

Einführung in die Numerik

Vorlesungsskript

Prof. Dr. Lutz Tobiska

Sommersemester 2013

Einleitung

Aufgabenstellung der numerischen Mathematik ist die Entwicklung von Methoden, mit deren Hilfe Lösungen mathematischer Problemstellungen effektiv berechnet bzw. möglichst mit Fehlerangabe approximiert werden können. Numerische Methoden sind der Schlüssel zur Simulation komplexer Näherungsvorgänge auf Rechneranlagen. Man möchte teure Experimente wie z.B. Windkanalversuche bei der Flugzeugkonstruktion oder Festigkeitstests bei Betonkonstruktionen durch beliebig oft und schnell wiederholbare Modellrechnungen ersetzen oder zumindest reduzieren. Dabei sind die verwendeten Verfahren aus einfachen Bausteinen zusammengesetzt (z.B. Integralbestimmung, Lösung von Gleichungssystemen, Nullstellenberechnung, usw.). Wir werden uns in der Vorlesung *Numerik* vor allem mit diesen einfachen Bausteinen befassen.

Zur numerischen Lösung eines Problems aus der Praxis gehört auch eine Information über den gemachten Fehler, um das Resultat richtig einschätzen zu können. Der Gesamtfehler setzt sich zusammen aus den

Modellfehlern

- Idealisierungsfehler: Zur Beschreibung eines physikalischen Sachverhaltes wird ein mathematisches Modell gebildet. Bei der Formulierung werden dabei Vereinfachungen angenommen, etwa Vernachlässigung kapillarer Kräfte oder linearisierte Materialgesetze.
- Datenfehler: Die Daten eines mathematischen Modells (z.B. Koeffizienten einer Differentialgleichung) sind aufgrund ungenauer Kenntnis der Materialeigenschaften notwendig mit Fehlern behaftet.

Numerischer Fehler

- Diskretisierungsfehler: Kontinuierliche Prozesse werden durch endliche Prozesse ersetzt (z.B. Approximation der Differentialgleichung durch eine Differenzgleichung)
- Abbruchfehler: Unendliche Algorithmen werden nach endlich vielen Schritten abgebrochen (z.B. Fixpunktberechnung $x_{n+1} = P(x_n)$)
- Rundungsfehler: Auf einer Rechneranlage müssen alle Rechnungen auf einem endlichen Zahlenbereich durchgeführt werden (z.B. $\frac{1}{3} \approx 0.3333$).

Inhaltsverzeichnis

1 Fehleranalyse	7
1.1 Zahldarstellung und Rundungsfehler	7
1.2 Weitere Beispiele	9
2 Lineare Gleichungssysteme I	11
2.1 Fehlerabschätzungen	12
2.2 Direkte Lösungsverfahren	21
2.3 Spezielle Gleichungssysteme	28
2.4 Nicht reguläre Systeme	34
3 Interpolation	37
3.1 Polynominterpolation	38
3.2 Interpolationsfehler	42
3.3 Hermite-Interpolation	46
3.4 Spline-Interpolation	49
4 Numerische Integration	59
4.1 Beispiele interpolatorischer Quadraturformeln	59
4.2 Newton-Cotes-Formeln	63
4.3 Gaußsche Quadraturformeln	65
5 Approximation	73
5.1 Gauß-Approximation	73
5.2 Tschebyscheff-Approximation	78
6 Nichtlineare Gleichungen	85
6.1 Nullstellen reeller Funktionen	85
6.2 Konvergenzverhalten iterativer Verfahren	90
6.3 Interpolationsverfahren	91
6.4 Newton-Verfahren im \mathbb{R}^d	95
7 Lineare Gleichungssysteme II	97
7.1 Einzelschritt- und Gesamtschrittverfahren	97
7.2 Abstiegsverfahren	103

Kapitel 1

Fehleranalyse

1.1 Zahldarstellung und Rundungsfehler

Die Verarbeitung numerischer Algorithmen auf *digitalen* Rechenanlagen führt zwangsweise zu Fehlern, die durch die Endlichkeit des Bereiches der auf einer solchen Maschine darstellbaren Zahlen bedingt ist. Sei $b \in \mathbb{N}, b \geq 2$ eine feste Basis. Dann besitzt jede reelle Zahl x bezüglich der Basis b eine Entwicklung der Form:

$$x = \pm b^E \sum_{i=1}^{\infty} m_i b^{-i}, \quad E = \sum_{j=0}^{s-1} e_j b^j$$

mit Koeffizienten $m_i, e_j \in \{0, 1, \dots, b-1\}$ und einem Exponenten $E \in \mathbb{Z}$. Die Eindeutigkeit der Zahlendarstellung folgt durch die Festlegung $m_1 \neq 0$ und $m_i < b-1$ für unendlich viele i und gilt für alle $x \in \mathbb{R} \setminus \{0\}$.

Zur Approximation reeller Zahlen werden auf Rechneranlagen sogenannte Gleitkommazahlen und Gleitkommaoperationen verwendet. Eine normalisierte Gleitkommazahl (zur Basis b) ist eine reelle Zahl a in der Form

$$a = \pm M \cdot b^E$$

mit der Mantisse $M = 0.m_1 \dots m_r$ und dem Exponenten $E \in \mathbb{Z}$, wobei $m_i \in \{0, \dots, b-1\}$. Für $a \neq 0$ ist die Darstellung durch die Normierung $m_1 \neq 0$ eindeutig. Für $a = 0$ setzt man $M = 0$ und E beliebig.

Zur Darstellung solcher normalisierten Gleitkommazahlen brauchen wir also

r Ziffern + 1 Vorzeichen für die Mantisse

s Ziffern + 1 Vorzeichen für den Exponenten.

Gebräuchliche Basen sind $b = 2$ (Dualsystem), $b = 10$ (Dezimalsystem) und $b = 16$ (Hexadezimalsystem). Die auf dem Rechner darstellbaren Gleitkommazahlen

heißen auch Maschinenzahlen, wegen der Endlichkeit gibt es eine größte (kleinste darstellbare Zahl)

$$\pm(b-1)\{b^{-1} + b^{-2} + \dots + b^{-r}\} \cdot b^{(b-1)(b^{s-1} + \dots + b^0)} = \pm(1 - b^{-r})b^{b^s - 1}$$

sowie eine kleinste positive und eine größte negative Zahl

$$a_{posmin} = b^{-1} \cdot b^{-(b^s - 1)} = b^{-b^s} \quad , \quad a_{negmax} = -b^{-b^s}$$

MATLAB verwendet intern $b = 2, r = 53$, der Exponent variiert zwischen -1021 und $+1024$. Betrachten wir das IEEE-Format, das 64 Bits verwendet:

$$a = \pm M \cdot 2^{c-1022}$$

Bits	Verwendung
1	Vorzeichen
52	Mantisse $M = 2^{-1} + m_2 2^{-2} + \dots + m_{53} 2^{-53}$ (die erste Mantissenstelle ist immer 1 aus Normierungsgründen, mit Ausnahme der Null)
11	Charakteristik $C = c_0 2^0 + c_1 2^1 + \dots + c_{10} 2^{10} \in (0, 2047)$

Die ausgenommenen Werte $C = 0$ und $C = 2047$ der Charakteristik werden zur Darstellung der Null ($m_2 = \dots = m_{53} = 0, c_0 = \dots = c_{10} = 0$) sowie der Sondergröße $NaN = \text{Not a Number}$ verwendet. Zahlen außerhalb des zulässigen Bereiches

$$D := [a_{min}, a_{negmin}] \cup \{0\} \cup [a_{posmin}, a_{max}]$$

werden auf NaN abgebildet bzw. als overflow registriert. Für $x \in D$ wird eine Rundungsoperation durchgeführt,

$$rd : D \rightarrow (\text{Menge der normierten Gleitkommazahlen}) = \mathbb{F}$$

$$|rd(x) - x| = \min_{f \in \mathbb{F}} |x - f| \quad \forall x \in D.$$

Im IEEE-Format bedeutet dies

$$rd(x) = \text{sgn}(x) \begin{cases} 0.m_1 \dots m_{53} \cdot 2^E & \text{falls } m_{54} = 0, \\ (0.m_1 \dots m_{53} + 2^{-53})2^E & \text{falls } m_{54} = 1. \end{cases}$$

Damit ist der absolute Fehler

$$|rd(x) - x| \leq \frac{b^{-r}}{2} b^E = 2^{-54} \cdot 2^E$$

vom Exponenten E abhängig, der relative Fehler

$$\left| \frac{rd(x) - x}{x} \right| \leq \frac{1}{2} \frac{b^{-r} b^E}{|M| b^E} \leq \frac{1}{2} \frac{b^{-r} b^E}{b^{-1} b^E} = \frac{1}{2} b^{1-r}$$

ist beschränkt für $x \in D$, $x \neq 0$ durch die Maschinengenauigkeit

$$eps = \frac{1}{2}b^{-r+1}.$$

Für $x \in D$ ist dann

$$rd(x) = x \cdot (1 + \varepsilon) \text{ mit } |\varepsilon| \leq eps.$$

Das IEEE-Format liefert

$$eps \leq \frac{1}{2}2^{-52} \approx 10^{-16}.$$

Arithmetische Grundoperationen $\{+, -, \cdot, \backslash\}$ werden auf der Rechanlage durch entsprechende Maschinenoperationen $\oplus, \ominus, \odot, \oslash$ ersetzt, welche Maschinenzahlen in Maschinenzahlen überführen. Dazu werden die Operationen maschinenintern (oft unter einer erhöhten Stellenzahl für die Mantisse) ausgeführt, in die normalisierte Form überführt und dann gerundet. Liegt das Ergebnis nicht in D wird eine Fehlermeldung ausgegeben. Man beachte, dass Assoziativ- bzw. Distributivgesetz im allgemeinen für Maschinenoperationen nicht gelten.

1.2 Weitere Beispiele

a) Betrachten wir das lineare Gleichungssystem

$$\begin{aligned} 0.780x_1 + 0.563x_2 &= 0.217 \\ 0.913x_1 + 0.659x_2 &= 0.254 \end{aligned}$$

mit der eindeutigen Lösung $(x_1, x_2) = (1, -1)$. Wir nehmen die Basis $b = 10$ und eine Mantissenlänge von $r = 3$ an. Die Cramersche Regel ist nicht anwendbar, denn $\det(A) = 0$.

b) Wir wollen $\ln 2$ berechnen und erinnern uns, dass

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}x^n}{n} \quad \forall x \in (-1, 1].$$

Somit ist

$$\ln 2 = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \sum_{n=1}^m \frac{(-1)^{n-1}}{n} + R_m.$$

Nach dem Leibnizschen Konvergenzkriterium bzw. der Fehlerabschätzung für alternierende Reihen ist $|R_m| \leq 1/(m+1) < 1/m$. Somit folgt

$$\begin{aligned} & \left| \ln 2 - rd \left(\sum_{n=1}^m \frac{(-1)^{n-1}}{n} \right) \right| \\ & \leq \underbrace{|R_m|}_{\text{Abbruchfehler}} + \underbrace{\left| \sum_{n=1}^m \frac{(-1)^{n-1}}{n} - rd \left(\sum_{n=1}^m \frac{(-1)^{n-1}}{n} \right) \right|}_{\text{Rundungsfehler}} \\ & \leq \frac{1}{m} + m \cdot eps \end{aligned}$$

Wir beobachten eine für die Numerik typische Situation; mit zunehmenden m nimmt der Abbruchfehler ab, der Rundungsfehler aber zu. Das Minimum wird für $m^2 = eps^{-1}$ angenommen. Der Gesamtfehler bleibt damit nur auf

$$\frac{1}{m} + m \cdot eps \approx 2\sqrt{eps}$$

beschränkt (kann also nicht auf eine beliebig kleine Schranke gedrückt werden). Für die numerische Bestimmung von $\ln 2$ ist die folgende Darstellung besser geeignet. Aus

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^n}{n}, \quad \ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n} \quad x \in (-1, +1)$$

folgt durch Subtraktion

$$\ln \frac{1+x}{1-x} = \sum_{n=1}^{\infty} \frac{[(-1)^{n-1} + 1]x^n}{n} = 2 \sum_{k=1}^{\infty} \frac{x^{2k-1}}{2k-1}.$$

Für $x = \frac{1}{3}$ erhalten wir

$$\ln 2 = 2 \sum_{k=1}^{\infty} \frac{1}{2k-1} \left(\frac{1}{3}\right)^{2k-1} = 2 \sum_{k=1}^m \frac{1}{2k-1} \left(\frac{1}{3}\right)^{2k-1} + \tilde{R}_m$$

mit der Fehlerschranke

$$\begin{aligned} |\tilde{R}_m| & \leq \left(\frac{2}{2m+1}\right) \left(\frac{1}{3}\right)^{2m+1} \sum_{n=0}^{\infty} \left(\frac{1}{3}\right)^{2n} = \left(\frac{2}{2m+1}\right) \left(\frac{1}{3}\right)^{2m+1} \frac{1}{1-1/9} \\ & = \frac{2}{3} \cdot \frac{9}{8} \cdot \frac{1}{2m+1} \cdot 9^{-m} \leq 10^{-7} \quad \text{für } m \geq 6. \end{aligned}$$

Wir sehen, dass die neue Reihendarstellung wesentlich schneller gegen $\ln 2$ konvergiert.

Kapitel 2

Lösung linearer Gleichungssysteme I

Seien A eine Matrix und b ein Vektor

$$A = (a_{jk})_{\substack{j=1,\dots,m \\ k=1,\dots,n}} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad b = (b_j)_{j=1,\dots,m} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

Gesucht ist ein Vektor $x = (x_k)_{k=1,\dots,n}$ mit der Eigenschaft

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

oder abgekürzt geschrieben: $Ax = b$. Das lineare Gleichungssystem $Ax = b$ heißt *unterbestimmt* im Fall $m < n$, *quadratisch* im Fall $m = n$ und *überbestimmt* im Fall $m > n$. Es ist genau dann lösbar, wenn $\text{Rang}(A) = \text{Rang}[A, b]$, mit der zusammengesetzten Matrix

$$[A, b] = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array} \right].$$

Im *quadratischen* Fall sind die folgenden Aussagen äquivalent:

- i) $Ax = b$ ist für jedes b eindeutig lösbar.
- ii) $\text{Rang}(A) = n$.
- iii) $\det(A) \neq 0$

- iv) Alle Eigenwerte von A sind ungleich Null.

Wir beschäftigen uns im folgenden hauptsächlich mit der Lösung von quadratischen Gleichungssystemen. Die dazu verwendeten Verfahren lassen sich grob in zwei Klassen einteilen: Ein *direktes* Verfahren zur Lösung des Gleichungssystems $Ax = b$ ist ein Algorithmus, der (bei Vernachlässigung von Rundungsfehlern) in endlich vielen Schritten die Lösung x liefert. Im Gegensatz dazu erzeugen die *iterativen* Verfahren sukzessive eine Folge von Vektoren $\{x^{(t)}\}_{t=1,2,\dots}$, die im Limes für $t \rightarrow \infty$ immer bessere Approximationen zur Lösung x sind.

2.1 Fehlerabschätzungen

Wir beschäftigen uns zunächst mit dem Problem der *Konditionierung* von quadratischen linearen Gleichungssystemen. Bei der Lösung eines Gleichungssystems $Ax = b$ treten zwei Fehlereinflüsse ein:

- Fehler in der *theoretischen* Lösung aufgrund von Eingangsfehlern in den Elementen von A und b ,
- Fehler in der *numerischen* Lösung aufgrund des Rundungsfehlers im Verlaufe des Lösungsprozesses.

Zur Erfassung dieser Fehler benötigen wir ein Maß für die *Größe* von Vektoren und Matrizen. Dazu dienen üblicherweise *Normen* auf dem n -dimensionalen Zahlenraum \mathbb{K}^n , $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. (Im Hinblick auf spätere Anwendungen lassen wir im folgenden auch komplexe Vektoren bzw. Matrizen zu.) Eine Abbildung $\|\cdot\| : \mathbb{K}^n \rightarrow \mathbb{R}_+$ heißt *Norm*, wenn sie folgende Eigenschaften besitzt:

- (N1) $\|x\| > 0$, $x \in \mathbb{K}^n \setminus \{0\}$ (Definitheit),
 (N2) $\|\alpha x\| = |\alpha| \cdot \|x\|$, $x \in \mathbb{K}^n$, $\alpha \in \mathbb{K}$ (positive Homogenität),
 (N3) $\|x + y\| \leq \|x\| + \|y\|$, $x, y \in \mathbb{K}^n$ (Subadditivität).

Beispiele:

$$\|x\|_2 := \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}} \quad \text{euklidische Norm (} l_2\text{-Norm)}$$

$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i| \quad \text{Maximumnorm (} l_\infty\text{-Norm)}$$

$$\|x\|_1 := \sum_{i=1}^n |x_i| \quad \text{Manhattanorm (} l_1\text{-Norm)}$$

Mit Hilfe einer Norm $\|\cdot\|$ auf \mathbb{K}^n lässt sich die *Konvergenz* einer Folge von Vektoren gegen einen Vektor erklären durch

$$x^{(t)} \rightarrow x \quad (t \rightarrow \infty) : \Leftrightarrow \|x^{(t)} - x\| \rightarrow 0 \quad (t \rightarrow \infty).$$

Aus der Dreiecksungleichung (N3) folgt über die Beziehung $\|x\| = \|x - y + y\|$ die wichtige Ungleichung

$$\|x - y\| \geq \left| \|x\| - \|y\| \right|, \quad x, y \in \mathbb{K}^n \quad (2.1.1)$$

welche u.a. die Stetigkeit von Normen als Funktionen von \mathbb{K}^n in \mathbb{R} impliziert.

Theorem 2.1.1 *Auf dem endlichdimensionalen Vektorraum \mathbb{K}^n sind alle Normen äquivalent, d.h.: Zu je zwei Normen $\|\cdot\|$, $\|\cdot\|'$ gibt es positive Konstanten m , M , mit denen gilt:*

$$m\|x\| \leq \|x\|' \leq M\|x\|, \quad x \in \mathbb{K}^n \quad (2.1.2)$$

Beweis: Es genügt, die Behauptung für den Fall zu zeigen, dass eine der beiden Normen die Maximumnorm $\|\cdot\|_\infty$ ist. Sei $\|\cdot\|$ irgendeine zweite Norm. Bezüglich der kartesischen Einheitsvektoren e_1, \dots, e_n hat jeder Vektor $x \in \mathbb{K}^n$ die Darstellung

$$x = \sum_{i=1}^n x_i e_i.$$

Folglich gilt

$$\|x\| \leq \gamma \|x\|_\infty, \quad \gamma := \sum_{i=1}^n \|e_i\|.$$

Die Norm $\|\cdot\|$ ist also auch stetig bezüglich der komponentenweisen Konvergenz von Vektoren. Die Punktmenge

$$S \equiv \{x \in \mathbb{K}^n, \|x\|_\infty = 1\} \subset \mathbb{K}^n$$

ist beschränkt und abgeschlossen. Die Norm $\|\cdot\|$ nimmt also auf S ihr Minimum und Maximum an. Es existieren also $x_0, x_1 \in S$, so dass

$$0 < \|x_0\| \leq \|x\| \leq \|x_1\| < \infty \quad \forall x \in S.$$

Für beliebiges $y \in \mathbb{K}^n \setminus \{0\}$ ist $y/\|y\|_\infty \in S$ und folglich

$$\|x_0\| \leq \frac{\|y\|}{\|y\|_\infty} \leq \|x_1\|.$$

Mit $m \equiv \|x_0\|$ und $M \equiv \|x_1\|$ gilt daher

$$m\|y\|_\infty \leq \|y\| \leq M\|y\|_\infty \quad \forall y \in \mathbb{K}^n,$$

die Norm $\|\cdot\|$ ist also zur Maximumnorm äquivalent. \square

Die Beziehung 2.1.2 impliziert, dass die durch eine beliebige Norm induzierte Konvergenz von Vektoren stets äquivalent zur *komponentenweisen* Konvergenz ist.

Wir betrachten nun den Vektorraum der $n \times n$ -Matrizen $A \in \mathbb{K}^{n \times n}$. Offenbar kann dieser mit dem Vektorraum der $n * n$ -Vektoren identifiziert werden. Somit übertragen sich alle Aussagen für Vektornormen auf Normen für Matrizen. Insbesondere sind alle Normen für $n \times n$ -Matrizen äquivalent, und die Konvergenz von Matrizen ist die komponentenweise Konvergenz:

$$A^{(t)} \rightarrow A \quad (t \rightarrow \infty) \quad \Leftrightarrow \quad a_{jk}^{(t)} \rightarrow a_{jk} \quad (t \rightarrow \infty), \quad j, k = 1, \dots, n.$$

Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n \times n}$ heißt *verträglich* mit einer Vektornorm $\|\cdot\|$ auf \mathbb{K}^n , wenn gilt:

$$\|Ax\| \leq \|A\| \cdot \|x\|, \quad x \in \mathbb{K}^n, \quad A \in \mathbb{K}^{n \times n}.$$

Sie heißt *Matrizennorm*, wenn sie submultiplikativ ist:

$$\|AB\| \leq \|A\| \cdot \|B\|, \quad A, B \in \mathbb{K}^{n \times n}.$$

Beispielsweise ist die Quadratsummennorm (*Frobeniusnorm*)

$$\|A\|_F := \left(\sum_{j,k=1}^n |a_{jk}|^2 \right)^{\frac{1}{2}}$$

eine mit der euklidischen Vektornorm verträgliche Matrizennorm. Für eine beliebige Vektornorm $\|\cdot\|$ auf \mathbb{K}^n wird durch

$$\|A\| := \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{K}^n \\ \|x\|=1}} \|Ax\|$$

eine mit $\|\cdot\|$ verträgliche Matrizennorm erklärt. Diese heißt die von $\|\cdot\|$ erzeugte *natürliche* Matrizennorm. Für natürliche Matrizennormen gilt $\|I\| = 1$.

