig mit $|J| \geq n$ $(= dim\ \mathbb{P}_n)$. Dabei ist $|J|$ die Anzahl der Punkte in J.
(b) $V = span(\{1, cos(t), sin(t), cos(2t), sin(2t), \ldots, cos(nt), sin(nt)\})$, und
$J \subset [a, b]$ mit $(b - a) < 2\pi$ (und natürlich $|J| \geq 2n + 1$ $(= dim V)$) .

Beweis:

Man prüfe jeweils die Nullstellenbedingung aus (2.1.3)(b) nach.

□

Verwendbarkeit der Interpolation :

Nur bedingt brauchbar ist die Interpolierende als gute Näherung bei großer Zahl der Interpolationspunkte oder bei großem Intervall.

Brauchbar ist die Interpolation im Zusammenhang mit speziellen Werten der Funktion f *bei kleinem Intervall* $[\min x_i, \max x_i]$ und nicht zu hoher Zahl der Interpolationspunkte.

Seien f bzw. die Funktionswerte von f in den Punkten $x_1, x_2, ..., x_n$ gegeben und sei $\tilde{f}$ die Interpolierende der gerade betrachteten Bauart zu den Werten $\{(x_i, f(x_i)) | 1 \leq i \leq n\}$.

Gesucht ist		ersetzt durch	Name der Vorgehensweise
$\int_a^b f(x)dx$	$\doteq$	$\int_a^b \tilde{f}(x)dx$	*Interpolations–Quadratur* , s. (4.1) und (4.2)
$f(x_0)$, $x_0 \neq x_1, \ldots, x_n$	$\doteq$	$\tilde{f}(x_0)$	*Interpolation* in einer Wertetafel, falls $x_0 \in]\min_{1\leq i\leq n} x_i\ ,\ \max_{1\leq i\leq n} x_i[$
			Extrapolation , falls $x_0 \notin]\min_{1\leq i\leq n} x_i, \max_{1\leq i\leq n} x_i[$
$f'(x)$	$\doteq$	$\tilde{f}'(x)$	*numerische Differentiation*
$\min_{x\in[a,b]} f(x)$	$\doteq$	$\min_{x\in[a,b]} \tilde{f}(x)$	*näherungsweise (grobe) Minimierung* , s. (2.2.8)

Die Präzisierung von 'großem' bzw. 'kleinem' Intervall in diesem Zusammenhang ergibt sich aus den *Fehlerabschätzungen* der folgenden Abschnitte: $|f(x) - \tilde{f}(x)| = ?$

2.2 Polynominterpolation

Polynominterpolation ist *gut bzw. brauchbar* bei *wenig Interpolationspunkten* und *kleinem Intervall*. Häufig kann man diese Situation erreichen durch *stückweise Polynominterpolation* (vgl. die zusammengesetzten Interpolations–Quadraturformeln in Abschnitt 4.1).

(2.2.1) Satz (Lagrange-Form des Restglieds/Fehlerabschätzung)
Sei $J = [a,b] \in \mathbb{R}, a \le x_1 < x_2 < ... < x_n \le b, f \in C^n(J)$ und $p \in \mathbb{P}_n : p(x_i) = f(x_i), 1 \le i \le n$. Dann gibt es zu jedem $x \in J$ ein $\xi, \min(x,x_1) < \xi < \max(x,x_n)$, mit

$$f(x) - p(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i) \ .$$

Beweis:

o.E. $x \notin \{x_1, ..., x_n\}$. Dann betrachte die Hilfsfunktion $z \longrightarrow d(z) := f(z) - p(z) - \frac{f(x)-p(x)}{\Phi(x)} \Phi(z)$ mit $\Phi(z) := \prod_{i=1}^{n} (z - x_i)$. d hat $n+1$ verschiedene Nullstellen $x_1, ..., x_n$ und x. Nach dem Mittelwertsatz hat $d^{(n)}$ mindestens eine Nullstelle $\xi \in]\min\{x_1, .., x_n, x\}, \max\{x_1, .., x_n, x\}[$. Da $p^{(n)} \equiv 0$ wegen $deg\ p \le n-1$, und $\Phi^{(n)} \equiv n!$ wegen $\Phi(z) = z^n + ...$. Auflösen nach $f(x) - p(x)$ ergibt die Behauptung.

□

Das Interpolationspolynom hat also eine ähnliche Güte wie das Taylorpolynom

$$q \ : \ q^{(i)}(x_1) = f^{(i)}(x_1) \ , \ i = 0, ..., n-1 \ ;$$

für das ja gilt

$$f(x) - q(x) = \frac{f^{(n)}(\xi)}{n!} (x - x_1)^n \ .$$

Zwischen diesen beiden 'Extremen'

- interpoliere die 0–te bis $(n-1)$–te Ableitung in einem Punkt
- interpoliere die 0–te Ableitung in n verschiedenen Punkten

sind alle 'Zwischen'–Aufgaben lösbar, z.B. gilt

(2.2.2) Hermite–Interpolation
Zu n verschiedenen Punkten $x_1, x_2, ..., x_n \in I = [a,b] \subset \mathbb{R}$ existiert zu $y_1, y_2, ..., y_n, y'_1, y'_2, ..., y'_n \in \mathbb{R}$ genau ein $p \in \mathbb{P}_{2n}$ mit $p(x_i) = y_i \ , \ p'(x_i) = y'_i \ , \ i = 1, ..., n$.

□

Beweisskizze zu (2.2.2) analog zur folgenden *Lagrange*–Darstellung des Interpolationspolynoms:

(2.2.3) Lagrange–Darstellung (des Interpolationspolynom)

$$p(x) = \sum_{j=1}^{n} y_j \ l_j(x) \quad \text{mit} \quad l_j(x) \ = \ \frac{(x - x_1) \cdot ... \cdot (x - x_{j-1})(x - x_{j+1}) \cdot ... \cdot (x - x_n)}{(x_j - x_1) \cdot ... \cdot (x_j - x_{j-1})(x_j - x_{j+1}) \cdot ... \cdot (x_j - x_n)} \ .$$

Denn offensichtlich gilt $l_j \in \mathbb{P}_n \ , \ l_j(x_i) = \delta_{ij} \ , \ 1 \le i,j \le n,$ und damit
$p \in \mathbb{P}_n \ , \ p(x_i) = y_i \ , \ i = 1, ..., n$.

□

Beweisskizze für Hermite–Interpolation (2.2.2):

Finde $l_{j0} \in \mathbb{P}_{2n}$ mit $l_{j0}(x_i) = \delta_{ij}$, $i = 1, ..., n,$ und $l_{j0}'(x_i) = 0$, $i = 1, ..., n$.

Finde $l_{j1} \in \mathbb{P}_{2n}$ mit $l_{j1}(x_i) = 0$, $i = 1, ..., n,$ und $l_{j1}'(x_i) = \delta_{ij}$, $i = 1, ..., n$.

Dann löst offensichtlich

$$p(x) = \sum_{j=1}^{n} y_j \, l_{j0}(x) \; + \; \sum_{j=1}^{n} y_j' \, l_{j1}(x)$$

die Aufgabe (2.2.2).

◇

Wie findet man l_{j0} und l_{j1}?
l_{j1} hat die doppelten Nullstellen x_i , $i = 1, .., j-1, j+1, .., n,$ und die einfache Nullstelle x_j .
Bis auf einen Faktor c – den man so wählt, daß $l_{j1}'(x_j) = 1$ gilt – hat l_{j1} also die Gestalt

$$l_{j1}(x) \; = \; c \, (x - x_1)^2 \cdot ... \cdot (x - x_{j-1})^2 \cdot (x - x_j) \cdot (x - x_{j+1})^2 \cdot ... \cdot (x - x_n)^2 \; .$$

l_{j0} konstruiert man ähnlich.

□

Zum Nachweis der Existenz – und für viele andere prinzipielle theoretische Fragestellungen – ist die *Lagrange–Darstellung des Interpolationspolynoms* geeignet. *Der Nachteil* der Lagrange–Darstellung: eine Erhöhung der Genauigkeit durch zusätzliche Interpolationspunkte erfordert eine völlige Neuberechnung der l_j. Diesen Nachteil vermeidet die

(2.2.4) Newton–Basis (zur Polynom–Interpolation in $\boldsymbol{n+1}$ Punkten $\boldsymbol{x_0, ..., x_n}$)
Die Newton–Basis $\{N_0, ..., N_n\}$ ist gegeben durch $N_0(x) \equiv 1$ und $N_j(x) := (x - x_0) \cdot ... \cdot (x - x_{j-1})$ für $j \geq 1$.

Wie bestimmt man die Koeffizienten c_j in der Linearkombination des Interpolationspolynoms

$$p(x) \; = \; \sum_{j=0}^{n-1} c_j \, N_j \quad , \quad p(x_i) = y_i \; , \; i = 0, ..., n-1 \; ?$$

Dies geschieht über *dividierte Differenzen*, deren *rekursive Definition* sich aus der folgenden Bemerkung sofort ergibt. Zur etwas weniger schwerfälligen Darstellung indiziere ich die Punkte $x_0, ..., x_n$ und die N_j nach ihrem *Grad*.

(2.2.5) Bemerkung
Der Höchstkoeffizient des Interpolationspolynoms in Newtonform genügt der Rekursionsformel

$$c_n(x_0 ... x_n) = \frac{c_{n-1}(x_1 ... x_n) - c_{n-1}(x_0 ... x_{n-1})}{(x_n - x_0)} \; , \; n \geq 1 \; .$$

Beweis:

Zur Verdeutlichung versehe ich die Interpolationpolynome und ihre Höchstkoeffizienten mit ihrem Grad und den Punkten, in denen sie interpolieren:

$$p_{n-1}(x_0...x_{n-1})(.) \text{ interpoliert in } x_0,...x_{n-1};$$

$$p_{n-1}(x_1.....x_n)(.) \text{ interpoliert in } x_1,...x_n;$$

Wie erhält man daraus $p_n(x_0.....x_n)(.)$? Es gilt

$$p_n(x_0.....x_n)(x) = \frac{x_n - x}{x_n - x_0} p_{n-1}(x_0...x_{n-1})(x) + \frac{x - x_0}{x_n - x_0} p_{n-1}(x_1...x_n)(x)$$

denn das rechts stehende Polynom ist vom Grad n und interpoliert wegen $\frac{x_n - x}{x_n - x_0} + \frac{x - x_0}{x_n - x_0} = 1$ in $x_1, ..., x_{n-1}$, und offensichtlich auch in x_0 und x_n. Der Höchstkoeffizient $c_n(x_0...x_n)$ von p ist aber gleich

$$\frac{1}{x_n - x_0} \left(-c_{n-1}(x_0...x_{n-1}) + c_{n-1}(x_1...x_n) \right) .$$

□

Wie in (2.2.5) definiert man rekursiv dividierte Differenzen n–ter Ordnung

(2.2.6) Dividierte Differenzen
Zu Paaren (x_i, y_i) , $i = 0, ..., n$ mit paarweise verschiedenen $\{x_i \ , \ i = 0, ..., n\}$ ist für $n = 0$ die dividierte Differenz 0–ter Ordnung $[x_i] := y_i$, und für $n \geq 1$ die dividierte Differenz n–ter Ordnung

$$[x_0...x_n] := \frac{[x_1...x_n] - [x_0...x_{n-1}]}{(x_n - x_0)} .$$

□

(2.2.7) Newtonform des Interpolationspolynoms
(a) Seien $\{x_0, ..., x_{n-1}\}$ paarweise verschieden. Dann gilt für das Interpolationspolynom $p \in \mathbb{P}_n$ mit $p(x_i) = f(x_i)$, $i = 0, ..., n-1$,

$$p(x) \ = \ \sum_{j=0}^{n-1} \ [x_0...x_j] \ \prod_{i=0}^{j-1} (x - x_i) .$$

Für $x \notin \{x_0, ..., x_{n-1}\}$ gilt die Fehlerdarstellung

$$f(x) - p(x) \ = \ [x_0...x_{n-1}x] \prod_{i=0}^{n-1} (x - x_i) .$$

(b) Seien $\{x_0, ..., x_n\}$ paarweise verschieden, und $f \in C^n\big([\min_{0 \leq i \leq n} x_i \, , \, \max_{0 \leq i \leq n} x_i \,]\big)$. Dann existiert $\xi \in \ ] \min_{0 \leq i \leq n} x_i \, , \, \max_{0 \leq i \leq n} x_i \, [$, so daß der verallgemeinerte Mittelwertsatz gilt

$$[x_0...x_n] = \frac{f^{(n)}(\xi)}{n!} .$$

Beweis:

Die Darstellung von p folgt wegen (2.2.5) durch Induktion.
Sei $p \in \mathbb{P}_n$ mit $p(x_i) = f(x_i)\ ,\ i = 0, ..., n-1$. Dann gilt nach der Newtonform des Interpolationspolynoms für den zusätzlichen Interpolationspunkt x

$$p(x) + [x_0 ... x_{n-1} x] \prod_{i=0}^{n-1} (x - x_i) = f(x)\ .$$

Für $x = x_n$ gilt andererseits nach (2.2.1)

$$f(x_n) - p(x_n) \ = \ \frac{f^{(n)}(\xi)}{n!} \prod_{i=0}^{n-1} (x_n - x_i)\ .$$

□

Wo kann man Hermite-Interpolation gebrauchen? *Eine* Anwendung ist die sog. *eindimensionale Suche* als Teilaufgabe von Minimierungsalgorithmen für $f : \mathbb{R}^k \longrightarrow \mathbb{R}$:

Gesucht ist $x^* \in \mathbb{R}^k$ mit $f(x^*) = \min\limits_{x \in \mathbb{R}^k} f(x)$.

Gegeben $x_{alt} \in \mathbb{R}^k$ und eine Abstiegsrichtung $s_{alt} \in \mathbb{R}^k$ für $f \in C^1(\mathbb{R}^k)$, d.h. es gilt $\lim\limits_{h \to +0} \big(f(x_{alt} + h s_{alt}) - f(x_{alt})\big)/h < 0$.
Gesucht $x_{neu} = x_{alt} + t^* s_{alt}$ mit $f(x_{alt} + t^* s_{alt}) = \min\limits_{t>0} f(x_{alt} + t\, s_{alt})$.

(2.2.8) Näherungsweise Minimierung
Gegeben $\varphi :]-\varepsilon, \infty[\longrightarrow \mathbb{R}\ ,\ \varepsilon > 0\ ,$ und $\varphi \in C^1\,(]-\varepsilon, \infty[)$ mit $\varphi'(0) < 0$.
Gesucht $t^* > 0$ (Existenz vorausgesetzt) mit

$$\varphi(t^*) = \min_{t>0} \varphi(t)\ .$$

1. Schritt: Gewählt sei $\Delta_1 > 0$. Finde $p \in \mathbb{P}_3$ mit

$$p(0) = \varphi(0)\ ,\ p'(0) = \varphi'(0)\ ,\ p(\Delta_1) = \varphi(\Delta_1)\ ;$$

$$\textit{bestimme}\ \ \Delta_2\ \ \textit{mit}\ p(\Delta_2) \ = \ \min_{\Delta>0} p(\Delta)\ (\text{ d.h. }\ p'(\Delta_2) \ = \ 0)\ .$$

2. Schritt: Finde $p \in \mathbb{P}_4$ mit

$$p(0) = \varphi(0)\ ,\ p'(0) = \varphi'(0)\ ,\ p(\Delta_1) = \varphi(\Delta_1)\ ,\ p(\Delta_2) = \varphi(\Delta_2)\ ;$$

$$\textit{bestimme}\ \ \Delta^* \in\ ]0, \Delta_1[\ \textit{mit}\ \ p(\Delta^*) \ = \ \min_{0<\Delta<\Delta_1} p(\Delta)\ (\textit{d.h. } p'(\Delta_2) \ = \ 0)\ .$$

Wiederhole 2. Schritt mit $\Delta_1 := \Delta_2\ ,\ \Delta_2 := \Delta^*$, bis ein geeignetes Abbruchkriterium (z.B. $\varphi(\Delta^*) < \varphi(0) + \beta\varphi'(0)\Delta^*$ mit einer Konstanten $0 < \beta < 1$) erfüllt ist.

□

2.3 Splines: Interpolation und andere Anwendungen

Splines bestehen aus Polynomstücken, die glatt aneinandergeheftet sind.

(2.3.1) Definition (Polynomsplines)
Sei $m \in \mathbb{N}$, $I = [a,b]$, $\underline{\tau} = (\tau_0, \tau_1, ..., \tau_{N+1})$ eine Zerlegung von I : $a = \tau_0 < \tau_1 < ... < \tau_{N+1} = b$.
Dann heißt $s : I \longrightarrow \mathbb{R}$ *(Polynom–)Spline* der *Ordnung* m mit den *Knoten* $\{\tau_i \mid 0 \le i \le N+1\}$, wenn gilt

(a) $s\Big|_{]\tau_{j-1},\tau_j[} \in \mathbb{P}_m \ , \ 1 \le j \le N+1;$

(b) $s \in C^{m-2}(I)$.

Dabei ist (für $m = 1$) $C^{-1}(I)$ die Menge der stückweise (rechtsseitig) stetigen Funktionen.

$\mathcal{S}_m(\underline{\tau})$ bezeichnet die Menge aller Splines der Ordnung m zum Knotenvektor $\underline{\tau}$.

□

Splines der Ordnung m sind für

$m = 1$: *Treppenfunktionen* mit Sprungstellen in den Knoten;
$m = 2$: *Polygonzüge* mit Knicken in den Knoten;
$m = 3$: *quadratische Splines* $\in C^1$;
$m = 4$: *kubische Splines* $\in C^2$; *die Splines* schlechthin, vgl. (2.3.2);
$m = 6$: *quintische Splines* $\in C^4$.

Motivation für kubische Splines resp. Splines gerader Ordnung ist die folgende

(2.3.2) Minimaleigenschaft interpolierender Splines (gerader Ordnung)
Sei $m = 2k$ und $\underline{\tau} = (\tau_0, \tau_1, ..., \tau_{N+1})$. Dann minimiert $s \in \mathcal{S}_m(\underline{\tau})$ mit

$$s^{(j)}(\tau_0) = 0 \ , \ s^{(j)}(\tau_{N+1}) = 0 \ , \ j = k, .., 2k-2 \ ; \ \text{(sog. } \textit{natürliche Randbedingungen}\text{)}$$

und den Interpolationsbedingungen

$$s(\tau_j) = y_j \ , \ j = 0, .., N+1 \ ;$$

unter allen $C^k([\tau_0, \tau_{N+1}])$–Funktionen f, die ebenfalls die Interpolationsbedingungen $f(\tau_j) = y_j \ , \ j = 0, .., N+1$; erfüllen, das Funktional

$$\int_{\tau_0}^{\tau_{N+1}} |f^{(k)}(x)|^2 dx \ \longrightarrow \ min!$$

□

Speziell der kubische interpolierende Spline (mit natürlichen Randbedingungen – oder auch mit allgemeinen Randbedingungen, s. Übungen) minimiert also

$$\int_{\tau_0}^{\tau_{N+1}} |f''(x)|^2 dx \ \longrightarrow \ min!$$

welches für kleines oder kaum variierendes $|f'|$ eine gute Näherung für die mittlere quadratische Krümmung $\int\limits_{\tau_0}^{\tau_{N+1}} \Big(\frac{f''(x)}{\left[1+f'(x)^2\right]^{3/2}}\Big)^2 \, dx$ ist. Aus diesem Grund wurden früher elastische

Lineale, sog. *splines*, im Schiffsbau zum Entwerfen glatter Schiffsformen benutzt. Diese Funktion übernehmen heutzutage (mathematische) Splinekurven z.B. beim Karosserieentwurf in der Kraftfahrzeugindustrie (die Historie dieser Entwicklung wird von *Bézier* – einem Mathematiker beim Kfz-Hersteller *Renault* – beschrieben in dem Buch von *Farin, G.E.*: Kurven und Flächen im Computer Aided Geometric Design : eine praktische Einführung. Vieweg Verlag 1994, ISBN 3-528-16542-1).

Beweis der Minimaleigenschaft (2.3.2) für $k = 2$:

Wir zeigen die Orthogonalitätseigenschaft

$$(2.3.2') \qquad \int_{\tau_0}^{\tau_{N+1}} (f-s)''s''dx \;=\; 0.$$

Daraus folgt (2.3.2), denn:

$$\begin{aligned}\int (f'')^2dx - \int (s'')^2dx &= \int (f''-s'')(f''+s'')dx \\ &= \int (f''-s'')(f''-s'')dx \;, \quad \text{da nach } (2.3.2') \int (f''-s'')s''dx = 0 \;; \\ &> 0 \;\text{ für } f'' \not\equiv s''.\end{aligned}$$

Aus $f'' = s''$ f.ü. folgt aber $f = s$ auf Grund der Randbedingungen $f(\tau_0) = s(\tau_0)$ und $f(\tau_{N+1}) = s(\tau_{N+1})$.

◇

Nachweis von (2.3.2'): Durch partielle Integration folgt

$$\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (f-s)''s''dx \;=\; \sum_{i=0}^{N} (f-s)'s''\Big|_{\tau_i}^{\tau_{i+1}} \;-\; \sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (f-s)'s^{(3)}dx$$

$$= -(f-s)'s''\Big|_{\tau_0+0} + \sum_{i=1}^{N} (f-s)'s''\Big|_{\tau_i+0}^{\tau_i-0} + (f-s)'s''\Big|^{\tau_{N+1}-0} \;+\; \sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (f-s)'s^{(3)}dx$$

Der 1. und 3. Summand verschwindet auf Grund der natürlichen Randbedingungen, der 2. Summand auf Grund der Stetigkeit von $(f-s)'s''$. Nochmalige partielle Integration ergibt wegen $s^{(3)}\big|_{]\tau_i,\tau_{i+1}[} = const$ und $s^{(4)} \equiv 0$ f.ü.

$$= \sum_{i=0}^{N} s^{(3)}(\frac{\tau_i+\tau_{i+1}}{2})(f-s)\Big|_{\tau_i}^{\tau_{i+1}} \;=\; 0$$

auf Grund der Interpolationsbedingungen $(f-s)(\tau_i) = 0, i = 0, ..., N+1$.