Theorem 2.1.2 Die natürlichen Matrizennormen zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind die maximale Zeilensumme

$$\|A\|_\infty := \max_{1 \leq j \leq n} \sum_{k=1}^n |a_{jk}|$$

bzw. die maximale Spaltensumme

$$\|A\|_1 := \max_{1 \leq k \leq n} \sum_{j=1}^n |a_{jk}|.$$

Beweis: Wir führen den Beweis nur für $\|\cdot\|_\infty$. Offenbar ist die maximale Zeilensumme $\|\cdot\|_\infty$ eine Matrizenorm. Wegen

$$\|Ax\|_\infty = \max_{1 \leq j \leq n} \left| \sum_{k=1}^n a_{jk} x_k \right| \leq \max_{1 \leq j \leq n} \sum_{k=1}^n |a_{jk}| \max_{1 \leq k \leq n} |x_k| = \|A\|_\infty \|x\|_\infty$$

ist sie verträglich mit $\|\cdot\|_\infty$. Im Falle $\|A\|_\infty = 0$ ist $A = 0$, d.h. es gilt trivialerweise

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty.$$

Sei also $\|A\|_\infty > 0$ und $m \in \{1, \dots, n\}$ ein Index mit der Eigenschaft

$$\|A\|_\infty = \max_{1 \leq j \leq n} \sum_{k=1}^n |a_{jk}| = \sum_{k=1}^n |a_{mk}|.$$

Wir setzen für $k = 1, \dots, n$: $z_k \equiv |a_{mk}|/a_{mk}$ für $a_{mk} \neq 0$ und $z_k \equiv 0$ sonst, somit gilt:

$z = (z_k)_k \in \mathbb{K}^n$, $\|z\|_\infty = 1$. Für $v := Az$ gilt dann

$$v_m = \sum_{k=1}^n a_{mk} z_k = \sum_{k=1}^n |a_{mk}| = \|A\|_\infty.$$

Folglich ist

$$\|A\|_\infty = v_m \leq \|v\|_\infty = \|Az\|_\infty \leq \sup_{\|y\|_\infty=1} \|Ay\|_\infty.$$

□

Die *Eigenwerte* $\lambda \in \mathbb{K}$ einer Matrix $A \in \mathbb{K}^{n \times n}$ sind die Nullstellen ihres charakteristischen Polynoms $p(\lambda) = \det(A - \lambda I)$. Folglich existieren genau n (ihrer Vielfachheit als Nullstelle entsprechend oft gezählte) Eigenwerte λ , und zu jedem λ existiert mindestens ein *Eigenvektor* $w \in \mathbb{K}^n \setminus \{0\} : Aw = \lambda w$. Sei nun $\|\cdot\|$ eine beliebige Vektornorm und $\|\cdot\|$ eine damit verträgliche Matrizenorm, (wobei die beiden Normen der Einfachheit halber gleich bezeichnet werden). Mit einem normierten Eigenvektor zum Eigenwert λ gilt

$$|\lambda| = |\lambda| \|w\| = \|\lambda w\| = \|Aw\| \leq \|A\| \|w\| = \|A\|,$$

d.h. alle Eigenwerte von A liegen in einer Kreisscheibe in \mathbb{C} mit Mittelpunkt Null und Radius $\|A\|$. Speziell mit $\|A\|_\infty$ erhält man die Abschätzung

$$\max |\lambda| \leq \|A\|_\infty = \max_{1 \leq j \leq n} \sum_{k=1}^n |a_{jk}|.$$

Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt *hermitesch*, wenn gilt:

$$A = \overline{A}^T \quad \text{bzw.} \quad a_{jk} = \overline{a_{kj}}, \quad j, k = 1, \dots, n.$$

Reelle hermitesche Matrizen werden *symmetrisch* genannt. Der Begriff der Symmetrie ist eng verknüpft mit dem des Skalarprodukts. Eine Abbildung $(\cdot, \cdot): \mathbb{K}^n \times \mathbb{K}^n \rightarrow \mathbb{K}$ wird ein *Skalarprodukt* genannt, wenn sie folgende Eigenschaften hat:

$$(S1) \quad (x, y) = \overline{(y, x)}, \quad x, y \in \mathbb{K}^n \quad (\text{Symmetrie}),$$

$$(S2) \quad (\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z), \quad x, y, z \in \mathbb{K}^n, \alpha, \beta \in \mathbb{K} \quad (\text{Linearität}),$$

$$(S3) \quad (x, x) > 0, \quad x \in \mathbb{K}^n \setminus \{0\} \quad (\text{Definitheit}).$$

Jedes Skalarprodukt auf $\mathbb{K}^n \times \mathbb{K}^n$ erzeugt durch

$$\|x\| := (x, x)^{\frac{1}{2}}, \quad x \in \mathbb{K}^n,$$

eine zugehörige Vektornorm. Im folgenden wird fast ausschließlich das *euklidische Skalarprodukt* verwendet:

$$(x, y)_2 = \sum_{j=1}^n x_j \overline{y_j}, \quad (x, x)_2 = \|x\|_2^2.$$

Mit Hilfe des euklidischen Skalarprodukts läßt sich die Eigenschaft einer Matrix, hermitesch zu sein, äquivalent ausdrücken durch:

$$A = \overline{A}^T \Leftrightarrow (Ax, y)_2 = (x, Ay)_2, \quad x, y \in \mathbb{K}^n.$$

Die von der euklidischen Vektornorm erzeugte natürliche Matrizennorm heißt die *Spektralnorm* und wird mit $\|\cdot\|_2$ bezeichnet.

Theorem 2.1.3 *Für hermitesche Matrizen gilt*

$$\|A\|_2 = \max\{|\lambda|, \lambda \text{ Eigenwert von } A\}. \quad (2.1.3)$$

Beweis: Bekanntlich besitzt eine hermitesche Matrix $A \in \mathbb{K}^{n \times n}$ nur reelle Eigenwerte und zwar genau n (ihrer Vielfachheit entsprechend oft gezählt), $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Ferner existiert ein zugehöriges *Orthonormalsystem* von Eigenvektoren

$$\{w_1, \dots, w_n\} \subset \mathbb{K}^n : A w_i = \lambda w_i, \quad (w_i, w_j)_2 = \delta_{ij}, \quad i, j = 1, \dots, n.$$

Jedes $x \in \mathbb{K}^n$ besitzt eine Darstellung der Form

$$x = \sum_{i=1}^n \alpha_i w_i, \quad \alpha_i = (x, w_i)_2,$$

und es gilt

$$\|x\|_2^2 = (x, x)_2 = \sum_{i,j=1}^n \alpha_i \bar{\alpha}_j (w_i, w_j)_2 = \sum_{i=1}^n |\alpha_i|^2,$$

$$\|Ax\|_2^2 = (Ax, Ax)_2 = \sum_{i,j=1}^n \lambda_i \alpha_i \bar{\lambda}_j \alpha_j (w_i, w_j)_2 = \sum_{i=1}^n \lambda_i^2 |\alpha_i|^2.$$

Hiermit folgt

$$\|A\|_2 \leq \max_{1 \leq i \leq n} |\lambda_i|.$$

Wegen der allgemeinen Eigenwertschranke $\max |\lambda| \leq \|A\|$ für beliebige verträgliche Matrizennormen folgt damit die Behauptung. \square

Für allgemeine Matrizen $A \in \mathbb{K}^{n \times n}$ gilt

$$\|A\|_2 = \max\{|\lambda|^{\frac{1}{2}}, \lambda \text{ Eigenwert von } \overline{A}^T A\}.$$

Lemma 2.1.4 *Eine Matrix $A \in \mathbb{C}^{n \times n}$ ist genau dann hermitisch, wenn (Ax, x) für alle $x \in \mathbb{C}^n$ reell ist.*

Beweis: Wir nutzen die Tatsache, dass für jede Matrix $A \in \mathbb{C}^{n \times n}$

$$(Ax, y) = (x, \overline{A}^T y) \quad \forall x, y \in \mathbb{C}^n \quad \text{gilt.}$$

a) Sei nun A hermitisch, d.h. $\overline{A}^T = A$. Setze $y = x$ und nutze die Eigenschaft des Skalarproduktes $(x, y) = \overline{(y, x)}$. Dann folgt

$$(Ax, x) = (x, Ax) = \overline{(Ax, x)}.$$

b) Sei nun (Ax, x) reell, d.h.

$$(Ax, x) = \overline{(Ax, x)} = \overline{(x, \overline{A}^T x)} = (\overline{A}^T x, x)$$

bzw. $((A - \overline{A}^T)x, x) = 0$ für alle $x \in \mathbb{C}^n$. Wir nutzen ein Hilfsresultat für beliebige $B \in \mathbb{C}^{n \times n}$

$$(Bx, x) = 0 \quad \forall x \in \mathbb{C}^n \quad \Rightarrow B = 0$$

woraus $A = \overline{A}^T$ folgen würde. Wählt man zunächst

$$x = (0, \dots, 0, \underbrace{1}_{k\text{-te Stelle}}, 0, \dots, 0),$$

so folgt das Verschwinden der Diagonale von B :

$$(Bx, x) = \sum_{i,j=1}^n b_{ij} x_j \overline{x_i} = b_{kk} = 0 \quad \forall k.$$

Die Wahl

$$x = (0, \dots, 0, \underbrace{1}_{k\text{-te}}, 0, \dots, 0, \underbrace{i}_{l\text{-te Stelle}}, 0, \dots, 0) \leftarrow \text{imaginäre Einheit}$$

ergibt

$$(Bx, x) = b_{kk} + b_{kl}i - b_{lk}i - i^2 b_{ll} = (b_{kl} - b_{lk})i = 0.$$

Schließlich führt die Wahl

$$x = (0, \dots, 0, \underbrace{1}_{k\text{-te}}, 0, \dots, 0, \underbrace{1}_{l\text{-te Stelle}}, 0, \dots, 0)$$

auf

$$(Bx, x) = b_{kk} + b_{kl} + b_{lk} + b_{ll} = b_{kl} + b_{lk} = 0.$$

Die Lösung des Gleichungssystems für b_{kl} und b_{lk} ergibt $b_{kl} = b_{lk} = 0$ für alle $l \neq k, l, k \in \{1, \dots, n\}$. \square

Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt *positiv definit*, wenn gilt:

$$(Ax, x)_2 \in \mathbb{R}, \quad (Ax, x)_2 > 0 \quad \forall x \in \mathbb{K}^n \setminus \{0\}.$$

Eine hermitesche Matrix ist genau dann positiv definit, wenn alle ihre (reellen) Eigenwerte positiv sind. Wir werden später sehen, daß lineare Gleichungssysteme mit positiv definiten Koeffizientenmatrizen besonders günstige Lösbarkeitseigenschaften besitzen.

Wir kommen nun zur Fehleranalyse für lineare Gleichungssysteme

$$Ax = b$$

mit regulärer Koeffizientenmatrix $A \in \mathbb{K}^{n \times n}$. Die Matrix A und der Vektor b seien mit Fehlern δA bzw. δb behaftet, so dass ein gestörtes System

$$\tilde{A}\tilde{x} = \tilde{b},$$

mit $\tilde{A} = A + \delta A$, $\tilde{b} = b + \delta b$ und $\tilde{x} = x + \delta x$ gelöst wird. Wir wollen den Fehler δx in Abhängigkeit von δA und δb abschätzen. Dazu sei im folgenden $\|\cdot\|$ eine beliebige Vektornorm und entsprechend $\|\cdot\|$ die zugehörige Matrixnorm.

Lemma 2.1.5 *Die Matrix $B \in \mathbb{K}^{n \times n}$ habe die Norm $\|B\| < 1$. Dann ist die Matrix $I + B$ regulär, und es gilt*

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \quad (2.1.4)$$

Beweis: Für alle $x \in \mathbb{K}^n$ gilt

$$\|(I + B)x\| \geq \|x\| - \|Bx\| \geq (1 - \|B\|)\|x\|.$$

Wegen $1 - \|B\| > 0$ ist also $I + B$ injektiv und folglich regulär. Mit der Abschätzung

$$\begin{aligned} 1 &= \|I\| = \|(I + B)(I + B)^{-1}\| = \|(I + B)^{-1} + B(I + B)^{-1}\| \\ &\geq \|(I + B)^{-1}\| - \|B\|\|(I + B)^{-1}\| = \|(I + B)^{-1}\|(1 - \|B\|) > 0. \end{aligned}$$

erhält man die behauptete Ungleichung. \square

Nach diesen Vorbereitungen können wir den folgenden allgemeinen Störungssatz für lineare Gleichungssysteme beweisen:

Theorem 2.1.6 *Die Matrix $A \in \mathbb{K}^{n \times n}$ sei regulär, und es sei*

$$\|\delta A\| < \|A^{-1}\|^{-1}. \quad (2.1.5)$$

Dann ist die gestörte Matrix $\tilde{A} = A + \delta A$ ebenfalls regulär, und für den relativen Fehler der Lösung gilt:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\|\delta A\|\|A\|} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \quad (2.1.6)$$

mit der Konditionszahl $\text{cond}(A)$ von A ,

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

Beweis: Aufgrund der Voraussetzungen ist

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1,$$

so dass auch $A + \delta A = A[I + A^{-1}\delta A]$ nach Lemma 2.1.5 regulär ist. Aus

$$(A + \delta A)\tilde{x} = b + \delta b, \quad (A + \delta A)x = b + \delta Ax$$

folgt für $\delta x = \tilde{x} - x$ zunächst

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

Somit haben wir

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

Wegen $\|b\| = \|Ax\| \leq \|A\| \|x\|$ folgt schließlich

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

und damit die Behauptung des Satzes. \square

Die Konditionszahl $\text{cond}(A)$ hängt offenbar von der bei ihrer Definition zugrundegelegten Vektornorm ab. Meistens verwendet man die Maximumnorm $\|\cdot\|_\infty$ oder die euklidische Norm $\|\cdot\|_2$. Im ersten Fall ist

$$\text{cond}_\infty(A) := \|A\|_\infty \|A^{-1}\|_\infty$$

mit der maximalen Zeilensumme $\|\cdot\|_\infty$. Speziell für *hermitesche* Matrizen gilt nach Lemma 2.1.2

$$\text{cond}_2(A) := \|A\|_2 \|A^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

mit den betragsmäßig größten bzw. kleinsten Eigenwerten λ_{\max} und λ_{\min} von A ; die Größe $\text{cond}_2(A)$ wird auch die *Spektralkonditionszahl* von A genannt.

Ist $\text{cond}(A) \|\delta A\| \|A\|^{-1} \ll 1$, so wird in Theorem 2.1.6

$$\frac{\|\delta x\|}{\|x\|} \lesssim \text{cond}(A) \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\},$$

d.h. die Konditionszahl $\text{cond}(A)$ ist gerade der Verstärkungsfaktor, mit dem sich die relativen Fehler in A und b auf den relativen Fehler in x auswirken. Diese

Fehlerabschätzung erlaubt folgenden Schluss:

Regel: Die Kondition von A sei $\text{cond}(A) \sim 10^s$. Sind die Elemente von A und b mit einem relativen Fehler der Art

$$\|\delta A\| \backslash \|A\| \sim 10^{-k}, \quad \|\delta b\| \backslash \|b\| \sim 10^{-k} (k > s)$$

behaftet, so muss mit einem relativen Fehler im Ergebnis der Größenordnung

$$\|\delta x\| \backslash \|x\| \sim 10^{s-k}$$

gerechnet werden, d.h. man verliert im ungünstigsten Fall s Stellen an Genauigkeit.

Beispiel:

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \cdot \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|A\|_\infty = 2.1617, \quad \|A^{-1}\|_\infty = 1.513 \cdot 10^8 \Rightarrow \text{cond}(A) \approx 3.3 \cdot 10^8.$$

Bei der Lösung des Gleichungssystems $Ax = b$ gehen also im ungünstigen Fall 8 wesentliche Stellen an der Genauigkeit, mit der die Elemente a_{jk} und b_j gegeben sind, verloren. Dieses System ist extrem *schlecht konditioniert*.

Wir demonstrieren anhand der Spektralkondition, dass die Abschätzung in Theorem 2.1.6 im wesentlichen scharf ist. Sei A eine positiv definite $n \times n$ -Matrix mit kleinstem und größtem Eigenwert λ_1 bzw. λ_n sowie zugehörigen normierten Eigenvektoren w_1 bzw. w_n . Wir wählen $\delta A \equiv 0$, $b \equiv w_n$, $\delta b \equiv \epsilon w_1$ ($\epsilon \neq 0$). Dann haben die Gleichungen $Ax = b$ und $A\tilde{x} = b + \delta b$ die Lösungen

$$x = \frac{1}{\lambda_n} w_n, \quad \tilde{x} = \frac{1}{\lambda_n} w_n + \epsilon \frac{1}{\lambda_1} w_1.$$

Folglich ist für $\delta x = \tilde{x} - x$

$$\frac{\|\delta x\|_2}{\|x\|_2} = \epsilon \frac{\lambda_n}{\lambda_1} \frac{\|w_1\|_2}{\|w_n\|_2} = \text{cond}_2(A) \frac{\|\delta b\|_2}{\|b\|_2}.$$

2.2 Direkte Lösungsverfahren

Im folgenden diskutieren wir direkte Lösungsmethoden für (reelle) quadratische lineare Gleichungssysteme der Form

$$Ax = b. \tag{2.2.7}$$

Besonders leicht lösbar sind gestaffelte Systeme, z.B. solche mit einer oberen Dreiecksmatrix $A = (a_{jk})$ als Koeffizientenmatrix

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

Im Falle $a_{jj} \neq 0, j = 1, \dots, n$, erhält man die Lösung durch sog. *sukzessives Rückwärtseinsetzen*

$$x_n = \frac{b_n}{a_{nn}}, \quad j = n-1, \dots, 1: \quad x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^n a_{jk}x_k \right).$$

Dazu sind offensichtlich $n^2/2 + \mathbf{O}(n)$ *arithmetische Operationen* erforderlich (1 arithmetische Operation := 1 Multiplikation (+1 Addition) oder 1 Division).

Das *klassische* direkte Verfahren zur Lösung (regulärer) Gleichungssysteme ist das *Gaußsche Eliminationsverfahren*. Dabei wird das gegebene System $Ax = b$ schrittweise in ein oberes Dreieckssystem $Rx = c$ umgeformt, welches dieselbe Lösung x besitzt und dann durch Rückwärtseinsetzen gelöst wird. Dazu stehen die folgenden elementaren Umformungen zur Verfügung:

- (i) Vertauschung zweier Gleichungen,
- (ii) Addition des Vielfachen einer Gleichung zu einer anderen.

(Die Vertauschung zweier Spalten von A ist ebenfalls zulässig, wenn die Unbekannten x_i entsprechend unnummeriert werden.)

In der praktischen Durchführung des Gaußschen Eliminationsverfahrens wendet man die elementaren Umformungen auf die zusammengesetzte Matrix $[A, b]$ an. Im folgenden wird A als regulär angenommen.

Zunächst setzt man $A^{(0)} \equiv A, b^{(0)} \equiv b$. Bestimme $a_{r1}^{(0)} \neq 0, r \in \{1, \dots, n\}$. (Solch ein Element existiert, da A sonst singular wäre). Vertausche die 1-te und die r -te Zeile. Das Resultat sei $[\tilde{A}^{(0)}, \tilde{b}^{(0)}]$. Subtrahiere für $j = 2, \dots, n$ das q_{j1} -fache,

$$q_{j1} \equiv \tilde{a}_{j1}^{(0)} \backslash \tilde{a}_{11}^{(0)} (= a_{r1}^{(0)} \backslash a_{rr}^{(0)}),$$

der 1-ten Zeile von der j -ten Zeile, das Resultat ist

$$[A^{(1)}, b^{(1)}] = \left[\begin{array}{cccc|c} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \dots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array} \right].$$

Ein Zeilen- oder Spaltentausch wird durch Multiplikation mit einer *Permutationsmatrix*

$$P_{kl} = \begin{bmatrix} I & & & \\ & 0 & \dots & 1 \\ & \vdots & I & \vdots \\ & 1 & \dots & 0 \\ & & & & I \end{bmatrix}$$

beschrieben, wobei I Einheitsmatrizen der Dimensionen $k-1$, $l-1-k$, und $n-l$ bezeichnen. Linksmultiplikation PA vertauscht die Zeilen k und l , Rechtsmultiplikation AP die Spalten k und l . Eine Permutationsmatrix ist eine durch Zeilenpermutation (Spaltenpermutation) aus der Einheitsmatrix entstehende Matrix. Die obigen Matrix P_{kl} ist der Permutation $(1, \dots, k-1, l, k+1, \dots, l-1, k, l+1, \dots, n)$ zugeordnet. Die Determinante einer Permutation ist $\det P = \pm 1$, je nachdem ob die Permutation gerade (+1) oder ungerade (-1) ist. Ferner gilt $P^{-1} = P$. Die Elimination der Unbekannten x_k in den Zeilen $k+1, \dots, n$ mittels der k -ten Zeile kann als Linksmultiplikation mit der *Frobeniusmatrix*

$$G_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -q_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -q_{n,k} & & & 1 \end{bmatrix}, \quad \det G_k = 1$$

beschrieben werden. Man rechnet nach, dass für die inverse Matrix

$$G_k^{-1} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & q_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & q_{n,k} & & & 1 \end{bmatrix} = 2I - G_k$$

gilt. Den oben beschriebenen Übergang $[A^{(0)}, b^{(0)}] \rightarrow [A^{(1)}, b^{(1)}]$ kann man nun mit Hilfe von Matrizenmultiplikationen beschreiben

$$[\tilde{A}^{(0)}, \tilde{b}^{(0)}] = P_1[A^{(0)}, b^{(0)}], \quad [A^{(1)}, b^{(1)}] = G_1[\tilde{A}^{(0)}, \tilde{b}^{(0)}],$$

wobei P_1 eine Permutationsmatrix und G_1 eine Frobenius-Matrix sind.

Die Gleichungssysteme $Ax = b$ und $A^{(1)}x = b^{(1)}$ haben offenbar dieselbe Lösung:

$$Ax = b \Leftrightarrow A^{(1)}x = G_1P_1Ax = G_1P_1b = b^{(1)}$$

Das Element $a_{r1} = \tilde{a}_{11}^{(0)}$ heißt "Pivotelement" und der ganze Teilschritt seiner Bestimmung "Pivotsuche". Aus Gründen der numerischen Stabilität trifft man gewöhnlich die Wahl

$$|a_{r1}| = \max_{1 \leq j \leq n} |a_{j1}|. \quad (2.2.8)$$

Der ganze Prozeß inkl. Zeilenvertauschung wird dann "*Spaltenpivotierung*" genannt. Sind die Elemente der Matrix A von sehr unterschiedlicher Größenordnung, so empfiehlt es sich, *totale* Pivotierung vorzunehmen. Diese besteht aus der Wahl

$$|a_{rs}| = \max_{1 \leq j, k \leq n} |a_{jk}| \quad (2.2.9)$$

und anschließender Vertauschung der ersten mit der r -ten Zeile und der ersten mit der s -ten Spalte. Entsprechend der Spaltenvertauschung müssen dann die Unbekannten x_k umnummeriert werden. Bei großen Gleichungssystemen ist die totale Pivotsierung meist zu aufwendig, so dass man sich mit der Spaltenpivotierung begnügt.