□

Mit Hilfe der sog. *abgeschnittenen Potenzen* ,

$$(x-\tau_i)_+{}^{m-1} \;:=\; \begin{cases} 0 & , \quad x < \tau_i \\ (x-\tau_i)^{m-1} & , \quad \tau_i \le x \end{cases} \;,$$

erhält man eine Basis von $\mathcal{S}_m(\underline{\tau})$:

(2.3.3) Satz
Sei $\underline{\tau} = (\tau_0, \tau_1, ..., \tau_{N+1})$. Dann hat $\mathcal{S}_m(\underline{\tau})$ die Basis

$$(x-\tau_0)^j \ , \ j = 0, .., m-1 \ ; \quad (x-\tau_i)_+^{m-1} \ , \ i = 1, ..., N \ .$$

Insbesondere gilt

$$dim \ \mathcal{S}_m(\underline{\tau}) = m + N \, .$$

Die Dimension ist also die *Splineordnung + Anzahl der inneren Knoten* .
Beweis:

Man zeigt durch Induktion für $k = 1, ..., N+1$, daß für $s \in \mathcal{S}_m(\underline{\tau})$ die folgende Darstellung gilt

$$s(x) = \sum_{j=1}^{m} a_j (x-\tau_0)^{j-1} + \sum_{i=1}^{k-1} c_i (x-\tau_i)_+^{m-1} \ , \ \tau_0 \le x < \tau_k \ .$$

Für $k = 1$ ist dies klar wegen $s\Big|_{[\tau_0, \tau_1[} \in \mathbb{P}_m$. Sei diese Darstellung richtig für $\tau_0 \le x < \tau_k$. Setze

$$s_k(x) = \sum_{j=1}^{m} a_j (x-\tau_0)^{j-1} + \sum_{i=1}^{k-1} c_i (x-\tau_i)_+^{m-1} \ , \ \tau_0 \le x < \infty \ \text{ mit } s_k\Big|_{[\tau_{k-1}, \infty[} \in \mathbb{P}_m .$$

Taylorentwicklung von $s - s_k$ um τ_k ergibt wegen $(s-s_k)^{(l)}(\tau_k + 0) \equiv 0$, $l \ge m$,

$$(s-s_k)(x) = \sum_{l=0}^{m-1} \frac{(s-s_k)^{(l)}(\tau_k+0)}{l!}(x-\tau_k)^l \ , \ \tau_k \le x < \tau_{k+1} \ .$$

Wegen $s \in C^{m-2}$ gilt

$$(s-s_k)^{(l)}(\tau_k+0) \ = \ s^{(l)}(\tau_k+0) - s^{(l)}(\tau_k-0) \ = \ 0 \ , \ l \le m-2 \ .$$

Damit folgt für $\tau_k \le x < \tau_{k+1}$ mit $c_k := \dfrac{s^{(m-1)}(\tau_k+0) - s^{(m-1)}(\tau_k-0)}{(m-1)!}$

$$s(x) = \begin{cases} s_k(x) & \text{für} \quad x < \tau_k \\ \\ s_k(x) + c_k(x-\tau_k)^{m-1} & \text{für} \quad \tau_k \le x < \tau_{k+1} \ , \end{cases}$$

$$= \quad s_k(x) + c_k(x-\tau_k)_+^{m-1} \quad \text{für} \quad \tau_0 \le x < \tau_{k+1} \ .$$

Daß dieses Erzeugendensystem linear unabhängig ist, zeigt man ebenso Intervallweise von links nach rechts.

□

Für $f : D \subset \mathbb{R}^k \longrightarrow \mathbb{R}$ ist der Träger (oder *support*)

$$supp(f) = cl(\{x \in D \mid f(x) \neq 0\}).$$

Die im Satz (2.3.3) angegebene Basis hat den Nachteil 'nichtlokale' Träger zu haben:

$$supp\big((x-\tau_0)^j\big) \cap [\tau_0, \tau_{N+1}] = [\tau_0, \tau_{N+1}] \ , \ j = 0, .., m-1 \ ;$$

$$supp\big((x-\tau_i)_+^{m-1}\big) \cap [\tau_0, \tau_{N+1}] = [\tau_i, \tau_{N+1}], \ i = 1, ..., N \ .$$

Spline-Basen mit lokalem Träger – B-Splines

(2.3.4) B-Spline m–ter Ordnung

Der B-Spline m–ter Ordnung zu den $m+1$ Knoten $\tau_0 < \tau_1 < ... < \tau_m$, $B_m(\tau_0, ..., \tau_m)$, ist definiert für $m = 1$ durch

$$B_1(\tau_0, \tau_1)(x) = \begin{cases} 0 & , \quad x < \tau_0 \\ 1 & , \quad \tau_0 \leq x < \tau_1 \\ 0 & , \quad \tau_1 \leq x \end{cases} ;$$

und rekursiv für $m > 1$ durch

$$B_m(\tau_0, ..., \tau_m)(x) = \frac{x - \tau_0}{\tau_{m-1} - \tau_0} B_{m-1}(\tau_0, ..., \tau_{m-1})(x) + \frac{\tau_m - x}{\tau_m - \tau_1} B_{m-1}(\tau_1, ..., \tau_m)(x) \ .$$

□

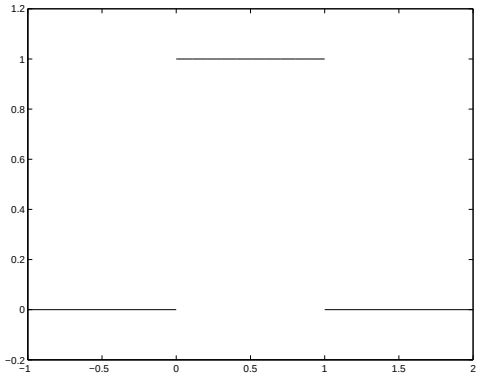

Abbildung 2.1: B–Spline 1. Ordnung

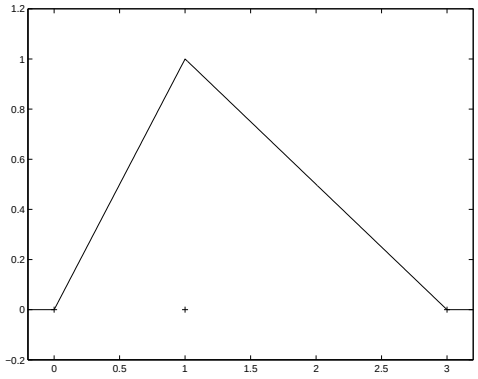

Abbildung 2.2: linearer B–Spline, d.h. 2. Ordnung – nichtäquidistante Knoten

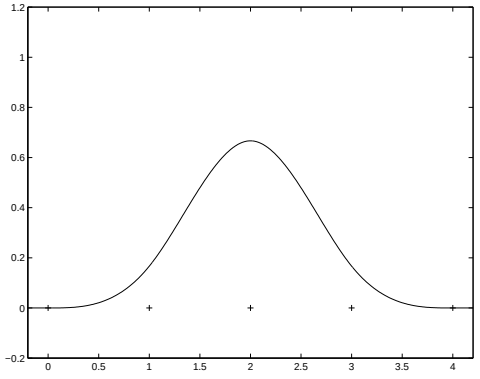

Abbildung 2.3: kubischer B–Spline, d.h. 4. Ordnung – äquidistante Knoten

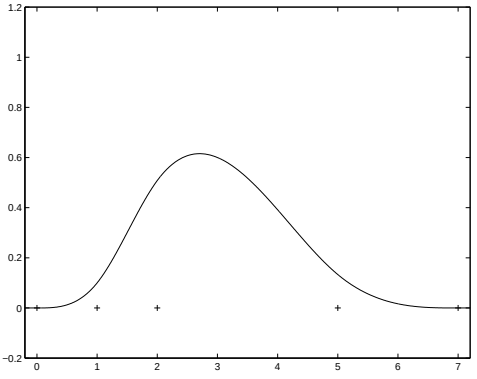

Abbildung 2.4: kubischer B–Spline, d.h. 4. Ordnung – nichtäquidistante Knoten

Diese rekursive Definition wird veranschaulicht durch die Abbildungen 2.1– 2.4.
Für $m = 3$ ist die C^1–Glattheit bei τ_0 und τ_3 klar, da auf Grund des Faktors $\frac{x - \tau_0}{\tau_2 - \tau_0}$ resp.

$\frac{\tau_3 - x}{\tau_3 - \tau_1}$ sowohl τ_0 als auch τ_3 doppelte Nullstelle von $B_3(\tau_0, .., \tau_3)$ ist. Für τ_1 und τ_2 hat man das 'Wunder' der C^1–Glattheit – wie *C. de Boor* sagte (vgl. auch de Boor, C.: A practical guide to Splines, Springer Verlag 1978).
Um nicht durch fehlende Knoten für 'Randindizes' in der Formulierung beeinträchtigt zu sein, setze ich für die folgende Bemerkung einen Knotenvektor $(\tau_j \mid j \in \mathbb{Z})$ voraus:

(2.3.5) Eigenschaften der B–Splines
Sei $\underline{\tau} := (\tau_j \mid j \in \mathbb{Z})$ mit $\tau_j < \tau_{j+1},\ j \in \mathbb{Z}$. Dann gilt für $m \in \mathbb{N}$

(a) $B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})$ ist Spline der Ordnung m zum Knotenvektor $\underline{\tau}$;

(b) $supp(B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})) = [\tau_i, \tau_{i+m}]$;

(c) $\sum_{i=l-m+1}^{r-1} B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})(x) = 1$, $\tau_l \le x < \tau_r$, d.h. auf dem Intervall $[\tau_l, \tau_r[$ bilden die $\left\{B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}}), l - m + 1 \le i < r\right\}$ eine sog. *'Zerlegung der Eins'.*

(d) $\left\{B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})\ ,\ l - m + 1 \le i < r\right\}$ bilden eine Basis von $\mathcal{S}_m([\tau_l, \tau_r])$ zur Zerlegung $\tau_l < ... < \tau_r$.

Beweis:

(a) Klar ist nach der rekursiven Definition der B–Splines die Eigenschaft

$$B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})\big|_{]\tau_j, \tau_{j+1}[} \in \mathbb{P}_m \ .$$

Zum Nachweis der C^{m-2}–Glattheit zeigt man die Identität

$$B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})(x) = (\tau_{i+m} - \tau_i)[\tau_i ... \tau_{i+m}](\cdot - x)_+^{m-1} \ .$$

Dabei wird auf der rechten Seite die dividierte Differenz m-ter Ordnung der Funktion $t \longmapsto (t - x)_+^{m-1}$ bzgl. der Punkte $\tau_i, ..., \tau_{i+m}$ gebildet. Daß die rechte Seite als Linearkombination der Funktionen $\{(\tau_j - x)_+^{m-1} \mid i \le j \le i + m\}$ eine C^{m-2}–Funktion ist, wissen wir bereits.

Zum Nachweis obiger Identität zeigt man, daß die rechte Seite derselben Rekursionsformel genügt wie (per def.) die B–Splines. Für $m = 1$ ist die Identität klar für $x \neq \tau_0, \tau_1$. Für $m \ge 2$ gilt diese Identität auch für $x = \tau_0, ..., \tau_m$ auf Grund der Stetigkeit von linker und rechter Seite. ◇

Bemerkung: Häufig wird die rechte Seite als Definition für B–Splines genommen.

(b) Durch Rekursion bzgl. m folgt

$$B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})(x) > 0 \quad \text{für } \tau_i < x < \tau_{i+m} \ ;$$

$$B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})(x) = 0 \quad \text{für } x < \tau_i$$
$$B_{\mathrm{m}}(\tau_{\mathrm{i}} ... \tau_{\mathrm{i+m}})(x) = 0 \quad \text{für } \tau_{i+m} < x \ .$$

◇

(*c*) und (*d*) zeigt man rekursiv wie (*b*). □

Wann kann man interpolieren? Offensichtlich unmöglich ist die Interpolation für $\tau_i \leq x_k < x_{k+1} < ... < x_{k+m} < \tau_{i+1}$. Entscheidend für die Interpolationsmöglichkeit ist ein 'wechselseitiges Trennen' von Interpolationspunkten und Splineknoten.

(2.3.6) Satz (Schoenberg-Witney)
Gegeben seien Knoten $(\tau_j \mid j = l, ..., r+m)$, $\tau_j < \tau_{j+1}$, und Interpolationspunkte $(x_j \mid j = l, ..., r+m)$, $x_j < x_{j+1}$. Dann gilt

$$det\left(\left[B_{\mathrm{m}}(\tau_{\mathrm{j}}...\tau_{\mathrm{j+m}})(x_i)\right]_{l \leq i,j \leq r}\right) \neq 0,$$

falls kein Diagonalelement verschwindet, d.h.

$$B_{\mathrm{m}}(\tau_{\mathrm{i}}...\tau_{\mathrm{i+m}})(x_i) \neq 0 \ , \ i = l, ..., r.$$

□

Für $m \geq 2$, d.h. mindestens stetige Splines, bedeutet dies

$$\tau_i < x_i < \tau_{i+m} \ .$$

Diesen Satz kann man zeigen durch Rekursion bezüglich der Splineordnung (siehe z.B. *Hämmerlin/Hoffmann*).
Statt für Splines auf $[a, b] = [\tau_0, \tau_{N+1}]$ zusätzliche Funktionswerte zu interpolieren wie im Satz von Schoenberg und Witney, kann man in den Knoten interpolieren und zusätzliche Bedingungen in den Randpunkten stellen (ich betrachte nur kubische Splines – es geht auch allgemeiner):

(2.3.7) Kubische Splines mit Randbedingungen
Gesucht ist zu $\underline{\tau} : \tau_0 < \tau_1 < ... < \tau_{N+1}$ ein Spline $s \in \mathcal{S}_4(\underline{\tau})$, der eine gegeben Funktion f in den Knoten interpoliert:

$$s(\tau_i) = f(\tau_i) \ , \ 0 \leq i \leq N+1 \ ;$$

und

(a) $s'(\tau_0) = f'(\tau_0)$, $s'(\tau_{N+1}) = f'(\tau_{N+1})$ (*allgemeine Randbedingungen* für $f \in C^1$)

(b) $s''(\tau_0) = 0$, $s''(\tau_{N+1}) = 0$ *(natürliche Randbedingungen)*

(c) $s'(\tau_0) = s'(\tau_{N+1})$; $s''(\tau_0) = s''(\tau_{N+1})$ (*periodischeRandbedingungen*)
für periodisches f mit $f(\tau_0) = f(\tau_{N+1})$.

□

Die Existenz wird nicht durch den Satz von Schoenberg–Witney abgedeckt. Diese interpolierenden kubischen Splines mit Randbedingungen existieren ebenfalls – ihre Existenz muß aber gesondert gezeigt werden (vgl. Übungen).

Freiformkurven mittels B–Splines

(2.3.8) B–Spline–Kurven
Sei $m \geq 2$ und $(\tau_j : j = l-m+1, ..., r+m-1)$ ein Knotenvektor. Zu $d_j \in \mathbb{R}^k$, $j = l-m+1, ..., r-1$, heißt

$$[\tau_l, \tau_r] \ni x \longmapsto \sum_{j=l-m+1}^{r-1} B_{\mathrm{m}}(\tau_{\mathrm{j}}...\tau_{\mathrm{j+m}})(x)\ d_j \ \in \ \mathbb{R}^k$$

B–Spline–Kurve der Ordnung m mit den Knoten (τ_j) und den *Kontrollpunkten* (d_j). □

Bei solchen Kurven sind für $m \geq 3$ also a priori Knicke ausgeschlossen. Man kann aber der rekursiven Definition der B-Splines auch bei zusammenfallenden Knoten, z.B. $\tau_0 < \tau_1 = \tau_2 < \tau_3$ für $m = 3$, vgl. Abb. 2.6, einen Sinn geben, wenn man nur die Division durch 0 vermeidet:

(2.3.4') B-Splines (mit mehrfachen Knoten)
Der B-Spline m–ter Ordnung zu den Knoten $\tau_0 \leq \tau_1 \leq ... \leq \tau_m$ ist definiert für $m = 1$ durch

$$B_1(\tau_0, \tau_1)(x) = \begin{cases} 1 & , \quad \tau_0 \leq x < \tau_1 \\ 0 & , \quad sonst \end{cases} ;$$

(d.h. $B_1 \equiv 0$ für $\tau_0 = \tau_1$) und mit

$$\omega_{j,\mu}(x) := \begin{cases} \dfrac{x - \tau_j}{\tau_{j+\mu} - \tau_j} & , \quad falls \ \tau_j < \tau_{j+\mu} \\ \\ 0 & , \quad sonst \end{cases}$$

rekursiv für $m > 1$ durch

$$B_m(\tau_0, ..., \tau_m)(x) = \omega_{0,m-1}(x) B_{m-1}(\tau_0, ..., \tau_{m-1})(x) + \big(1 - \omega_{1,m-1}(x)\big) B_{m-1}(\tau_1, ..., \tau_m)(x) \ .$$

□

Offensichtlich erhält man für $\tau_0 < \tau_{m-1}$ und $\tau_1 < \tau_m$ die früheren Vorfaktoren
$\dfrac{x - \tau_j}{\tau_{j+m-1} - \tau_j}\Big|_{j=0} = \dfrac{x - \tau_0}{\tau_{m-1} - \tau_0}$ und $1 - \dfrac{x - \tau_j}{\tau_{j+m-1} - \tau_j}\Big|_{j=1} = \dfrac{\tau_m - x}{\tau_m - \tau_1}$.
Genau so offensichtlich ist $B_m(\tau_j, ..., \tau_j + m) \not\equiv 0$, falls

$$\tau_j < \tau_{j+m} \ ,$$

ist – *was wir von jetzt an immer voraussetzen.* Durch Rekursion bzgl. der Splineordnung kann man zeigen

(2.3.9) Satz
Fallen k der Knoten $\tau_j, ..., \tau_{j+m}$ zusammen, $1 \leq k \leq m$, so sinkt die Differenzierbarkeit von $B_m(\tau_j, ..., \tau_{j+m})$ in diesem k–fachen Knoten auf C^{m-1-k}.
□

Für einfache Knoten, i.e. $k = 1$, erhält man wieder die C^{m-2}–Glattheit, für einen m–fachen Knoten die C^{-1}–'Glattheit'.
Ich veranschauliche für Splineordnung $m = 3$ die möglichen Fälle mehrfacher Knoten in den Abbildungen 2.5 – 2.8.

Bemerkung (erweiterter Knotenvektor)
Seien N innere (höchstens $(m-1)$–fache) Knoten $\tau_0 < \tau_1 \leq \tau_2 \leq ... \leq \tau_N < \tau_{N+1}$ gegeben. Dann benötigt man nach (2.3.5) für die B–Splines als Basis von $\mathcal{S}_m([\tau_0, \tau_{N+1}])$ zusätzliche Knoten

$$\tau_{-m+1} \leq \tau_{-m+2} \leq ... \leq \tau_0 \ ; \ \tau_{N+1} \leq \tau_{N+2} \leq ... \leq \tau_{N+m} \ .$$

Häufig 'erweitert' man den gegebenen Knotenvektor um die m–fachen Knoten τ_0 und τ_{N+1}

$$\underbrace{\tau_0 = \tau_0 = ... = \tau_0}_{\text{m-fach}} < \tau_1 \leq \tau_2 \leq ... \leq \tau_N < \underbrace{\tau_{N+1} = \tau_{N+2} = ... = \tau_{N+m}}_{\text{m-fach}} \ .$$

□

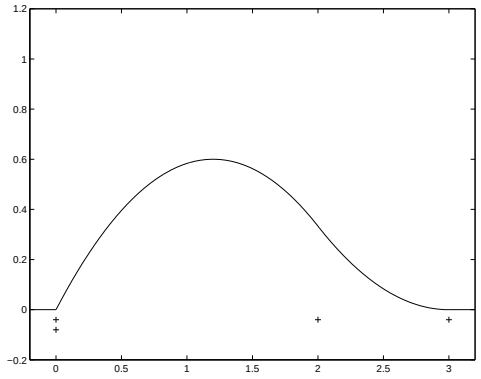

Abbildung 2.5: quadratischer B–Spline, 2–facher Randknoten

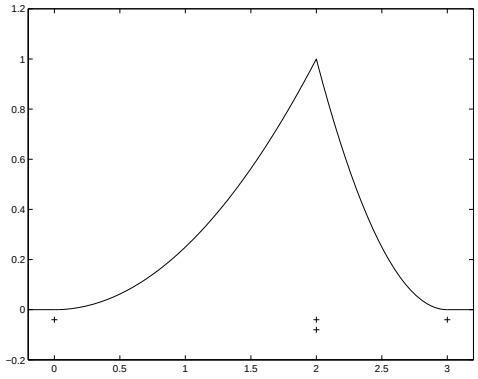

Abbildung 2.6: quadratischer B–Spline, 2–facher innerer Knoten

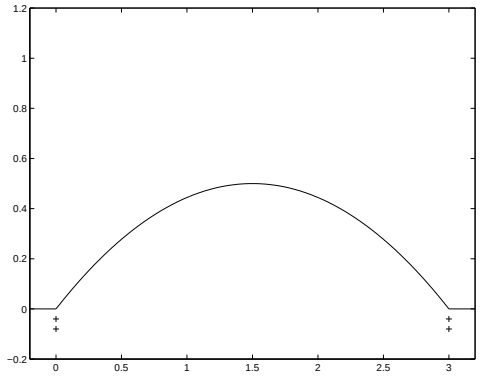

Abbildung 2.7: quadratischer B–Spline, zwei 2–fache (Rand–)Knoten

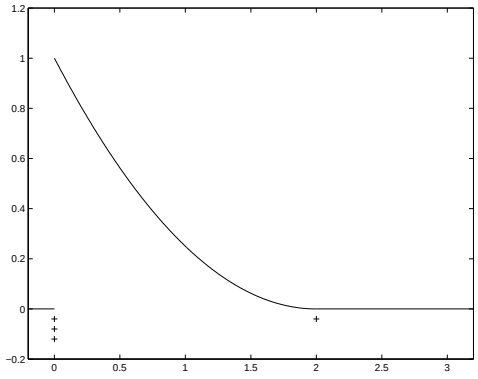

Abbildung 2.8: quadratischer B–Spline, 3–facher (Rand–)Knoten

(2.3.8') B–Spline–Kurven (mit eventuell mehrfachen Knoten)
Sei $m \geq 2$ und $(\tau_j : j = 0, ..., N)$ ein Knotenvektor, wobei $\tau_j \ , \ j = 1, ..., N$ höchstens $(m-1)$–fach, τ_0 und τ_{N+1} m–fach sind. Zu $d_j \in \mathbb{R}^k \ , \ j = -m+1, ..., N$, heißt

$$[\tau_0, \tau_{N+1}[\ni x \longmapsto \sum_{j=-m+1}^{N} B_{\mathrm{m}}(\tau_{\mathrm{j}} ... \tau_{\mathrm{j+m}})(x) \ d_j \in \mathbb{R}^k$$

B–Spline–Kurve der Ordnung m mit den Knoten (τ_j) und den *Kontrollpunkten* (d_j). □

Da die inneren Knoten höchstens $(m-1)$–fach sind, ist obige B–Spline–Kurve (2.3.8') mindestens stetig auf $[\tau_0, \tau_{N+1}]$ (hier habe ich o.E. $\tau_l = \tau_0$ und $\tau_r = \tau_{N+1}$ gesetzt).

Die rekursive Definition der B-Splines erlaubt für Linearkombinationen von B-Splines die Berechnung von Funktionswerten, Werten von Ableitungen, und Werten von Stammfunktionen über eine 'kurze' Rekursion.
Ich veranschauliche dies am folgenden *Algorithmus von de Boor*, der den Algorithmus von *de Casteljau*, der rekursiv Polynomkurven auswertet, auf B-Spline-Kurven überträgt. de Casteljau war beim Kfz-Hersteller *Citroën* etwa zur selben Zeit wie Beziér bei Renault.

(2.3.10) Satz

Sei $s(x) = \sum_{j=l-m+1}^{r-1} B_{\mathrm{m}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m}})(x)\ d_j$. Dann kann man für $\tau_k \le x < \tau_{k+1}$, $l \le k < r$ den Wert $s(x)$ berechnen durch die Rekursion

$$
\begin{aligned}
&\textit{für}\ \ j = k-m+1, ..., k \\
&\qquad d_j^{(1)} := d_j\ ; \\
&\textit{für}\ \ l = 2, ..., m \\
&\qquad \textit{für}\ \ j = k-m+l, ..., k \\
&\qquad\qquad d_j^{(l)} := (1 - \omega_{j,m-l+1}(x))\, d_{j-1}^{(l-1)} + \omega_{j,m-l+1}(x) d_j^{(l-1)}\ ; \\
&s(x) = d_k^{(m)}\ .
\end{aligned}
$$

Beweis:

Man wende für $l = 1, ..., m-1$ die rekursive Definition (2.3.4') an:

$$
\begin{aligned}
&\sum_j d_j B_{\mathrm{m}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m}})(x) \\
&= \sum_j d_j \left[\omega_{j,m-1}(x) B_{\mathrm{m\text{-}1}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m\text{-}1}})(x) + (1 - \omega_{j+1,m-1}(x)) B_{\mathrm{m\text{-}1}}(\tau_{\mathrm{j+1}} \cdots \tau_{\mathrm{j+1+m\text{-}1}})(x)\right] \\
&= \sum_j B_{\mathrm{m\text{-}1}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m\text{-}1}})(x) \left[\omega_{j,m-1}(x) d_j + (1 - \omega_{j,m-1}(x)) d_{j-1}\right] \\
&= \sum_{j=k-m+2}^{k} B_{\mathrm{m\text{-}1}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m\text{-}1}})(x) \underbrace{\left[\omega_{j,m-1}(x) d_j^{(1)} + (1 - \omega_{j,m-1}(x)) d_{j-1}^{(1)}\right]}_{=:\, d_j^{(2)}} \quad \text{für } \tau_k \le x < \tau_{k+1}\ , \\
&= \sum_{j=k-m+3}^{k} B_{\mathrm{m\text{-}2}}(\tau_{\mathrm{j}} \cdots \tau_{\mathrm{j+m\text{-}2}})(x) \underbrace{\left[\omega_{j,m-2}(x) d_j^{(2)} + (1 - \omega_{j,m-2}(x)) d_{j-1}^{(2)}\right]}_{=:\, d_j^{(3)}} \quad \text{für } \tau_k \le x < \tau_{k+1}\ ,
\end{aligned}
$$

Nach insgesamt $(m-1)$–maliger Anwendung der rekursiven Definition der B–Splines erhält man

$$
\begin{aligned}
&= \sum_{j=k}^{k} B_1(\tau_j \tau_{j+1})(x) d_j^{(m)} \qquad \text{für } \tau_k \le x < \tau_{k+1} \\
&= d_k^{(m)} \qquad \text{für } \tau_k \le x < \tau_{k+1}\ .
\end{aligned}
$$

□

Die Rekursion in (2.3.10) kann man durch Umindizierung auch schreiben als

$$
\begin{aligned}
&\textit{für}\ \ j = 1, ..., m \\
&\qquad d_j^{(1)} := d_{k-m+j}\ ; \\
&\textit{für}\ \ l = 2, ..., m \\
&\qquad \textit{für}\ \ j = l, ..., m \\
&\qquad\qquad d_j^{(l)} := (1 - \omega_{k-m+j,m-l+1}(x))\, d_{j-1}^{(l-1)} + \omega_{k-m+j,m-l+1}(x) d_j^{(l-1)}\ ; \\
&s(x) = d_m^{(m)}\ .
\end{aligned}
$$

□

Da die Rekursion in (2.3.10) gerade Konvexkombinationen aus $d_{j-1}^{(l-1)}$ und $d_j^{(l-1)}$ bildet, folgt sofort

(2.3.11) Satz

Sei $s(x) = \sum_{j=l-m+1}^{r} B_{\mathrm{m}}(\tau_{\mathrm{j}} ... \tau_{\mathrm{j+m}})(x)\ d_j$ B–Spline–Kurve. Dann gilt für $\tau_k \leq x < \tau_{k+1}$ mit $l \leq k \leq r$,

$$s(x) \in \textit{konvexe Hülle}\ \left(\{d_{k-m+1}, ..., d_k\}\right) .$$

□

Kapitel 3

Bestapproximation

3.1 Existenz– und Eindeutigkeitsfragen

(3.1.1) Aufgabe
Gegeben ist ein normierter Raum $(U, \|.\|)$, eine Teilmenge $\emptyset \neq V \subset U$, und $u \in U$.

(a) $\inf_{v \in V} \|u - v\| =: dist(u, V)$ heißt *Abstand von u zu V*.

(b) $v_o \in V$ heißt *beste Approximation aus V für u*, wenn gilt

$$\|u - v_o\| = dist(u, V) \ .$$

□

Typische Beispiele sind $(U, \|.\|) = (\mathbb{R}^N, l_p\text{-Norm})$ und $(U, \|.\|) = (C(J), L_p\text{-Norm})$, wobei z.B. $J \subset \mathbb{R}^d$ kompakt ist. V ist häufig n-dimensionaler Teilraum von $C(J)$, z.B. $V = \mathbb{P}_n$ für $J = [a, b] \subset \mathbb{R}$.
Die *Existenzfrage* bedeutet

analytisch: $\inf_{v \in V} \|u - v\| \stackrel{?}{=} \min_{v \in V} \|u - v\|$;

geometrisch: $Rd\ K[u; dist(u, V)] \cap V \stackrel{?}{\neq} \emptyset$.

Bemerkung
Notwendige Bedingung dafür, daß zu jedem $u \in U$ eine beste Approximation existiert, ist die Abgeschlossenheit von V.

Denn ein Randpunkt, der nicht zu V gehört, hat den Abstand 0, der jedoch durch kein $v \in V$ angenommen wird.

Für endlichdimensionale Teilmengen V ist die Bedingung, daß V abgeschlossen ist, auch hinreichend. Es gilt

(3.1.2) Satz
Sei U normierter Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$, $V \subset U$ Teilraum endlicher Dimension. Dann existiert zu jedem $u \in U$ eine beste Approximation aus V für u.

Beweis:

Sei $dim\ V = n$, $V = \{\sum_{i=1}^n \alpha_i v_i\}$ und $\varphi : (\mathbb{K}^n, \|\ \|_\infty) \longrightarrow \mathbb{R}$, $\varphi([\alpha_i]) := \|u - \sum_i \alpha_i v_i\|$. Nach der Dreiecksungleichung gilt

$$|\varphi([\alpha_i]) - \varphi([\beta_i])| \leq \left(\sum_i \|v_i\|\right) \|[\alpha_i] - [\beta_i]\|_\infty;$$

also ist φ stetig.

Weiter gilt mit $c := \min\{\ \|\sum_i \alpha_i v_i\| \ :\ \|[\alpha_i]\|_\infty = 1\ \} > 0$ (die Annahme $c = 0$ ergibt einen Widerspruch zur linearen Unabhängigkeit der $\{v_i\}$)

$$\begin{aligned}\inf\{\ \varphi(\alpha) \mid \alpha \in \mathbb{K}^n\ \} &= \inf\{\ \varphi(\alpha) \mid \alpha \in \mathbb{K}^n : \|\alpha\|_\infty \leq 2\|u\|/c\ \} \\ &= \min\{\ \varphi(\alpha) \mid \alpha \in \mathbb{K}^n : \|\alpha\|_\infty \leq 2\|u\|/c\ \}\ .\end{aligned}$$

Zum Nachweis der ersten Identität beachte man, daß für α mit $\|\alpha\|_\infty > 2\|u\|/c$ gilt

$$\varphi(\alpha) \geq \|\sum_i \alpha_i v_i\| - \|u\| \geq c\|\alpha\|_\infty - \|u\| > \|u\| = \varphi(0)\ .$$

□

Mit identischem Beweis gilt

(3.1.2') Folgerung
Sei $W \subset V$, W abgeschlossen und V endlichdimensionaler Teilraum von $(U, \|.\|)$. Dann existiert zu jedem $u \in U$ eine beste Approximation aus W für u.

□

Generelle *Eindeutigkeit* für beliebige konvexe Teilmengen V gilt, wenn die Norm *strikt konvex* ist:

(3.1.3) Definition (strikt konvex)
$(U, \|.\|)$ heißt *strikt konvex*, falls für alle $x, y \in U$ mit $x \neq y,\ \|x\| = \|y\| = 1$ gilt

$$]x, y[\cap \{z \in U : \|z\| = 1\} = \emptyset\ .$$

Dabei ist $]x, y[:= \{\lambda x + (1-\lambda)y\ :\ \lambda \in\ ]0, 1[\ \}$.

□

Eine offensichtlich äquivalente Bedingung ist

$$\|\frac{1}{2}x + \frac{1}{2}y\| < \frac{1}{2}\|x\| + \frac{1}{2}\|y\| \quad \text{für} \quad x, y \in U,\ x \neq y,\ \|x\| = \|y\|\ .$$

(3.1.4) Satz
Sei $(U, \|.\|)$ strikt konvex, und $V \subset U$ konvex. Dann gibt es zu jedem $u \in U$ höchstens eine beste Approximation aus V.

Beweis:

Seien $v, w \in V$ beste Approximationen aus V für u. Dann gilt

$$\|u - [\lambda v + (1-\lambda)w]\| \leq \lambda\|u - v\| + (1-\lambda)\|u - w\| = dist(u; V) \quad \text{für} \quad \lambda \in \ ]0,1[\ .$$

Da V konvex ist, gilt $[\lambda v + (1-\lambda)w] \in V$ und damit $\|u - [\lambda v + (1-\lambda)w]\| = dist(u; V)$. Wegen U strikt konvex muß notwendig $v = w$ gelten.

□

(3.1.5) Beispiele (strikt konvexer Räume)
(a) $(\mathbb{K}^n, \|.\|_p)$ ist strikt konvex für $1 < p < \infty$.

(b) $(C(J) \cap L_p(J), \|.\|_{L_p(J)})$ ist strikt konvex für $1 < p < \infty$; dabei $J \subset \mathbb{R}^d$ meßbar.

In beiden Fällen ist für $p = 1$ und $p = \infty$ die Norm nicht strikt konvex.

(c) Jede durch ein Skalarprodukt induzierte Norm ist strikt konvex.

□

3.2 Bestimmung linearer Quadratmittel–Approximationen

In diesem Abschnitt sei immer $(U, <\ ,\ >^{1/2})$ ein Skalarproduktraum und $V \subset U$ ein Teilraum, $V = span(\{v_1, ..., v_n\})$ mit $dim\ V = n$ – diese Situation ist die sog. lineare Quadratmittel–Approximation (least squares approximation).

(3.2.1) Beispiele
Sei $J \subset \mathbb{R}^d$ meßbar (meist kompakt), $V = span(\{v_1, ..., v_n\}) \subset C(J) \cap L_2(J)$, $dim V = n$.

(a) Zu $u \in C(J) \cap L_2(J)$ ist $v_o \in V$ gesucht mit

$$\int_J |u(x) - v_o(x)|^2 dx < \int_J |u(x) - v(x)|^2 dx \quad \text{für } v \in V \ , \ v \neq v_o \ ; \ '(kontinuierliche)\ L_2 - (Best)Approximation'$$

(b) Seien $\{x_1, ..., x_N\} =: I \subset J$ paarweise verschiedene Punkte. Zu $u : I \longrightarrow \mathbb{R}$ ist $v_o \in V$ gesucht mit

$$\sum_{i=1}^{N} |u(x_i) - v_o(x_i)|^2 < \sum_{i=1}^{N} |u(x_i) - v(x_i)|^2 \quad \text{für } v \in V \ , \ v \neq v_o \ ; \ 'diskrete\ L_2 - (Best)Approximation'$$

auf $I := \{x_1, ..., x_N\} \subset J$.

(c) Zu $[y_i \in \mathbb{R} \ : \ 1 \leq i \leq N]$ und nicht notwendig verschiedenen Punkten $x_i \in J \ , \ 1 \leq i \leq N$, ist $v_o \in V$ gesucht mit

$$\sum_{i=1}^{N} |y_i - v_o(x_i)|^2 < \sum_{i=1}^{N} |y_i - v(x_i)|^2 \quad \text{für } v \in V \ , \ v \neq v_o \ , \ 'Ausgleichs - (Best)Approximation'$$

auf $I := \{x_1, ..., x_N\} \subset J$.

□

Der folgende Satz gibt eine Charakterisierung der besten Approximation und damit eine Berechnungsmöglichkeit:

(3.2.2) Satz (Charakterisierung der besten linearen Approximation)
Sei $(U, <\ ,\ >^{1/2})$ ein Skalarproduktraum über $\mathbb{K}$, und $V \subset U$ ein Teilraum, $V = span(\{v_1, ..., v_n\}) \subset U$ mit $dim\ V \leq n$ (d.h. $\{v_1, ..., v_n\}$ nicht notwendig linear unabhängig). Dann gilt für $u \in U$

$\sum_{i=1}^n a_i v_i$ ist beste Approximation aus V für u genau dann, wenn gilt

$[a_i]_{1 \leq i \leq n}$ löst das lineare Gleichungssystem $Ga = b$.

Dabei ist $G \equiv [G_{ij}] := [< v_j, v_i >]_{1 \leq i,j \leq n}$ die Gramsche Matrix der Elemente $v_1, ..., v_n$ und

$b \equiv [b_i]_{1 \leq i \leq n} := [< u, v_i >]_{1 \leq i \leq n}$ die rechte Seite.

Beweis:

Durch Ausmultiplizieren erhält man für $w_t = v_o + t(v - v_o)\ ,\ t > 0$,

$$(*) \quad \|u - w_t\|^2 = \|u - v_o\|^2 + t^2 \|v - v_o\|^2 - Re\ < u - v_o, v - v_o > \ .$$

Damit folgt für $v \in V$ und $t > 0, t \to 0$,

$$\|u - w_t\|^2 \geq \|u - v_o\|^2, \text{ genau dann, wenn } -2Re < u - v_o, v - v_o > \ \geq\ 0\ ,\ v \in V\ .$$

Da mit $v - v_o$ auch $-(v - v_o) \in V$ (bzw. für $\mathbb{K} = \mathbb{C}$ auch $e^{i\phi}(v - v_o) \in V$) ist, gilt

$$\|u - v\|^2 \geq \|u - v_o\|^2, \text{ genau dann, wenn } < u - v_o, v - v_o > \ =\ 0\ ,\ v \in V\ .$$

Wegen $v - v_o \in span(\{v_1, ..., v_n\})$ ist also notwendig und hinreichend

$$< u - v_o, v_i > \ =\ 0\ ,\ 1 \leq i \leq n\ .$$

Für $v_o = \sum_{j=1}^n a_j v_j$ ist also notwendig und hinreichend

$$< u, v_i > \ =\ \sum_{j=1}^n a_j < v_j, v_i >\ ,\ 1 \leq i \leq n\ .$$

□

Bemerkung
Falls $\{v_1, ..., v_n\}$ nicht linear unabhängig sind, gibt es viele Lösungen von $Ga = b$. Sie stellen aber alle dasselbe Element $\sum_{j=1}^n a_j v_j \in V$ dar.

□

Will man Satz (3.2.2) auf die Beispiele (3.2.1) anwenden, ergeben sich – abhängig von der jeweiligen Situation – unterschiedliche Probleme bezüglich der Voraussetzungen und Eigenschaften:

(3.2.3) Bemerkungen
(a) Ist $\{v_1, ..., v_n\}$ Orthonormalbasis von V, d.h. es gilt $< v_j, v_i >= \delta_{ij}$, so ist

$$v_0 := \sum_{j=1}^{n} < u, v_j > v_j$$

beste Approximation aus V für u.

(b) Für $I = [a, b] \subset \mathbb{R}$ und $V = span(\{B_m(\tau_j ... \tau_{j+m})\}$ ist G 'dünn' besetzt, denn es gilt $G_{ij} = 0$, $|i - j| \geq m$.

(c) Z.B. ist $\{x \longmapsto x,\ x \longmapsto 1 - x^2, x \longmapsto x^3\}$ nicht linear unabhängig auf $I = \{-1, 0, +1\}$.

□

Um zusätzliche Konditionsverschlechterung zu vermeiden, verwendet man zum Lösen der sog. *Normalgleichungen* $Ga = b$ in (3.2.2) besser die QR-Zerlegeung statt der LR-Zerlegeung.

Es gibt einige Situationen, wo Orthogonalbasen explizit bekannt sind. Beispiele sind

(3.2.4) Beispiele (für Orthonormal-/Orthogonalbasen)
(a) In $U = C([0, 2\pi])$, $< f, g >= \int_0^{2\pi} f(x)g(x)dx$
sind die trigomometrischen Funktionen ein Orthonormalsystem:

$$V = span\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos(x), \frac{1}{\sqrt{\pi}}\sin(x), ..., \frac{1}{\sqrt{\pi}}\cos(nx), \frac{1}{\sqrt{\pi}}\sin(nx)\} \ .$$

(b) In $[0, 2\pi]$ sei $x_k := \frac{2k\pi}{N}$, $k = 0, 1, ..., N - 1$, $N = 2m \in 2\mathbb{N}$.
Dann ist bzgl. dem 'diskreten L_2–Skalarprodukt' (im $\mathbb{R}^N$)

$$< f, g >:= \frac{2\pi}{N} \sum_{k=0}^{N-1} f(x_k)g(x_k)$$

ein Orthonormalsystem gegeben durch

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos(x), ..., \frac{1}{\sqrt{\pi}}\cos((m-1)x), \frac{1}{\sqrt{2\pi}}\cos(mx),$$

$$\frac{1}{\sqrt{\pi}}\sin(x), ..., \frac{1}{\sqrt{\pi}}\sin((m-1)x) \ .$$

(c) In $C([-1, 1])$ bilden die Tschebyschew-Polynome 1. Art

$$T_n(x) := \cos(n(\arccos x)) \ , \ n \in \mathbb{N}_0 \ ;$$

ein Orthogonalsystem bzgl. dem 'gewichteten L_2–Skalarprodukt'

$$< f, g > := \int_{-1}^{+1} \frac{1}{\sqrt{1 - x^2}} f(x)g(x)dx \ .$$

$w(x) := \frac{1}{\sqrt{1 - x^2}}$ ist eine sog. 'Gewichtsfunktion'.

□

3.3 Bestimmung linearer Tschebyschew–Approximationen

In diesem Abschnitt betrachte ich die Beispiele (3.3.1) in der Maximum–Norm.

(3.3.1) Beispiele
Sei $J = [a,b] \subset \mathbb{R}$ kompaktes Intervall, $V = span(\{v_1, ..., v_n\}) \subset C(J)$, $dimV = n$.

(a) Zu $u \in C(J)$ ist $v_o \in V$ gesucht mit

$$\max_{x\in J} |u(x) - v_o(x)| \leq \max_{x\in J} |u(x) - v(x)| \quad \text{für } v \in V \; ; \; '(kontinuierliche)\, Tschebyschew-(Best)Approximation'$$

(b) Seien $\{x_1, ..., x_N\} =: I \subset J$ paarweise verschiedene Punkte. Zu $u : I \longrightarrow \mathbb{R}$ ist $v_o \in V$ gesucht mit

$$\max_{1\leq i\leq N} |u(x_i) - v_o(x_i)| \leq \max_{1\leq i\leq N} |u(x_i) - v(x_i)| \quad \text{für } v \in V \; ; \; 'diskrete\, Tschebyschew-(Best)Approximation'.$$

□

Der folgende Satz gibt eine hinreichende Charakterisierung der besten Approximation und damit eine Berechnungsmöglichkeit.

(3.3.2) Satz (Charakterisierung der Best–Approximation in Haarschen Teilräumen)
Für $J = [a,b]$ sei $V = span(\{v_1, ..., v_n\}) \subset C(J)$ Haarscher Teilraum der Dimension n und $u \in C(J) \setminus V$. Dann ist $v_0 \in V$ beste Approximation aus V für u in der Maximumnorm, wenn gilt:

Es existieren $n+1$ Punkte $x_0 < x_1 < ... < x_n$ mit

$$|(u - v_0)(x_i)| = \max_{x\in J} |(u - v_0)(x)| \; , \; 0 \leq i \leq n \; ;$$

und

$$sign(u - v_0)(x_{i-1}) = -sign(u - v_0)(x_i) \; , \; 1 \leq i \leq n \; .$$

Beweis:

Durch Widerspruch. Angenommen, es gibt eine bessere Approximation $v \in V$ als v_0. Dann gilt

$$|(u - v)(x_i)| \leq \|u - v\|_\infty < \|u - v_0\|_\infty = |(u - v_0)(x_i)| \; , \; 0 \leq i \leq n \; ,$$

und damit

$$sign\big(u - v_0 - (u - v)\big)(x_i) = sign(u - v_0)(x_i), 0 \leq i \leq n.$$

Wegen $(v - v_0)(x_i) = \big(u - v_0 - (u - v)\big)(x_i)$ folgt, daß

$$sign(v - v_0)(x_{i-1}) = -sign(v - v_0)(x_i) \; , \; 1 \leq i \leq n \; .$$

D.h. $v - v_0$ hat mindestens n verschiedene Nullstellen. Da V Haarsch ist, folgt $v = v_0$ Widerspruch ! .