Die im ersten Schritt erzeugte Matrix $A^{(1)}$ ist wieder regulär. Dasselbe gilt auch für die um die erste Zeile und Spalte reduzierte Teilmatrix, so dass auf sie der Eliminationsprozeß analog zu Schritt 1 angewendet werden kann. Durch Weiterführung dieses Eliminationsprozesses erhält man in $n - 1$ Schritten eine Kette von Matrizen

$$[A, b] \rightarrow [A^{(1)}, b^{(1)}] \rightarrow \dots \rightarrow [A^{(n-1)}, b^{(n-1)}] =: [R, c],$$

wobei

$$[A^{(i)}, b^{(i)}] = G_i P_i [A^{(i-1)}, b^{(i-1)}], \quad [A^{(0)}, b^{(0)}] := [A, b],$$

mit Permutationsmatrizen P_i und (regulären) Frobenius-Matrizen G_i sind.

Das Endresultat

$$[R, c] = G_{n-1} P_{n-1} \cdots G_1 P_1 [A, b] \quad (2.2.10)$$

entspricht einem oberen Dreieckssystem $Rx = c$, welches dieselbe Lösung wie das Ausgangssystem $Ax = b$ besitzt.

Im i -ten Eliminationsschritt $[A^{(i-1)}, b^{(i-1)}] \rightarrow [A^{(i)}, b^{(i)}]$ werden in der i -ten Spalte die Elemente unterhalb der Diagonalen annulliert. Den frei gewordenen Platz benutzt man zur Abspeicherung der wesentlichen Elemente $q_{i+1,i}, \dots, q_{n,i}$ der Frobenius-Matrizen G_i^{-1} ($i = 1, \dots, n - 1$). Da im i -ten Eliminationsschritt die vorausgehenden Zeilen 1 bis i nicht verändert werden, arbeitet man also mit

Matrizen der Form

$$\left[\begin{array}{ccccccc|c} r_{11} & r_{12} & \dots & r_{1i} & r_{1,i+1} & \dots & r_{1n} & c_1 \\ \lambda_{21} & r_{22} & \dots & r_{2i} & r_{2,i+1} & \dots & r_{2n} & c_2 \\ \lambda_{31} & \lambda_{32} & & r_{3i} & r_{3,i+1} & \dots & r_{3n} & c_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{i1} & \lambda_{i2} & & r_{ii} & r_{i,i+1} & \dots & r_{in} & c_i \\ \lambda_{i+1,1} & \lambda_{i+1,2} & & \lambda_{i+1,i} & a_{i+1,i+1}^{(i)} & \dots & a_{i+1,n}^{(i)} & b_{i+1}^{(i)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \dots & \lambda_{n,i} & a_{n,i+1}^{(i)} & \dots & a_{n,n}^{(i)} & b_n^{(i)} \end{array} \right]$$

Dabei sind die Subdiagonalelemente $\lambda_{k+1,k}, \dots, \lambda_{nk}$ der k -ten Spalte gewisse Permutationen der wesentlichen Elemente $q_{k+1,k}, \dots, q_{nk}$ von G_k^{-1} , da die Zeilvertauschungen (nur diese!) an der gesamten Matrix vorgenommen werden. Als Endresultat erhält man eine Matrix

$$\left[\begin{array}{cccc|c} r_{11} & & \dots & r_{1n} & c_1 \\ l_{21} & r_{22} & & r_{2n} & c_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ l_{n1} & \dots & l_{n,n-1} & r_{nn} & c_n \end{array} \right].$$

Theorem 2.2.1 Die Matrizen

$$L = \left[\begin{array}{cccc} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \dots & l_{n,n-1} & 1 \end{array} \right] \quad \text{und} \quad R = \left[\begin{array}{cccc} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{array} \right]$$

bilden eine sog. LR-Zerlegung der Matrix PA :

$$PA = LR, \quad P = P_{n-1} \cdots P_1. \quad (2.2.11)$$

Beweis: Sei P eine Permutationsmatrix, die nur Zeilen mit einem Index größer oder gleich $k+1$ tausche. Dann gilt

$$\widehat{G}_k = PG_kP^{-1} = \left[\begin{array}{cccc} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -\tilde{q}_{k+1,k} & 1 \\ & & \vdots & \ddots \\ & & -\tilde{q}_{n,k} & & 1 \end{array} \right],$$

dies bedeutet, dass beim Übergang von G_k auf \widehat{G}_k nur die entsprechenden Zeileneinträge der q_{ij} geändert werden. Wir schreiben nun

$$R = G_{n-1}P_{n-1}G_{n-2}P_{n-1}^{-1}P_{n-1}P_{n-2}G_{n-3}(P_{n-1}P_{n-2})^{-1}(P_{n-1}P_{n-2}P_{n-3})\cdots$$

Die Permutationsmatrizen P_{n-1} , $P_{n-1}P_{n-2}$, $P_{n-1}P_{n-2}P_{n-3}$ tauschen höchstens die letzten 2, 3, bzw. 4 Zeilen, somit gilt

$$R = G_{n-1}\widehat{G}_{n-2}\cdots\widehat{G}_1P_{n-1}P_{n-2}\cdots P_1A.$$

Nun sind das Produkt $G_{n-1}\widehat{G}_{n-2}\cdots\widehat{G}_1$ eine untere Dreiecksmatrix mit Diagonaleinträgen Null und $P = P_{n-1}P_{n-2}\cdots P_1$ eine geeignete Permutationsmatrix. Setzen wir

$$L = \left(G_{n-1}\widehat{G}_{n-2}\cdots\widehat{G}_1\right)^{-1} = \widehat{G}_1^{-1}\widehat{G}_2^{-1}\cdots\widehat{G}_{n-1}^{-1},$$

so erhalten wir die behauptete Darstellung $LR = PA$. \square

Theorem 2.2.2 Die LR-Zerlegung (nach Gauß) einer regulären Matrix ist, wenn sie existiert, eindeutig bestimmt.

Beweis: Sei $A = L_1R_1 = L_2R_2$. Dann ist

$$L_2^{-1}L_1 = R_2R_1^{-1} = I, \quad L_1^{-1}L_2 = R_1R_2^{-1} = I,$$

da $L_2^{-1}L_1, L_1^{-1}L_2$ untere Dreiecksmatrizen mit Einsen auf der Hauptdiagonalen und $R_2R_1^{-1}, R_1R_2^{-1}$ obere Dreiecksmatrizen sind. Folglich ist $L_1 = L_2$ und $R_1 = R_2$. \square

Das Gaußsche Verfahren liefert $Rx = c$ aus $Ax = b$. Äquivalent hierzu ist die Lösung zweier Dreieckssysteme, falls man die Dreieckszerlegung $PA = LR$ bereits hat:

$$PAx = LRx = Pb \quad \Leftrightarrow \quad Ly = Pb \text{ und } Rx = y.$$

Dies ist vor allem dann sinnvoll, wenn man mehrere Gleichungssysteme mit verschiedenen rechten Seiten und gleicher Koeffizientenmatrix A hat.

Die Lösung eines $n \times n$ Gleichungssystems $Ax = b$ mit Hilfe des Gaußschen Verfahrens erfordert

$$\frac{n^3}{3} + \mathbf{O}(n^2) \text{ arithm. Operationen.}$$

Dasselbe gilt für die Bestimmung der Dreieckszerlegung $PA = LR$.

Hinweis: MATLAB-Routinen

$$\begin{aligned} [L, R, P] = lu(A) & \text{ liefert } L, R \text{ und } P \text{ einer gegebenen Matrix } A \\ [L^*, R] = lu(A) & \text{ liefert } L^*, R \text{ gemäß } L^*R = P^{-1}LR = A \end{aligned}$$

Varianten und Anwendungen der Gaußelimination

Nachiteration. Rundungsfehler liefern in der Regel $\tilde{L}\tilde{R} \neq PA$ und damit nur eine Näherung x^0 mit dem Defekt $d^0 := b - Ax^0 \neq 0$. Unter Verwendung der bereits erstellten Dreieckszerlegung $\tilde{L}\tilde{R} \sim PA$ löst man nun (näherungsweise) die sog. Defektgleichung

$$A\tilde{x} = d^0$$

und erhält daraus eine Korrektur \tilde{x} für x^0 :

$$x^1 := x^0 + \tilde{x}.$$

Bei exakter Lösung der Defektgleichung wäre x^1 tatsächlich die exakte Lösung des Gleichungssystems, im allgemeinen kann man auch bei nur näherungsweise Lösung der Defektgleichung eine bessere Näherung als x^0 erwarten. Dazu muss allerdings der Defekt mit erhöhter Genauigkeit berechnet werden. Diese Überlegungen stützen sich auf die folgende Fehleranalyse.

Wir nehmen an, dass sich die relative Störung der Matrix A durch eine kleine Zahl ε abschätzen lässt. Nach dem Störungstheorem 2.1.6 gilt

$$\frac{\|x^0 - x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}} \underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim \varepsilon}$$

Der Verlust an Stellen entspricht der Größe von $\text{cond}(A)$. Zusätzlich auftretende Rundungsfehler werden vernachlässigt. Der exakte Defekt werde durch $b - \bar{A}x^0$ ersetzt, wobei \bar{A} eine genauere Approximation für A ist,

$$\frac{\|A - \bar{A}\|}{\|A\|} \leq \bar{\varepsilon} \ll \varepsilon.$$

Nach Konstruktion gilt dann

$$\begin{aligned} x^1 &= x^0 + \tilde{x} = x^0 + (\tilde{L}\tilde{R})^{-1}[b - Ax^0] \\ &= x^0 + (\tilde{L}\tilde{R})^{-1}[Ax - Ax^0 + (A - \bar{A})x^0]. \end{aligned}$$

Nun haben wir

$$\begin{aligned} x^1 - x &= x^0 - x + (\tilde{L}\tilde{R})^{-1}A(x - x^0) + (\tilde{L}\tilde{R})^{-1}(A - \bar{A})x^0 \\ &= (\tilde{L}\tilde{R})^{-1}(\tilde{L}\tilde{R} - A)(x^0 - x) + (\tilde{L}\tilde{R})^{-1}(A - \bar{A})x^0. \end{aligned}$$

Wegen

$$\begin{aligned} (\tilde{L}\tilde{R})^{-1} &= (A - A + \tilde{L}\tilde{R})^{-1} = \left[A(I - A^{-1}(A - \tilde{L}\tilde{R})) \right]^{-1} \\ &= \left(I - A^{-1}(A - \tilde{L}\tilde{R}) \right)^{-1} A^{-1} \end{aligned}$$

folgt die Abschätzung

$$\|(\tilde{L}\tilde{R})^{-1}\| \leq \|A^{-1}\| \frac{1}{1 - \|A^{-1}\| \|A - \tilde{L}\tilde{R}\|} = \|A^{-1}\| \left(1 - \text{cond}(A) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}\right)^{-1}.$$

Dies führt letztlich zu

$$\frac{\|x^1 - x\|}{\|x\|} \leq \text{cond}(A) \left(\underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim \varepsilon} \underbrace{\frac{\|x^0 - x\|}{\|x\|}}_{\sim \text{cond}(A)\varepsilon} + \underbrace{\frac{\|A - \bar{A}\|}{\|A\|}}_{\sim \bar{\varepsilon}} \frac{\|x^0\|}{\|x\|} \right).$$

In der Praxis reichen oft schon 2-3 Nachiterationen aus, um den Fehler in x auf die Größenordnung der Genauigkeit der Defektauswertung zu drücken.

Determinantenbestimmung. Ausgehend von der LR -Zerlegung $PA = LR$ erhalten wir mit $\det P = 1$

$$\det A = \det P \det A = \det L \det R = \prod_{i=1}^n r_{ii}.$$

Rangbestimmung. Ist der Gaußalgorithmus mit Spaltenpivotierung durchführbar, d.h. lassen sich immer ein nicht verschwindendes Pivotelement finden und ist auch das letzte Diagonalelement $a_{nn}^{(n-1)} \neq 0$, so ist $\text{Rang}(A) = n$. Gilt jedoch im i -ten Eliminationsschritt für alle Elemente in der i -ten Spalte

$$a_{ji}^{(i-1)} = 0, \quad j = i, \dots, n,$$

so ist A singulär. In diesem Fall wird Totalpivotierung vorgenommen (Spalten und Zeilentauch ändern den Rang einer Matrix nicht!). Gilt nun nach dem i -ten Iterationschritt

$$a_{jk}^{(i)} = 0, \quad j, k = i + 1, \dots, n,$$

so ist $\text{Rang}(A) = i$.

2.3 Spezielle Gleichungssysteme

Die Lösung sehr großer Gleichungssysteme mit dem Gaußschen Eliminationsverfahren ist mit Schwierigkeiten verbunden, insbesondere wenn der auf dem Computer verfügbare Hauptspeicher zur Speicherung nicht ausreicht. Die teilweise Auslagerung und Verwendung externer Speicher treibt die Rechenzeit aufgrund des erforderlichen Datentransfers in die Höhe und ist deshalb keine echte Alternative. Viele in der Praxis vorkommende Matrizen besitzen jedoch eine besondere Struktur, die es erlaubt, bei der Durchführung des Gaußschen Verfahrens Speicherplatz zu sparen.

Bandmatrizen

Definition 2.3.1 Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt Bandmatrix vom Bandtyp (m_l, m_r) , $0 \leq m_l, m_r \leq l - 1$, wenn

$$a_{jk} = 0 \quad \text{für } k < j - m_l \quad \text{oder } k > j + m_r, \quad j, k = 1, \dots, n.$$

Die Elemente von A sind also bis auf die Hauptdiagonale und höchstens $m_l + m_r$ Nebendiagonalen gleich Null. Die Größe $m = m_l + m_r + 1$ heißt Bandbreite.

Untere Dreiecksmatrizen sind vom Typ $(n - 1, 0)$, obere Dreiecksmatrizen vom Typ $(0, n - 1)$ und Tridiagonalmatrizen vom Typ $(1, 1)$. Ein Beispiel einer (16×16) -Matrix vom Bandtyp $(4, 4)$ ist:

$$A = \begin{bmatrix} B & I & & & \\ I & B & I & & \\ & I & B & I & \\ & & I & B & \\ & & & I & B \end{bmatrix} \quad \text{mit } B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{bmatrix}$$

Die Speicherung einer Bandmatrix $A \in \mathbb{R}^{n \times n}$ vom Bandtyp (m_l, m_r) erfolgt üblicherweise in der Form einer $(m_l + m_r + 1) \times n$ -Matrix \tilde{A} durch die Zuordnung

$$\tilde{a}_{i-j+m_r+1, j} = a_{i, j}.$$

Die obige Matrix A in kompakter Speicherung also als

$$\tilde{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ -1 & -1 & -1 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Die Bänder werden also in der oberen linken und unteren rechten Ecke mit Nullen aufgefüllt und dann "vertikal zusammengeschoben".

Die Inverse einer Bandmatrix ist im allgemeinen voll besetzt. So gilt für die obige (4×4) -Matrix B

$$B^{-1} = \frac{1}{209} \begin{bmatrix} 56 & 15 & 4 & 1 \\ 15 & 60 & 16 & 4 \\ 4 & 16 & 60 & 15 \\ 1 & 4 & 15 & 56 \end{bmatrix}.$$

Für die LR -Zerlegung benötigt man jedoch keinen zusätzlichen Speicherplatz ausserhalb des Bandes, wir haben zum Beispiel $B = LR$ mit

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -0.2500 & 1 & 0 & 0 \\ 0 & -0.2667 & 1 & 0 \\ 0 & 0 & -0.2679 & 1 \end{bmatrix} \quad R = \begin{bmatrix} 4 & -1 & 0 & 0 \\ 0 & 3.7500 & -1 & 0 \\ 0 & 0 & 3.7333 & -1 \\ 0 & 0 & 0 & 3.7321 \end{bmatrix}$$

Theorem 2.3.1 *Ist $A \in \mathbb{R}^{n \times n}$ eine Bandmatrix vom Typ (m_l, m_r) , für die das Gaußsche Eliminationsverfahren ohne Zeilenvertauschung durchführbar ist, dann sind auch alle reduzierten Matrizen Bandmatrizen desselben Typs, und die Faktoren der Dreieckszerlegung von A sind Bandmatrizen vom Typ $(m_l, 0)$ bzw. $(0, m_r)$.*

Beweis: Man überlegt sich leicht, dass die Frobeniusmatrizen Bandmatrizen vom Typ $(m_l, 0)$ sind. \square

Zur Durchführung der Gaußschen Elimination genügt es also das "Band" zu speichern, bei Größenordnungen von $n \sim 10000$ und $m \sim 100$ bedeutet dies bereits eine beträchtliche Reduktion des Speicherplatzbedarfes. Eine extreme Ersparnis ergibt sich bei Tridiagonalmatrizen

$$\begin{bmatrix} a_1 & b_1 & & & \\ c_2 & \ddots & \ddots & & \\ & \ddots & \ddots & b_{n-1} & \\ & & c_n & a_n & \end{bmatrix}.$$

Hier lassen sich die Elemente der LR -Zerlegung

$$L = \begin{bmatrix} 1 & & & & \\ \gamma_2 & \ddots & & & \\ & \ddots & 1 & & \\ & & & \ddots & \\ & & & & \gamma_n & 1 \end{bmatrix} \quad R = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \ddots & \ddots & & \\ & & \alpha_{n-1} & \beta_{n-1} & \\ & & & & \alpha_n \end{bmatrix}$$

durch einfache, rekursive Beziehungen bestimmen:

$$\begin{aligned} i = 2, \dots, n-1 : \quad & \gamma_i = c_i / \alpha_{i-1}, & \alpha_i &= a_i - \gamma_i \beta_{i-1}, & \beta_i &= b_i \\ & \gamma_n = c_n / \alpha_{n-1}, & \alpha_n &= a_n - \gamma_n \beta_{n-1} \end{aligned}$$

Hierzu sind $3n - 2$ Speicherplätze und $2n - 2$ Operationen erforderlich.

Wesentlich für Theorem 2.3.1 war, dass das Gaußsche Verfahren ohne Zeilenvertauschungen durchgeführt werden kann, da andernfalls die Bandbreite anwächst. Wir betrachten im folgenden zwei Klassen von Matrizen, bei denen dies der Fall ist.

Diagonaldominante Matrizen

Definition 2.3.2 Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt *diagonaldominant*, wenn

$$\sum_{j=1, j \neq i}^n |a_{ij}| \leq |a_{ii}|, \quad i = 1, \dots, n.$$

Theorem 2.3.2 Die Matrix $A \in \mathbb{R}^{n \times n}$ sei regulär und diagonaldominant. Dann existiert eine LR-Zerlegung $A = LR$, die mit Gaußscher Elimination ohne Pivotierung bestimmt werden kann.

Beweis: Wir stellen zunächst fest, dass $a_{11} \neq 0$ ist, da andernfalls aus der Diagonaldominanz

$$\sum_{j=2}^n |a_{1j}| \leq |a_{11}| = 0$$

folgt, also die erste Zeile nur aus Nullelementen besteht und die Matrix nicht regulär wäre. Der erste Eliminationsschritt $A \rightarrow A^{(1)}$ kann somit *ohne* Pivotierung durchgeführt werden. Die Elemente $a_{ij}^{(1)}$ werden aus bestimmt aus:

$$\begin{aligned} j = 1, \dots, n : & \quad a_{1j}^{(1)} = a_{1j} \\ i = 2, \dots, n, j = 1, \dots, n : & \quad a_{ij}^{(1)} = a_{ij} - q_{i,1}a_{1j}, \quad q_{i,1} = \frac{a_{i,1}}{a_{11}}. \end{aligned}$$

Für die neuen Zeilen $i = 2, \dots, n$ gilt:

$$\begin{aligned} \sum_{j=2, j \neq i}^n |a_{ij}^{(1)}| & \leq \sum_{j=2, j \neq i}^n |a_{ij}| + |q_{i,1}| \sum_{j=2, j \neq i}^n |a_{1j}| \\ & \leq \sum_{j=1, j \neq i}^n |a_{ij}| - |a_{i1}| + |q_{i,1}| \sum_{j=2}^n |a_{1j}| - |q_{i,1}| |a_{1i}| \\ & \leq |a_{ii}| - |q_{i,1}a_{1i}| \leq |a_{ii}^{(1)}| \end{aligned}$$

Die Matrix $A^{(1)}$ ist regulär und wieder diagonaldominant, und folglich ist $a_{22}^{(1)} \neq 0$. Die Eigenschaft bleibt also in jedem Schritt der Gaußschen Elimination erhalten. Wir können die Elimination folglich ohne Zeilenvertauschungen durchführen. \square

Positiv definite Matrizen

Theorem 2.3.3 Für positiv definite Matrizen $A \in \mathbb{R}^{n \times n}$ ist das Gaußsche Eliminationsverfahren ohne Zeilenvertauschung durchführbar, und die dabei auftretenden Pivotelemente $a_{ii}^{(i)}$ sind positiv.

Beweis: Da eine positiv definite Matrix positive Diagonalelemente hat, ist insbesondere $a_{11} > 0$. Die Beziehung

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} = a_{ij} - \frac{a_{1i}a_{j1}}{a_{11}} = a_{ij}^{(1)}, \quad i, j = 2, \dots, n, \quad (2.3.12)$$

zeigt, dass die im ersten Eliminationsschritt erzeugte $(n-1) \times (n-1)$ -Matrix $\tilde{A}^{(1)} = \left(a_{ij}^{(1)} \right)_{i,j=2,\dots,n}$ symmetrisch ist. Wir wollen zeigen, dass sie auch positiv definit ist. Dann folgt wieder $a_{22}^{(1)} > 0$ und der Eliminationsprozeß kann mit positivem Pivotelement fortgesetzt werden. Die positive Definitheit zeigen wir durch Induktion. Dazu sei $\tilde{x} = (x_2, \dots, x_n)^T \in \mathbb{R}^{n-1} \setminus \{0\}$ und $x = (x_1, \tilde{x})^T \in \mathbb{R}^n$ mit

$$x_1 = -\frac{1}{a_{11}} \sum_{k=2}^n a_{1k} x_k.$$

Nun ist

$$0 < \sum_{j,k=1}^n a_{jk} x_j x_k = \sum_{j,k=2}^n a_{jk} x_j x_k + 2x_1 \sum_{k=2}^n a_{1k} x_k + a_{11} x_1^2.$$

Die Nullergänzung

$$0 = -\frac{1}{a_{11}} \sum_{j,k=2}^n a_{k1} a_{1j} x_k x_j + \frac{1}{a_{11}} \left\{ \sum_{k=2}^n a_{1k} x_k \right\}^2$$

führt zu

$$\begin{aligned} 0 < \sum_{j,k=1}^n a_{jk} x_j x_k &= \sum_{j,k=2}^n \left\{ a_{jk} - \frac{a_{k1} a_{1j}}{a_{11}} \right\} x_j x_k + a_{11} \left\{ x_1 + \frac{1}{a_{11}} \sum_{k=2}^n a_{1k} x_k \right\}^2 \\ &= \sum_{j,k=2}^n a_{jk}^{(1)} x_j x_k \end{aligned}$$

und damit zu $\tilde{x}^T \tilde{A}^{(1)} \tilde{x} > 0$. □

Für positiv definite Matrizen existiert also stets eine LR -Zerlegung $A = LR$ mit positiven Pivotelementen

$$r_{ii} = a_{ii}^{(i)} > 0, \quad i = 1, \dots, n.$$

Wegen $A = A^T$ gilt aber auch

$$A = A^T = (LR)^T = (LDR\tilde{R})^T = \tilde{R}^T DL^T$$

mit den Matrizen

$$\tilde{R} = \begin{bmatrix} 1 & r_{12}/r_{11} & \dots & r_{1n}/r_{11} \\ & \ddots & & \vdots \\ & & 1 & r_{n-1,n}/r_{n-1,n-1} \\ & & & 1 \end{bmatrix}, \quad D = \begin{bmatrix} r_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & r_{nn} \end{bmatrix}.$$

Wegen der Eindeutigkeit der LR -Zerlegung folgt aus

$$A = LR = \tilde{R}^T DL^T$$

notwendig $L = \tilde{R}^T$ bzw. $R = DL^T$. Positiv definite Matrizen gestatten also eine sog. *Cholesky-Zerlegung*

$$A = LDL^T = \tilde{L}\tilde{L}^T$$

mit der Matrix $\tilde{L} = LD^{1/2}$. Bei der Berechnung der Cholesky-Zerlegung genügt es, die Matrizen D und L zu bestimmen. Dies reduziert den Speicherplatzbedarf auf $n(n+1)/2$ und die benötigten Operationen auf $n^3/6 + \mathbf{O}(n^2)$.