□

(3.3.3) Beispiel
Die beste Approximation aus $\mathbb{P}_2$ für $x \longmapsto u(x) = x^2 + 1 \in C\big([0,1]\big)$ in der Maximumnorm ist

$$p_0(x) \;=\; x + \frac{7}{8}\;,\; x \in [0,1]\;;\quad \text{es gilt}\quad dist(u, \mathbb{P}_2) \;=\; \frac{1}{8}\;.$$

Beweis:

$$x_0 = 0\;,\; x_2 = 1\;,\; x_1 \in]0,1[\text{ mit } (u-p_0)'(x_1) = 0\;;$$

liefert für die Koeffizienten a, b von $p_0(x) = ax + b$ und die Extremum–Stelle x_1 von $u - p_0$ drei Gleichungen

$$(u-p_0)(0) = -(u-p_0)(x_1)\;;\; (u-p_0)(x_1) = -(u-p_0)(1)\;;\; (u-p_0)'(x_1) = 0\;.$$

Daraus folgt $x_1 = 1/2$, $a = 1$, und $b = 7/8$. Damit gelten auch die hinreichenden Bedingungen von Satz (3.3.2).

□

Die hinreichenden Bedingungen aus Satz (3.3.2) sind auch notwendig – was aber viel aufwändiger zu zeigen ist. Im allgemeinen kann man iterativ näherungsweise eine sog. Alternante $x_0 < x_1 < \ldots < x_n$ mit den Eigenschaften von (3.3.2) bestimmen. Dies macht gerade das Verfahren von *Remez* .

Kapitel 4

Numerische Integration

4.1 Newton-Cotes-Formel

Die meisten Methoden Quadraturformeln (kurz QF'n) zu erzeugen, sind sog. Interpolations–Quadraturformeln:

$$\widetilde{f}(x) \doteq f(x), x \in I \quad \Longrightarrow \quad Q(f) :\equiv \int_I \widetilde{f}(x)dx \doteq \int_I f(x)dx \ .$$

typisches Beispiel: *Sehnentrapezregel*

$\widetilde{f} \in \mathbb{P}_2 \ : \ \widetilde{f}(s_1) = f(s_1)$ und $\widetilde{f}(s_2) = f(s_2)$.

Häufigste Bauart für Quadraturformeln; $Q_{\widehat{I}}(f) = \sum_{i=1}^{m} w_i f(s_i)$

$\{w_i \ : \ 1 \le i \le m\}$ sog. *Gewichte*

$\{s_i \ : \ 1 \le i \le m\} \subset \widehat{I}$ sog. *Stützstellen* oder *Knoten* der Quadraturformel auf dem (Standard–) Intervall $\widehat{I}$. Meist ist $\widehat{I} = [0,1]$ oder $\widehat{I} = [-1,1]$.

Zur Erhöhung der Genauigkeit verwendet man die Quadraturformel nur auf Intervallen kleiner Länge $I_k = [x_{k-1}, x_k]$ und betrachtet mit $h_k = |I_k| = x_k - x_{k-1}$, $Q_{I_k}(f) := h_k Q_{\widehat{I}}(f \circ \varphi_k)$ mit $\varphi_k : \widehat{I} \longrightarrow I_k$ bijektiv, z.B. ist für $\widehat{I} = [0,1] \ni t \longmapsto \varphi_k(t) := x_{k-1} + h_k t \in I_k$.
Für die Zerlegung $I = \bigcup_{k=1}^{N} I_k$ mit $I = [a,b]$, $a = x_0 < x_1 < ... < x_N = b$, $I_k = [x_{k-1}, x_k]$, $k = 1, ..., N$, erhält man als Quadraturformel

(4.1.1) Zusammengesetzte Sehnentrapezregel
Für beliebig verteilte Teilpunkte $\{x_k\}$ ist mit diesen Bezeichnungen

$$ST(f) = \sum_{k=1}^{N} \frac{h_k}{2} \left[f(x_{k-1}) + f(x_k)\right],$$

und für äquidistante Teilpunkte $x_k = a + kh, k = 0, 1, .., N, h := (b-a)/N$,

$$ST(f) = h\left[0.5f(x_0) + f(x_1) + f(x_2) + ... + f(x_{N-1}) + 0.5f(x_N)\right].$$

□

Wie genau ist die Sehnentrapezregel?

(4.1.2) Satz
Sei $f \in C^2([a,b])$. Dann gilt für $a = x_0 < x_1 < ... < x_N = b$ mit $h := \max_{1 \le k \le N} h_k$, $h_k = x_k - x_{k-1}$, die Fehlerabschätzung

$$(a) \qquad |\int_a^b f(x)dx - ST_h(f)| \le \frac{b-a}{12} \|f^{(2)}\|_{\infty,[a,b]}\ h^2 \ ,$$

und für äquidistante Teilpunkte $x_k = a + kh, k = 0, 1, .., N, h := (b-a)/N$, die Fehler*darstellung* mit einem $\xi \in]a,b[$

$$(b) \qquad \int_a^b f(x)dx - ST_h(f) = -\frac{b-a}{12} f^{(2)}(\xi)\ h^2 \ .$$

Beweis:

$$Quadraturfehler \ = \ \sum_{k=1}^{N} \int_{I_k} (f(x) - p_k(x))\, dx$$

$$= \ \sum_{k=1}^{N} \int_{I_k} [x_{k-1}x_k x](f)(x - x_{k-1})(x - x_k)dx.$$

Beachtet man $[x_{k-1}x_k x](f) = \frac{1}{2!} f^{(2)}(\xi(x))$ und $\int_{x_{k-1}}^{x_k} (x - x_{k-1})(x - x_k)dx = -(x_k - x_{k-1})^3/6$, so folgt (a).

Zum Nachweis von (b) beachte man

$$\int_{I_k} [x_{k-1}x_k x](f) \underbrace{(x - x_{k-1})(x - x_k)}_{\le 0} dx = \underbrace{[x_{k-1}x_k \widetilde{\xi_k}](f)}_{\frac{1}{2!} f^{(2)}(\xi_k)} \underbrace{\int_{x_{k-1}}^{x_k} (x - x_{k-1})(x - x_k)dx}_{= -h^3/6}$$

nach dem 2. Mittelwertsatz der Integralrechnung. Damit folgt die Behauptung, da für arithmetische Mittel stetiger Funktionen der Zwischenwertsatz gilt:

$$\sum_{k=1}^{N} \frac{1}{2!} f^{(2)}(\xi_k) \frac{1}{N} = \frac{1}{2!} f^{(2)}(\xi).$$

□

Fehlerabschätzungen sind oft mühsam auszuwerten, und häufig zu pessimistisch. Sie sind jedoch hilfreich bei der Auswahl einer Formel, denn sie geben zumindest die *Konvergenzordnung* an. *Fehlerdarstellungen* sind wichtig für eine adaptive Wahl der Gitterbreite h_k mit $x_k = x_{k-1} + h_k$ in Abhängigkeit einer gewünschten Fehlertoleranz, d.h. (eine eventuell nur näherungsweise Auswertung einer Fehlerdarstellung ergibt eine) $Fehlerschätzung \doteq Toleranz$.

Durch Verwendung von Interpolationspolynomen höherer Approximationsordnung erhält man Quadraturformeln höherer Konvergenzordnung – zumindest für glatte Integranden. Ich beschreibe dies exemplarisch für $m = 3$:

(4.1.3) *Keplersche* Faßregel (Johannes Kepler (1571–1630))
Sei $\widehat{I} = [0,1]$, $s_1 = 0 < s_2 = 1/2 < s_3 = 1$, $p_3 \in \mathbb{P}_3$ mit $p_3(s_i) = f(s_i), i = 1,2,3$. Dann gilt für $Q_{\widehat{I}}(f) \equiv \int_{\widehat{I}} p_3(x)dx$

$$Q_{\widehat{I}}(f) = \frac{1}{6}f(s_1) + \frac{2}{3}f(s_2) + \frac{1}{6}f(s_3) \ .$$

Für $f \in C^4(\widehat{I})$ gilt mit einem $\xi \in]s_1, s_3[$ die Fehlerdarstellung

$$\int_{\widehat{I}} f(x)dx - Q_{\widehat{I}}(f) = -\frac{f^{(4)}(\xi)}{4!\ 120} \ .$$

Beweis:

$p_3 \in \mathbb{P}_3$ mit $p_3(s_i) = f(s_i), i = 1,2,3$, in Lagrangedarstellung ergibt

$$\int_{\widehat{I}} p_3(x)dx = \int_{\widehat{I}} \sum_{i=1}^{3} f(s_i) l_i(x)dx = \sum_{i=1}^{3} f(s_i) \underbrace{\int_{\widehat{I}} l_i(x)dx}_{=: w_i} \ .$$

Es gilt $w_1 \equiv \int_0^1 \frac{(x-s_2)(x-s_3)}{(s_1-s_2)(s_1-s_3)}dx = 1/6$ und analog $w_2 = 2/3, w_3 = 1/6$.

Zur Herleitung der Fehlerdarstellung sei $\sigma \in \widehat{I}$ ein weiterer Interpolationspunkt, $\sigma \neq s_i, i = 1,2,3$, und $p_4 \in \mathbb{P}_4$ mit $p_4(s_i) = f(s_i), i = 1,2,3,\ p_4(\sigma) = f(\sigma)$. Dann gilt

$$p_4(x) = p_3(x) + [s_1 s_2 s_3 \sigma](f)(x-s_1)(x-s_2)(x-s_3),$$

und wegen $\int_{\widehat{I}}(x-s_1)(x-s_2)(x-s_3)dx = 0$ folgt mit $f - p_4$ in Newton-Form

$$\int_{\widehat{I}} f(x)dx - \int_{\widehat{I}} p_3(x)dx = \int_{\widehat{I}} f(x)dx - \int_{\widehat{I}} p_4(x)dx = \int_{\widehat{I}} [s_1 s_2 s_3\, \sigma x](f)(x-s_1)(x-s_2)(x-s_3)(x-\sigma)dx \ .$$

Für C^1–Funktionen f existiert der Grenzwert für $\sigma \to s_2$, wobei Differenzenquotienten durch Differentialquotienten ersetzt werden. Für $\sigma = s_2$ darf man wieder den 2. Mittelwertsatz der Integralrechnung anwenden, und erhält ein $\xi \in]s_1, s_3[$ mit

$$\int_{\widehat{I}} [s_1 s_2 s_3 \sigma x](f)(x-s_1)(x-s_2)(x-s_3)(x-\sigma)dx = \frac{f^{(4)}(\xi)}{4!} \int_{\widehat{I}} (x-s_1)(x-s_2)^2(x-s_3)dx = \frac{f^{(4)}(\xi)}{4!}\frac{-1}{120} \ .$$

□

Bemerkung
Da man die Keplersche Faßregel durch Integration des Interpolationspolynoms $p_3 \in \mathbb{P}_3$ mit $p_3(s_i) = f(s_i), i = 1,2,3$, erhält, gilt damit für die zusammengesetzte Regel, die sog. *Simpson–Regel* (Thomas Simpson (1710-1761)), mit $h = \max_k |I_k|$ für die stückweise $\mathbb{P}_3$–Interpolierende $\widetilde{f}$ zum (mindestens C^3–) Integranden f die

Approximationsordnung für den Integranden $\quad \|f - \widetilde{f}\|_{\infty,I} \quad = \quad O(h^3).$

Eine direkte Folgerung daraus ist eine

Mindest–Konvergenzordnung der Integrale $|\int_I f dx - \int_I \widetilde{f} dx| \quad = \quad |I|\ O(h^3)$.

Erstaunlicherweise ist für C^4–Integranden f die

Konvergenzordnung der Integrale $|\int_I f dx - \int_I \widetilde{f} dx| \quad = \quad O(h^4)$,

also von *höherer Ordnung* als die Approximationsordnung der Integranden. Der Grund dafür sind die Symmetrieeigenschaften der Quadraturformel. Bei symmetrisch zur Intervallmitte verteilten Stützstellen $\{s_i\}$ tritt dieser Effekt immer für ungerade Polynomordnung auf, vgl. (4.1.5). Den Extremfall für diesen Effekt, die Gauß-Quadraturformeln, lernen wir in Abschnitt (4.2) kennen.

(4.1.4) Satz (Fehler der Simpsonregel)
Für die Zerlegung $I = \bigcup_{k=1}^N I_k$ mit $I = [a,b], a = x_0 < x_1 < ... < x_N = b, I_k = [x_{k-1}, x_k], h_k = x_k - x_{k-1}, \widehat{I} = [0,1] \ni t \longmapsto \varphi_k(t) := x_{k-1} + h_k t \in I_k, k = 1, ..., N$, ist mit $s_{2k} := x_k, k = 0, ..., N, s_{2k-1} := (x_{k-1} + x_k)/2, k = 1, ..., N$,

$$Q_{h,Simpson}(f) = \sum_{k=1}^N h_k Q_{Kepler}(f \circ \varphi_k) \ .$$

Für $f \in C^4([a,b])$ gilt mit $h := \max\limits_{1 \le k \le 2N} (s_k - s_{k-1}) \quad \left(= \max\limits_{1 \le k \le N} h_k/2 \right)$ die Fehlerabschätzung

$$|\int_a^b f(x)dx - Q_{h,Simpson}(f)| \le \frac{b-a}{180} \|f^{(4)}\|_{\infty,[a,b]}\ h^4 \ .$$

Beweis:

$$\begin{aligned}
\int_a^b f(x)dx - Q_{h,Simpson}(f) &= \sum_{k=1}^N h_k \left(\int_{\widehat{I}} (f \circ \varphi_k)(t)dt - Q_{Kepler}(f \circ \varphi_k) \right) \\
&\le \sum_{k=1}^N h_k \left| \int_{\widehat{I}} (f \circ \varphi_k)(t)dt - Q_{Kepler}(f \circ \varphi_k) \right| \\
&\le \sum_{k=1}^N h_k \frac{1}{4!\ 120} \left| (f \circ \varphi_k)^{(4)}(\tau_k) \right| \quad \text{mit } \tau_k \in \widehat{I}, \\
&= \sum_{k=1}^N h_k \frac{1}{4!\ 120} |f^{(4)}(\xi_k)| {h_k}^4 \quad \text{mit } \xi_k \in I_k, \\
&\le \frac{1}{3 \cdot 60} \|f^{(4)}\|_{\infty,I} h^4 \sum_{k=1}^N h_k \quad \text{wegen } h_k \le 2h \ .
\end{aligned}$$

□

(4.1.5) Tabelle (einiger Newton–Cotes–Formeln / Sir Isaac Newton (1643–1727) , Roger Cotes (1682–1716))

Pol.-ord. m	s_1	.	.	.	s_m	w_1	.	.	.	w_m	*Name der Regel*	*Konv. Ord.*	*Diff. Ord f*
1	0					1					Rechteck	1	1
1	1/2					1					Mittelpunkt	2	2
2	0	1				1/2	1/2				Sehnentrapez	2	2
3	0	1/2	1			1/6	2/3	1/6			Simpson (1710-1761)	4	4
4	0	1/3	2/3	1		1/8	3/8	3/8	1/8		Newtons 3/8	4	4
5	0	1/4	1/2	3/4	1	7/90	32/90	12/90	32/90	7/90	Milne (1896-1950)	6	6

□

4.2 Gauß-Quadratur

Johann Carl Friedrich Gauß (1777–1855)

Für hochdifferenzierbare Integranden eignen sich Gauß–Quadraturformeln, die auf der Idee beruhen: Bestimme sowohl die Gewichte $\{w_i \,:\, 1 \le i \le m\}$ *als auch die Stützstellen* $\{s_i \,:\, 1 \le i \le m\}$ in der Quadraturformel

$$\int_I w(x)f(x)dx \;\doteq\; \sum_{i=1}^{m} w_i f(s_i)$$

derart, daß die *polynomiale Exaktheitsordnung* maximal ist, d.h. für den Fehler $E(f) := \int_I w(x)f(x)dx - \sum_1^m w_i f(s_i)$ gilt

$$E(p) = 0 \;,\; p \in \mathbb{P}_{m^*} \quad \text{mit} \quad m^* \longrightarrow \textit{maximal} \;!$$

Da die Anzahl der anzupassenden Parameter $2m$ ist, erwartet man $m^* = 2m$ – was der nächste Satz bestätigt. Ich möchte kurz motivieren, wieso im folgenden Satz Orthogonalpolynome ins Spiel kommen.

Orthogonalität (als notwendige Bedingung)
Für $Q(f) = \sum_{i=1}^m w_i f(s_i)$ sei $E(p) = 0 \;,\; p \in \mathbb{P}_{2m}$. Dann gilt für $p^\star(x) := (x - s_1) \cdot \ldots \cdot (x - s_m)$ die Orthogonalitätsbedingung

$$\int_I w(x)p^\star(x)q(x)dx \;=\; 0 \;,\; q \in \mathbb{P}_m \;.$$

Denn wegen $deg(p^\star \cdot q) \le 2m - 1$ ist $p^\star \cdot q \in \mathbb{P}_{2m}$. Wegen der Exaktheitsordnung $2m$ folgt

$$\int_I w(x)p^\star(x)q(x)dx \;=\; Q(p^\star \cdot q) = \sum_{i=1}^{m} w_i \underbrace{p^\star(s_i)}_{=0} q(s_i) \;=\; 0 \;.$$

◇

Im folgenden Satz betrachte ich eine allgemeinere Situation als bei den Newton–Cotes–Formeln:

– I nicht notwendig kompaktes Intervall, z.B. $I =]0, \infty[$, $w(x) = e^{-x}$,

– $w : \overset{o}{I} \longrightarrow]0, \infty[$ stetig mit eventuellen Randsingularitäten, z.B. $I = [-1, +1]$, $w(x) = \frac{1}{\sqrt{1-x^2}}$.

Den passenden Rahmen liefert

$$L_p(I; w) := \left\{ u : I \longrightarrow \mathbb{R} \quad | \quad \int_I w(x)|u(x)|^p dx < \infty \right\} \quad ,$$

$$< f, g >_{L_2(I;w)} \equiv < f, g >_w := \int_I w(x) f(x) g(x) dx \ .$$

Dann gilt

(4.2.1) Satz (Existenz und Charakterisierung der $\{s_i\}, \{w_i\}$)
Sei $I \subset \mathbb{R}$ ein Intervall, $w : \overset{o}{I} \longrightarrow]0, \infty[$ stetig mit $x^j \in L_1(I; w)$, $j \in \mathbb{N}_0$. Dann gilt

(a) Es existiert eine Folge $p_0^*, p_1^*, p_2^*, ...$ von in $L_2(I; w)$ orthonormalen Polynomen

$$p_j^* \in \mathbb{P}_{j+1} \setminus \mathbb{P}_j \quad , \quad < p_i^*, p_j^* >_w = \delta_{ij} \ , \ i, j \in \mathbb{N}_0 \ ;$$

(b) p_j^* hat j einfache Nullstellen in $\overset{o}{I}$.

Seien $s_1 < s_2 < ... < s_m$ die Nullstellen von p_m^* und $w_i = \int_I w(x) l_i(x) dx$, $1 \le i \le m$, mit dem i–ten Lagrangepolynom $l_i \in \mathbb{P}_m$: $l_i(s_j) = \delta_{ij}$, $1 \le i, j \le m$. Dann gilt für $Q(f) = \sum_{i=1}^m w_i f(s_i)$ für das *Fehlerfunktional* $E(f) = \int_I w(x) f(x) dx - \sum_{i=1}^m w_i f(s_i)$

(c) $E(p) = 0$, $p \in \mathbb{P}_{2m}$;

(d) die Gewichte $\{w_i : 1 \le i \le m\}$ sind positiv ;

(e) die Stützstellen und Gewichte sind durch die Forderung $E(p) = 0$, $p \in \mathbb{P}_{2m}$, eindeutig bestimmt ;

(f) es gibt keine Quadraturformel $\widetilde{Q}$ der Bauart $\widetilde{Q}(f) \equiv \sum_{i=1}^m \widetilde{w}_i f(\widetilde{s}_i)$ derart, daß für den Fehler $\widetilde{E}(f) = \int_I w(x) f(x) dx - \widetilde{Q}(f)$ gilt: $\widetilde{E}(p) = 0$, $p \in \mathbb{P}_{2m+1}$.

Beweis:

(a) rekursiv nach dem Gram-Schmidtschen Orthonormalisierungsverfahren.

◇

(b) indirekt: angenommen, p_j^* wechselt das Vorzeichen in den k verschiedenen Nullstellen $\{x_1, ..., x_k\}$ mit $k < j$. Dieselben Vorzeichenwechsel hat

$$q(x) := \begin{cases} 1 & , \ k = 0 \\ \prod_{i=1}^k (x - x_i) & , \ k \ge 1 \end{cases} \in \mathbb{P}_{k+1} \subset \mathbb{P}_j \ \text{ wegen } \ k + 1 \le j.$$

Damit folgt

$$0 \neq \int_I w(x) q(x) p_j^*(x) dx \quad \text{ im Widerspruch zu } \quad < q, p_j^* >_w = 0 \ .$$

Also hat p_j^* in $\overset{o}{I}$ j einfache Nullstellen.

◇

(c) Zu $p \in \mathbb{P}_{2m}$ gibt es $q, r \in \mathbb{P}_m$ mit $p = p_m^* \cdot q + r$. Auf Grund der Orthogonalität gilt (vgl. obige Motivation)

$$\int_I w(x) p_m^\star(x) q(x) dx \; - \; \sum_{i=1}^{m} w_i p_m^\star(s_i) q(s_i) \; = \; 0 \; .$$

Außerdem gilt wie für alle m–punktigen Interpolations–Quadraturformeln Q natürlich

$$E(r) = 0 \; , \; r \in \mathbb{P}_m \; .$$

$\diamond$

(d) $w_i = \int_I w(x) l_i(x) dx$ mit $l_i \in \mathbb{P}m$, $l_i(s_j) = \delta_{ij}$. Dann folgt wegen $\{0,1\} \ni l_i(s_j) = {l_i}^2(s_j)$

$$\int_I w(x) l_i(x) dx = Q(l_i) = Q({l_i}^2) \; = \; \int_I w(x) {l_i}^2(x) dx \; > \; 0 \; ,$$

wobei für l_i , ${l_i}^2 \in \mathbb{P}_{2m}$ die Exaktheit ausgenutzt wurde.

$\diamond$

(e) und (f) als Übungen. $\square$

Die Stützstellen und Gewichte sind nach (4.2.1) eindeutige Lösung des nichtlinearen Gleichungssystems

$$(*) \quad \sum_{i=1}^{m} w_i x^j(s_i) = \int_I w(x) x^j dx \; , \; j = 0, 1, .., 2m-1,$$

$(*)$ ist i.Allg. sehr schlecht konditioniert. Daher erfolgt die tatsächliche numerische Bestimmung der Stützstellen und Gewichte über ein gut konditioniertes Eigenwertproblem. Dies ist eine Folgerung aus dem nächsten Satz (vgl. auch die Übungen)

(4.2.2) Satz (dreigliedrige Rekursion orthogonaler Polynome)
Unter den Voraussetzungen von Satz (4.2.1) erfüllen die auf Höchstkoeffizienten 1 normierten w–orthogonalen Polynome $p_n, \; n \in \mathbb{N}_0$,

$$p_1(x) = (x - \delta_0) p_0(x) \; ,$$

und für $n \geq 1$ die dreigliedrige Rekursionsformel

$$p_{n+1}(x) = (x - \delta_n) p_n(x) - {\gamma_n}^2 p_{n-1}(x) \; , \; n = 1, 2, ...$$

mit $\delta_n = < x \cdot p_n, p_n >_w \; / \; < p_n, p_n >_w, \; n \geq 0$, und ${\gamma_n}^2 = < p_n, p_n >_w \; / \; < p_{n-1}, p_{n-1} >_w, \; n \geq 1$.

Beweis:

Die Existenz der w–orthogonalen $(p_n)_{n \in \mathbb{N}_0}$ ist klar: $p_n = {c_n}^{-1} p_n^*$, wobei c_n der Höchstkoeffizient von p_n^* ist. Damit ist

$$(*) \quad p_{n+1}(x) = x \; p_n(x) + \sum_{k=0}^{n} \alpha_k p_k(x).$$

Für $n = 1$ erfüllt $p_1(x) = (x - \delta_0) p_0(x)$ für $\delta_0 = < x p_0, p_0 >_w \; / \; < p_0, p_0 >_w$ die Bedingung $0 = < p_1, p_0 >_w$.
Zum Nachweis der dreigliedrigen Rekursion zeige ich, daß in $(*)$ gilt: $\alpha_j = 0, \;\; 0 \leq j \leq n-2$.

Beweis:

$$0 =< p_{n+1}, p_j >_w \overset{(*)}{=} \underbrace{< x\ p_n, p_j >_w}_{=< p_n, x\ p_j >_w = 0\ ,\ j+1<n;} + \sum_{k=0}^{n} \alpha_k < p_k, p_j >_w = \alpha_j < p_j, p_j >_w \ .$$

Wegen $< p_j, p_j >_w > 0$ ist $\alpha_j = 0,\ j < n-1$. Mit $(*)$ folgt aus $0 =< p_{n+1}, p_n >_w$ die Darstellung für δ_n, und aus $0 =< p_{n+1}, p_{n-1} >_w$ die Darstellung für ${\gamma_n}^2$.

□

(4.2.3) Tabelle einiger Orthogonalpolynome (mit einigen Eigenschaften)

***Legendre*–Polynome $P_n(x)$**

Intervall: $[-1,+1]$ Gewicht: $w(x) = 1$; Normalisierung: $P_n(1) = 1$;

$L_2(I;w)$–Norm: $\int_{-1}^{+1} P_n(x)^2 dx = \frac{2}{2n+1}$;

Rekursion: $(n+1)P_{n+1} = (2n+1)xP_n(x) - nP_{n-1}(x)\ ,\ P_0(x) = 1\ ,\ P_1(x) = x$.

◇

***Tschebyschew*–Polynome 1. Art $T_n(x)$**

Intervall: $]-1,+1[$ Gewicht: $w(x) = (1-x^2)^{-1/2}$; Normalisierung: $T_n(1) = 1$;

$L_2(I;w)$–Norm: $\int_{-1}^{+1}(1-x^2)^{-1/2}T_n(x)^2dx = \begin{cases} \pi\ , & n\ = 0\ , \\ \pi/2\ , & n \neq 0\ , \end{cases}$

Rekursion: $T_{n+1} = 2xT_n(x) - T_{n-1}(x)\ ,\ T_0(x) = 1\ ,\ T_1(x) = x$.

◇

***Tschebyschew*–Polynome 2. Art $U_n(x)$**

Intervall: $[-1,+1]$ Gewicht: $w(x) = (1-x^2)^{1/2}$; Normalisierung: $U_n(1) = n+1$;

$L_2(I;w)$–Norm: $\int_{-1}^{+1}(1-x^2)^{1/2}U_n(x)^2dx = \pi/2$,

Rekursion: $U_{n+1} = 2xU_n(x) - U_{n-1}(x)\ ,\ U_0(x) = 1\ ,\ U_1(x) = 2x$.

◇

***Laguerre*–Polynome $L_n(x)$**

Intervall: $[0,\infty[$ Gewicht: $w(x) = e^{-x}$; Normalisierung: $L_n(x) = \frac{(-1)^n}{n!}x^n + \ldots$;

$L_2(I;w)$–Norm: $\int\limits_0^\infty e^{-x}L_n(x)^2dx = 1$;

Rekursion: $(n+1)L_{n+1} = [(2n+1) - x]L_n(x) - nL_{n-1}(x)\ ,\ L_0(x) = 1\ ,\ L_1(x) = -x+1$.

◇

***Hermite*–Polynome $H_n(x)$**

Intervall: $]-\infty,\infty[$ Gewicht: $w(x) = e^{-x^2}$; Normalisierung: $H_n(x) = 2^n x^n + \ldots$;

$L_2(I;w)$–Norm: $\int\limits_{-\infty}^{\infty} e^{-x^2} H_n(x)^2 dx = \sqrt{\pi} 2^n n!$;

Rekursion: $H_{n+1} = 2xH_n(x) - 2nH_{n-1}(x)$, $H_0(x) = 1$, $H_1(x) = 2x$.

□

Diese und weitere Eigenschaften, sowie andere Orthogonalpolynome, findet man z.B. aufgelistet im Appendix I–VI des Buches von *Davis, Ph.J.*: Interpolation and Approximation. Blaisdell 1963, pp. 365-368.

So schön der optimal hohe Exaktheitsgrad von Gauß–Quadraturformeln auch ist, so haben diese Quadraturformeln bzgl. Fehlerschätzungen durch Vergleich verschieden genauer Formeln den *Nachteil* , daß – anders als bei Newton–Cotes– oder den Extrapolationsverfahren, die ich nicht behandle und für die ich auf die Literatur verweise – Funktionswerte des Integranden in disjunkten Stützstellenmengen genommen werden:

$$\{x \in I \ : \ p_m^*(x) = 0\} \cap \{x \in I \ : \ p_n^*(x) = 0\} \ = \ \emptyset \ , \quad m \neq n \ .$$

Damit sind solche Fehlerschätzungen nicht sehr verläßlich. Die Behebung dieses Mangels ergibt verallgemeinerte Gauß–Quadraturformeln, die sog. *Gauß–Kronrod–Quadraturformeln* . Diese heute gegenüber Gaußformeln bevorzugten Formeln haben natürlich nicht mehr den optimalen Exaktheitsgrad der Gaußformeln.

4.3 Quadratur von Anfangswertaufgaben

Ich betrachte folgende Form von Anfangswertaufgaben

(4.3.1) Anfangswertaufgabe
Gegeben $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $I = [0,T]$, und $x_0 \in \mathbb{R}^n$. Gesucht $x : I \longrightarrow \mathbb{R}^n, x \in C^1(I)$, so daß gilt

$$x'(t) = f(t, x(t)) \ , \ t \in I \quad ; \quad x(0) = x_0 \ .$$

□

Dabei ist

$$x'(t) := \begin{bmatrix} \frac{dx_1}{dt}(t) \\ \vdots \\ \frac{dx_n}{dt}(t) \end{bmatrix} \quad \text{für} \quad x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

komponentenweise erklärt.
Beispiele für Anfangswertaufgaben sind etwa Populationsmodelle

(4.3.2) Räuber–Beute–Modell[1]

$y(t)$ Anzahl der Individuen der 'Beute'–Bevölkerung zur Zeit t, $z(t)$ Anzahl der Individuen der 'Räuber'–Bevölkerung zur Zeit t
$\alpha > 0$ Vermehrungsrate (=Differenz von Geburts– und Sterberate) der isolierten Beute–Population
$\gamma > 0$ Sterberate der isolierten Räuber–Population
Dann gilt

$$y'(t) = \alpha y(t) - \beta y(t)z(t)$$

Dabei sind folgende Modellannahmen für die Beute–Population zugrundegelegt:
Vermehrung $\sim$ Anzahl ; Voraussetzung bei dieser Modellannahme ist Nahrungsüberfluß;
Verminderung $\sim$ Anzahl der Kontakte zwischen Beute– und Räubertieren:
Anzahl der Kontakte eines y–Individuums $\sim z$, Anzahl der Kontakte aller y–Individuen $\sim yz$.
Für die Räuber–Population ergibt sich analog:

$$z'(t) = -\gamma z(t) + \delta y(t)z(t)$$

◇

Mit

$$x_1 := y \quad , \qquad x_2 := z \quad ,$$
$$f_1(t, x_1, x_2) := \alpha x_1 - \beta x_1 x_2 \, ,$$
$$f_2(t, x_1, x_2) := -\gamma x_2 + \delta x_1 x_2 \, ,$$

erhält man die Form (4.3.1)

$$x'(t + t_0) = f(t + t_0, x(t + t_0)), \quad t > 0,$$
$$x(t_0) = x_0;$$

wobei $x_0 = [y_0, z_0]^t$ ein zum Zeitpunkt t_0 beobachteter Zustand der Populationen ist.

□

Formal äquivalent zur Lösung von (4.3.1) ist die Lösung der Integralgleichung

$$x(t) = x_0 + \int_0^t f(\tau, x(\tau)) \; d\tau \; , \; t \in I \; .$$

Dabei ist für $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ das Integral komponentenweise erklärt. Dieser Zusammenhang ist Grundlage und Motivation für einige Verfahren: Setzt man $h = T/N > 0$, $t_0 = 0$ und $t_j = t_0 + jh, j = 1, ..., N$, so ist für $t_1 = t_0 + h$

$$x(t_1) = x_0 + \int_{t_0}^{t_1} f(\tau, x(\tau)) \; d\tau \;\doteq\; x_0 + hf(t_0, x_0) =: x_1 \; ,$$

wenn man das Integral durch die Rechteckregel approximiert. Wiederholt man dieses Vorgehen, wobei man die unbekannten Werte $x(t_j)$ der Lösung durch die berechneten Werte x_j ersetzt, so erhält man das

(4.3.3) Polygonzugverfahren von Euler

$$x_{j+1} = x_j + hf\big(t_j, x_j\big) \; , \; j = 0, ..., N - 1 \; .$$

□

[1] vgl. z.B. Braun, M.: Differentialgleichungen und ihre Anwendungen. 2. Aufl. 1991, SS. 474.

Der Fehler für einen Schritt – o.E. von x_0 zu x_1 – ist für $f \in C^1$ und damit $x \in C^2$

$$x_1 - x(t_1) = (x_1 - x_0) - (x(t_1) - x_0) = \int_{t_0}^{t_1} f(t_0, x_0)\ d\tau - \int_{t_0}^{t_1} f(\tau, x(\tau))\ d\tau$$

$$= \int_{t_0}^{t_1} \big(x'(t_0) - x'(\tau)\big)\ d\tau \ = \ h\, O(h).$$

Eine *Abschätzung für den Fehler* $\delta_j := \|x_j - x(t_j)\|$ nach j Schritten erhält man aus der Darstellung

$$x(t_{j+1}) = x(t_j) + \underbrace{hf\big(t_j, x(t_j)\big) - \int_{t_j}^{t_{j+1}} \big(x'(t_j) - x'(\tau)\big)\ d\tau}_{=0}\ ,$$

für die Lösung durch Subtraktion von der Polygonzuggleichung

$$x_{j+1} = x_j + hf\big(t_j, x_j\big)\ .$$

Damit ergibt sich

$$\delta_{j+1} \ \le \ \delta_j + h\|f\big(t_j, x_j\big) - f\big(t_j, x(t_j)\big)\| + \frac{h^2}{2}\|x''\|_{\infty,[t_j,t_{j+1}]} \le \delta_j(1 + hL) + \frac{h^2}{2}\|x''\|_{\infty,[t_j,t_{j+1}]}.$$

Dabei ist L die Lipschitzkonstante von f bezüglich dem zweiten Argument. Wendet man die Abschätzung $\delta_{j+1} \le \delta_j(1 + hL) + \frac{h^2}{2}\|x''\|_{\infty,[t_j,t_{j+1}]}$ rekursiv an, erhält man wegen $\delta_0 = 0$

$$\delta_j \ \le \ \delta_0(1 + hL)^j \ + \ \frac{h^2}{2}\|x''\|_{\infty,[t_0,t_j]} \sum_{k=0}^{j-1} (1 + hL)^k \ = \ \frac{h^2}{2}\|x''\|_{\infty,[t_0,t_j]} \frac{(1 + hL)^j - 1}{hL}.$$

Wegen $(1 + hL) < e^{hL}$ folgt unter Beachtung von $jh \le T$

$$\max_{j\ :\ t_j \in [0,T]} \delta_j \ \le \ \frac{e^{TL} - 1}{L}\, \frac{1}{2}\, \|x''\|_{\infty,[0,T]}\, h \ = \ O(h).$$

□

Zur Idee der Runge–Kutta–Verfahren möchte ich an die Fehlerordnung der Rechteckregel mit Funktionswert im Mittelpunkt erinnern: $O(h^2) \cdot$ *Intervallänge* für $\tau \longmapsto f\big(\tau, x(\tau)\big) \in C^2$, also

$$x(t_{j+1}) - x(t_j) \ = \ \int_{t_j}^{t_j+h} f\big(\tau, x(\tau)\big) d\tau \ = \ h\, f\big(t_j + \frac{h}{2},\ \underbrace{x(t_j + \frac{h}{2})}_{\text{unbekannt}}\ \big) + O(h^3)$$

$$f\big(t_j + \frac{h}{2},\ x(t_j + \frac{h}{2})\ \big) =$$

$$f\big(t_j + \frac{h}{2},\ x(t_j) + \frac{h}{2}\underbrace{f\big(t_j, x(t_j)\big)}_{\| \ x'(t_j)}\ \big) + O_L\Big(\ \underbrace{\|x(t_j + \frac{h}{2}) - \big(x(t_j) + \frac{h}{2}x'(t_j)\big)\|}_{= O_{\|x''\|_{[t_j,t_{j+1}]}}(h^2)}\ \Big)\ .$$

Insgesamt folgt nach dieser Beweisskizze für den Fehler nach einem Schritt mit beliebigem Startwert $(t_0, x_0) \in B :\equiv [0,T] \times \{x_0 \in \mathbb{R}^n \ : \ \|x_0 - x(t_0)\| \leq \varrho\}$ für die sog. *Konsistenzordnung* p mit

$$\max_{(t_0,x_0) \in B} \|x_1 - y(t_0+h)\| =: hO(h^p)$$

wobei für $(t_0,x_0) \in B \quad x_1 = x_0 + hf\big(t_0 + \frac{h}{2}, x_0 + \frac{h}{2}f(t_0,x_0)\big)$ und $y'(s) = f\big(s, y(s)\big), s \in [t_0, t_0+h]$; $y(t_0) = x_0$ ist.

(4.3.4) Beispiel
Die 'verbesserte Polygonzugmethode'

$$x_{j+1} := x_j + hf\big(t_j + \frac{h}{2}, x_j + \frac{h}{2}f(t_j,x_j)\big) \ , \ t_{j+1} := t_j + h \ ;$$

hat die Konsistenzordnung $p = 2$, falls $f \in C^2$ ist. Wie beim Polygonzugverfahren existiert $c > 0$ mit

$$\max_{j \ : \ t_j \in [0,T]} \|x_j - x(t_j)\| \ \leq \ c\,h^2 \ .$$

□

Anstelle dieser Integralnäherung unter Verwendung des Integranden in den zwei Punkten t_j und $t_j + h/2$ kann man natürlich auch das Mittel aus mehreren Punkten nehmen. *Verwendung von s Näherungen für $x'(.) = f(.,x(.))$* im Intervall $[t_j, t_j + h]$ liefert sog.
Explizite Runge–Kutta–Verfahren der Stufe s:
Sei $(t_0,x_0) \in I \times \mathbb{R}^n$ und $h > 0$ gegeben:

$$\begin{aligned}
k_1 &= f(t_0 \ , \ x_0) \\
k_2 &= f(t_0 + \alpha_2 h \ , \ x_0 + h\beta_{21}k_1) \\
&\ \ \vdots \\
k_s &= f\big(t_0 + \alpha_s h \ , \ x_0 + h(\beta_{s1}k_1 + \beta_{s2}k_2 + \ldots + \beta_{s,s-1}k_{s-1})\big) \ ,
\end{aligned}$$

und

$$x_1 = x_0 \ + \ h \ (\ \gamma_1 k_1 + \gamma_2 k_2 + \ldots + \gamma_s k_s \ ) \ .$$

Ordnet man die α_i , $\beta_{i,j}$ und γ_i nach folgendem Schema

$$\begin{array}{r|cccc}
0 = \alpha_1 & 0 & & \cdots & 0 \\
\alpha_2 & \beta_{21} & 0 & & \vdots \\
\vdots & \vdots & & \ddots & \vdots \\
\alpha_m & \beta_{m1} & \cdots & \beta_{m,m-1} & 0 \\
\hline
 & \gamma_1 & \cdots & \gamma_{m-1} & \gamma_m
\end{array}$$

so sind von besonderer Attraktivität 4–stufige Runge–Kutta–Verfahren

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Standard–Runge–Kutta–Formel

0				
1/3	1/3			
2/3	−1/3	1		
1	1	−1	1	
	1/8	3/8	3/8	1/8

3/8–*Formel*

Diese beiden Formeln haben die Konsistenzordnung $p = 4$. Sie konvergieren für $f \in C^4$ von der Ordnung $O(h^4)$ (vgl. dazu *Numerik 2*).

Kapitel 5

Iterationsverfahren für Gleichungssysteme

5.1 Newtonverfahren zur Lösung nichtlinearer Gleichungssysteme

Für nichtlineare Gleichungssysteme gibt es zwei Standardformen der Aufgabe: als *Nullstellen–Gleichung* und als *Fixpunkt–Gleichung*.

(5.1.1) Nullstellengleichung
Gegeben $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Gesucht ist eine *Nullstelle* $x^* = [x^*{}_i]_{1 \le i \le n} \in D$ mit $F(x^*) = 0 \in \mathbb{R}^n$.

□

(5.1.1') Fixpunktgleichung
Gegeben $G : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Gesucht ist ein *Fixpunkt* $x^* \in D$ mit $G(x^*) = x^*$.

□

Mit $F(x) = [F_j(x)]_{1 \le j \le n}$ bedeutet (5.1.1) ausgeschrieben

$$F_j(x_1^*, x_2^*, ..., x_n^*) = 0 \ , \ 1 \le j \le n \ .$$

Dabei ist $F_j : D \subset \mathbb{R}^n \longrightarrow \mathbb{R} \ , \ 1 \le j \le n$. Diese Aufgabenstellung kommt in vielen innermathematischen und außermathematischen Anwendungen vor, z.B. ist ja für $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$

$$\frac{\partial f}{\partial x_j}(x^*) = 0 \ , \ 1 \le j \le n \ ;$$

eine notwendige Bedingung für ein lokales inneres Extremum x^* von $f \in C^1$. Offensichtlich gilt

(5.1.2) Bemerkung
Sei $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ und eine Schar linearer Abbildungen $D \ni x \longmapsto A^-(x) \in \mathbb{R}^{n \times n}$ gegeben. Dann gilt für

$$G(x) := x - A^-(x)F(x) \ , \ x \in D \ ,$$

(a) x^* ist Nullstelle von $F \Longrightarrow x^*$ ist Fixpunkt von G;

(b) x^* ist Fixpunkt von G und $A^-(x^*)$ nichtsingulär $\Longrightarrow x^*$ ist Nullstelle von F.

□

Idee der Newton(ähnlichen) Verfahren:
'Verbessere' iterativ eine gegebene Näherung x_{alt} für x^* zu einer neuen Näherung x_{neu}. Dazu

— ersetze F ('lokal' in einer Umgebung von x_{alt}) durch eine affin-lineare Näherung $\widetilde{F}$, z.B.

$$\widetilde{F}(x_{alt}+s) = F(x_{alt}) + F'(x_{alt})s \ , \ s \in \mathbb{R}^n \ ,$$

wobei $F'(x_{alt}) \equiv \left[\frac{\partial F_i}{\partial x_j}(x_{alt})\right]_{1\le i,j\le n}$ die Jacobische Funktionalmatrix von F im Punkt x_{alt} ist. Mit $A(x_{alt}) \doteq F'(x_{alt})$ kann man allgemeiner betrachten

$$\widetilde{F}(x_{alt}+s) = F(x_{alt}) + A(x_{alt})s \ , \ s \in \mathbb{R}^n \ .$$

— als neue (hoffentlich bessere) Näherung x_{neu} für die Nullstelle x^* von F nimmt man die Nullstelle von $\widetilde{F}$:

$$\widetilde{F}(x_{alt}+s) = 0 \ , \ x_{neu} = x_{alt} + s \ .$$

s ist also die Lösung des linearen Gleichungssystems $A(x_{alt})s = -F(x_{alt})$.

Damit erhält man (ab jetzt verwende ich $x_a :\equiv x_{alt}$ und $x_n :\equiv x_{neu}$ – $'n'$ wie *'neu'* sollte zu keinen Verwechslungen mit $'n'$ in $\mathbb{R}^n$ führen! – und auch $A_a :\equiv A(x_{alt})$ und $A_n :\equiv A(x_{neu})$)

(5.1.3) Newtonähnliche Verfahren (prinzipieller Algorithmus)
(1) bestimme Startnäherung $x_a \in D$ und $A_a \in \mathbb{R}^{n\times n}$.

Solange für x_a das Abbruchkriterium nicht erfüllt ist, wiederhole

(2) bestimme $s_a \in \mathbb{R}^n$ mit $A_a s_a = -F(x_a)$, $x_n = x_a + s_a$ sog. *Quasi–Newton–Schritt* ;

(3) bestimme A_n ;

setze $x_a := x_n$ und $A_a := A_n$.