Der *Algorithmus von Cholesky* zur Berechnung der Zerlegungsmatrix

$$\tilde{L} = \begin{bmatrix} \tilde{l}_{11} & & & \\ \vdots & \ddots & & \\ \tilde{l}_{n1} & \dots & \tilde{l}_{nn} & \end{bmatrix}$$

geht direkt von der Beziehung $A = \tilde{L}\tilde{L}^T$ aus, die als System von $n(n+1)/2$ Gleichungen für die Größen \tilde{l}_{jk} , $k \leq j$, auffassen kann. Ausmultiplizieren von

$$\begin{bmatrix} \tilde{l}_{11} & & & \\ \vdots & \ddots & & \\ \tilde{l}_{n1} & \dots & \tilde{l}_{nn} & \end{bmatrix} \begin{bmatrix} \tilde{l}_{11} & \dots & \tilde{l}_{n1} \\ & \ddots & \vdots \\ & & \tilde{l}_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

ergibt in der ersten Spalte von \tilde{L} :

$$\tilde{l}_{11}^2 = a_{11}, \quad \tilde{l}_{21}\tilde{l}_{11} = a_{21}, \quad \dots, \quad \tilde{l}_{n1}\tilde{l}_{11} = a_{n1}$$

woraus sich

$$\tilde{l}_{11} = \sqrt{a_{11}}, \quad j = 2, \dots, n: \quad \tilde{l}_{j1} = \frac{a_{j1}}{\tilde{l}_{11}},$$

berechnen. Seien für ein $i \in \{2, \dots, n\}$ die Elemente \tilde{l}_{jk} , $k = 1, \dots, i-1$, $j = k, \dots, n$, schon bekannt. Dann erhält man aus

$$\begin{aligned} \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \dots + \tilde{l}_{ii}^2 &= a_{ii}, \quad \tilde{l}_{ii} > 0 \\ \tilde{l}_{j1}\tilde{l}_{i1} + \tilde{l}_{j2}\tilde{l}_{i2} + \dots + \tilde{l}_{ji}\tilde{l}_{ii} &= a_{ji} \end{aligned}$$

die nächsten Elemente \tilde{l}_{ii} und \tilde{l}_{ji} , $j = i+1, \dots, n$.

2.4 Nicht reguläre Systeme

Wir betrachten nun allgemeinere Gleichungssysteme der Form $Ax = b$, bei denen $A \in \mathbb{R}^{m \times n}$ eine gegebene, nicht notwendig quadratische Matrix und $b \in \mathbb{R}^m$ die rechte Seite bezeichnen. Wir lassen auch den Fall $\text{Rang}(A) < \text{Rang}[A, b]$ zu, d. h. das System muß nicht im eigentlichen Sinne lösbar sein. Wir verallgemeinern den Lösungsbegriff in dem Sinne, dass ein Vektor $\bar{x} \in \mathbb{R}^n$ gesucht wird, dessen Defekt $d \equiv b - A\bar{x}$ die kleinste euklidische Norm besitzt. Im Falle $\text{Rang}(A) = \text{Rang}[A, b]$ fällt dieser Lösungsbegriff mit dem üblichen zusammen.

Theorem 2.4.1 *Es existiert eine verallgemeinerte Lösung $\bar{x} \in \mathbb{R}^n$ von $Ax = b$ mit kleinsten Fehlerquadraten*

$$\|A\bar{x} - b\|_2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2.$$

Dies ist äquivalent dazu, dass \bar{x} Lösung der Normalgleichung

$$A^T Ax = A^T b$$

ist. Im Fall $\text{Rang}(A) = n$ ist \bar{x} eindeutig bestimmt, andernfalls ist jede Lösung von der Form $\bar{x} + y$, mit $y \in \text{Kern}(A)$.

Beweis. Für eine Minimallösung \bar{x} gilt notwendig

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_i} \|Ax - b\|_2^2 \Big|_{x=\bar{x}} = \frac{\partial}{\partial x_i} \left(\sum_{j=1}^n \left| \sum_{k=1}^m a_{jk} x_k - b_j \right|^2 \right) \Big|_{x=\bar{x}} \\ &= 2 \sum_{j=1}^n a_{ji} \left(\sum_{k=1}^m a_{jk} \bar{x}_k - b_j \right) = 2(A^T A \bar{x} - A^T b)_i, \end{aligned}$$

somit löst \bar{x} die Normalgleichung. Sei umgekehrt \bar{x} Lösung der Normalgleichung. Für beliebiges $x \in \mathbb{R}^n$ gilt dann

$$\begin{aligned} \|b - Ax\|_2^2 &= \|b - A\bar{x} + A(\bar{x} - x)\|_2^2 \\ &= \|b - A\bar{x}\|_2^2 + 2(b - A\bar{x}, A(\bar{x} - x)) + \|A(\bar{x} - x)\|_2^2. \end{aligned}$$

Das orthogonale Komplement von $\text{Bild}(A)$ in \mathbb{R}^m ist $\text{Kern}(A^T)$. Nun liegt aber $b - A\bar{x}$ im Kern von A^T , falls \bar{x} eine Lösung der Normalgleichung ist. Folglich

$$\|b - Ax\|_2^2 = \|b - A\bar{x}\|_2^2 + \|A(\bar{x} - x)\|_2^2 \geq \|b - A\bar{x}\|_2^2.$$

Bleibt die Lösbarkeit des Normalgleichungssystems zu untersuchen. Wegen $\mathbb{R}^m = \text{Bild}(A) \oplus \text{Kern}(A^T)$ kann b eindeutig in

$$b = s + r, \quad s \in \text{Bild}(A), \quad r \in \text{Kern}(A^T)$$

zerlegt werden. Für ein $\bar{x} \in \mathbb{R}^n$ mit $A\bar{x} = s$ gilt dann

$$A^T A\bar{x} = A^T s = A^T s + A^T r = A^T b,$$

d.h. \bar{x} löst das Normalgleichungssystem. Im Fall $\text{Rang}(A) = n$ ist $\text{Kern}(A) = \{0\}$ und $\text{Bild}(A) = \mathbb{R}^n$. Aus $A^T Ax = 0$ folgt wegen $\text{Kern}(A^T) \perp \text{Bild}(A)$ notwendig $Ax = 0$ bzw. $x = 0$. Die Matrix $A^T A \in \mathbb{R}^{n \times n}$ ist also regulär und \bar{x} eindeutig bestimmt. Im Fall $\text{Rang}(A) < n$ gilt für jede weitere Lösung x_1 der Normalgleichung

$$b = Ax_1 + (b - Ax_1) \in \text{Bild}(A) + \text{Kern}(A^T).$$

Aus der Eindeutigkeit dieser orthogonalen Zerlegung schliessen wir $Ax_1 = s = A\bar{x}$ und $\bar{x} - x_1 \in \text{Kern}(A)$. \square

Die klassische Anwendung des Satzes ist die sog. Gaußsche Ausgleichsrechnung. Die Aufgabenstellung besteht dabei in folgendem: Zu gegebenen Funktionen $\varphi_1, \dots, \varphi_n$ und Punkten $(x_j, y_j) \in \mathbb{R}^2$, $j = 1, \dots, m$, $m > n$, ist eine Linearkombination

$$\varphi(x) = \sum_{k=1}^n c_k \varphi_k(x)$$

derart zu bestimmen, dass die mittlere Abweichung

$$\left(\sum_{j=1}^m |\varphi(x_j) - y_j|^2 \right)^{1/2}$$

möglichst klein wird. Zur Lösung dieser Aufgabe setzen wir

$$y := (y_1, \dots, y_m)^T, \quad c := (c_1, \dots, c_n)^T, \\ a_k := (\varphi_k(x_1), \dots, \varphi_k(x_m))^T, \quad k = 1, \dots, n, \quad A := [a_1, \dots, a_n]$$

Dann ist das Funktional

$$F(c) = \|Ac - y\|_2$$

bezüglich $c \in \mathbb{R}^n$ zu minimieren. Dies ist gleichbedeutend damit, für das überbestimmte ($m > n$) Gleichungssystem $Ac = y$ eine verallgemeinerte Lösung mit kleinsten Fehlerquadraten zu bestimmen. Im Fall $\text{Rang}(A) = n$ ist diese eindeutig als Lösung der Normalgleichung

$$A^T Ac = A^T y$$

gegeben. Für den Spezialfall polynomialer Funktionen $\varphi_k(x) = x^{k-1}$ ist die Matrix A gegeben durch

$$A = \begin{bmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ \vdots & & \ddots & \\ 1 & x_m & \dots & x_m^{n-1} \end{bmatrix}, \quad m > n.$$

Wegen der Regularität der *Vandermondeschen Determinante*

$$\begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ \vdots & & & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{vmatrix} = \prod_{\substack{j,k=1 \\ j < k}}^n (x_k - x_j) \neq 0$$

für paarweise verschiedene Stützstellen x_j ist dann $\text{Rang}(A) = n$ und die Gaußsche Ausgleichsaufgabe eindeutig lösbar.

Beispiel. Zu den Meßdaten

$$\begin{array}{c|cccccc} x_j & -2 & -1 & 0 & 1 & 2 \\ \hline y_j & 0.5 & 0.5 & 2 & 3.5 & 3.5 \end{array}$$

soll mit Hilfe der Gaußschen Ausgleichsrechnung eine lineare Funktion $y(x) = a + bx$ angepasst werden. Dies ist äquivalent zur Lösung des überbestimmten Gleichungssystems

$$\begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 2 \\ 3.5 \\ 3.5 \end{bmatrix}.$$

Das zugehörige Normalgleichungssystem lautet

$$\begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 10 \\ 9 \end{bmatrix}$$

mit der Lösung $a = 2$ und $b = 0.9$.

Basierend auf den folgenden Satz gelingt die Berechnung von \bar{x} auch ohne explizite Aufstellung der Normalgleichung.

Theorem 2.4.2 Sei $A \in \mathbb{R}^{m \times n}$ eine rechteckige Matrix mit $m \geq n$ und $\text{Rang}(A) = n$. Dann existiert eine eindeutig bestimmte (orthogonale) Matrix $Q \in \mathbb{R}^{m \times n}$ mit der Eigenschaft $Q^T Q = I$ und eine eindeutig bestimmte obere Dreiecksmatrix $R \in \mathbb{R}^{n \times n}$ mit positiven Diagonalelementen $r_{ii} > 0$, $i = 1, \dots, n$, so dass $A = QR$.

Wenden wir die QR -Zerlegung auf die Normalgleichung an, so folgt

$$A^T A x = R^T Q^T Q R x = R^T R x = R^T Q^T b$$

und wegen der Regularität von $R^T R x = Q^T b$. Dieses System kann durch Rückwärtseinsetzen gelöst werden. Wegen

$$A^T A = R^T R$$

ist mit R also die Cholesky-Zerlegung von $A^T A$ bestimmt, ohne $A^T A$ explizit berechnen zu müssen. Ein stabiles Verfahren zur Berechnung der QR -Zerlegung ist das *Householder-Verfahren*.

Kapitel 3

Interpolation

Ein Grundproblem der Numerik ist die Darstellung und Auswertung von Funktionen. Dabei ergeben sich folgende Aufgabenstellungen:

- Eine Funktion f ist nur auf einer diskreten Menge von Argumenten x_0, \dots, x_n bekannt und soll mit dieser Information rekonstruiert werden, um z.B. die Funktion grafisch darzustellen oder um die Funktion an gewissen Zwischenstellen auszuwerten.
- Eine analytisch gegebene Funktion f soll auf einer Rechenanlage so dargestellt werden, dass Funktionswerte $f(x)$ zu beliebigen x leicht innerhalb einer vorgegebene Toleranz berechnet werden können.

In beiden Fällen hat man ein System mit unendlich vielen Freiheitsgraden, nämlich die funktionale Abhängigkeit $y = f(x)$, durch einen endlichen Datensatz zu simulieren. Hierzu bedient man sich gewisser Klassen P einfach strukturierter Funktionen; z.B.

Polynome:	$p(x) = c_0 + c_1x + c_2x^2 + \dots + c_nx^n$
Rationale Funktionen:	$r(x) = \frac{c_0 + c_1x + c_2x^2 + \dots + c_nx^n}{b_0 + b_1x + b_2x^2 + \dots + b_mx^m}$
Trigonometrische Polynome:	$t(x) = \frac{a_0}{2} + \sum_{k=1}^n \{a_k \cos(kx) + b_k \sin(kx)\}$
Exponentialsummen:	$e(x) = \sum_{k=1}^n a_k \exp(b_kx).$

Geschieht die Zuordnung eines Elementes $g \in P$ zur Funktion f durch Fixieren von Funktionswerten, etwa

$$g(x_i) = y_i := f(x_i), \quad i = 0, \dots, n,$$

so spricht man von *Interpolation*. Ist $g \in P$ als in einem gewissen Sinn beste Darstellung von f zu bestimmen, etwa als

$$\max_{a \leq x \leq b} |f(x) - g(x)| \quad \text{minimal für } g \in P,$$

so spricht man von *Approximation*. Die jeweilige Wahl der Konstruktion von $g \in P$ hängt von der zu erfüllenden Aufgabe ab. In diesem Abschnitt behandeln wir Interpolationsaufgaben, im Abschnitt 5 wenden wir uns der Frage der Approximation zu.

3.1 Polynominterpolation

Wir bezeichnen mit P_n den Vektorraum der Polynome vom Grad kleiner oder gleich n :

$$P_n := \{p(x) = c_0 + c_1x + \dots + c_nx^n : c_i \in \mathbb{R}, i = 0, \dots, n\}.$$

Die *Lagrange'sche Interpolationsaufgabe* besteht darin, zu $n + 1$ paarweise verschiedenen Stützstellen (Knoten) $x_0, x_1, \dots, x_n \in \mathbb{R}$ ein Polynom $p \in P_n$ zu bestimmen, das in x_i vorgegebene Werte y_i annimmt, für das also $p(x_i) = y_i$ gilt.

Theorem 3.1.1 *Die Lagrange'sche Interpolationsaufgabe besitzt eine eindeutig bestimmte Lösung.*

Beweis: Wir zeigen zunächst die Eindeutigkeit. Angenommen es gäbe zwei Lösungen $p_1, p_2 \in P_n$, dann verschwindet das Polynom $p := p_1 - p_2 \in P_n$ in den $n + 1$ paarweise verschiedenen Punkten x_0, \dots, x_n , und ist folglich Null. Zur Existenz betrachten wir die Forderungen $p(x_i) = y_i$, $i = 0, \dots, n$, als $n + 1$ Gleichungen zur Bestimmung der $n + 1$ unbekanntenen Koeffizienten c_i , $i = 0, \dots, n$. Da das homogene System nur die Nulllösung hat (Eindeutigkeit!), hat das inhomogene System stets eine Lösung. \square

Zur expliziten Konstruktion des Interpolationspolynoms $p \in P_n$ verwendet man die *Lagrange'schen Basispolynome*

$$L_i^{(n)}(x) = \prod_{k=0, k \neq i}^n \frac{x - x_k}{x_i - x_k} \in P_n, \quad i = 0, \dots, n.$$

Eine wichtige Eigenschaft der Lagrange'schen Basispolynome ist

$$L_i^{(n)}(x_j) = \delta_{ij}.$$

Damit kann das gesuchte Polynom in der Form

$$p = \sum_{i=0}^n y_i L_i^{(n)} \in P_n$$

geschrieben werden. Das Polynom heißt Lagrange'sches Interpolationspolynom zu den Stützpunkten (x_i, y_i) , $i = 0, \dots, n$.

Tip: In MATLAB können die $n + 1$ Stützpunkte (x_i, y_i) in zwei Vektoren x und y gespeichert werden. Dann liefert

$$c = \text{polyfit}(x, y, n)$$

im Vektor $c = (c(1), \dots, c(n + 1))$ die Koeffizienten des zugeordneten Interpolationspolynoms

$$p(x) = \sum_{i=1}^{n+1} c(i)x^{n+1-i}.$$

Benötigt man die Ableitung des Interpolationspolynoms, so kann man mit

$$d = \text{polyder}(c)$$

den Koeffizientenvektor der Ableitung bestimmen.

Das Lagrange'sche Interpolationspolynom hat den Nachteil, dass sich die verwendeten Basisfunktionen von P_n bei Hinzunahme einer weiteren Stützstelle völlig ändern, vgl. Abb. 3.1. Dieser Effekt kann durch Übergang zu einer an-

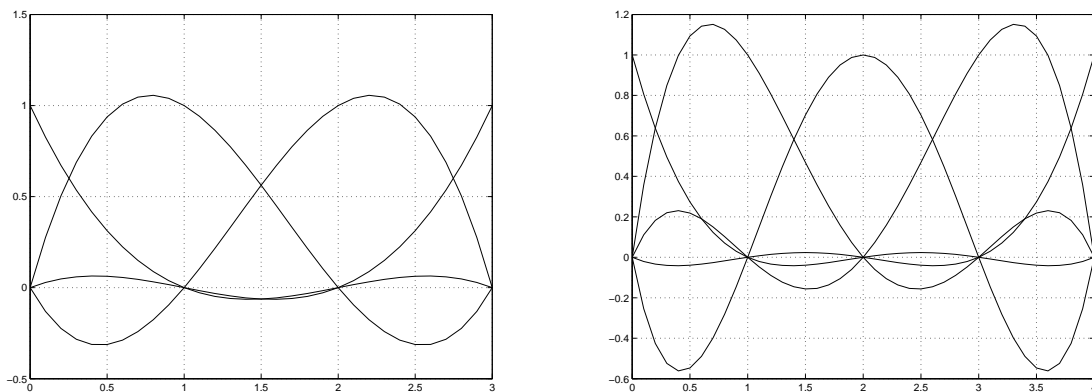


Abbildung 3.1: Basispolynome $L_i^{(3)}$ auf der Knotenmenge $\{0, 1, 2, 3\}$ (links) und $L_i^{(4)}$ auf der Knotenmenge $\{0, 1, 2, 3, 4\}$ (rechts).

deren (hierarchischen) Basis in P_n vermieden werden. Dazu verwendet man die Basisfunktionen

$$N_0(x) := 1, \quad N_{i+1}(x) := (x - x_i)N_i(x), \quad i = 0, \dots, n - 1.$$

Die rekursive Definition führt auf

$$N_i(x) = \prod_{j=0}^{i-1} (x - x_j), \quad i = 1, \dots, n.$$

Aus dem Ansatz

$$p(x) = \sum_{i=0}^n a_i N_i(x)$$

folgt dann das gestaffelte Gleichungssystem

$$\begin{aligned} y_0 &= p(x_0) = a_0 \\ y_1 &= p(x_1) = a_0 + a_1(x_1 - x_0) \\ &\vdots \\ y_n &= p(x_n) = a_0 + a_1(x_n - x_0) + \cdots + a_n(x_n - x_0) \cdots (x_n - x_{n-1}), \end{aligned}$$

aus dem sukzessive die a_i ermittelt werden können. Insbesondere ist die Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) problemlos möglich, denn die bislang berechneten a_i , $i = 0, \dots, n$, ändern sich bei Hinzunahme der neuen Basisfunktion N_{n+1} nicht. Praktisch bestimmt man die Koeffizienten a_i jedoch auf eine andere Weise, die im folgenden beschrieben wird.

Sei $p_{i,i+k} \in P_k$ das Polynom, das die Stützpunkte $(x_i, y_i), \dots, (x_{i+k}, y_{i+k})$ interpoliert. Zu den Punkten (x_i, y_i) definieren wir die *dividierten Differenzen* $y[x_i, \dots, x_{i+k}]$ rekursiv durch

$$\begin{aligned} i = 0, \dots, n : & \quad y[x_i] := y_i \\ k = 1, \dots, n - i : & \quad y[x_i, \dots, x_{i+k}] := \frac{y[x_{i+1}, \dots, x_{i+k}] - y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i} \end{aligned}$$

Theorem 3.1.2 Für $i = 0, \dots, n$ und $k = 0, \dots, n - i$ gilt

$$\begin{aligned} p_{i,i+k}(x) &= y[x_i] + y[x_i, x_{i+1}](x - x_i) + \dots \\ &\quad \cdots + y[x_i, \dots, x_{i+k}](x - x_i) \cdots (x - x_{i+k-1}). \end{aligned}$$

Beweis: Induktion bezüglich der Indexdifferenz $k = (i+k) - i$. Für $k = 0$ ist $p_{i,i} = y_i = y[x_i]$, $i = 0, \dots, n$. Sei die Behauptung nun für $k - 1$ richtig. Da $p_{i,i+k} \in P_k$ das Polynom bezeichnet, das die Stützpunkte $(x_i, y_i), \dots, (x_{i+k}, y_{i+k})$ interpoliert, gilt mit gewissem $a \in \mathbb{R}$

$$p_{i,i+k}(x) = p_{i,i+k-1}(x) + a(x - x_i) \cdots (x - x_{i+k-1}).$$

Zu zeigen ist also $a = y[x_i, \dots, x_{i+k}]$. Nach Induktionsannahme ist

$$\begin{aligned} p_{i,i+k-1}(x) &= \cdots + y[x_i, \dots, x_{i+k-1}]x^{k-1}, \\ p_{i+1,i+k}(x) &= \cdots + y[x_{i+1}, \dots, x_{i+k}]x^{k-1}, \end{aligned}$$

wobei “...” für Polynomanteile vom Grad kleiner oder gleich $k - 2$ steht. Das Polynom

$$q(x) := \frac{(x - x_i) p_{i+1,i+k}(x) - (x - x_{i+k}) p_{i,i+k-1}(x)}{x_{i+k} - x_i}$$

interpoliert die $k + 1$ Stützpunkte (x_j, y_j) , $j = i, \dots, i + k$. In der Tat haben wir für $j = i + 1, \dots, i + k - 1$

$$\begin{aligned} q(x_i) &= p_{i,i+k-1}(x_i) = y_i, \\ q(x_j) &= \frac{(x_j - x_i) p_{i+1,i+k}(x_j) - (x_j - x_{i+k}) p_{i,i+k-1}(x_j)}{x_{i+k} - x_i} \\ &= \frac{(x_j - x_i) y_j - (x_j - x_{i+k}) y_j}{x_{i+k} - x_i} = y_j, \\ q(x_{i+k}) &= p_{i+1,i+k}(x_{i+k}) = y_{i+k}. \end{aligned}$$

Aus der Eindeutigkeit des Interpolationspolynoms folgt $q = p_{i,i+k}$. Für den Koeffizienten bei x^k in q beziehungsweise $p_{i,i+k}$ gilt somit

$$a = \frac{y[x_{i+1}, \dots, x_{i+k}] - y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i} = y[x_i, \dots, x_{i+k}],$$

was zu zeigen war. □

Setzt man $i = 0$ und $k = n$, so erhält man die *Newton'sche Darstellung* des Interpolationspolynoms zu den Stützpunkten $(x_0, y_0), \dots, (x_n, y_n)$

$$p(x) = \sum_{i=0}^n y[x_0, \dots, x_i] N_i(x).$$

Zur Berechnung der Koeffizienten $a_i = y[x_0, \dots, x_i]$ nutzt man folgendes Schema

$x_n - x_0$	\dots	$x_2 - x_0$	$x_1 - x_0$	x_0	y_0	$y[x_0, x_1]$	$y[x_0, x_1, x_2]$	\dots	$y[x_0, \dots, x_n]$
		$x_3 - x_1$	$x_2 - x_1$	x_1	y_1	$y[x_1, x_2]$	$y[x_1, x_2, x_3]$		
			$x_3 - x_2$	x_2	y_2	$y[x_2, x_3]$			
				\vdots	\vdots				
				x_n	y_n				

Bei Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) berechnet man den Koeffizienten $y[x_0, \dots, x_{n+1}]$ im Newton'schen Interpolationspolynom einfach durch Berechnung einer weiteren Diagonalen im obigen Schema.

Die im Beweis des Theorem 3.1.2 verwendete Beziehung zwischen den Polynomen $p_{i,i+k}$ kann direkt zur rekursiven Berechnung des Interpolationspolynoms $p = p_{0,n}$ verwendet werden. Das durch

$$\begin{aligned} p_{i,i}(x) &= y_i & i &= 0, \dots, n, \\ p_{i,i+k}(x) &= p_{i,i+k-1}(x) + (x - x_i) \frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i} & k &= 0, \dots, n - i, \end{aligned}$$

erzeugte Polynom $p_{0,n}$ ist die *Neville'sche Darstellung* des Interpolationspolynoms zu den Stützstellen $(x_0, y_0), \dots, (x_n, y_n)$ (kurz: Neville'sches Interpolationspolynom). Die praktische Berechnung erfolgt nach folgendem Schema:

x_0	y_0	$p_{0,1}(x)$	$p_{0,2}(x)$	$p_{0,3}(x)$	\dots	$p_{0,n-1}(x)$	$p_{0,n}(x)$
x_1	y_1	$p_{1,2}(x)$	$p_{1,3}(x)$	$p_{1,4}(x)$	\dots	$p_{1,n}(x)$	
x_2	y_2	$p_{2,3}(x)$	$p_{2,4}(x)$	$p_{2,5}(x)$	\dots		
\vdots	\vdots	\vdots	\vdots	\vdots			
x_{n-1}	y_{n-1}	$p_{n-1,n}(x)$					
x_n	y_n						

Auch hier ist die Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) problemlos möglich. Die Neville'sche Darstellung des Interpolationspolynoms bietet eine sehr effiziente Möglichkeit zur Berechnung einzelner Funktionswerte $p(\xi)$ ($\xi \neq x_i$) ohne vorherige Bestimmung der Koeffizienten in der Newtonschen Darstellung. Dazu setzt man im obigen Neville-Schema einfach $x = \xi$ und verwendet zur Berechnung von $p_{i,k} := p_{i,k}(\xi)$ die Rekursionsformeln

$$\begin{aligned}
 i &= 0, \dots, n & p_{i,i} &= y_i \\
 k &= 1, \dots, n-i & p_{i,i+k} &= p_{i,i+k-1} + \frac{p_{i,i+k-1} - p_{i+1,i+k}}{(\xi - x_{i+k})/(\xi - x_i) - 1}.
 \end{aligned}$$

3.2 Interpolationsfehler

Wir wollen den Interpolationsfehler abschätzen, der bei Ersetzung einer gegebenen Funktion f durch ihr Interpolationspolynom p_n mit den Knoten x_0, x_1, \dots, x_n entsteht. Dazu bezeichne $[x_0, \dots, x_n] \subset [a, b]$ das kleinste Intervall, dass alle in der Klammer eingeschlossenen Punkte enthält.

Theorem 3.2.1 *Sei $f \in C^{n+1}[a, b]$. Dann gibt es zu jedem $x \in [a, b]$ ein $\xi_x \in [x_0, \dots, x_n, x]$, so dass*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^n (x - x_j).$$

Beweis: Für $x \in \{x_0, \dots, x_n\}$ ist die Behauptung aufgrund der Interpolationseigenschaft des Polynoms p_n trivial. Sei nun $x \in [a, b] \setminus \{x_0, \dots, x_n\}$. Wir setzen

$$l(t) := \prod_{j=0}^n (t - x_j), \quad c(x) := \frac{f(x) - p_n(x)}{l(x)}.$$

Die Funktion $t \mapsto F(t) := f(t) - p_n(t) - c(x)l(t)$ besitzt in $[a, b]$ mindestens die $n+2$ Nullstellen x_0, x_1, \dots, x_n, x . Aus der wiederholten Anwendung des Satzes

von Rolle folgt, dass die Ableitung $t \mapsto F^{(n+1)}$ eine Nullstelle $\xi_x \in [x_0, \dots, x_n, x]$ hat. Wegen

$$\begin{aligned} 0 &= F^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - p_n^{(n+1)}(\xi_x) - c(x)l^{(n+1)}(\xi_x) \\ &= f^{(n+1)}(\xi_x) - c(x)(n+1)! \end{aligned}$$

folgt die Behauptung des Satzes. \square

Wir wollen den Fehler bei der Lagrange'schen Interpolation diskutieren. Für großes n wird $1/(n+1)!$ sehr klein. Das Produkt wird klein, wenn die Stützstellen immer stärker zusammenrücken. Sind also die Ableitungen von f auf $[a, b]$ beschränkt, so gilt

$$\max_{a \leq x \leq b} |f(x) - p_n(x)| \rightarrow 0 \quad n \rightarrow \infty.$$

Oft sind jedoch die Ableitung der zu interpolierenden Funktionen nicht beschränkt, z.B.

$$f(x) = \frac{1}{1+x^2}, \quad |f^{(n)}(x)| \approx 2^n n! \mathcal{O}(|x|^{-(n+2)}),$$

so dass gleichmäßige Konvergenz nicht zu erwarten ist. Der Weierstraß'sche Approximationssatz besagt, dass jede auf $[a, b]$ stetige Funktion beliebig genau durch Polynome approximiert werden kann. Die Vermutung, dass dies mit Lagrange'schen Interpolationspolynomen auf äquidistanten Stützstellen geschehen kann, ist jedoch im allgemeinen falsch.

Beispiel: Seien $f(x) = |x|$, $x \in [-1, +1]$, Stützstellen $x_i = -1 + ih$, $i = 0, \dots, 2m$, $h = 1/m$; $x \notin \{-1, 0, 1\}$. In Abb. 3.2 sind die Interpolationspolynome der Betragsfunktion für $m = 4$, $m = 8$, $m = 12$ und $m = 16$ dargestellt. Man erkennt an den Intervallgrenzen deutlich einen Trend zum Überschwingen. Eine wesentlich Verbesserung kann durch Übergang zu nicht äquidistanten Knoten erreicht werden. Abb. 3.3 zeigt das Interpolationspolynom vom Grade $m = 16$ bei Verwendung der Tschebyscheff-Knoten $x_i = -\cos(\pi i/16)$, $i = 0, \dots, 16$. Die Tschebyscheff-Knoten auf dem Intervall $[a, b]$ sind durch

$$x_i = \frac{a+b}{2} - \frac{b-a}{2} \cos \frac{\pi i}{n}, \quad i = 0, \dots, n,$$

gegeben. \square

Für die mit den Stützpunkten $(x_i, f(x_i))$ gebildeten dividierten Differenzen schreiben wir $f[x_i, \dots, x_{i+k}] = y[x_i, \dots, x_{i+k}]$.

Theorem 3.2.2 Sei $f \in C^{n+1}[a, b]$. Dann gilt für $x \in [a, b] \setminus \{x_0, \dots, x_n\}$ die Darstellung

$$f(x) - p_n(x) = f[x_0, \dots, x_n, x] \prod_{j=0}^n (x - x_j)$$

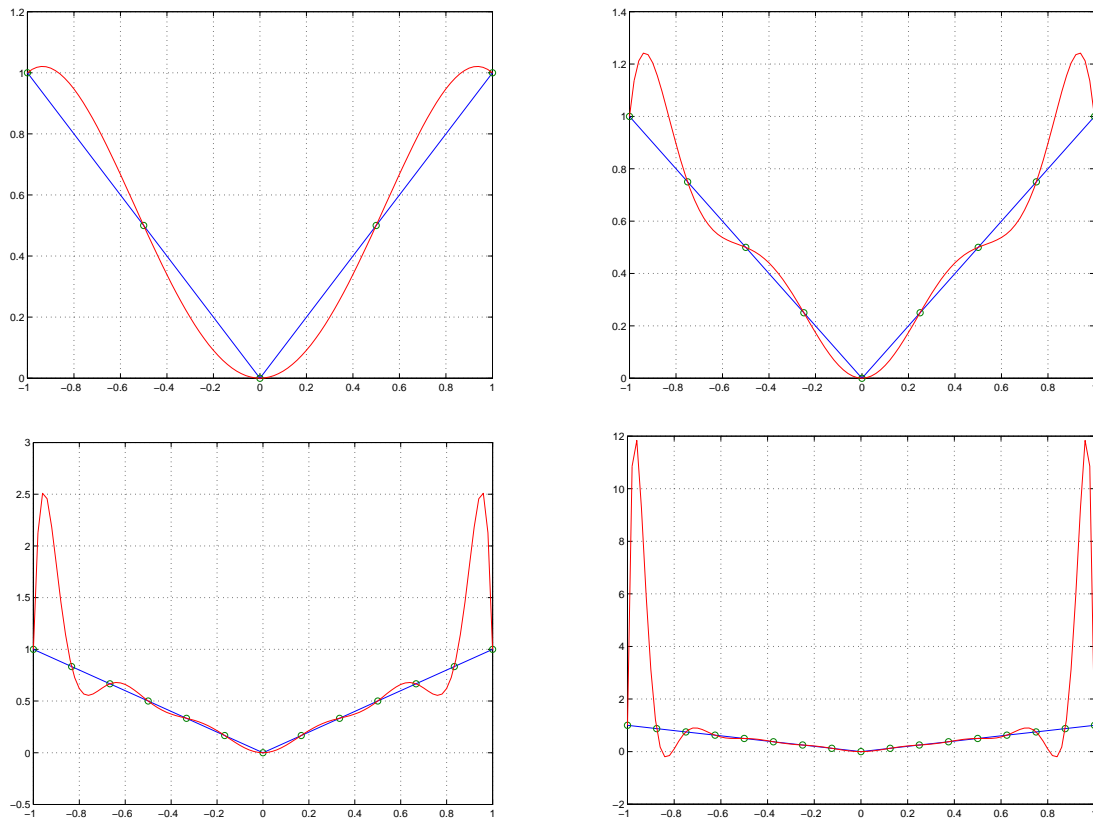


Abbildung 3.2: Interpolationspolynome der Betragsfunktion der Ordnung 4, 8, 12 und 16 bei Verwendung äquidistanter Stützstellen.

und es ist

$$f[x_0, \dots, x_n, x] = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} \int_0^{t_n} f^{(n+1)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t(x - x_n)) dt dt_n \cdots dt_2 dt_1.$$

Beweis: Wir führen den Beweis durch vollständige Induktion nach der Anzahl der Stützstellen (in der Reihung x_0, \dots, x_n). Für $n = 0$ gilt

$$\begin{aligned} f(x) - p_0(x) &= f(x) - f(x_0) = f[x_0, x](x - x_0) \\ &= (x - x_0) \int_0^1 f'(x_0 + t(x - x_0)) dt. \end{aligned}$$

Gelte die Behauptung nun für $n - 1 \geq 0$, wir zeigen die Richtigkeit für n . Aus der Newton'schen Darstellung des Interpolationspolynoms und der Richtigkeit

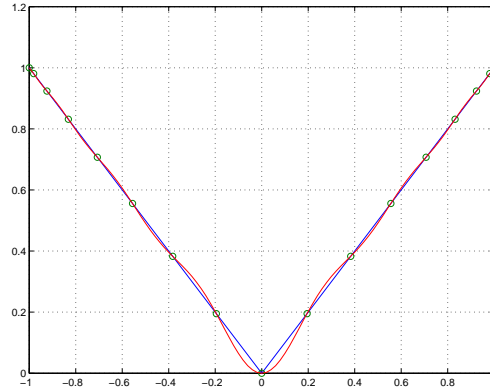


Abbildung 3.3: Interpolationspolynom der Betragsfunktion der Ordnung 16 bei Verwendung der Tschebyscheff-Knoten.

für $n - 1$ folgt

$$\begin{aligned}
 f(x) - p_n(x) &= f(x) - \sum_{i=0}^n f[x_0, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j) \\
 &= f(x) - p_{n-1}(x) - f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x - x_j) \\
 &= f[x_0, \dots, x_{n-1}, x] \prod_{j=0}^{n-1} (x - x_j) - f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x - x_j).
 \end{aligned}$$

Wegen $f[x_0, \dots, x_{n-1}, x] = f[x, x_0, \dots, x_{n-1}]$ (Übungsaufgabe) und der Definition dividierter Differenzen erhalten wir

$$f(x) - p_n(x) = f[x_0, \dots, x_n, x] \prod_{j=0}^n (x - x_j).$$

Ferner ist nach Induktionsvoraussetzung (Änderung der Notationen im ersten Term $n \mapsto n - 1$, $t \mapsto t_n$ und im zweiten Term $n \mapsto n - 1$, $x = x_n$)

$$\begin{aligned}
 &f[x_0, \dots, x_{n-1}, x] - f[x_0, \dots, x_n] \\
 &= \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} \{f^{(n)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x - x_{n-1})) \\
 &\quad - f^{(n)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}))\} dt_n \cdots dt_1 \\
 &= \int_0^1 \int_0^{t_1} \cdots \int_0^{t_n} \frac{d}{dt} f^{(n)}(x_0 + \cdots + t_n(x_n - x_{n-1}) + t(x - x_n)) dt dt_n \cdots dt_1,
 \end{aligned}$$

woraus wegen

$$\begin{aligned} \frac{d}{dt} f^{(n)}(x_0 + \dots + t_n(x_n - x_{n-1}) + t(x - x_n)) \\ = f^{(n+1)}(x_0 + \dots + t_n(x_n - x_{n-1}) + t(x - x_n)) (x - x_n) \end{aligned}$$

und der Definition dividierter Differenzen die Behauptung folgt. \square

Die obige Darstellung dividierter Differenzen gestattet für differenzierbare Funktionen die stetige Fortsetzung für den Fall, dass einige der Stützstellen zusammenfallen:

$$f[x_0, \dots, x_r, x_r, \dots, x_n] := \lim_{\varepsilon \rightarrow 0} f[x_0, \dots, x_r, x_r + \varepsilon, \dots, x_n].$$

Im Extremfall fallen alle Stützstellen zusammen, wir haben

$$f[x_0, \dots, x_0] = \int_0^1 \int_0^{t_1} \dots \int_0^{t_{n-1}} f^{(n)}(x_0) dt_n \dots dt_1 = \frac{1}{n!} f^{(n)}(x_0)$$

und die Newton'sche Interpolationsformel geht in das Taylor-Polynom n -ten Grades über

$$p_n(x) = \sum_{i=0}^n f[x_0, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j) = \sum_{i=0}^n \frac{1}{i!} f^{(i)}(x_0) (x - x_0)^i.$$

3.3 Hermite-Interpolation

Die Lagrange'sche Interpolation kann auf den Fall erweitert werden, in dem neben Funktionswerten auch Werte der Ableitungen einer Funktion f in gewissen (oder allen) Knoten bekannt sind. Die Hermite-Interpolation kann wie folgt charakterisiert werden:

Gegeben seien paarweise verschiedene Knoten x_i , $i = 0, \dots, n$, der Funktionswert und die Ableitungen $y_i^{(k)}$ bis zur Ordnung $k = m_i$ im Knoten x_i , $i = 0, \dots, n$. Gesucht ist ein Polynom N -ten Grades,

$$N = \sum_{i=0}^n (1 + m_i) - 1,$$

mit der Eigenschaft $p^{(k)}(x_i) = y_i^{(k)}$, $k = 0, \dots, m_i$, $i = 0, \dots, n$. Die Knoten x_i werden auch als $m_i + 1$ -fache Stützstellen bezeichnet.

Theorem 3.3.1 *Die Hermite'sche Interpolationsaufgabe ist eindeutig lösbar.*

Beweis: Für die $N + 1$ unbekanntenen Koeffizienten c_i des gesuchten Polynoms ergeben sich aus den Interpolationsbedingungen $N + 1$ lineare Bestimmungsgleichungen. Dieses Gleichungssystem ist genau dann eindeutig lösbar, wenn das zugeordnete homogene System nur die Nulllösung besitzt. Sei also $p^{(k)}(x_i) = 0$ für $k = 0, \dots, m_i$, $i = 0, \dots, n$. Da p_N in x_0, \dots, x_n verschwindet, können wir das gesuchte Polynom in der Form

$$p_N(x) = r_0(x) \prod_{i=0}^n (x - x_i)$$

darstellen, wobei r_0 ein Polynom vom Grade kleiner oder gleich $N - n - 1$ ist. Die Ableitung

$$p'_N(x) = r'_0(x) \prod_{i=0}^n (x - x_i) + r_0(x) \sum_{j=0}^n \prod_{j \neq i} (x - x_j)$$

verschwindet in allen Knoten x_i für die $m_i \geq 1$, folglich verschwindet dort auch r_0 . Somit haben wir für ein gewisses Polynom r_1

$$p_N(x) = r_1(x) \prod_{i=0}^n \prod_{\substack{j=1 \\ m_i \geq 1}}^2 (x - x_i), \quad \deg r_1 \leq N - n - 1 - \sum_{\substack{i=0 \\ m_i \geq 1}}^n 1.$$

Sei $m = \max_{i=0, \dots, n} m_i$. Nach $m - 1$ Schritten haben wir

$$p_N(x) = r_{m-1}(x) \prod_{i=0}^n \prod_{j=1}^{\min(m, m_i+1)} (x - x_i),$$

$$\deg r_{m-1} \leq N - n - 1 - \sum_{\substack{i=0 \\ m_i \geq 1}}^n 1 - \dots - \sum_{\substack{i=0 \\ m_i \geq m-1}}^n 1 = \sum_{\substack{i=0 \\ m_i = m}}^n 1 - 1.$$

Das Polynom r_{m-1} verschwindet in allen x_i , für die $m_i = m$ gilt, d.h. in mehr paarweise verschiedenen Punkten als sein Grad angibt, kann somit nur das Nullpolynom sein. \square

Wir suchen ähnlich wie bei der Lagrange'schen Interpolation eine Darstellung des Hermite'schen Interpolationspolynoms in der Form

$$H_N(x) = \sum_{i=0}^n \sum_{k=0}^{m_i} y_i^{(k)} L_{ik}(x). \quad (3.3.1)$$

Betrachten wir zunächst die durch

$$l_{ij}(x) = \frac{(x - x_i)^j}{j!} \prod_{\substack{k=0 \\ k \neq i}}^n \left(\frac{x - x_k}{x_i - x_k} \right)^{m_k+1}, \quad i = 0, \dots, n, \quad j = 0, \dots, m_i,$$

definierten $N + 1$ Polynome vom Grade kleiner oder gleich N . Die Polynome haben die Eigenschaft

$$l_{ij}(x_i) = l_{ij}^{(1)}(x_i) = \dots = l_{ij}^{(j-1)}(x_i) = 0, \quad l_{ij}^{(j)}(x_i) = 1,$$

$$l_{ij}(x_k) = l_{ij}^{(1)}(x_k) = \dots = l_{ij}^{(m_k)}(x_k) = 0,$$

Die gesuchten Hermite'schen Basispolynome L_{ij} werden nun rekursiv wie folgt definiert:

$$L_{im_i}(x) = l_{im_i}(x), \quad i = 0, \dots, n$$

$$L_{ij}(x) = l_{ij}(x) - \sum_{k=j+1}^{m_i} l_{ij}^{(k)}(x_i) L_{ik}(x), \quad j = m_i - 1, m_i - 2, \dots, 0$$

Theorem 3.3.2 Die rekursiv definierten Hermite'schen Basispolynome L_{ij} erfüllen die Beziehungen

$$L_{ij}^{(r)}(x_k) = \begin{cases} 1 & \text{falls } i = k \text{ und } j = r \\ 0 & \text{andernfalls} \end{cases}$$

Hieraus folgt unmittelbar die Darstellung (3.3.1).