□

(5.1.4) Newton–Varianten
Mögliche Wahl von A_a in (5.1.3)*(1)* bzw. von A_n in (5.1.3)*(3)* ist etwa

(a) $A_a := F'(x_a)$; $A_n := F'(x_n)$; sog. *klassisches Newtonverfahren*;

(b) $A_a := F'(x_a)$; $A_n := A_a$; sog. *vereinfachtes Newtonverfahren*;

(c) $A_a := F'(x_a) + R(x_a)$; $A_n := F'(x_n) + R(x_n)$; sog. *Newtonverfahren mit näherungsweiser Ableitung*;

(d) Im $\mathbb{R}^1$ ist eine Möglichkeit für (c) das sog. *Sekantenverfahren* $A_a := F'(x_a)$ in (5.1.3)*(1)* bzw. $A_n := \big(F(x_n) - F(x_a)\big)/(x_n - x_a)$ in (5.1.3)*(3)*;

(e) $A_a := F'(x_a)$; A_n die sog. *Broyden–Rang 1–Aufdatierungsformel* (5.1.10).

□

(5.1.5) Beispiel
$F : \mathbb{R} \longrightarrow \mathbb{R}, F(x) := x^2 - 1, x^* = 1$. Dann ergeben die verschiedenen Varianten des Newtonverfahrens

k	$x^{(k)}$ *klassisches Newtonverfahren*	Konv. ord.	$x^{(k)}$ *Sekanten-verfahren*	Konv. ord.	$x^{(k)}$ *vereinfachtes Newtonverfahren*	Konv. ord.
0	2		2		2	
1	1.25		1.25		1.25	
2	1.025		1.07692307692308		1.109375	
3	1.00030487804878	2.67	1.00826446280992	1.89	1.05169677734375	1.76
4	1.00000004646115	2.19	1.00030487804878	1.88	1.02518024947494	1.47
5	1.00000000000000	2.09	1.00000125445174	1.69	1.01243161349657	1.35
6			1.00000000019120	1.68	1.00617717049475	1.28
7			1.00000000000000	1.65	1.00307904588855	1.24
8					1.00153715281338	1.21

□

Die Konvergenzgüte von Folgen misst man durch

(5.1.6) Konvergenzordnung von Folgen
$(x^{(k)} \mid k \in \mathbb{N}_0)$ mit $\lim\limits_{k\to\infty} x^{(k)} = x^*$ konvergiert gegen x^* asymptotisch (mindestens)

(a) *linear* oder *von der Ordnung 1* , wenn $c: \ 0 < c < 1$ existiert mit

$$\|x^{(k+1)} - x^*\| \leq c \ \|x^{(k)} - x^*\| \ ;$$

(b) *superlinear* , wenn

$$\|x^{(k+1)} - x^*\| \ = \ o\big(\|x^{(k)} - x^*\|\big) \ \text{ für } \ k \to \infty;$$

(c) *quadratisch* , wenn

$$\|x^{(k+1)} - x^*\| \ = \ O\Big(\|x^{(k)} - x^*\|^2\Big) \ \text{ für } \ k \to \infty;$$

(d) *von der Ordnung* $p \geq 2$, wenn

$$\|x^{(k+1)} - x^*\| \ = \ O\Big(\|x^{(k)} - x^*\|^p\Big) \ \text{ für } \ k \to \infty;$$

□

Eine aus den berechneten Näherungen 'numerisch beobachtete Konvergenzordnung' $\widetilde{p}$ mit

$$\|x^{(k+1)} - x^*\| \ \doteq \ c \ \|x^{(k)} - x^*\|^{\widetilde{p}}$$

ergibt sich wegen $x^{(k+2)} \ \doteq \ x^*$ nach der Formel

$$\widetilde{p} \ = \ log\Big(\frac{\|x^{(k+1)} - x^{(k+2)}\|}{\|x^{(k)} - x^{(k+2)}\|}\Big) \ / \ log\Big(\frac{\|x^{(k)} - x^{(k+2)}\|}{\|x^{(k-1)} - x^{(k+2)}\|}\Big) \ .$$

Dies sind die Einträge in den Spalten 'Konv.ord.' in obigem Beispiel (5.1.5). Das Sekantenverfahren konvergiert theoretisch von der Ordnung $(1+\sqrt{5})/2 \doteq 1.61$ (s. z.B. *Hämmerlin–Hoffmann*).

Zur Konvergenzanalyse des Newtonverfahrens betrachte ich die Fixpunktformulierung von (5.1.3): Ist $A(x_a)$ nichtsingulär für alle x_a, so gilt offensichtlich

$$x_n = G(x_a) \quad \text{mit} \quad G(x) = x - A(x)^{-1}F(x).$$

Die Konvergenz des Newtonverfahrens spielt man auf den von *Stefan Banach (1892–1945)* stammenden Konvergenzsatz für *kontrahierende Abbildungen* zurück. Eine sehr einfache, für unsere Zwecke ausreichende Version – da wir die Existenz von Nullstellen bzw. Fixpunkten immer voraussetzen – ist

(5.1.7) Lemma (Konvergenzsatz für kontrahierende Abbildungen)
Sei $G : D \subset \mathbb{R}^n \longrightarrow D \subset \mathbb{R}^n$, $G(x^*) = x^*$ für ein $x^* \in D$. Für die Folge $(x^{(k)})_{k\in\mathbb{N}_0}$ mit $x^{(k+1)} = G(x^{(k)})$, $k \in \mathbb{N}_0$, und $x^{(0)} \in D$, existiere $c : 0 < c < 1$, mit

$$\|G(x^{(k)}) - G(x^*)\| \leq c \, \|x^{(k)} - x^*\| \ , \ k \in \mathbb{N}_0.$$

Dann konvergiert $(x^{(k)})_{k\in\mathbb{N}_0}$ linear gegen x^*.

Beweis:

$$\|x^{(k)} - x^*\| = \|G(x^{(k-1)}) - G(x^*)\| \leq c \, \|x^{(k-1)} - x^*\| \leq \ \ldots \ \leq c^{k+1} \, \|x^{(0)} - x^*\| \longrightarrow 0 \ , \ k \to \infty.$$

□

Im folgenden möchte ich das unterschiedliche Konvergenzverhalten der verschiedenen Newton-Varianten (vgl. (5.1.5)) auch theoretisch erhärten. Dazu sei

$K[x;\varrho] := \{y \in \mathbb{R}^n \ : \ \|y - x\| \leq \ \varrho \ \}$

die abgeschlossene Kugel um x mit Radius ϱ.

(5.1.8) Satz (lokale Konvergenz für 'Newtonähnliche Verfahren')
Gegeben sei $D \subset \mathbb{R}^n$, D offen, und eine Norm $\|.\|$ auf $\mathbb{R}^n$. Sei $F : D \longrightarrow \mathbb{R}^n$ stetig differenzierbar, $x^* \in D$ mit $F(x^*) = 0$ und $F'(x^*) \in \mathbb{R}^{n\times n}$ nichtsingulär. Für $A : D \longrightarrow \mathbb{R}^{n\times n}$ sei $A(x)$ nichtsingulär für $x \in D$ und

$$\limsup_{x\to x^*} \| \, A(x)^{-1} \, \|_{ind} \ < \ \infty \ .$$

(a) Gilt

$$\limsup_{x\to x^*} \| \ A(x)^{-1} \, F'(x^*) \ - \ E \, \|_{ind} \ < \ 1 \ ,$$

so existiert $\varrho > 0$, so daß für die Iterationsfunktion $G : D \longrightarrow \mathbb{R}^n$, $G(x) := x - A(x)^{-1} \, F(x)$, gilt

$$G(K[x^*;\varrho]) \ \subset \ K[x^*;\varrho] \ .$$

Für beliebiges $x^{(0)} \in K[x^*;\varrho]$ ist die Folge $(x^{(k)} \ : \ k \in \mathbb{N}_0)$,

$$x^{(k+1)} \ := \ G(x^{(k)}) \ , \ k = 0, 1, 2, \ldots$$

wohldefiniert, und *konvergiert mindestens linear* gegen x^* .

(b) Gilt $\lim_{x\to x^*} \| \ A(x) \ - \ F'(x^*) \, \|_{ind} \ = \ 0$, so *konvergiert* die Folge $(x^{(k)} \ : \ k \in \mathbb{N}_0)$ aus (a) *superlinear* gegen x^* .

(c) Ist F' Lipschitzstetig mit Lipschitzkonstante $L > 0$, d.h. gilt

$$\| F'(x) - F'(y) \|_{ind} \leq L \| x - y \|, \; x, y \in D,$$

so gilt für $A(x) := F'(x)$

$$\| x^{(k+1)} - x^* \| = O\Big(\| x^{(k)} - x^* \|^2 \Big),$$

d.h. das *klassische Newtonverfahren* ist *quadratisch konvergent* für $F \in C^2$.

Beweis:

(a)

$$G(x) - x^* = x - x^* - A(x)^{-1}\big(F(x) - F(x^*) \big) \quad \text{nach Def. von } G, \text{ denn } F(x^*) = 0$$

$$= \big(E - A(x)^{-1} F'(x^*) \big) (x - x^*) - A(x)^{-1}\big(F(x) - F(x^*) - F'(x^*)(x - x^*) \big)$$

$\Longrightarrow$

$$\| G(x) - x^* \| \leq \Big(\|E - A(x)^{-1} F'(x^*)\| + \|A(x)^{-1}\| \frac{\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|}{\|x - x^*\|} \Big) \|x - x^*\|$$

(o.E. ist $\|x - x^*\| \neq 0$; für $\|x - x^*\| = 0$ gelten die unten gezeigten Abschätzungen trivialerweise). Man beachte nun

$$\|E - A(x)^{-1} F'(x^*)\| \leq (1 - 2\varepsilon_1), \text{ falls } \|x - x^*\| \leq \varrho_1,$$

wobei o.B.d.A. $0 < \varepsilon_1 < 1/2$, und ϱ_1 nach Vor. (a) geeignet gewählt sei.

$$\|A(x)^{-1}\| \leq c_1 + 1, \text{ falls } \|x - x^*\| \leq \varrho_2,$$

wobei ϱ_2 nach der Vor. $c_1 := \limsup_{x \to x^*} \big\| A(x)^{-1} \big\|_{ind} < \infty$ geeignet gewählt sei.

$$\frac{\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|}{\|x - x^*\|} \leq \frac{\varepsilon_1}{(c_1 + 1)}, \text{ falls } \|x - x^*\| \leq \varrho_3,$$

wobei ϱ_3 nach der Voraussetzung 'F differenzierbar in x^*' geeignet gewählt sei.

Für $x : \|x - x^*\| \leq \varrho := min(\varrho_1, \varrho_2, \varrho_3)$ folgt

$$\|G(x) - x^*\| \leq \underbrace{\left(1 - 2\varepsilon_1 + (c_1 + 1) \frac{\varepsilon_1}{(c_1 + 1)} \right)}_{= (1 - \varepsilon_1) =: c < 1} \|x - x^*\|$$

Insgesamt erhalten wir (da D offen ist, nach eventueller Verkleinerung von ϱ)

$$K[x^*; \varrho] \subset D;$$

und

$$(*) \quad \| G(x) - x^* \| \leq c \| x - x^* \|, \; x \in K[x^*; \varrho].$$

Behauptung: $(x^{(k)} : k \in \mathbb{N}_0)$ ist wohldefiniert für alle $x^{(0)} \in K[x^*; \varrho]$.

Beweis:

$$x^{(k)} \in K[x^*; \varrho] \overset{(*)}{\Longrightarrow} x^{(k+1)} = G(x^{(k)}) \in K[x^*; \varrho]$$

$\diamond$

Behauptung: $(x^{(k)} : k \in \mathbb{N}_0)$ konvergiert linear gegen x^* .
Beweis: Es gilt

$$\|x^{(k+1)} - x^*\| = \|G(x^{(k)}) - G(x^*)\| \overset{(*)}{\leq} c\, \|x^{(k)} - x^*\| \,,\ k \in \mathbb{N}_0 \,.$$

Damit folgt *mindestens lineare Konvergenz* wegen $c < 1$ nach (5.1.7).

$\diamond$

(b) Nach dem Beweisanfang von (a) gilt

$$x^{(k+1)} - x^* = A(x)^{-1}\big(F(x^*) + F'(x^*)\,(x - x^*) - F(x)\big) \;+\; A(x)^{-1}\big(\,F'(x^*) - A(x)\big)\,(x - x^*)$$

Wegen $\limsup_{x \to x^*} \| A(x)^{-1} \|_{ind} < \infty$ ist der zweite Summand $o(\|x^{(k)} - x^*\|)$ nach der Voraussetzung in (b) und der erste Summand $o(\|x^{(k)} - x^*\|)$, da $F \in C^1$ ist.

$\diamond$

(c) Für $A(x^{(k)}) := F'(x^{(k)})$ gilt

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}) - x^* \\ &= F'(x^{(k)})^{-1}\Big[\underbrace{F(x^*)}_{=0} - F(x^{(k)}) - F'(x^{(k)})\big(x^* - x^{(k)}\big)\Big] \,; \end{aligned}$$

Abschätzung der eckigen Klammer über den *Mittelwertsatz* liefert

$$\begin{aligned} &\Big[F(x^*) - F(x^{(k)}) - F'(x^{(k)})\big(x^* - x^{(k)}\big)\Big] \\ &= \int_0^1 F'\big(\,x^{(k)} + s(x^* - x^{(k)})\big)\big(\,x^* - x^{(k)}\,\big)\,ds \;-\; F'\big(x^{(k)}\big)\big(\,x^* - x^{(k)}\,\big) \\ &= \int_0^1 \Big(F'\big(\,x^{(k)} + s(x^* - x^{(k)})\big) \;-\; F'\big(x^{(k)}\big)\Big)\big(\,x^* - x^{(k)}\,\big)\,ds \end{aligned}$$

$\Longrightarrow$

$$\begin{aligned} &\Big\|\Big[F(x^*) - F(x^{(k)}) - F'(x^{(k)})\big(x^* - x^{(k)}\big)\Big]\Big\| \\ &\leq \int_0^1 \Big\| \Big(F'\big(\,x^{(k)} + s(x^* - x^{(k)})\big) \;-\; F'\big(x^{(k)}\big)\Big) \Big\|_{ind} \;\|\; x^* - x^{(k)} \,\|\, ds \\ &\leq \int_0^1 L\, s\, \|x^* - x^{(k)}\| \;\cdot\; \|x^* - x^{(k)}\|\, ds \,, \quad \text{da } F' \in Lip_L \\ &= \frac{L}{2}\, \|x^* - x^{(k)}\|^2 \,. \end{aligned}$$

$\square$

Diese sehr schönen Konvergenzaussagen gelten allerdings *nur lokal*. Zur Ernüchterung:

– weit entfernt von der Lösung ist, selbst globale Konvergenz vorausgesetzt, die *Konvergenzgeschwindigkeit nur linear*.

– u.U. hängt das Konvergenzverhalten 'chaotisch' vom Startwert ab.

Die Nachteile der bisherigen Varianten sind:

– hoher Aufwand zur Bestimmung von A_n;

– 'Konvergenzverhalten' empfindlich gegenüber starken Änderungen der aktuellen Näherung $\widetilde{F}$ für F.

(5.1.9) Motivation für das Broyden–Verfahren
$\widetilde{F}_a(x) = F(x_a) + A_a(x - x_a)$ liefert durch die Forderung $\widetilde{F}_a(x_n) = 0$, diesen sog. *Quasi-Newtonschritt*, die neue Näherung x_n mit

$$A_a(x_n - x_a) = -F(x_a) \ .$$

Wie bestimmt man A_n in der neuen Approximation $\widetilde{F}_n(x) = F(x_n) + A_n(x - x_n)$ an F ?

1. Bedingung: $\widetilde{F}_n(x_a) \stackrel{!}{=} \widetilde{F}_a(x_a)$.

Damit gilt

$$A_n(x_n - x_a) = F(x_n) - F(x_a) \ , \quad \text{die sog. } \textit{Sekantengleichung } (vgl.(5.1.4)(d) \text{ für } n = 1) \ ;$$

und daraus folgend die für das superlineare Konvergenzverhalten ausreichende (Fast–)Newtoneigenschaft

$$A_n(x_n - x_a) \doteq F'(x_n)(x_n - x_a) \ .$$

2. Bedingung: $\|A_n - A_a\|_F = \min\{ \|B - A_a\|_F \mid B \in \mathbb{R}^{n\times n} : B(x_n - x_a) = F(x_n) - F(x_a) \}$,
eine Bedingung, die offensichtlich starke Änderungen beim Übergang von $\widetilde{F_a}$ zu $\widetilde{F_n}$ vermeidet.

(5.1.10) Bemerkung
A_n ist durch die beiden Bedingungen $A_n(x_n - x_a) = F(x_n) - F(x_a)$ und
$\|A_n - A_a\|_F = \min\{ \|B - A_a\|_F \mid B \in \mathbb{R}^{n\times n} : B(x_n - x_a) = F(x_n) - F(x_a) \}$ eindeutig bestimmt. Mit $s_a = x_n - x_a$ gilt

$$A_n = A_a + \frac{1}{{s_a}^t s_a}\left(F(x_n) - F(x_a) - A_a s_a \right) {s_a}^t \ .$$

Beweis:

$F(x_n) - F(x_a) = B(x_n - x_a)$ eingesetzt ergibt

$$A_n - A_a = \frac{1}{{s_a}^t s_a} \left(B - A_a \right) s_a {s_a}^t \ ,$$

und damit

$$\|A_n - A_a\|_F \le \|B - A_a\|_F \quad \text{wegen} \quad \|s_a {s_a}^t\|_F = {s_a}^t s_a \ .$$

Der Minimumpunkt in diesem Bestapproximationsproblem in der Frobenius–Norm ist eindeutig bestimmt.

□

Wegen $rang\ y\ s^t = 1$ für $s \neq 0, y \neq 0$ heißt (5.1.10) *Broyden–Rang 1–Updating–Formel.* Das zugehörige Quasi–Newton–Verfahren ist *lokal superlinear konvergent* . Den allerdings deutlich schwierigeren Beweis (als für (5.1.7)) findet man z.B. in *Dennis, J.E., Schnabel, R.:* Numerical methods for unconstrained optimization and nonlinear equations. sect. 8.2. Academic Press 1983.

(5.1.11) Bemerkung (Bedeutung superlinearer Konvergenz)
Sei $\lim_{k\to\infty} x^{(k)} = x^*$ superlinear konvergent. Dann existieren $\underline{c} > 0$ und $\overline{c} > 0$ mit

$$\underline{c}\ \|x^{(k+1)} - x^{(k)}\| \leq \|x^{(k)} - x^*\| \leq \overline{c}\ \|x^{(k+1)} - x^{(k)}\|\ ,\ k \text{ hinreichend groß}\ .$$

Beweis als Übungsaufgabe. □

Für nur linear konvergente Iterationsverfahren kann $\|x^{(k+1)} - x^{(k)}\|$ sehr klein sein, obwohl $\|x^* - x^{(k)}\|$ noch sehr groß ist. Auch wenn man $\overline{c}$ in (5.1.11) nicht kennt und diese Konstante sehr groß sein kann, ist bei superlinearer Konvergenz $\|x^{(k+1)} - x^{(k)}\| \doteq tol$ ein – zumindest asymptotisch – 'vernünftiges' Abbruchkriterium (neben einem kleinen 'Defekt', etwa $\|F\big(x^{(k+1)}\big)\| < eps$ im Vergleich zu $\|F\big(x^*\big)\| = 0$) .

5.2 Iterationsverfahren für lineare Gleichungssyteme

Vorgehen und grundlegende Eigenschaften

Gegeben ist die quadratische Matrix $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$. Gesucht ist $x \in \mathbb{R}^n$ mit $Ax = b$. Es gibt viele Möglichkeiten dieses Problem in eine Fixpunktgleichung umzuwandeln. Betrachten wir die Zerlegung

$$(i) \quad A = L + D + R\ ,$$

wobei $D = diag(A_{11}, A_{22}, ..., A_{nn})$, L der strikt untere Teil von A ($L_{ij} = A_{ij}$ für $j < i$ und $L_{ij} = 0$ sonst) und R der strikt obere Teil von A ($R_{ij} = A_{ij}$ für $j > i$ und $R_{ij} = 0$ sonst) ist. Dann ist naheliegend

$$Dx = b - Lx - Rx \quad , \quad Dx_{neu} = b - Lx_{alt} - Rx_{alt} \quad bzw. \quad Dx_{neu} = b - Lx_{neu} - Rx_{alt}.$$

Auflösen nach x_{neu} kann man, falls $A_{ii} \neq 0, 1 \leq i \leq n$. Damit erhält man die beiden folgenden Verfahren

(5.2.1) Gesamtschrittverfahren (nach Jacobi (1804-1841))
Sei $A_{ii} \neq 0, 1 \leq i \leq n$, und A nach (i) zerlegt. Dann liefert das Gesamtschrittverfahren die Iteration

$$Dx_{neu} = b - Lx_{alt} - Rx_{alt}.$$

□

Setzt man die bereits neu berechneten (hoffentlich verbesserten) Komponenten ein, erhält man das sog.

(5.2.2) Einzelschrittverfahren (nach Gauß (1777–1855) und Seidel (1821–1896))
Sei $A_{ii} \neq 0, 1 \leq i \leq n$, und A nach (i) zerlegt. Dann liefert das Einzelschrittverfahren die Iteration

$$Dx_{neu} = b - Lx_{neu} - Rx_{alt},$$

die man rekursiv für $(x_{neu})_i$, $i = 1, 2, ..., n$ auflöst.

□

Analog kann man rekursiv für $(x_{neu})_i$, $i = n, n-1, ..., 1$

$$Dx_{neu} = b - Lx_{alt} - Rx_{neu}$$

lösen, und auch das ganze 'symmetrisieren' durch 'einmal vorwärts, einmal rückwärts'.

□

Zur Konvergenzanalyse betrachtet man die 'Fixpunktiteration'

(5.2.3)(i) $$x^{(k+1)} := Tx^{(k)} + y \ , \ k \in \mathbb{N}_0;$$

mit gegebenem $y \in \mathbb{R}^n$ und $T \in \mathbb{R}^{n \times n}$. Für das Gesamtschrittverfahren (5.2.1) gilt in dieser Notation

$$T_J = -D^{-1}(L + R) \quad , \quad y = D^{-1}b \ ,$$

für das Einzelschrittverfahren (5.2.2) gilt

$$T_{GS} = -(L + D)^{-1}R \quad , \quad y = (L + D)^{-1}b \ .$$

$x^* \stackrel{?}{=} \lim\limits_{k \to \infty} x^{(k)}$ ist dann notwendig Lösung der Gleichung

$$x^* = Tx^* + y \ .$$

(5.2.3) Satz (Konvergenz für Kontraktionen)
Betrachte $\left(x^{(k)}\right)_{k \in \mathbb{N}_0}$ nach (5.2.3)(i), und eine gegebene Norm $\|.\|$ auf $\mathbb{R}^n$. Dann gilt

(a) Ist $\|T\|_{ind} < 1$, so konvergiert $\left(x^{(k)}\right)_{k \in \mathbb{N}_0}$ gegen x^* mit $x^* = Tx^* + y$.