Beweis: Übungsaufgabe. □

Theorem 3.3.3 Sei $f \in C^{N+1}[a, b]$. Dann gibt es zu jedem $x \in [a, b]$ einen Punkt $\xi_x \in [x_0, \dots, x_n, x]$, so dass für den Fehler des Hermite'schen Interpolationspolynom p_N folgende Darstellung gilt:

$$f(x) - p_N(x) = \frac{1}{(N+1)!} f^{(N+1)}(\xi_x) \prod_{i=0}^n (x - x_i)^{m_i+1}$$

Beweis: Analog zum Fehler des Lagrange'schen Interpolationspolynoms. □

Beispiel: Bestimmen Sie das Hermite'sche Interpolationspolynom zu den vorgegebenen Werten $f(0) = 2$, $f'(0) = 1$, $f''(0) = 4$ und $f(1) = -1$. Der Ansatz

$$H_3(x) = Ax^3 + Bx^2 + Cx + D$$

führt auf das Gleichungssystem

$$\begin{aligned} 2 &= H_3(0) = D \\ 1 &= H_3'(0) = C \\ 4 &= H_3''(0) = 2B \\ -1 &= H_3(1) = A + B + C + D \end{aligned}$$

mit der Lösung $A = -6$, $B = 2$, $C = 1$ und $D = 2$. Das gesuchte Interpolationspolynom ist damit $H_3(x) = -6x^3 + 2x^2 + x + 2$. Für den oben beschriebenen allgemeinen Ansatz folgt

$$x_0 = 0, \quad m_0 = 2, \quad x_1 = 1, \quad m_1 = 0, \quad N = (m_0 + 1) + (m_1 + 1) - 1 = 3.$$

Mit den Hilfspolynomen

$$l_{00}(x) = -(x-1), \quad l_{01}(x) = -x(x-1), \quad l_{02}(x) = -\frac{x^2}{2}(x-1), \quad l_{10}(x) = x^3$$

erhalten wir die Hermite'schen Basispolynome L_{ij} rekursiv, zunächst L_{im_i} für $i = 0, 1$:

$$L_{02}(x) = l_{02}(x) = -\frac{x^2}{2}(x-1), \quad L_{10}(x) = l_{10}(x) = x^3.$$

Die Rekursionsbeziehung liefert

$$L_{01}(x) = l_{01}(x) - l_{01}''(0)L_{02}(x) = -x(x-1) - (-2)\left(-\frac{x^2}{2}(x-1)\right) = x(1-x^2)$$

$$L_{00}(x) = l_{00}(x) - l_{00}'(0)L_{01}(x) - l_{00}''(0)L_{02}(x) = 1 - x^3$$

Aus der allgemeinen Darstellung

$$H_N(x) = \sum_{i=0}^n \sum_{k=0}^{m_i} y_i^{(k)} L_{ik}(x)$$

folgt im vorliegenden Fall

$$H_3(x) = 2(1-x^3) + x(1-x^2) + 2x^2(1-x) - x^3 = -6x^3 + 2x^2 + x + 2.$$

Der Vorteil der Bestimmung der Hermite'schen Basispolynome besteht darin, dass wir die Hermite-Interpolation für beliebige Daten $f(0), f'(0), f''(0), f(1)$ gelöst haben, es gilt nämlich

$$H_3(x) = f(0)(1-x^3) + f'(0)x(1-x^2) + \frac{f''(0)}{2}x^2(1-x) + f(1)x^3.$$

□

3.4 Spline-Interpolation

Wie wir gesehen haben, eignen sich Lagrangesche Interpolationspolynome nicht besonders gut zur Approximation von (nicht glatten) Funktionen, da sie bei

Vermehrung der Stützstellenzahl dazu neigen, zwischen den Stützstellen immer größere Werte anzunehmen. Dies ist die Folge ihrer *Steifheit* bedingt durch die Forderung von C^∞ -Übergängen in den Knoten. Zur Reduzierung dieser Steifheit setzt man die interpolierende Funktion stückweise polynomial bezüglich einer Zerlegung $a = x_0 < x_1 < \dots < x_n = b$ an. In den Knoten werden dann geeignete Differenzierbarkeitseigenschaften vorausgesetzt. Wir bezeichnen die Länge des Teilintervalls $I_i = [x_{i-1}, x_i]$ durch $h_i = x_i - x_{i-1}$, die Größe $h = \max_{i=1, \dots, n} h_i$ charakterisiert die Feinheit der Intervallzerlegung. Auf einer solchen Intervallzerlegung werden Vektorräume von stückweise polynomialen Funktionen betrachtet

$$S_h^{(k,r)}[a, b] = \{p \in C^r[a, b] : p|_{I_i} \in P_k(I_i), i = 1, \dots, n\}, \quad k, r = 0, 1, 2, \dots$$

Zu einem Satz von Stützwerten in Punkten aus dem Intervall $[a, b]$ wird dann eine *Interpolierende* $p \in S_h^{(k,r)}[a, b]$ mit Hilfe geeigneter Interpolationsbedingungen bestimmt. Wir betrachten nun einige Beispiele.

Beispiel: Die *stetige, stückweise lineare Lagrange-Interpolation* (Fall $k = 1$, $r = 0$) approximiert eine gegebene Funktion f auf $[a, b]$ durch einen Polygonzug in den Stützstellen x_i , $i = 0, \dots, n$:

$$p \in S_h^{(1,0)}[a, b] = \{p \in C[a, b], p|_{I_i} \text{ linear}\}, \quad p(x_i) = f(x_i), \quad i = 0, \dots, n.$$

Die Anwendung der Fehlerabschätzung für die Lagrange-Interpolation separat auf jedem der Teilintervalle I_i

$$\begin{aligned} f(x) - p(x) &= \frac{f''(\xi_x)}{2}(x - x_{i-1})(x - x_i) \\ |f(x) - p(x)| &\leq \frac{1}{2} \max_{x \in I_i} |f''(x)| \frac{h_i^2}{4} \quad x \in I_i \end{aligned}$$

ergibt die globale Fehlerabschätzung

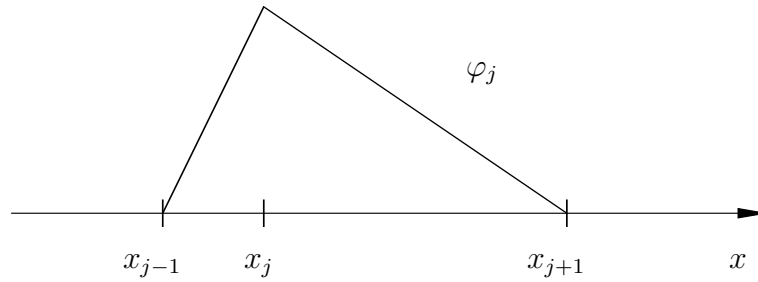
$$\max_{x \in [a, b]} |f(x) - p(x)| \leq \frac{h^2}{8} \max_{x \in [a, b]} |f''(x)|.$$

Für die Konstruktion der Interpolierenden verwendet man die Knotenfunktionale $N_i(f) = f(x_i)$, $i = 0, \dots, n$, die eine Knotenbasis $\varphi_j \in S_h^{(1,0)}[a, b]$, $j = 0, \dots, n$ eindeutig durch die Bedingung

$$N_i(\varphi_j) = \delta_{ij}, \quad i, j = 0, \dots, n$$

festlegen. Die Interpolierende kann dann in der Form

$$p(x) = \sum_{i=0}^n N_i(f) \varphi_i(x)$$

Abbildung 3.4: Lokale Basisfunktion φ_j zum Knotenfunktional N_j .

dargestellt werden. \square

Beispiel: Stetige, stückweise kubische Lagrange-Interpolation (Fall $k = 3$, $r = 0$). Zur Erzielung globaler Stetigkeit verwendet man die Knotenfunktionale

$$N_i(f) = f(x_i), \quad i = 0, \dots, n.$$

Diese werden in jedem Intervall I_i durch zwei weitere

$$N_{ij}(f) = f(x_{ij}), \quad x_{ij} \in (x_{i-1}, x_i), \quad j = 0, 1$$

ergänzt. Damit ist eindeutig eine global stetige Interpolierende $p \in S_h^{(3,0)}[a, b]$ festgelegt. Man erhält die Fehlerabschätzung

$$\max_{x \in [a, b]} |f(x) - p(x)| \leq \frac{h^4}{4!} \max_{x \in [a, b]} |f^{(4)}(x)|$$

intervallweise aus der Abschätzung für kubische Lagrange-Interpolation. \square

Beispiel: Stetig differenzierbare, stückweise kubische Hermite-Interpolation (Fall $k = 3$, $r = 1$). Zur Erzielung globaler stetiger Differenzierbarkeit verwendet man die Knotenfunktionale

$$N_i(f) = f(x_i), \quad N_{n+i+1}(f) = f'(x_i), \quad i = 0, \dots, n.$$

Damit ist eindeutig eine global stetig differenzierbare Interpolierende $p \in S_h^{(3,1)}[a, b]$ festgelegt. Man erhält für den Fehler

$$\max_{x \in [a, b]} |f(x) - p(x)| \leq \frac{h^4}{4!} \max_{x \in [a, b]} |f^{(4)}(x)|,$$

diesmal aus der intervallweisen Anwendung der Abschätzung für kubische Hermite-Interpolation. \square

Die Forderung nach höherer Glattheit, etwa eine Interpolation in $S_h^{(k, k-1)}[a, b]$ führt auf die *Spline-Interpolation*, die von großer praktischer Bedeutung, etwa in

der Computer-Graphik ist.

Unsere Aufgabe besteht nun darin, aus vorgegebenen Werten

$$(x_i, y_i), \quad i = 0, \dots, n,$$

eine global zweimal stetige, stückweise kubische Interpolation s_n zu bestimmen. Ein solche Funktion wird interpolierender kubischer Spline genannt.

Theorem 3.4.1 *Der interpolierende kubische Spline s_n existiert und ist eindeutig bestimmt durch zusätzliche Vorgabe von einer der folgenden Randbedingungen*

- (a) $s_n''(a) = s_n''(b) = 0$ (natürlicher kubischer Spline)
- (b) $s_n'(a) = s_n'(b)$ und $s_n''(a) = s_n''(b)$ (periodischer kubischer Spline)
- (c) $s_n'(a) = y'(a)$ und $s_n'(b) = y'(b)$ (gebundener kubischer Spline)

Beweis. Jeder kubische Spline hat bezüglich der Zerlegung $a = x_0 < x_1 < \dots < x_n = b$ die Form $s|_{I_i} = p_i$, $p_i \in P_3(I_i)$, $i = 1, \dots, n$. Jedes der kubischen Polynome p_i hat 4 unbestimmte Koeffizienten, dies ergibt $4n$ Freiheitsgrade. Zu ihrer Bestimmung stehen folgende lineare Beziehungen zur Verfügung:

$s(x_i) = y_i, i = 0, \dots, n$:	2n	Gleichungen
$s' \in C[a, b]$:	n-1	Gleichungen
$s'' \in C[a, b]$:	n-1	Gleichungen
Zusatzbedingungen	:	2	Gleichungen
Insgesamt	:	4n	Gleichungen

Zum Nachweis der Existenz einer Lösung des linearen Gleichungssystems von $4n$ Gleichungen mit $4n$ Unbekannten genügt es, wie üblich zu zeigen, dass das zugeordnete homogene System nur die Nulllösung besitzt.

Sei $f \in C^2[a, b]$ eine beliebige, die Daten (x_i, y_i) , $i = 0, \dots, n$, interpolierende Funktion mit

$$s_n''(x) (f'(x) - s_n'(x)) \Big|_{x=a}^b = 0.$$

Dann erhalten wir durch elementweise partielle Integration

$$\begin{aligned} \int_a^b s_n''(x) (f''(x) - s_n''(x)) dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} s_n''(x) (f''(x) - s_n''(x)) dx \\ &= \sum_{i=1}^n \left\{ s_n''(x) (f'(x) - s_n'(x)) \Big|_{x_{i-1}}^{x_i} - s_n'''(x) (f(x) - s_n(x)) \Big|_{x_{i-1}}^{x_i} \right. \\ &\quad \left. + \int_{x_{i-1}}^{x_i} s_n^{(iv)}(x) (f'(x) - s_n'(x)) dx \right\}. \end{aligned}$$

Nun sind $s^{(iv)}(x) = 0$ für $x \in I_i$, $i = 1, \dots, n$, und $f(x_i) - s_n(x_i) = 0$ für $i = 0, \dots, n$. Die verbleibende Summe reduziert sich wegen der Stetigkeit von s_n'' und $f' - s_n'$ auf die Differenz der Werte an den Intervallenden, es folgt

$$\int_a^b s_n''(x) (f''(x) - s_n''(x)) dx = s_n''(b) (f'(b) - s_n'(b)) - s_n''(a) (f'(a) - s_n'(a)) = 0.$$

Wir betrachten nun den Fall eines Null interpolierenden Splines, d.h. dass $s_n(x_i) = 0$ für $i = 0, \dots, n$ gilt. Für den gebundenen Spline sei zusätzlich $s_n'(b) = s_n'(a) = 0$. Dann genügt die Nullfunktion $f = 0$ allen oben getroffenen Voraussetzungen, unabhängig davon, ob ein natürlicher, periodischer oder gebundener Spline vorliegt. Wir bekommen

$$\int_a^b |s_n''(x)|^2 dx = 0,$$

somit s_n linear auf $[a, b]$. Wegen $s_n(a) = s_n(b) = 0$ folgt $s_n(x) \equiv 0$. \square

Die im Beweis beobachtete Orthogonalitätseigenschaft hat die interessante Konsequenz, dass sich der interpolierende kubische Spline durch besonders geringe Oszillation auszeichnet.

Theorem 3.4.2 *Unter allen Funktionen $f \in C^2[a, b]$, die den Interpolationsbedingungen $f(x_i) = y_i$, $i = 0, \dots, n$, und einer der Randbedingungen*

- (a) $f''(a) = f''(b) = 0$
- (b) $f'(a) = f'(b)$ und $f''(a) = f''(b)$
- (c) $f'(a) = y'(a)$ und $f'(b) = y'(b)$

genügen, besitzt die interpolierende kubische Splinefunktion s_n die kleinste Gesamtkrümmung in folgendem Sinne:

$$\int_a^b |s_n''(x)|^2 dx \leq \int_a^b |f''(x)|^2 dx.$$

Beweis. Aus der Identität $f = s_n + (f - s_n)$ folgt mit der oben beobachteten Orthogonalität der zweiten Ableitungen

$$\begin{aligned} \int_a^b |f''(x)|^2 dx - \int_a^b |s_n''(x)|^2 dx \\ = 2 \int_a^b s_n''(x) (f''(x) - s_n''(x)) dx + \int_a^b |f''(x) - s_n''(x)|^2 dx \geq 0 \end{aligned}$$

die Behauptung. \square

Bemerkung. Der Name “Spline” erklärt sich durch die physikalische Interpretation der obigen Minimaleigenschaft. Beschreibt $y = f(x)$ die Lage einer dünnen Holzlatte, so mißt

$$E = \int_a^b \left(\frac{y''(x)}{(1 + y'(x)^2)^{3/2}} \right)^2 dx$$

die “Biegeenergie” der Latte. Aufgrund des Hamiltonschen Prinzips stellt sich die Latte so ein, dass diese Energie minimiert wird. Für kleine Auslenkungen gilt näherungsweise

$$E \approx \int_a^b y''(x)^2 dx,$$

der interpolierende kubische Spline beschreibt also annähernd die Lage einer dünnen Holzlatte, die an den Knoten x_i fixiert ist. Bei der gebundenen Spline-Interpolation haben wir die Latte an den Randknoten unter zusätzlicher Vorgabe der Steigungen an den Randknoten eingespannt. Die natürlichen Randbedingungen entsprechen der Situation, dass die Latte ausserhalb des Intervalls $[a, b]$ gerade ist. Bei den periodischen Randbedingungen handelt es sich um ringförmig geschlossenen Latten. Derartige dünne Holzplatten wurde tatsächlich als Zeichenwerkzeug benutzt und tragen im Englischen den Namen “Spline”. \square

Zur expliziten Bestimmung des interpolierenden kubischen Spline s_n benutzen wir die Bezeichnungen

$$y_i = s_n(x_i), \quad M_i = s_n''(x_i), \quad i = 0, \dots, n.$$

Da $s_n|_{I_i} = p_i \in P_3(I_i)$, ist p_i'' auf I_i linear, somit

$$p_i''(x) = M_{i-1} \frac{x_i - x}{h_i} + M_i \frac{x - x_{i-1}}{h_i} \quad x \in I_i, \quad i = 1, \dots, n.$$

Zweimalige Integration ergibt

$$p_i(x) = M_{i-1} \frac{(x_i - x)^3}{6h_i} + M_i \frac{(x - x_{i-1})^3}{6h_i} + A_{i-1}(x - x_{i-1}) + B_{i-1}$$

mit noch zu bestimmenden Konstanten A_{i-1} und B_{i-1} , $i = 1, \dots, n$. Die Interpolationsbedingungen $p_i(x_{i-1}) = s_n(x_{i-1}) = y_{i-1}$ und $p_i(x_i) = s_n(x_i) = y_i$ liefern

$$y_{i-1} = M_{i-1} \frac{h_i^2}{6} + B_{i-1}, \quad y_i = M_i \frac{h_i^2}{6} + A_{i-1}h_i + B_{i-1}$$

woraus

$$B_{i-1} = y_{i-1} - M_{i-1} \frac{h_i^2}{6}, \quad A_{i-1} = \frac{y_i - y_{i-1}}{h_i} - \frac{h_i}{6}(M_i - M_{i-1})$$

folgt. Die Forderung der Stetigkeit der ersten Ableitungen in x_i , $i = 1, \dots, n-1$, führt auf

$$M_i \frac{h_i}{2} + A_{i-1} = p'_i(x_i) = p'_{i+1}(x_i) = -M_i \frac{h_{i+1}}{2} + A_i$$

woraus sich ein lineares Gleichungssystem von $n-1$ Gleichungen, $i = 1, \dots, n-1$:

$$\frac{h_i}{h_i + h_{i+1}} M_{i-1} + 2M_i + \frac{h_{i+1}}{h_i + h_{i+1}} M_{i+1} = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right)$$

für die $n+1$ gesuchten Größen M_i , $i = 0, \dots, n$, ergibt.

Im Fall der natürlichen Randbedingungen komplettieren die beiden Zusatzbedingungen $M_0 = s''_n(a) = M_n = s''_n(b) = 0$ das Gleichungssystem, das dann nach Elimination von M_0 und M_n die Form

$$\begin{bmatrix} 2 & \frac{h_2}{h_1 + h_2} & 0 & \dots & 0 \\ \frac{h_2}{h_2 + h_3} & 2 & \frac{h_3}{h_2 + h_3} & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \frac{h_{n-2}}{h_{n-2} + h_{n-1}} & 2 & \frac{h_{n-1}}{h_{n-2} + h_{n-1}} \\ 0 & \dots & 0 & \frac{h_{n-1}}{h_{n-1} + h_n} & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} \end{bmatrix}$$

annimmt. Hierbei wurde zur Abkürzung

$$d_i = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right), \quad i = 1, \dots, n-1,$$

benutzt.

Im Fall gebundener Splines ergänzt man die Gleichungen durch die Forderungen

$$\begin{aligned} y'(a) &= p'_1(x_0) = -M_0 \frac{h_1}{2} + A_0 = -M_0 \frac{2h_1}{6} - M_1 \frac{h_1}{6} + \frac{y_1 - y_0}{h_1} \\ y'(b) &= p'_n(x_n) = M_n \frac{h_n}{2} + A_{n-1} = M_n \frac{2h_n}{6} + M_{n-1} \frac{h_n}{6} + \frac{y_n - y_{n-1}}{h_n} \end{aligned}$$

und man erhält das System

$$\begin{bmatrix} 2 & 1 & 0 & \cdots & 0 \\ \frac{h_1}{h_1+h_2} & 2 & \frac{h_2}{h_1+h_2} & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & \frac{h_{n-1}}{h_{n-1}+h_n} & 2 & \frac{h_n}{h_{n-1}+h_n} \\ 0 & \cdots & 0 & 1 & \frac{h_n}{2} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}$$

mit

$$d_0 = \frac{6}{h_1} \left(\frac{y_1 - y_0}{h_1} - y'(a) \right) \quad d_n = \frac{6}{h_n} \left(y'(b) - \frac{y_n - y_{n-1}}{h_n} \right).$$

Beide Gleichungssysteme (natürliche und gebundene Splines) sind strikt diagonal dominant (keine Pivotierung erforderlich) und können effizient mit dem Thomas-Verfahren (Tridiagonalmatrix) gelöst werden.

Interpolierende Spline-Funktionen besitzen bessere Approximationseigenschaften für

$$h = \max_{i=1, \dots, n} h_i \rightarrow 0$$

als die Lagrangeschen Interpolationspolynome.

Theorem 3.4.3 Sei $f \in C^4[a, b]$. Für den interpolierenden kubischen Spline mit $s_n''(a) = f''(a) = s_n''(b) = f''(b) = 0$ gilt

$$\max_{a \leq x \leq b} |f^{(j)}(x) - s_n^{(j)}(x)| \leq C h^{4-j} \max_{a \leq x \leq b} |f^{(4)}(x)|, \quad j = 0, 1, 2.$$

Beweis. Siehe: H. Werner, R. Schaback, Praktische Mathematik II, Springer-Verlag, 1972 \square

Theorem 3.4.4 Sei $f \in C^4[a, b]$. Für den gebundenen interpolierenden kubischen Spline gilt

$$\max_{a \leq x \leq b} |f(x) - s_n(x)| \leq \frac{5}{384} h^4 \max_{a \leq x \leq b} |f^{(4)}(x)|.$$

Beweis. Siehe: Hall und Meyer, Optimal error bounds for cubic spline interpolation, J. Approx. Theory 16(1976), 105-122 \square

Tipp: In MATLAB kann der gebundene, interpolierende kubische Spline mit Hilfe des Komandos `yy=spline(x,y,xx)` berechnet werden. Dazu enthält der Vektor x

die Knotenwerte x_i , der Vektor y die zu interpolierenden Funktionswerte y_i sowie als erstes und letztes Element die Steigungen $y'(a)$, $y'(b)$, und xx die zu berechnenden Abzissen. Die zum Abzissenvektor gehörigen Werte der Splineapproximation stehen dann im Ausgabevektor yy . In dieser Variante des spline-Befehls gilt also $\text{size}(y) = \text{size}(x) + 2$.

Sind die Längen der Vektoren x und y gleich, so werden die zwei erforderlichen Zusatzbedingungen für die Eindeutigkeit des interpolierenden kubischen Splines aus der Forderung bestimmt, dass die dritten Ableitungen von s_n in den Knoten x_1 und x_{n-1} stetig sind ("not a knot" Bedingungen oder "Einheitlichkeitsbedingungen"). Das beinhaltet, dass s_n auf den Intervallen $[x_0, x_2]$ und $[x_{n-2}, x_n]$ nicht nur stückweise kubisch sondern kubisch ist. Man sagt auch, die Knoten x_1 und x_{n-1} sind keine aktiven Knoten. Die aus den Einheitlichkeitsbedingungen resultierenden Gleichungen sind

$$\frac{M_1 - M_0}{h_1} = \frac{M_2 - M_1}{h_2} \qquad \frac{M_n - M_{n-1}}{h_n} = \frac{M_{n-1} - M_{n-2}}{h_{n-1}}.$$

Durch diese zusätzlichen Gleichungen ist die Koeffizientenmatrix des Gleichungssystems zur Bestimmung der M_0, \dots, M_n , nicht mehr strikt diagonal dominant, eliminiert man jedoch M_0 und M_n aus diesem System, erhält man wieder eine strikt diagonal dominante Matrix und die eindeutige Lösbarkeit ist gesichert. Beispielsweise führt die Elimination von

$$M_0 = \left(\frac{h_1 + h_2}{h_2} \right) M_1 - \frac{h_1}{h_2} M_2$$

in der Gleichung

$$\frac{h_1}{h_1 + h_2} M_0 + 2M_1 + \frac{h_2}{h_1 + h_2} M_2 = d_1$$

auf

$$\left(2 + \frac{h_1}{h_2} \right) M_1 + \left(1 - \frac{h_1}{h_2} \right) M_2 = d_1$$

und die strikte Diagonaldominanz ist erfüllt. \square

Beispiel: In Abbildung 3.5 sind der kubische interpolierende Spline der Funktion $y = 1/(1 + x^2)$ auf dem Intervall $[-5, 5]$ sowie das Lagrange'sche Interpolationspolynom vom Grade 12 dargestellt. Hierzu wurden die MATLAB-Befehle `cs=spline(x,y,xx)` sowie `p=polyfit(x,y,12)` und `interp=polyval(p,xx)` verwendet. In den Vektoren x und y sind die 13 äquivalenten Stützstellen $x_i = -5 + 5(i-1)/6$ und die zugeordneten Funktionswerte $y(x_i)$, $i = 1, \dots, 13$, gespeichert. Deutlich ist die schlechte Approximation der Runge-Funktion durch das Interpolationspolynoms am Rande zu sehen. Die Abschwächung der C^∞ Glattheit im Innern des Intervalls $[-5, 5]$ und die Hinzunahme zweier Zusatzbedingungen am Rande (Einheitlichkeitsbedingungen) führen zu drastischen Verbesserungen der Approximationsgüte. \square

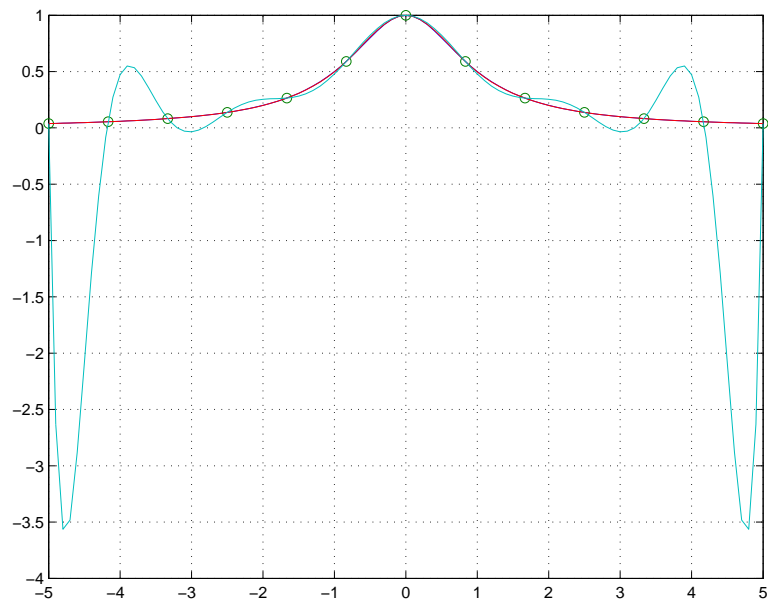


Abbildung 3.5: Stückweise kubische Spline-Approximation im Vergleich zum Lagrangeschen Interpolationspolynom vom Grade 12.

Kapitel 4

Numerische Integration

Die Berechnung bestimmter Integrale kann in der Praxis meist nur näherungsweise mit Hilfe von “Quadraturformeln” erfolgen. Dazu macht man für eine Funktion $f \in C[a, b]$ den Ansatz

$$I(f) = \int_a^b f(x) dx \approx I_n(f) = (b - a) \sum_{i=0}^n \omega_i f(x_i)$$

mit Stützstellen $a \leq x_0 < x_1 < \dots < x_n \leq b$ und Gewichten $\omega_i \in \mathbb{R}$. Ein typisches Beispiel ist die Trapezregel, bei der $n = 1$, $x_0 = a$, $x_1 = b$, $\omega_0 = \omega_1 = 1/2$ gilt und zu

$$\int_a^b f(x) dx \approx \frac{b - a}{2} (f(a) + f(b))$$

führt.

4.1 Beispiele interpolatorischer Quadraturformeln

Ein naheliegender Weg zur Konstruktion von Quadraturformeln ist der über die Polynominterpolation. Zu paarweise verschiedenen Stützstellen $a \leq x_0 < x_1 < \dots < x_n \leq b$ wird das Lagrangesche Interpolationspolynom gebildet

$$p_n(x) = \sum_{i=0}^n f(x_i) L_i^{(n)}(x)$$

und dann die Integration über f durch die über p_n ersetzt, wir setzen also

$$I_n(f) := \int_a^b p_n(x) dx = (b - a) \sum_{i=0}^n f(x_i) \frac{1}{b - a} \int_a^b L_i^{(n)}(x) dx = \sum_{i=0}^n \omega_i f(x_i).$$

Die Gewichte ω_i hängen offenbar nur von den Stützstellen x_0, \dots, x_n , insbesondere nicht von f ab. Der Quadraturfehler einer interpolatorischen Quadraturformel läßt sich leicht angeben:

Theorem 4.1.1 Für interpolatorische Quadraturformeln gilt:

$$I(f) - I_n(f) = \int_a^b f[x_0, x_1, \dots, x_n, x] \prod_{j=0}^n (x - x_j) dx.$$

Beweis. Folgt sofort aus der Fehlerdarstellung der Lagrange-Interpolation. \square

Aus der Fehlerdarstellung folgt, dass die interpolatorische Quadraturformel I_n exakt für Polynome vom Grade kleiner oder gleich n ist.

Mittelpunkts- oder Rechteckregel

Wir ersetzen f im Intervall $[a, b]$ durch den Funktionswert im Mittelpunkt des Intervalls und erhalten

$$I_n(f) = (b - a) f\left(\frac{a + b}{2}\right)$$

mit dem Gewicht $\omega_0 = 1$ und der Stützstelle $x_0 = (a + b)/2$. Für $f \in C^2[a, b]$ genügt der Quadraturfehler der Beziehung

$$E_0(f) = \int_a^b f(x) dx - (b - a) f\left(\frac{a + b}{2}\right) = \frac{(b - a)^3}{24} f''(\xi), \quad \xi \in (a, b).$$

Dies folgt aus der Taylorentwicklung

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0 + \theta(x - x_0)) \frac{(x - x_0)^2}{2}$$

durch Integration und Anwendung des Zwischenwertsatzes der Integralrechnung

$$\int_a^b f''(x_0 + \theta(x - x_0)) \frac{(x - x_0)^2}{2} dx = f''(\xi) \int_a^b \frac{(x - x_0)^2}{2} dx = \frac{(b - a)^3}{24} f''(\xi).$$

Die Mittelpunktsformel ist nach Konstruktion exakt für Polynome vom Grade Null, die Fehlerabschätzung zeigt, dass sie für Polynome vom Grade kleiner oder gleich 1 exakt ist. Zur Erhöhung der Genauigkeit, kann man das Intervall $[a, b]$ in Teilintervalle der Breite $H = (b - a)/m$ zerlegen und wendet die Mittelpunktsformel auf jedem der Teilintervalle an. Die Mittelpunkte der Teilintervalle sind gegeben durch $x_k = a + (2k + 1)H/2$, $k = 0, 1, \dots, m - 1$. Im Ergebnis erhält man die zusammengesetzte Mittelpunktsformel

$$I_{0,m}(f) = H \sum_{k=0}^{m-1} f(x_k), \quad m \geq 1.$$

Sei wieder $f \in C^2[a, b]$. Dann gilt für den Quadraturfehler der zusammengesetzten Mittelpunktsformel

$$\begin{aligned} E_{0,m}(f) &:= I(f) - I_{0,m}(f) = \frac{H^3}{24} \sum_{k=0}^{m-1} f''(\xi_k) \\ &= \frac{b-a}{24} H^2 \frac{1}{m} \sum_{k=0}^{m-1} f''(\xi_k) = \frac{b-a}{24} H^2 f''(\xi). \end{aligned}$$

Trapezregel

Ersetzt man f durch das Lagrangesche Interpolationspolynom bezüglich der Knoten $x_0 = a$ und $x_1 = b$, so erhält man die Trapezregel

$$I_1(f) = \frac{b-a}{2} (f(a) + f(b))$$

mit den Gewichten $\omega_0 = \omega_1 = 1/2$ und den Stützstellen $x_0 = a$ und $x_1 = b$. Ist $f \in C^2[a, b]$, so ist der Quadraturfehler gegeben durch

$$E_1(f) = I(f) - I_1(f) = -\frac{(b-a)^3}{12} f''(\xi), \quad \xi \in (a, b).$$

Tatsächlich erhält man aus der Formel für den Interpolationsfehler und dem Mittelwertsatz der Integralrechnung

$$\begin{aligned} E_1(f) &= \frac{1}{2} \int_a^b f''(\xi(x))(x-a)(x-b) dx \\ &= -\frac{f''(\xi)}{2} \int_a^b (x-a)(b-x) dx = -\frac{(b-a)^3}{12} f''(\xi). \end{aligned}$$

Zur Erhöhung der Genauigkeit, kann man das Intervall $[a, b]$ in Teilintervalle der Breite $H = (b-a)/m$ zerlegen und wendet die Trapezregel auf jedem der Teilintervalle getrennt an. Dazu seien $x_k = a + kH$, $k = 0, 1, \dots, m$, die Quadraturknoten und wir erhalten

$$\begin{aligned} I_{1,m}(f) &= \frac{H}{2} \sum_{k=0}^{m-1} (f(x_k) + f(x_{k+1})) \\ &= H \left(\frac{1}{2} f(x_0) + f(x_1) + \dots + f(x_{m-1}) + \frac{1}{2} f(x_m) \right). \end{aligned}$$

Cavalieri-Simpson-Formel

Ersetzt man f durch das Lagrangesche Interpolationspolynom bezüglich der Knoten $x_0 = a$, $x_1 = (a+b)/2$ und $x_2 = b$, so erhält man die Cavalieri-Simpson-Formel

$$I_2(f) = \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right)$$

mit den Gewichten $\omega_0 = \omega_2 = 1/6$ und $\omega_1 = 2/3$ sowie den Stützstellen $x_0 = a$, $x_1 = (a + b)/2$ und $x_2 = b$. Im vorliegenden Fall gilt für den Quadraturfehler

$$E_2(f) = I(f) - I_2(f) = \frac{1}{3!} \int_a^b f'''(\xi(x))(x - x_0)(x - x_1)(x - x_2) dx,$$

aber der Faktor $(x - x_1)$ ändert sein Vorzeichen im Integrationsintervall, so dass wir den Mittelwertsatz der Integralrechnung nicht direkt anwenden können. Wir schreiben unter Verwendung der Newtonschen Darstellung des Interpolationsfehlers den Quadraturfehler wie folgt um

$$\begin{aligned} E_2(f) &= \int_a^b f[x_0, x_1, x_2, x](x - x_0)(x - x_1)(x - x_2) dx \\ &= \int_a^b \frac{f[x_0, x_1, x_2, x] - f[x_0, x_1, x_2, x_1]}{x - x_1} (x - x_0)(x - x_1)^2(x - x_2) dx \\ &\quad + f[x_0, x_1, x_2, x_1] \int_a^b (x - x_0)(x - x_1)(x - x_2) dx. \end{aligned}$$

Das im zweiten Summanden stehende Integral

$$Q = \int_a^b (x - x_0)(x - x_1)(x - x_2) dx$$

verschwindet, denn die Transformation $x = x_0 + x_2 - t$ ergibt $Q = -Q$. Der unter dem Integralzeichen stehende Term $(x - x_0)(x - x_1)^2(x - x_2)$ wechselt in $[a, b]$ das Vorzeichen nicht; Anwendung des Mittelwertsatzes der Integralrechnung und die Darstellbarkeit der dividierten Differenzen in Integralform ergibt nun

$$\begin{aligned} E_2(f) &= \int_a^b f[x_0, x_1, x_2, x, x_1](x - x_0)(x - x_1)^2(x - x_2) dx \\ &= \int_a^b \frac{f^{(4)}}{4!}(\xi(x))(x - x_0)(x - x_1)^2(x - x_2) dx \\ &= \frac{f^{(4)}}{4!}(\xi) \int_a^b (x - x_0)(x - x_1)^2(x - x_2) dx. \end{aligned}$$

Ist $f \in C^4[a, b]$, so ist der Quadraturfehler gegeben durch

$$E_2(f) = -\frac{(b - a)^5}{2880} f^{(4)}(\xi).$$

Man beachte, dass nach Konstruktion Polynome vom Grade kleiner oder gleich 2 mit der Cavalieri-Simpson-Formel exakt integriert werden. Die Fehlerabschätzung zeigt nun, dass sogar Polynome vom Grade kleiner oder gleich 3 exakt integriert werden.

Zur Erhöhung der Genauigkeit, kann man das Intervall $[a, b]$ wieder in Teilintervalle der Breite $H = (b - a)/m$ zerlegen und wendet die Cavalieri-Simpson-Formel auf jedem der Teilintervalle getrennt an. Dazu seien $x_k = a + kH/2$, $k = 0, 1, \dots, 2m$, die Quadraturknoten und wir erhalten

$$I_{2,m}(f) = \frac{H}{6} \left(f(x_0) + 2 \sum_{r=1}^{m-1} f(x_{2r}) + 4 \sum_{s=0}^{m-1} f(x_{2s+1}) + f(x_{2m}) \right).$$

Für den Quadraturfehler der zusammengesetzten Cavalieri-Simpson-Formel folgt

$$E_{2,m}(f) = -\frac{b-a}{5760} H^4 f^{(4)}(\xi), \quad \xi \in (a, b).$$

4.2 Newton-Cotes-Formeln

Die Newton-Cotes-Formeln sind interpolatorische Quadraturformeln, die auf eine äquidistante Verteilung der Stützstellen basieren. Man unterscheidet

- (i) abgeschlossene Newton-Cotes-Formeln (a, b sind Stützstellen)

$$x_i = a + ih, \quad i = 0, \dots, n, \quad h = \frac{b-a}{n},$$

- (ii) offene Newton-Cotes-Formeln (a, b sind keine Stützstellen)

$$x_i = a + (i+1)h, \quad i = 0, \dots, n, \quad h = \frac{b-a}{n+2},$$

In beiden Fällen wird der Integrand f durch das entsprechende Interpolationspolynom vom Grade n ersetzt, also

$$I_n(f) = \sum_{i=0}^n f(x_i) \int_a^b L_i^{(n)}(x) dx.$$

Für die Gewichte der abgeschlossenen Formeln erhält man nach Koordinatentransformation $x = a + th$

$$\omega_i = \frac{1}{b-a} \int_a^b L_i^{(n)}(x) dx = \frac{1}{n} \int_0^n L_i^{(n)}(a+th) dt = \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^n \left(\frac{t-j}{i-j} \right) dt$$

Diese Gewichte werden ein für allemal berechnet und tabelliert. Im Fall der offenen Newton-Cotes-Formeln verfährt man analog. Man beachte, dass bei den abgeschlossenen Newton-Cotes-Formeln ab $n = 8$ und bei den offenen ab $n = 2$ negative Gewichte auftreten. Die für das Integral geltende Positivitätseigenschaft

$$f(x) \geq 0 \quad \text{für } x \in [a, b] \quad \Rightarrow \quad I(f) \geq 0$$

kann im Fall negativer Gewichte für die Quadraturformel I_n nicht mehr gesichert werden. Zusätzlich erhöht sich die Rundungsfehleranfälligkeit der Formeln (Auslöschungsgefahr). Um die Genauigkeit zu erhöhen, ist es daher ratsam zu zusammengesetzten Formeln und nicht zu höheren Werten von n überzugehen.

Tabelle 4.1: Gewichte abgeschlossener Newton-Cotes Formeln

n	$\omega_0 \dots, \omega_n$	Name
1	$\frac{1}{2} \quad \frac{1}{2}$	Trapezregel
2	$\frac{1}{6} \quad \frac{4}{6} \quad \frac{1}{6}$	Cavalieri-Simpson
3	$\frac{1}{8} \quad \frac{3}{8} \quad \frac{3}{8} \quad \frac{1}{8}$	Newtonsche 3/8-Regel
4	$\frac{7}{90} \quad \frac{32}{90} \quad \frac{12}{90} \quad \frac{32}{90} \quad \frac{7}{90}$	Milne-Regel
5	$\frac{19}{288} \quad \frac{75}{288} \quad \frac{50}{288} \quad \frac{50}{288} \quad \frac{75}{288} \quad \frac{19}{288}$	
6	$\frac{41}{840} \quad \frac{216}{840} \quad \frac{27}{840} \quad \frac{272}{840} \quad \frac{27}{840} \quad \frac{216}{840} \quad \frac{41}{840}$	
7	$\frac{751}{17280} \quad \frac{3577}{17280} \quad \frac{1323}{17280} \quad \frac{2989}{17280} \quad \frac{2989}{17280} \quad \frac{1323}{17280} \quad \frac{3577}{17280} \quad \frac{751}{17280}$	
8	$\frac{989}{28350} \quad \frac{5888}{28350} \quad -\frac{928}{28350} \quad \frac{10496}{28350} \quad -\frac{4540}{28350} \quad \frac{10496}{28350} \quad -\frac{928}{28350} \quad \frac{5888}{28350} \quad \frac{989}{28350}$	

Tabelle 4.2: Gewichte offener Newton-Cotes Formeln

n	$\omega_0 \dots, \omega_n$	Name
0	1	Mittelpunktsregel
1	$\frac{1}{2} \quad \frac{1}{2}$	
2	$\frac{2}{3} \quad -\frac{1}{3} \quad \frac{2}{3}$	
3	$\frac{11}{24} \quad \frac{1}{24} \quad \frac{1}{24} \quad \frac{11}{24}$	
4	$\frac{11}{20} \quad -\frac{14}{20} \quad \frac{26}{20} \quad -\frac{14}{20} \quad \frac{11}{20}$	

4.3 Gaußsche Quadraturformeln

Die interpolatorischen Quadraturformeln zu den Stützstellen x_0, x_1, \dots, x_n sind nach Konstruktion mindestens exakt für Polynome vom Grade kleiner oder gleich n . Wir haben gesehen, dass die Rechteckregel ($n = 0$) Polynome vom Grade kleiner oder gleich 1, die Cavalieri-Simpson-Formel ($n = 2$) Polynome vom Grade kleiner oder gleich 3 exakt integriert. Es stellt sich die Frage, die Stützstellen x_0, x_1, \dots, x_n und die Gewichte $\omega_0, \omega_1, \dots, \omega_n$ so zu wählen, dass Polynome möglichst hohen Grades exakt integriert werden.

Eine obere Schranke für den maximalen Grad der Polynome, die von der Quadraturformel

$$I_n(f) = (b - a) \sum_{i=0}^n \omega_i f(x_i)$$

exakt integriert werden, ergibt sich aus der Überlegung, dass für das Polynom vom Grade $2n + 2$

$$p(x) = \prod_{i=0}^n (x - x_i)^2$$

$I_n(p)$ verschwindet, jedoch $I(p)$ nicht.

Wir wollen im folgenden zeigen, dass es tatsächlich interpolatorische Quadraturformeln zu $n + 1$ Stützstellen gibt, die Polynome vom Grade kleiner oder gleich $2n + 1$ exakt integrieren. Sie heißen *Gaußsche Quadraturformeln*.

Seien $p_n \in P_n$ und $p_{2n+1} \in P_{2n+1}$ die Lagrangeschen Interpolationspolynome einer Funktion $f \in C[a, b]$ zu den $n + 1$ bzw. $2n + 2$ Stützstellen x_0, \dots, x_n bzw. $x_0, \dots, x_n, x_{n+1}, \dots, x_{2n+1} \in [a, b]$. Für die zugeordneten Quadraturformeln gilt dann

$$\begin{aligned} I(f) - I_{2n+1}(f) &= I(f) - \sum_{i=0}^{2n+1} f[x_0, \dots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j) dx \\ &= I(f) - I_n(f) - \sum_{i=n+1}^{2n+1} f[x_0, \dots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j) dx. \end{aligned}$$

Wir schreiben für $i = n + 1, \dots, 2n + 1$:

$$\int_a^b \prod_{j=0}^{i-1} (x - x_j) dx = \int_a^b \prod_{j=0}^n (x - x_j) \prod_{j=n+1}^{i-1} (x - x_j) dx$$

Die Polynome

$$1, \quad (x - x_{n+1}), \quad (x - x_{n+1})(x - x_{n+2}), \quad \dots, \quad \prod_{j=n+1}^{2n} (x - x_j)$$

bilden eine Basis des P_n . Wählen wir nun die ersten $n + 1$ Stützstellen x_0, \dots, x_n aus $[a, b]$ derart, dass

$$\int_a^b q(x) \prod_{j=0}^n (x - x_j) dx = 0 \quad \forall q \in P_n,$$

so folgt

$$I(f) - I_n(f) = I(f) - I_{2n+1}(f),$$

d.h. die interpolatorische Quadraturformel I_n ist exakt für Polynome aus P_{2n+1} . Damit ergibt sich die Frage nach der Existenz eines Polynoms $n + 1$ -ten Grades der Form

$$p(x) = x^{n+1} + r(x), \quad r \in P_n,$$

das im $L^2(a, b)$ orthogonal zum P_n ist und reelle Nullstellen besitzt, die im Intervall $[a, b]$ liegen.