(b) In (a) gelten mit $c := \|T\|_{ind}$ die Fehlerabschätzungen

$$\|x^{(k)} - x^*\| \ \leq \ \frac{c^k}{1-c} \, \|x^{(1)} - x^{(0)}\| \ , \ k \in \mathbb{N}_0 \ ,$$

$$\|x^{(k)} - x^*\| \ \leq \ \frac{c}{1-c} \, \|x^{(k)} - x^{(k-1)}\| \ , \ k \in \mathbb{N}_0 \ .$$

(c) Konvergiert $\left(x^{(k)}\right)_{k \in \mathbb{N}_0}$ nach (5.2.3)(i) für jedes beliebige $x^{(0)}$ und y, so ist dafür notwendig und hinreichend, daß der Spetralradius $\rho(T) < 1$ ist ($\rho(T) \equiv \max\limits_{\lambda \in S(T)} |\lambda|$).

Beweis:

Mit $G(x) := Tx + y$ gilt $x^{(k+1)} = G(x^{(k)})$ und $x^* = G(x^*)$. Wegen

$$\|G(u) - G(v)\| = \|T(u - v)\| \leq \|T\|_{ind} \|u - v\|$$

ist in (a) $\left(x^{(k)}\right)_{k \in \mathbb{N}_0}$ eine Cauchyfolge (vgl. die folgende Abschätzung für $\|x^{(k+m+1)} - x^{(k)}\|$), deren Limes x^* die Gleichung $x^* = G(x^*)$ erfüllt.

◇

Zum Nachweis der Abschätzungen in (b) beachte man, daß mit der Abschätzung $\|x^{(j+1)} - x^{(j)}\| \leq c^j \|x^{(1)} - x^{(0)}\|$ folgt

$$\|x^* - x^{(k)}\| = \lim_{m\to\infty} \|x^{(k+m+1)} - x^{(k)}\| = \lim_{m\to\infty} \| \sum_{j=k}^{k+m} \left(x^{(j+1)} - x^{(j)}\right) \| \leq$$

$$\leq \sum_{j=k}^{\infty} c^j \|x^{(1)} - x^{(0)}\| = c^k \frac{1}{1-c} \|x^{(1)} - x^{(0)}\|.$$

Aus der Abschätzung $\|x^{(k+j+1)} - x^{(k+j)}\| \leq c^j \|x^{(k+1)} - x^{(k)}\|$ folgt analog

$$\|x^* - x^{(k)}\| \leq \|x^{(k+1)} - x^{(k)}\| \sum_{j=0}^{\infty} c^j \leq c\|x^{(k)} - x^{(k-1)}\| \frac{1}{1-c}.$$

◇

(c) Daß $\rho(T) < 1$ hinreichend ist, folgt aus (a) auf Grund der Tatsache, daß

$$\rho(T) = \inf\{\|T\|_{ind} : \|.\| \text{ Vektornorm auf } \mathbb{R}^n\}$$

(vgl. z.B. *Hämmerlin, Hoffmann*). Gilt andererseits $\rho(T) \geq 1$, so kann man leicht geeignet konstruierte nicht konvergente Iterationsfolgen angeben.

□

(5.2.3)(c) klärt die Konvergenzsituation theoretisch. Natürlich kann $\|T\|_{ind}$ bzgl. einer Norm größer 1 sein und bzgl. einer anderen Norm kleiner 1:

$$T = \begin{bmatrix} 1/2 & 1/4 \\ 3/4 & 1/8 \end{bmatrix}, \quad \|T\|_1 = 1.25\ , \quad \|T\|_\infty = 0,875\ .$$

(5.2.4) Folgerung
Hinreichend für die Konvergenz des Gesamtschrittverfahrens ist jedes der beiden Kriterien

(a) $\sum_{j=1, j\neq i}^{n} |A_{ij}| < |A_{ii}|\ ,\ i = 1, ..., n$; sog. *starkes Zeilensummenkriterium* ;

(b) $\sum_{i=1, i\neq j}^{n} |A_{ij}| < |A_{jj}|\ ,\ j = 1, ..., n$; sog. *starkes Spaltensummenkriterium* .

Beweis:

(a) $T_{i.} = -A_{ii}^{-1}[A_{i1}...A_{i\,i-1}\,0\,A_{i\,i+1}...A_{in}] \implies$
$\|T\|_\infty = \max_i \|T_{i.}\|_1 = \max_i \left(\frac{1}{|A_{ii}|} \sum_{j=1, j\neq i}^{n} |A_{ij}|\right) < 1$;

(b) analog zu (a) gilt $\|T\|_1 = \max_j \|T.j\|_1 = \max_j \left(\frac{1}{|A_{jj}|} \sum_{i=1, i\neq j}^{n} |A_{ij}|\right) < 1$.

□

Bezeichnung
A mit (5.2.4)(a) oder (5.2.4)(b) heißt *diagonaldominant.*

□

Der arithmetische Aufwand für k Schritte eines iterativen Verfahrens beträgt für beliebige $(n \times n)$–Matrizen $O(k \cdot n^2)$ im Vergleich zu $O(n^3)$ für direkte Lösungsverfahren. Iterative Verfahren sind also von Vorteil, wenn der Iterationsschritt z.B. nur $O(n)$ kostet – was etwa bei Bandmatrizen der Fall ist. Eine zweite Möglichkeit ist die Kenntnis einer sehr guten Startnäherung, was k klein hält. Ein Beispiel für die letztgenannte Situation ist

Nachiteration

Dabei wird eine Näherungslösung $x^{(0)}$, die mit einem direkten Verfahren in Gleitpunktarithmetik berechnet wurde, verbessert durch

(5.2.5) Nachiteration (zur Genauigkeitsverbesserung)
Gegeben $x^{(0)}$ mit $\widetilde{A}x^{(0)} = \widetilde{b}$ und $A \doteq \widetilde{L}\widetilde{R} =: \widetilde{A}$. Dann führe man folgende Schritte durch:

(1) $r^{(0)} := b - Ax^{(0)}$ *doppelt genau* rechnen (wegen $Ax^{(0)} \doteq b$ ist Auslöschung zu erwarten);

(2) bestimme $d^{(0)}$ mit $\widetilde{L}\widetilde{R}d^{(0)} = r^{(0)}$;

(3) $x^{(1)} := x^{(0)} + d^{(0)}$;

Iteriere eventuell dieses Vorgehen mit $x^{(0)} := x^{(1)}$.

□

(5.2.5) ist eine Iteration der Form (5.2.3)(i) mit

$$T := E - \widetilde{A}^{-1}A \quad , \quad y := \widetilde{A}^{-1}b.$$

Es gilt

(5.2.6) Bemerkung
Für die Iterationsmatrix T in (5.2.5) gilt

$$\|T\| \leq cond(\widetilde{A})\ \varepsilon_0{}' \quad \text{, falls } \frac{\|\widetilde{A} - A\|}{\|\widetilde{A}\|} \leq \varepsilon_0{}' \,.\ \varepsilon_0{}' \text{ ist dabei klein analog (1.1.8).}$$

Beweis:

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

□

(5.2.5') Beispiel zu (5.2.5)

$$A = \begin{bmatrix} 7.000 & 6.990 \\ 4.000 & 4.000 \end{bmatrix}\ ,\ b = \begin{bmatrix} 34.97 \\ 20.00 \end{bmatrix}\ ,\ x^* = \begin{bmatrix} 2 \\ 3 \end{bmatrix}\ .$$

Die LR–Zerlegung in 4–stelliger Gleitpunktarithmetik ergibt

$$\widetilde{L}\widetilde{R} = \begin{bmatrix} 1.000 & \\ 0.5714 & 1.000 \end{bmatrix}\begin{bmatrix} 7.000 & 6.990 \\ & 0.006 \end{bmatrix}\ ;$$

Durchführung der Nachiteration (5.2.5) (mit $r^{(k)}$ doppelt genau) als Übungsaufgabe.

Zum *Vergleich von Gesamt(Jacobi)–und Einzelschrittverfahren(Gauß–Seidel)* die folgende

(5.2.7) Bemerkung
Gilt das starke Zeilensummenkriterium (5.2.4)(a), so gilt

$$\|T_{GS}\|_\infty \leq \|T_J\|_\infty \ < \ 1 \ .$$

□

Denn unter Beachtung von $\|T_J\|_\infty < 1$ folgt rekursiv für $k = 1, 2, ..., n$ die Abschätzung $|(T_{GS}\, x)_k| \leq \|T_J\|_\infty \|x\|_\infty$. Damit ist das starke Zeilensummenkriterium auch hinreichend für die Konvergenz des Einzelschrittverfahrens.

Konvergenzverbesserung durch Relaxation

Zum Abschluß möchte ich für das Einzelschrittverfahren noch eine Strategie vorstellen, die Norm der Iterationsmatrix zu verkleinern durch sog.

(5.2.8) Relaxation (beim Einzelschrittverfahren[1])
Die Voraussetzungen und Bezeichnungen seien wie in (5.2.3). Sei ein sog. *Relaxationsparameter* $\omega > 0$ gewählt. Mit gegebenem x_{alt} lautet dann die Iteration:

Bestimme x_{neu} rekursiv für $i = 1, ..., n$ durch Lösen der Gleichung

$A_{ii}\hat{x}_i = b_i - (Lx_{neu})_i - (Rx_{alt})_i$ *Gauß–Seidel–Schritt*;

und setze

$x_{neu,i} := x_{alt,i} + \omega(\hat{x}_i - x_{alt,i})$ *(sukzessiver) Relaxationsschritt.*

□

Bemerkung: $x_{neu}\big|_{\omega=1} = \hat{x}$ ist gerade die neue Iterierte des Einzelschrittverfahrens. Ist $\omega > 1$, spricht man daher von *Überrelaxation* – auch *SOR*–Verfahren genannt, für $\omega < 1$ von *Unterrelaxation.*

Wegen $Lx_{neu} + D\hat{x} + Rx_{alt} = b$ und $x_{neu} = x_{alt} + \omega(\hat{x} - x_{alt})$ hat das Verfahren (5.2.8) (nach Elimination von $\hat{x}$) in der Form (5.2.3)(i) die Gestalt

$$(5.2.9)(i) \qquad T_\omega = (D + \omega L)^{-1}\big((1-\omega)D - \omega R\big) \quad ; \quad y = \omega(D + \omega L)^{-1} b \ .$$

Ein *optimaler Relaxationsparameter* $\omega^* > 0$ ist einer mit dem kleinsten Spektralradius der zugehörigen Iterationsmatrix

$$\rho(T_{\omega^*}) = \min_{\omega > 0} \rho(T_\omega) \ .$$

Einen ersten Hinweis auf die Lage von ω^* liefert

(5.2.9) Satz
Sei $A = L + D + R$ wie oben und $A_{ii} \neq 0, i = 1, ..., n$. Dann gilt für T_ω in (5.2.9)(i)

$$\rho(T_\omega) \geq |1 - \omega| \ ;$$

d.h. Relaxation ist sinnvoll höchstens für $0 < \omega < 2$.[2]

[1] für das Gesamtschrittverfahren wird die Relaxation z.B. in *Schaback, Wendland* besprochen
[2] deshalb wurde $\omega \leq 0$ garnicht betrachtet

Beweis:

$\det(T_\omega) = \det\left((D+\omega L)^{-1}\right)\det\left((1-\omega)D - \omega R\right) = \frac{1}{\det(D)}\det((1-\omega)D) = (1-\omega)^n$, da L bzw. R strikt untere bzw. obere Dreiecksmatrizen sind. Wegen $\det(T_\omega) = \prod_{i=1}^n \lambda_i$, $\lambda_i \in S(T_\omega)$ in algebraischer Vielfachheit gezählt, ergibt die Annahme $\rho(T_\omega) < |1-\omega|$ einen Widerspruch.

□

Unter den Voraussetzungen von (5.2.7) gilt zumindest in einer Umgebung von $\omega = 1$ die Ungleichung $\rho(T_\omega) < 1$. Unter speziellen Voraussetzungen an A sind genauere qualitative Untersuchungen dieser Abhängigkeit möglich. Häufig hat

$$\omega \longrightarrow \rho(T_\omega)$$

ein Aussehen wie in Abbildung 5.1 (für einen Beweis im Fall 'A symmetrisch positiv definit' s.

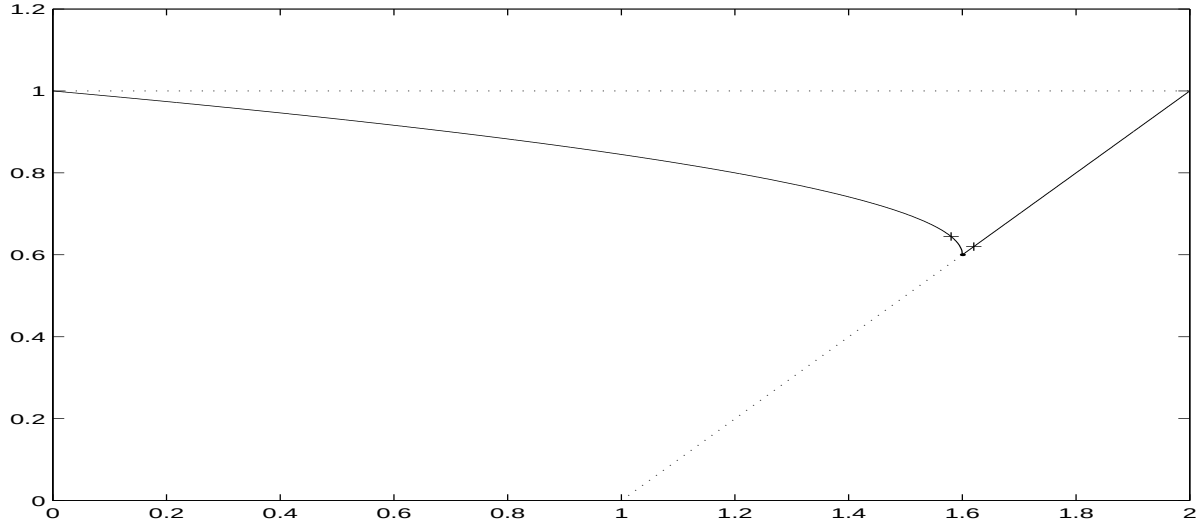

Abbildung 5.1: Spektralradius vs. Relax.parameter ω: A symmetrisch

z.B. *Hämmerlin, Hoffmann*; eine klassische Monografie zum Thema von Abschnitt 5.2 ist *Varga, R.S.:* Matrix iterative Analysis, 2. rev. expanded ed., Series in computational mathematics Vol. 27, Springer Verlag 2000).
Für 'schiefsymmetrisches' A, d.h. in der Zerlegung $A = L + D + R$ gilt $L = -R^t$, ergibt sich die Situation von Abbildung 5.2 Die klaren Konsequenzen sind für

A symmetrisch: Überrelaxation mit besser über- als unterschätztem (quasi–) 'optimalem' Relaxationsparameter;

A 'schiefsymmetrisch': Unterrelaxation mit besser unter- als überschätztem (quasi–) 'optimalem' Relaxationsparameter.

Noch ein Hinweis zur Notation: häufig wird in der Literatur A zerlegt in $A = -L + D - R$ mit dem offensichtlichen Vorteil, daß dann die Iterationsmatrix nicht wie bei mir negative Vorzeichen hat; für das Programmieren ist wohl die von mir gewählte Notation 'natürlicher'.

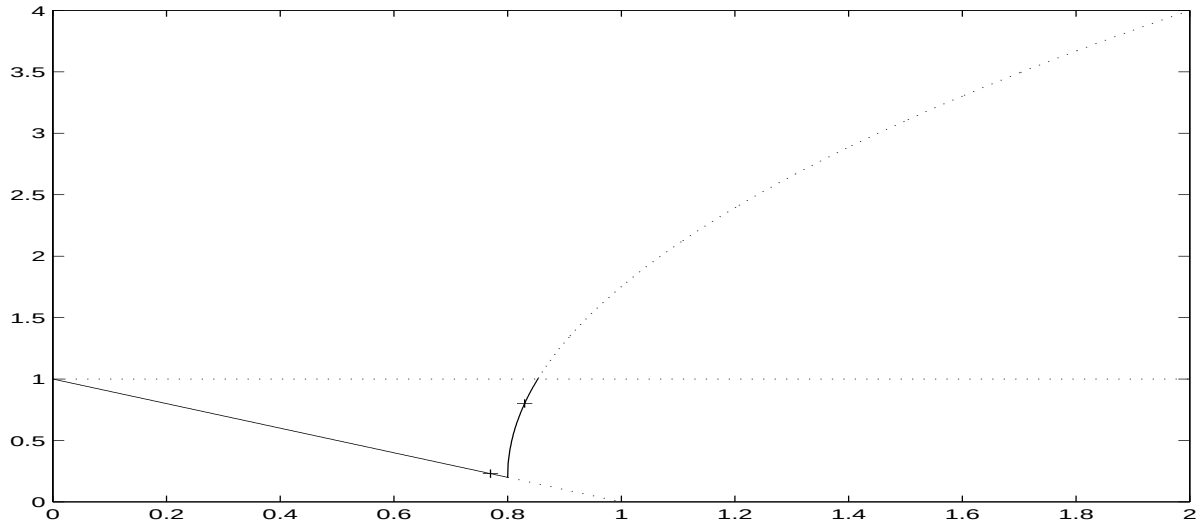

Abbildung 5.2: Spektralradius vs. Relax.parameter ω: A schiefsymmetrisch

Verfahren der konjugierten Gradienten

Das Verfahren der konjugierten Gradienten beruht auf der Tatsache, daß für $q(x) \equiv \frac{1}{2}x^{\mathrm{t}}Qx + c^{\mathrm{t}}x$ mit einer symmetrischen positiv definiten Matrix Q wegen $grad\ q(x) = Qx + c$ notwendig und hinreichend für $\min\limits_{x \in \mathbb{R}^n} q(x) = q(x^\star)$ ist, daß $Qx^\star + c = 0$ gilt. Das cg –Verfahren bestimmt iterativ eine Folge von Näherungen an dieses Minimum durch
– Wahl einer Abstiegsrichtung s_{alt} im aktuellen Iterationspunkt x_{alt}, und
– anschließende 'eindimensionale' Minimierung von q in dieser Richtung.

(5.2.10) *cg* –Verfahren *(zur Lösung großer symmetrischer positiv definiter linearer Gleichungssysteme)*
Gegeben: $Q \in \mathbb{R}^{n \times n}$, $c \in \mathbb{R}^n$ *symmetrisch, positiv definit*;

Gesucht: $x^\star \in \mathbb{R}^n$ *mit* $Q\,x^\star + c = 0$;
bzw. $x^\star = arg\ min\ q(x)$, *wobei* $q(x) \equiv \frac{1}{2}x^t Qx + c^t x$.

Bestimme eine Startnäherung x_{alt} *(gewöhnlich* $x_{alt} = 0$ *) , und berechne*

(1) $g_{alt} := Qx_{alt} + c$ ($g_{alt} \equiv grad\ q(x_{alt})$);
$s_{alt} := -g_{alt}$;

Solange $g_{alt} \neq 0$ (*i.e.* ${g_{alt}}^t g_{alt} = \|g_{alt}\|_2^{\,2} \overset{?}{>} 0$) , *bestimme neue Näherung* x_{neu} *durch*

(2) $\alpha := ({g_{alt}}^t\, g_{alt})/({s_{alt}}^t\, Qs_{alt})$ (*theoretisch* $\alpha := -({g_{alt}}^t\, s_{alt})/({s_{alt}}^t\, Qs_{alt})$)
(3) $x_{neu} := x_{alt} + \alpha\, s_{alt}$
(4) $g_{neu} := g_{alt} + \alpha\, Qs_{alt}$ (*theoretisch* $g_{neu} := grad\ q(x_{neu})$)
(5) $\beta := ({g_{neu}}^t\, g_{neu})/({g_{alt}}^t\, g_{alt})$ (*theoretisch* $\beta := ({g_{neu}}^t\, Qs_{alt})/({s_{alt}}^t\, Qs_{alt})$)
(6) $s_{neu} := -g_{neu} + \beta\, s_{alt}$;
$(x_{alt}, g_{alt}, s_{alt}) := (x_{neu}, g_{neu}, s_{neu})$.

Andernfalls ist $x^\star := x_{alt}$ *Lösung von* $Qx + c = 0$.

□

(5.2.11) Bemerkung
In Algorithmus (5.2.10) ergibt sich

(a) die Wahl von α aus der Bedingung $q(x_{alt} + \alpha s_{alt}) = \min\limits_{t \in \mathbb{R}} q(x_{alt} + t s_{alt})$, der sog. 'eindimensionalen' Minimierung,

(b) die Wahl von β in $s_{neu} = -grad\ q(x_{neu}) + \beta s_{alt}$ aus der Bedingung ${s_{neu}}^t Q s_{alt} = 0$, der sog. 'Q-Konjugiertheit' aller Abstiegsrichtungen s.

Beweis:

(a) Mit $\varphi(t) := q(x_{alt} + t s_{alt})$ gilt notwendig (und hier auch hinreichend)

$$0 = \varphi(t)'\big|_{t=\alpha} =< grad\ q(x_{alt}+ts_{alt})\big|_{t=\alpha}, s_{alt} >=< +\alpha s_{alt}), s_{alt} >=< grad\ q(x_{alt}, s_{alt} > +\alpha < s_{alt}, s_{alt} > .$$

(b) Für $s_{neu} = -grad\ q(x_{neu}) + \beta s_{alt}$ gilt

$$< s_{neu}, Q s_{alt} >= 0 \Longleftrightarrow \beta =< grad\ q(x_{neu}), Q s_{alt} > / < s_{alt}, Q s_{alt} >$$

□

Seien die von der iterativen Anwendung des *cg* –Verfahrens erzeugten Größen entsprechend mit
$\{x_0, x_1, x_2, ...\}$, $\{g_0, g_1, s_2, ...\}$ und $\{s_0, s_1, s_2, ...\}$
bezeichnet. Dann gelten die folgenden

(5.2.12) Eigenschaften des *cg* –Verfahrens
Für alle $k \in \mathbb{N}_0$ mit $g_i \neq 0,\ 0 \leq i \leq k$, gilt

(a) ${g_k}^t s_i = 0\ ,\ 0 \leq i \leq k-1$;

(b) ${s_k}^t Q s_i = 0\ ,\ 0 \leq i \leq k-1$;

(c) $span\big(\{g_0, g_1, ..., g_k\}\big) = span\big(\{g_0, Qg_0, Q^2 g_0, .., Q^k g_0\}\big)$;

(d) $span\big(\{s_0, s_1, ..., s_k\}\big) = span\big(\{g_0, Qg_0, Q^2 g_0, .., Q^k g_0\}\big)$.

Beweis:

Wir zeigen (a) – (d) gleichzeitig durch Induktion. $k = 0$ ist klar wegen $s_0 = -g_0$.
$k \rightsquigarrow k+1$:
(a) ${g_{k+1}}^t s_k = 0$ nach (2) in (5.2.10); für $i \leq k-1$ gilt
${g_{k+1}}^t s_i = (g_k + \alpha_k Q s_k)^t s_i = {g_k}^t s_i + \alpha_k {s_k}^t Q s_i = 0$
nach Induktionsvoraussetzung (a) und (b).

(b) ${s_{k+1}}^t Q s_k = -{g_{k+1}}^t Q s_k + \beta_k {s_k}^t Q s_k = 0$ nach (5) in (5.2.10); für $i \leq k-1$ gilt nach Induktionsvoraussetzung (d)
$Q s_i \in span\big(\{Q Q^j g_0 \ :\ 0 \leq j \leq k-1\}\big) \subset span\big(\{s_0, ..., s_k\}\big)$.
Damit folgt für $i \leq k-1$ nach Induktionsvoraussetzung (b) ${s_k}^t Q s_i = 0$
${s_{k+1}}^t Q s_i = -{g_{k+1}}^t Q s_i + \beta_k {s_k}^t Q s_i = -{g_{k+1}}^t (Q s_i) = -{g_{k+1}}^t (\sum_{j \leq k} c_j s_j) = 0$ nach (a).

(c) $g_{k+1} = Q x_{k+1} + c = Q x_k + \alpha_k Q s_k + c = g_k + \alpha_k Q s_k$. Also gilt nach (c) und (d) der Induktionsvoraussetzung $g_{k+1} \in span\big(\{Q^j g_0 \ :\ 0 \leq j \leq k\}\big) + span\big(\{Q Q^j g_0 \ :\ 0 \leq j \leq k\}\big)$.
Damit gilt

$span\big(\{g_0, g_1, ..., g_{k+1}\}\big) \subset span\big(\{Q^j g_0 \ : \ 0 \le j \le k+1\}\big)$.
Aus der Annahme, daß die beiden Mengen nicht gleich sind, folgt aus Dimensiondgründen, daß $g_{k+1} \in span\big(\{g_0, g_1, ..., g_k\}\big) = span\big(\{s_0, s_1, ..., s_k\}\big)$. Da (a) für $k+1$ bereits gezeigt wurde, folgt $g_{k+1} = 0$ im Widerspruch zur Voraussetzung.

(d) $s_{k+1} = -g_{k+1} + \beta_k s_k$ ist nach (c) linear unabhängig von $\{s_0, ..., s_k\}$. Wieder aus Dimensionsgründen gilt damit in $span\big(\{s_0, s_1, ..., s_{k+1}\}\big) \subset span\big(\{Q^j g_0 \ : \ 0 \le j \le k+1\}\big)$ die Gleichheit.

□

$K(Q; g_0) = span\big(\{g_0, Qg_0, Q^2 g_0, .., Q^k g_0, ..\}\big)$ heißt der *Krylov–Raum* von Q zu g_0.

(5.2.13) Folgerung
Sei dim $K(Q; g_0) = m$. Dann gilt $g_m = 0$, d.h. es ist $x_m = x^\star$ nach $m \le n$ Schritten.

Beweis:

$g_m \in span\big(\{Q^j g_0 \ : \ 0 \le j \le m\}\big) = span\big(\{Q^j g_0 \ : \ 0 \le j \le m-1\}\big)$, da die Dimension von $K(Q; g_0)$ gleich m ist. Wegen ${g_m}^t s_j = 0, 0 \le j \le m-1$, folgt $g_m = 0$.

□

Besonders wirkungsvoll wird das *cg* –Verfahren durch *Vorkonditionierung* : an Stelle von

$$\arg\min q(x) \quad , \quad q(x) \equiv \frac{1}{2} x^{\mathrm{t}} Q x + c^{\mathrm{t}} x$$

sucht man äquivalent das Minimum von

$$\arg\min \widetilde{q}(x) \quad , \quad \widetilde{q}(x) \equiv \frac{1}{2} x^{\mathrm{t}} \widetilde{Q} x + \widetilde{c}^{\mathrm{t}} x$$

wobei wir mit $C = HH^{\mathrm{t}}$, H nichtsingulär, setzen: $\widetilde{Q} := H^{-1} Q H^{-t}$, $\widetilde{c} = H^{-1} c$.

Ich verweise auf weiterführende Literatur (s. Ende des nächsten Kapitels), wie man C sinnvoll wählen kann/soll.

Kapitel 6

Eigenwertaufgaben bei Matrizen

6.1 Einführung; Vektoriteration

In Kapitel 6 betrachte ich folgende

(6.1.1) Matrix–EWA
Gegeben $A \in \mathbb{K}^{n\times n}$. Gesucht ist $(x,\lambda) \in \mathbb{K}^n \times \mathbb{K}$, $x \neq 0$ mit $Ax = \lambda x$. λ heißt Eigenwert von A, x zugehöriger Eigenvektor; $S(A)$ bezeichnet die Menge aller Eigenwerte.

□

Wenn eine Basis des $\mathbb{K}^n$ von Eigenvektoren $\{x_1, ..., x_n\}$ existiert, heißt A *diagonalisierbar oder digonalähnlich.* Dann gilt mit $Ax_i = \lambda_i x_i, i = 1, .., n$, und $X = [x_1|x_2|...|x_n]$

$$X^{-1}AX = diag(\lambda_1, ..., \lambda_n).$$

(6.1.2) Vektoriteration (für eine Matrix A)
Gegeben $A \in \mathbb{K}^{n\times n}$, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. Wähle $y^{(0)} \in \mathbb{K}^n$ (als Startnäherung für einen Eigenvektor von A) mit $\|y^{(0)}\|_\infty = 1$, und berechne

$$\widetilde{y}^{(k)} := Ay^{(k-1)} \quad , \quad y^{(k)} := \widetilde{y}^{(k)}/\|\widetilde{y}^{(k)}\|_\infty \quad \text{(oder irgendeine andere Normierung)} .$$

□

(6.1.3) Satz
$A \in \mathbb{K}^{n\times n}$ sei diagonalähnlich mit einem einfachen dominanten Eigenwert λ_1, d.h. es gelte

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| ,$$

wobei $\{\lambda_1 \ldots \lambda_n\}$ die Eigenwerte von A und $x_1, \ldots x_n$ eine zugehörige Basis von Eigenvektoren ist. Dann gilt für jedes $y^{(0)} = \sum_{i=1}^n c_i x_i$, $\|y^{(0)}\|_\infty = 1$, mit $c_1 \neq 0$, für die Folge $(y^{(k)})_{k\in\mathbb{N}_0}$ aus (6.1.2)

$$\lim_{k\to\infty} \frac{\langle \widetilde{y}^{(k)}, e_i\rangle}{\langle y^{(k-1)}, e_i\rangle} = \lambda_1 .$$

für alle i : $< x_1, e_i > \neq 0$.

Beweis:

$$\Big(\prod_{j=1}^{k-1}\|\widetilde{y}^{(j)}\|_\infty\Big)\langle\widetilde{y}^{(k)},e_i\rangle = \langle A^k y^{(0)},e_i\rangle$$

$$= {\lambda_1}^k c_1 < x_1,e_i > \Big\{1+\sum_{j=2}^{n}\frac{c_j}{c_1}\Big(\frac{\lambda_j}{\lambda_1}\Big)^k\Big(\frac{< x_j,e_i >}{< x_1,e_i >}\Big)\Big\}.$$

Nach Voraussetzung gilt $\lim\limits_{k\to\infty}\Big(\frac{\lambda_j}{\lambda_1}\Big)^k = 0\ ,\ 2\le j\le n$, und daher $\langle\widetilde{y}^{(k)},e_i\rangle \neq 0$ und folglich $\langle y^{(k)},e_i\rangle \neq 0$ für hinreichend großes k. Für hinreichend großes k folgt weiter

$$\lim_{k\longrightarrow\infty}\frac{\Big(\prod\limits_{j=1}^{k-1}\|\widetilde{y}^{(j)}\|_\infty\Big)\langle\widetilde{y}^{(k)},e_i\rangle}{\Big(\prod\limits_{j=1}^{k-2}\|\widetilde{y}^{(j)}\|_\infty\Big)\langle\widetilde{y}^{(k-1)},e_i\rangle} = \lambda_1.$$

□

Aus dem Beweis von (6.1.2) erkennt man, daß

$$\frac{\langle\widetilde{y}^{(k)},e_i\rangle}{\langle y^{(k-1)},e_i\rangle} = \lambda_1 + O\Big(\Big|\frac{\lambda_2}{\lambda_1}\Big|^k\Big)$$

gilt. Die Vektoriteration ist also nur günstig, falls $\Big|\frac{\lambda_2}{\lambda_1}\Big| << 1$ gilt. Gute Verfahren – wie das *QR–Verfahren mit shift* – versuchen diese Voraussetzung implizit zu erfüllen. Explizit gegeben ist diese Situation bei der sog. *Inversen Iteration*

(6.1.4) Inverse Iteration (von Wielandt)
Sei A diagonalisierbar, und eine Näherung $\widetilde{\lambda}$ von $\lambda\in S(A)$ gegeben mit

$$0 \;<\; |\widetilde{\lambda}-\lambda| \;<\; \frac{1}{2}\min_{\mu\in S(A)\setminus\{\lambda\}}|\mu-\lambda|\ .$$

Mit $\lambda_{alt} := \widetilde{\lambda}$ und (irgendeinem) y_{alt} mit $\|y_{alt}\|_2 = 1$ wiederhole man die Iteration

(1) bestimme y_{neu} $(\ A-\lambda_{alt}E\ )\ y_{neu} = y_{alt}\ ,\ \|y_{neu}\|_2 = 1$;

(2) berechne die neue (hoffentlich bessere) Näherung λ_{neu} von λ

$$\lambda_{neu} := \frac{\langle Ay_{neu},y_{neu}\rangle}{\langle y_{neu},y_{neu}\rangle} = \lambda_{alt}+\frac{\langle y_{alt},y_{neu}\rangle}{\langle y_{neu},y_{neu}\rangle}\ .$$

$(\lambda_{alt},y_{alt}) := (\lambda_{neu},y_{neu})$.

□

(6.1.5) Bemerkung
Die iterative Durchführung von (6.1.4)(1) ist gerade die Vektoriteration für die Matrix $(A - \widetilde{\lambda}E)^{-1}$ – daher der Name 'inverse Iteration'. Ist $\widetilde{\lambda}$ eine gute Näherung für $\lambda \in S(A)$, d.h. gilt

$$0 \;<\; |\lambda - \widetilde{\lambda}| \;<<\; \min_{\mu \,\in\, S(A) \setminus \{\lambda\}} |\mu - \widetilde{\lambda}| \;.$$

so folgt

$$\max_{\nu \,\in\, S(\,(A - \widetilde{\lambda}E)^{-1}\,) \setminus (\lambda - \widetilde{\lambda})^{-1}} \left|\frac{\nu}{(\lambda - \widetilde{\lambda})^{-1}}\right| \;<<\; 1 \;;$$

d.h. die inverse Iteration konvergiert 'recht gut'.

□

Zur Beurteilung sowohl der Güte berechneter Näherungen als auch des Einflußes von Störungen z.B. durch Gleitpunktarithmetik dient

(6.1.6) Satz (Störungssatz für Eigenwerte)
Sei $A \in \mathbb{K}^{n\times n}$ diagonalisierbar, etwa $X^{-1}AX = diag(\lambda_1, ..., \lambda_n) \equiv: \Lambda$, und $\|.\| : \mathbb{K}^n \longrightarrow \mathbb{R}$ 'monoton', i.e. $\|D\|_{ind} = \varrho(D) \equiv \max_{\lambda\in S(D)} |\lambda|$ für jede Diagonalmatrix D.

(a) Für $\widetilde{x} \in \mathbb{K}^n, \widetilde{x} \neq 0$, und $\widetilde{\lambda} \in \mathbb{K}$ gilt

$$\min_{\lambda \,\in\, S(A)} |\widetilde{\lambda} - \lambda| \;\leq\; cond_{ind}(X) \frac{\|A\widetilde{x} - \widetilde{\lambda}\widetilde{x}\|}{\|\widetilde{x}\|} \;.$$

(b) Sei $\Delta A \in \mathbb{K}^{n\times n}$. Dann gilt für jedes $\widetilde{\lambda} \in S(A + \Delta A)$

$$\min_{\lambda \,\in\, S(A)} |\widetilde{\lambda} - \lambda| \;\leq\; cond_{ind}(X)\|\Delta A\|_{ind} \;.$$

Beweis:

(a) Für $\widetilde{\lambda} \in S(A)$ ist die Behauptung klar. Sei also o.E. $\widetilde{\lambda} \notin S(A)$, d.h. $(A - \widetilde{\lambda}E)$ nicht singulär, und $\min_{\lambda\in S(A)} |\widetilde{\lambda} - \lambda| > 0$.

$$\Longrightarrow \quad A\widetilde{x} - \widetilde{\lambda}\widetilde{x} \;=\; (X\Lambda X^{-1} - \widetilde{\lambda}E)\widetilde{x} \;=\; X(\Lambda - \widetilde{\lambda}E)X^{-1}\widetilde{x}.$$

$$\begin{aligned}\Longrightarrow \quad \|\widetilde{x}\| \;&=\; \|[X(\Lambda - \widetilde{\lambda}E)X^{-1}]^{-1}(A\widetilde{x} - \widetilde{\lambda}\widetilde{x})\| \\ &\leq\; \|X\|\;\|(\Lambda - \widetilde{\lambda}E)^{-1}\|\;\|X^{-1}\|\;\|A\widetilde{x} - \widetilde{\lambda}\widetilde{x}\| \;.\end{aligned}$$

Da die Norm monoton ist, gilt

$$\|(\Lambda - \widetilde{\lambda}E)^{-1}\| \;=\; \|diag\big((\lambda_1 - \widetilde{\lambda})^{-1}, ..., (\lambda_n - \widetilde{\lambda})^{-1}\big)\| \;=\; \max_i |\lambda_i - \widetilde{\lambda}|^{-1} \;=\; \frac{1}{\min_i |\lambda_i - \widetilde{\lambda}|}$$

Insgesamt folgt

$$\min_i |\lambda_i - \widetilde{\lambda}| \;\leq\; \|X\|\;\|X^{-1}\|\;\frac{\|A\widetilde{x} - \widetilde{\lambda}\widetilde{x}\|}{\|\widetilde{x}\|}$$

◇

(b) Zu $\widetilde{\lambda} \in S(A + \Delta A)$ existiert $y \neq 0$ mit $(A + \Delta A)y = \widetilde{\lambda}y$. Damit gilt für den Defekt von $(\widetilde{\lambda}, y)$ bzgl. der Eigenwertgleichung von A

$$\|Ay - \widetilde{\lambda}y\| \;=\; \|\Delta A\, y\| \;\leq\; \|\Delta A\|\, \|y\| \,.$$

Nach (a) folgt die Behauptung.

□

Wann ist der Verstärkungsfaktor für die absoluten Fehler in (6.1.6) klein?

Bemerkung
(a) $A \in \mathbb{C}^{n\times n}$ heißt *normal*, wenn U unitär existiert mit $U^h A U = \Lambda$ Diagonalmatrix.

(b) A ist normal genau dann, wenn $A^h A = A A^h$ gilt.

(c) Jede symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ ist normal.

□

(6.1.7) Folgerung
(a) Für normale Matrizen A ist die Eigen*wert*bestimmung gut konditioniert bzgl. absoluter Fehler ΔA.
(b) Für normale Matrizen A liefert
(6.1.6)(a) ein Abbruchkriterium für die Vektoriteration;
(6.1.6)(b) ein Abbruchkriterium für das QR–Verfahren (s. (6.2.1)) für Hessenbergmatrizen.

Beweis:

$cond_2(U) = 1 = cond_2(Q)$.

□

6.2 QR–Verfahren

Dieses Verfahren löst die 'volle Eigenwertaufgabe', nämlich alle Eigenwerte und zugehörigen Eigenvektoren zu bestimmen. Erreicht wird dies durch eine Folge von Ähnlichkeitstransformationen

$$A_{alt} \quad \longmapsto \quad X^{-1} A_{alt} X =: A_{neu}$$

mit dem Ziel im Limes eine (obere) Dreiecksmatrix zu erzeugen.

(6.2.1) QR–Verfahren (prinzipielles Vorgehen)
Gegeben $A_{alt} := A \in \mathbb{R}^{n\times n}$. Wiederhole bis zum Abbruch

(a) Bestimme QR–Zerlegung $A_{alt} = QR$ mit Q orthogonal, R obere Dreiecksmatrix ;
(b) $A_{neu} = RQ$; $A_{alt} := A_{neu}$.

□

Klar gilt: $Q^{-1}A_{alt}Q = Q^t(QR)Q = RQ = A_{neu}$.

Eine Steigerung der Konvergenzgüte erreicht man durch Nullpunktsverschiebungen, sog. *shifts*.

(6.2.2) QR–Verfahren mit shift

Gegeben $A_{alt} := A \in \mathbb{R}^{n\times n}$. Wiederhole bis zum Abbruch

(a) Wähle $s_{alt} \in \mathbb{R}$, z.B. $s_{alt} = (A_{alt})_{nn}$ (besser sind abwechselnd die beiden Eigenwerte von $\begin{bmatrix} (A_{alt})_{n-1\ n-1} & (A_{alt})_{n-1\ n} \\ (A_{alt})_{n\ n-1} & (A_{alt})_{n\ n} \end{bmatrix}$, die man auch bei konjugiert komplexen Eigenwerten zu einem reellen sog. Doppelschritt zusammenfassen kann).

(b) Bestimme QR–Zerlegung $(\ A_{alt} - s_{alt}E\ ) = QR$ mit Q orthogonsl, R obere Dreiecksmatrix.

(c) $A_{neu} = RQ + s_{alt}E \quad ; \quad A_{alt} := A_{neu}$.

□

Wieder gilt: $Q^{-1}A_{alt}Q = Q^t(QR + s_{alt}E)Q = RQ + s_{alt}E = A_{neu}$.

Für die arithmetische Komplexität eines Iterationsschrittes in (6.2.1) und (6.2.2) gilt

$O(n^3)$ für beliebiges A,

$O(n^2)$ für (obere) Hessenbergmatrix A.

Deshalb transformiert man die *EWA* für A in einem einmaligen Vorbereitungsschritt in eine äquivalente *EWA* für eine obere Hessenbergmatrix H. H ist obere Hessenbergmatrix, wenn $H_{ij} = 0$ für $j+2 \le i \le n$, $1 \le j \le n$; gilt, d.h. es ist nur eine untere Nebendiagonale $\neq 0$:

$$Q\,A\,Q^t = H \ ,\ Q \text{ orthogonal} \ ,\ H \text{ obere Hessenbergmatrix} .$$

(6.2.3) Vorbereitungsschritt

Bestimme Q als Produkt von Matrizen H_m der Form $H_m = \left[\begin{array}{c|c} E_m & 0 \\ \hline 0 & \widetilde{H}_m \end{array}\right]$ mit $E_m \in \mathbb{R}^{m\times m}$ Einheitsmatrix, wobei die Householdertransformation $\widetilde{H}_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ den 1. Restspaltenvektor $\widetilde{A}_{.1} = [A_{21}A_{31}...A_{n1}]^t$ zu einem Vielfachen von $[\ 1\ 0\ ...\ 0\ ]^t \in \mathbb{R}^{n-1}$ macht. Die anschließende Rechtsmultiplikation mit $\left[\begin{array}{c|c} E_1 & 0 \\ \hline 0 & \widetilde{H}_1 \end{array}\right]^t$ ändert die 1. Spalte nicht. Analog wirkt $\left[\begin{array}{c|c} E_2 & 0 \\ \hline 0 & \widetilde{H}_2 \end{array}\right]$ (von links) und $\left[\begin{array}{c|c} E_2 & 0 \\ \hline 0 & \widetilde{H}_2 \end{array}\right]^t$ (von rechts) auf den 2. Restspaltenvektor, usw. Schließlich ist

$$H_{n-2}H_{n-3}\cdots H_1 A {H_1}^t \cdots {H_{n-3}}^t {H_{n-2}}^t$$

eine zur Ausgangsmatrix A ähnliche obere Hessenbergmatrix.

□

Da die obere Hessenbergform erhalten bleibt bei jedem Iterationsschritt des QR–Verfahrens (mit und ohne shift), betrachte ich den Iterationschritt (6.2.2) für obere Hessenbergmatrizen H_{alt} und H_{neu}. Dann gilt

Bemerkung
Unter geeigneten Voraussetzungen hat das QR–Verfahren mit shift folgende asymptotische Konvergenzgeschwindigkeit:

– für beliebige Matrizen A gilt $\left|\left(H_{neu}\right)_{n\ n-1}\right| \ = \ O\Big(\ \left|\left(H_{alt}\right)_{n\ n-1}\right|^2 \ \Big)$

– für symmetrische Matrizen A gilt $\left|\left(H_{neu}\right)_{n\ n-1}\right| \ = \ O\Big(\ \left|\left(H_{alt}\right)_{n\ n-1}\right|^3 \ \Big)$.

□

Damit unterscheiden sich $\widehat{H}_{neu}$, das aus H_{neu} entsteht durch Abänderung von $\left(H_{neu}\right)_{n\ n-1}$ zu $\left(\widehat{H}_{neu}\right)_{n\ n-1} = 0$, und H_{neu} nur sehr wenig. Nach (6.1.7)(b) ist $\left(H_{neu}\right)_{n\ n}$ also eine sehr gute Näherung für einen Eigenwert von H_{neu}.

Literatur zu diesem Kapitel ist insbesondere

GOLUB, G.H., van LOAN, C.F.: Matrix Computations. Johns Hopkins University Press, versch. Aufl. seit 1983, ISBN 0-8018-3011-7 Pbk.

BUNSE, W., BUNSE–GERSTNER, A.: Numerische Lineare Algebra. Teubner 1985, ISBN 3-519-02067-X.