Wir betrachten im folgenden die Aufgabe etwas allgemeiner und legen ein gewichtetes Skalarprodukt der Form

$$(f, g)_\omega := \int_a^b \omega(x) f(x) g(x) dx$$

mit einer integrierbaren Gewichtsfunktion $\omega > 0$, $x \in (a, b)$ zugrunde. Wie üblich sei dann

$$\|f\|_\omega := \left(\int_a^b \omega(x) f^2(x) dx \right)^{1/2}.$$

Mit Hilfe des *Schmidtschen Orthogonalisierungsverfahrens* gewinnen wir aus den Polynomen $\{1, x, x^2, \dots\}$ Orthogonalpolynome \tilde{p}_n , $n = 0, 1, \dots$:

$$\begin{aligned} \tilde{p}_0(x) &= 1 & p_0(x) &= \|\tilde{p}_0\|_\omega^{-1} \tilde{p}_0(x), \\ \tilde{p}_k(x) &= x^k - \sum_{j=0}^{k-1} (x^k, p_j)_\omega p_j(x), & p_k(x) &= \|\tilde{p}_k\|_\omega^{-1} \tilde{p}_k(x), \quad k = 1, \dots, n+1. \end{aligned}$$

Dann ist $\{\tilde{p}_0, \dots, \tilde{p}_{n+1}\}$ ein *Orthogonalsystem* und $\{p_0, \dots, p_{n+1}\}$ ein *Orthonormalsystem* in P_{n+1} .

Theorem 4.3.1 Die bezüglich des gewichteten Skalarproduktes $(\cdot, \cdot)_\omega$ orthogonalen Polynome \tilde{p}_n , $n \geq 1$, besitzen reelle, einfache Nullstellen, die alle im Innern des Intervalls $[a, b]$ liegen.

Beweis. Wir definieren die Menge

$$N_n := \{\lambda \in (a, b) : \lambda \text{ Nullstelle ungerader Vielfachheit von } \tilde{p}_n\}$$

und setzen

$$\begin{aligned} q(x) &:= 1 && \text{für } N_n = \emptyset \\ q(x) &:= \prod_{i=1}^m (x - \lambda_i) && \text{für } N_n = \{\lambda_1, \dots, \lambda_m\}. \end{aligned}$$

Dann ist das Polynom $q \cdot \tilde{p}_n \in P_{n+m}$ reell und hat in (a, b) keinen Vorzeichenwechsel. Somit gilt

$$(\tilde{p}_n, q) = \int_a^b \omega(x) \tilde{p}_n(x) q(x) dx \neq 0.$$

Angenommen die Anzahl m der reellen Nullstellen in (a, b) ist kleiner als n . Dann ist $q \in P_{n-1}$ und $(\tilde{p}_n, q) = 0$ im Widerspruch zur obigen Beziehung. \square

Die orthogonalen Polynome \tilde{p}_n bezüglich des Gewichtes $\omega(x) = 1$ auf $[a, b]$ heißen *Legendre-Polynome*; üblicherweise betrachtet man diese auf dem Referenzintervall $[-1, +1]$ und bezeichnet sie mit L_n . Wir können nun die Nullstellen $\lambda_0, \dots, \lambda_n$ des $(n+1)$ -ten Legendre-Polynoms L_{n+1} als Stützstellen einer interpolierenden Quadraturformel auf dem Intervall $[-1, +1]$ verwenden:

$$\int_{-1}^{+1} f(x) dx \approx I_n(f) := \sum_{i=0}^n \alpha_i f(\lambda_i), \quad \alpha_i = \int_{-1}^{+1} \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - \lambda_j}{\lambda_i - \lambda_j} dx$$

Theorem 4.3.2 *Es gibt genau eine interpolatorische Quadraturformel zu $n+1$ paarweise verschiedenen Stützstellen über dem Intervall $[-1, 1]$, die Polynome vom Grade kleiner oder gleich $2n+1$ exakt integriert. Ihre Stützstellen sind die Nullstellen $\lambda_0, \dots, \lambda_n \in (-1, 1)$ des $(n+1)$ -ten Legendre-Polynoms $L_{n+1} \in P_{n+1}$, und ihre Gewichte genügen der Beziehung*

$$\alpha_i = \int_{-1}^{+1} \prod_{\substack{j=0 \\ j \neq i}}^n \left(\frac{x - \lambda_j}{\lambda_i - \lambda_j} \right)^2 dx > 0, \quad i = 0, \dots, n.$$

Für $f \in C^{2n+2}[-1, 1]$ gilt die Restglieddarstellung

$$R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{+1} \prod_{j=0}^n (x - \lambda_j)^2 dx, \quad \xi \in (-1, 1).$$

Beweis. Das Legendre-Polynom L_{n+1} ist orthogonal zu P_n und hat mit seinen (reellen) Nullstellen $\lambda_0, \dots, \lambda_n$ die Darstellung

$$L_{n+1}(x) = \prod_{i=0}^n (x - \lambda_i).$$

Nach den obigen Vorbetrachtungen integriert die interpolatorische Quadraturformel dann Polynome vom Grade kleiner gleich $2n + 1$ exakt. Zur Bestimmung der Gewichte α_i betrachten wir die Polynome

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \left(\frac{x - \lambda_j}{\lambda_i - \lambda_j} \right), \quad i = 0, \dots, n.$$

Da $l_i^2 \in P_{2n}$ folgt

$$0 < \int_{-1}^1 l_i^2(x) dx = \sum_{j=0}^n \alpha_j l_i(\lambda_j)^2 = \alpha_i.$$

Zur Eindeutigkeit der Gaußschen Quadraturformel sei angenommen, es gäbe eine zweite Formel

$$I_n^*(f) = \sum_{i=0}^n \alpha_i^* f(\lambda_i^*),$$

die Polynome vom Grade kleiner oder gleich $2n + 1$ exakt integriert. Ersetzt man in der Definition von l_i die Nullstellen λ_i durch λ_i^* erhält man Polynome $l_i^* \in P_n$, mit denen man wie oben $\alpha_i^* > 0$ zeigen kann. Somit wäre

$$0 = \int_{-1}^1 \frac{1}{\alpha_i^*} l_i^*(x) L_{n+1}(x) dx = \sum_{j=0}^n \frac{\alpha_j^*}{\alpha_i^*} l_i^*(\lambda_j^*) L_{n+1}(\lambda_j^*) = L_{n+1}(\lambda_i^*).$$

Aus der eindeutigen Bestimmtheit der Nullstellen λ_i von L_{n+1} folgt damit $\lambda_i = \lambda_i^*$ und $\alpha_i = \alpha_i^*$.

Wir müssen noch die Darstellung für das Restglied zeigen. Nach den Aussagen zur Hermite-Interpolation gibt es ein eindeutig bestimmtes Polynom $H \in P_{2n+1}$, das die Hermitesche Interpolationsaufgabe

$$H(\lambda_i) = f(\lambda_i), \quad H'(\lambda_i) = f'(\lambda_i), \quad i = 0, \dots, n,$$

löst und für $f \in C^{2n+2}[-1, 1]$ die Restglieddarstellung

$$f(x) - H(x) = f[\lambda_0, \lambda_0, \dots, \lambda_n, \lambda_n, x] \prod_{i=0}^n (x - \lambda_i)^2$$

hat. Die Anwendung der Gaußschen Quadraturformel auf H ergibt dann

$$\begin{aligned} I(f) - I_n(f) &= I(f - H) - I_n(f - H) \\ &= \int_{-1}^1 f[\lambda_0, \lambda_0, \dots, \lambda_n, \lambda_n, x] \prod_{i=0}^n (x - \lambda_i)^2 dx - \sum_{i=0}^n \alpha_i (f(\lambda_i) - h(\lambda_i)) \\ &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^1 \prod_{i=0}^n (x - \lambda_i)^2 dx. \end{aligned}$$

Im letzten Schritt haben wir die Integaldarstellung dividierter Differenzen und den Mittelwertsatz der Integralrechnung benutzt. \square

Die Legendre-Polynome $L_n \in P_n$ lassen sich auf $[-1, 1]$ in der Form (Formel von Rodriguez)

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, \dots,$$

darstellen und genügen der rekursiven Beziehung

$$L_0(x) = 1, \quad L_1(x) = x, \quad L_{n+1}(x) = \frac{2n+1}{n+1} x L_n(x) - \frac{n}{n+1} L_{n-1}(x), \quad n \geq 1.$$

Ihre Nullstellen werden analytisch beziehungsweise (für $n > 3$) numerisch bestimmt und können der Tabelle 4.3 entnommen werden. Die Skalierung ergibt sich aus $L_n(1) = 1$ für alle $n \in \mathbb{N}$. Wir haben insbesondere

$$\begin{aligned} L_2(x) &= \frac{3x^2 - 1}{2} & \lambda_0 &= -\frac{1}{\sqrt{3}}, & \lambda_1 &= \frac{1}{\sqrt{3}}, \\ L_3(x) &= \frac{5x^3 - 3x}{2} & \lambda_0 &= -\sqrt{\frac{3}{5}}, & \lambda_1 &= 0, & \lambda_2 &= \sqrt{\frac{3}{5}}. \end{aligned}$$

Zur Bestimmung der Gewichte α_i nutzen wir die Eigenschaft, dass

$$\int_{-1}^1 p(x) dx = \sum_{j=0}^n \alpha_j p(\lambda_j) \quad \forall p \in P_{2n+1}.$$

Setzen wir

$$p(x) = \frac{L_{n+1}(x)}{x - \lambda_i} L_n(x) \in P_{2n},$$

so folgt wegen $p(\lambda_j) = 0$ für $j \neq i$ und $p(\lambda_i) = L'_{n+1}(\lambda_i) L_n(\lambda_i)$

$$\int_{-1}^1 \frac{L_{n+1}(x)}{x - \lambda_i} L_n(x) dx = \alpha_i L'_{n+1}(\lambda_i) L_n(\lambda_i).$$

Wir haben mit einem gewissen Polynom $Q_{n-1}(x)$ vom Grade kleiner gleich $n - 1$

$$\frac{L_{n+1}(x)}{x - \lambda_i} = \frac{2n+1}{n+1} L_n(x) + Q_{n-1}(x),$$

woraus wegen der Orthogonalität $L_n \perp P_{n-1}$ die Beziehung

$$\frac{2n+1}{n+1} (L_n, L_n) = \alpha_i L'_{n+1}(\lambda_i) L_n(\lambda_i)$$

Tabelle 4.3: Stützstellen und Gewichte der Gaußquadratur auf $[-1, 1]$.

n	$\lambda_0, \dots, \lambda_{n-1}$	$\alpha_0, \dots, \alpha_{n-1}$
1	0	2
2	$\pm 1/\sqrt{3}$	1
3	$\pm \sqrt{3/5}$	5/9
	0	8/9
4	± 0.8611363116	0.3478548451
	± 0.3399810436	0.6521451549

folgt. Aus der Formel von Rodriguez folgt durch mehrfache partielle Integration

$$\begin{aligned}
 (L_n, L_n) &= \frac{(-1)^n}{2^{2n}(n!)^2} \int_{-1}^1 (x^2 - 1)^n \frac{d^{2n}}{dx^{2n}} (x^2 - 1)^n dx \\
 &= \frac{(-1)^n (2n)!}{2^{2n}(n!)^2} \int_{-1}^1 (x-1)^n (x+1)^n dx \\
 &= \frac{(2n)!}{2^{2n}(n!)^2} \frac{n!}{(n+1) \cdots (2n)} \int_{-1}^1 (x+1)^{2n} dx \\
 &= \frac{2}{2n+1}.
 \end{aligned}$$

Für die Gewichte folgt damit

$$\alpha_i = \frac{1}{L'_{n+1}(\lambda_i) L_n(\lambda_i)} \cdot \frac{2}{n+1},$$

wobei λ_j , $j = 0, 1, \dots, n$, die Nullstellen des Legendre-Polynoms L_{n+1} sind. Für das Restglied gilt

$$R_n(f) = \frac{(n+1)!^4 2^{2n+3}}{(2n+2)!^3 (2n+3)} f^{(2n+2)}(\xi).$$

Die Gaußschen Quadraturformeln über einem beliebigen Intervall $[a, b]$ gewinnt man durch Anwendung einer Koordinatentransformation, die das Referenzintervall $[-1, 1]$ vermöge $t \mapsto x = (b-a)t/2 + (b+a)/2$ auf $[a, b]$ abbildet.

Theorem 4.3.3 Sei $I_n(f)$ die $n+1$ -punktige Gauß-Legendre-Formel zur Interpolation von f auf $[-1, 1]$. Dann gilt für jede Funktion $f \in C[-1, 1]$

$$\int_{-1}^1 f(x) dx - I_n(f) \rightarrow 0 \quad \text{für } n \rightarrow \infty.$$

Beweis. Betrachten wir die Gauß-Legendre-Formeln für beliebiges n ,

$$I_n(f) = \sum_{i=0}^n \alpha_i^{(n)} f(\lambda_i^{(n)}), \quad \alpha_i^{(n)} > 0, \quad \sum_{i=0}^n \alpha_i^{(n)} = 2.$$

Nach dem Weierstraßschen Approximationssatz, gibt es zu jeder vorgegebenen Toleranzschranke $\varepsilon > 0$ und jeder auf $[-1, 1]$ stetigen Funktion f ein Polynom p_ε mit

$$\|f - p_\varepsilon\| = \max_{x \in [-1, 1]} |f(x) - p_\varepsilon(x)| < \frac{\varepsilon}{4}.$$

Wir splitten den Fehler in die drei Anteile

$$|I(f) - I_n(f)| \leq |I(f - p_\varepsilon)| + |I(p_\varepsilon) - I_n(p_\varepsilon)| + |I_n(p_\varepsilon - f)|.$$

Der erste wird mittels Integralabschätzung (Länge des Integrationsintervalls \times Betragsmaximum des Integranden) abgeschätzt, also

$$|I(f - p_\varepsilon)| \leq 2 \frac{\varepsilon}{4}.$$

Der zweite verschwindet für hinreichend grosse n , da dann das Polynom p_ε exakt integriert wird. Für den dritten gilt

$$|I_n(p_\varepsilon - f)| \leq \sum_{i=0}^n \alpha_i^{(n)} \|p_\varepsilon - f\| \leq \frac{\varepsilon}{4} \sum_{i=0}^n \alpha_i^{(n)} = \frac{\varepsilon}{2}.$$

Fassen wir die drei Abschätzungen zusammen, erhalten wir die Konvergenz der Folge $I_n(f)$ gegen $I(f)$. \square

Die Methode der Konstruktion der Gauß-Legendre-Formeln zur optimalen Berechnung von $I(f)$ kann auf den Fall von Integralen

$$I_\omega(f) = \int_a^b \omega(x) f(x) dx$$

mit einer integrierbaren Gewichtsfunktion $\omega(x) > 0$ auf (a, b) übertragen werden. Hierzu verwendet man als Stützstellen gerade die Nullstellen der bezüglich des gewichteten Skalarproduktes

$$(f, g)_\omega = \int_a^b \omega(x) f(x) g(x) dx$$

orthogonalen Polynome.

Beispiel: Wir betrachten $[a, b] = [-1, 1]$ und $\omega(x) = (1 - x^2)^{-1/2}$. Die orthogonalen Polynome sind in diesem Fall die *Tschebyscheff-Polynome* $T_n \in P_n$, die durch die rekursive Beziehung

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1,$$

bestimmt sind. Die Stützstellen und Gewichte der zugehörigen Quadraturformeln sind

$$\lambda_i = \cos\left(\frac{2i+1}{n+1} \cdot \frac{\pi}{2}\right), \quad \alpha_i = \frac{\pi}{n+1}, \quad i = 0, \dots, n.$$

Die Restglieder haben die Form

$$R_n(f) = \frac{2\pi}{(2n+2)!} \left(\frac{1}{2}\right)^{2n+2} f^{(2n+2)}(\xi), \quad \xi \in (-1, 1).$$

Der Fall $n = 2$ ergibt

$$\int_{-1}^1 \omega(x) f(x) dx = \frac{\pi}{3} \left\{ f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right) \right\} + \frac{\pi}{23040} f^{(4)}(\xi).$$

Kapitel 5

Approximation

Bei der Interpolationsaufgabe wurde eine stetige Funktion durch ein Interpolationspolynom angenähert. Die Forderung war dabei, dass Funktion und Interpolationspolynom in ausgewählten Stützstellen – den Interpolationsknoten – übereinstimmen. Wir betrachten in diesem Abschnitt die Aufgabe, dass eine stetige Funktion in gewissem Sinne bestmöglich approximiert wird. Dazu sei im folgenden die Menge der auf $[a, b]$ stetigen reell- bzw. komplexwertigen Funktionen als Vektorraum über den Zahlkörper $\mathbb{K} = \mathbb{R}$ bzw. $\mathbb{K} = \mathbb{C}$ aufgefasst. Gegeben sei eine Funktion $f \in C[a, b]$ sowie eine Teilmenge $S \subset C[a, b]$, deren Elemente zur Approximation von f dienen sollen.

Die Aufgabe der besten Approximation einer Funktion $f \in C[a, b]$ durch Elemente aus S lautet:

$$\text{Finde } p \in S \text{ mit } \|f - p\| = \inf_{q \in S} \|f - q\|.$$

Die Güte der Approximation wird dabei in der Norm $\|\cdot\|$ gemessen.

5.1 Gauß-Approximation

Bei der Gauß-Approximation bzw. der Approximation im quadratischen Mittel verwendet man die durch das Skalarprodukt

$$(f, g) = \int_a^b f(x) \overline{g(x)} dx$$

erzeugte Norm

$$\|f\| = \left(\int_a^b |f(x)|^2 dx \right)^{1/2}.$$

Versehen mit dem Skalarprodukt (\cdot, \cdot) wird $C[a, b]$ zu einem unitären Raum.

Theorem 5.1.1 *Seien H ein unitärer Raum und $S \subset H$ ein endlich dimensionaler Teilraum. Dann existiert zu jedem $f \in H$ eine eindeutig bestimmte beste Approximation $p \in S$.*

Beweis. Sei $\{\psi_1, \dots, \psi_n\}$ eine Basis von S , also $n = \dim S$. Mit dem Schmidtschen Orthogonalisierungsverfahren

$$\begin{aligned}\tilde{\varphi}_1 &= \psi_1, & \varphi_1 &= \|\tilde{\varphi}_1\|^{-1}\tilde{\varphi}_1, \\ \tilde{\varphi}_i &= \psi_i - \sum_{j=1}^{i-1} (\psi_i, \varphi_j) \varphi_j, & \varphi_i &= \|\tilde{\varphi}_i\|^{-1}\tilde{\varphi}_i,\end{aligned}$$

erzeugt man ein Orthonormalsystem $\{\varphi_1, \dots, \varphi_n\}$ in S , d.h. wir haben

$$(\varphi_i, \varphi_j) = \delta_{ij}, \quad i, j = 1, \dots, n.$$

In der Tat ist φ_1 wegen $\psi_1 \neq 0$ wohldefiniert. Seien nun $\{\varphi_1, \dots, \varphi_{m-1}\}$ wohldefiniert und orthonormal. Wäre nun $\tilde{\varphi}_m = 0$, so würde

$$\psi_m = \sum_{j=1}^{m-1} (\psi_m, \varphi_j) \varphi_j$$

gelten, d.h. ψ_m wäre linear von $\psi_1, \dots, \psi_{m-1}$ linear abhängig im Widerspruch zur Annahme, dass $\{\psi_1, \dots, \psi_n\}$ eine Basis von S sei. Damit ist auch $\varphi_m = \|\tilde{\varphi}_m\|^{-1}\tilde{\varphi}_m$ wohldefiniert und für $k = 1, \dots, i-1$ gilt

$$(\varphi_m, \varphi_k) = (\psi_i, \varphi_k) - \sum_{j=1}^{i-1} (\psi_i, \varphi_j) (\varphi_j, \varphi_k) = (\psi_i, \varphi_k) - (\psi_i, \varphi_k) = 0.$$

Als Orthonormalsystem ist $\{\varphi_1, \dots, \varphi_n\}$ automatisch eine Basis von S , denn

$$\sum_{j=1}^n \alpha_j \varphi_j = 0 \quad \Rightarrow \quad \alpha_k = \sum_{j=1}^n \alpha_j (\varphi_j, \varphi_k) = 0.$$

Jedes $p \in S$ kann damit eindeutig in der Form

$$p = \sum_{j=1}^n \alpha_j \varphi_j$$

dargestellt werden, die Koeffizienten α_j sind bestimmt durch

$$(p, \varphi_k) = \sum_{j=1}^n \alpha_j (\varphi_j, \varphi_k) = \alpha_k.$$

Für beliebiges

$$q = \sum_{j=1}^n \beta_j \varphi_j \in S$$

gilt nun

$$\begin{aligned} \|f - q\|^2 &= (f - q, f - q) = \left(f - \sum_{j=1}^n \beta_j \varphi_j, f - \sum_{j=1}^n \beta_j \varphi_j \right) \\ &= (f, f) - \sum_{j=1}^n \beta_j (\varphi_j, f) - \sum_{j=1}^n \overline{\beta_j} (f, \varphi_j) + \sum_{i,j=1}^n \beta_i \overline{\beta_j} (\varphi_i, \varphi_j) \\ &= \|f\|^2 - 2 \sum_{j=1}^n \operatorname{Re}[\beta_j (\varphi_j, f)] + \sum_{j=1}^n |\beta_j|^2. \end{aligned}$$

Wegen

$$|\beta_j - (f, \varphi_j)|^2 = [\beta_j - (f, \varphi_j)] [\overline{\beta_j} - \overline{(f, \varphi_j)}] = |\beta_j|^2 - 2\operatorname{Re}[\beta_j (\varphi_j, f)] + |(f, \varphi_j)|^2$$

folgt

$$\|f - q\|^2 = \|f\|^2 - \sum_{j=1}^n |(f, \varphi_j)|^2 + \sum_{j=1}^n |\beta_j - (f, \varphi_j)|^2,$$

somit ist $q \in S$ genau dann beste Approximation von f , wenn für die Koeffizienten β_j gilt $\beta_j = (f, \varphi_j)$, $j = 1, \dots, n$. \square

Die beste Approximation $p \in S$ von $f \in C[a, b]$ im quadratischen Mittel kann also in der Form

$$p = \sum_{j=1}^n (f, \varphi_j) \varphi_j$$

dargestellt werden, wobei $\{\varphi_1, \dots, \varphi_n\}$ ein Orthonormalsystem von S ist. Der Fehler der besten Approximation berechnet sich demnach als

$$\|f - p\| = \left(\|f\|^2 - \sum_{j=1}^n |(f, \varphi_j)|^2 \right)^{1/2}.$$

Theorem 5.1.2 Die beste Approximation $p \in S$ von f kann äquivalent durch die Orthogonalitätseigenschaft $f - p \perp S$, i.e.

$$(f - p, \varphi) = 0 \quad \forall \varphi \in S$$

charakterisiert werden.

Beweis. Sei $\{\varphi_1, \dots, \varphi_n\}$ ein Orthonormalsystem in S . Dann ist q genau dann beste Approximation von f , wenn

$$(f - q, \varphi_i) = \left(f - \sum_{j=1}^n \beta_j \varphi_j, \varphi_i \right) = (f, \varphi_i) - \beta_i = 0, \quad i = 1, \dots, n.$$

Andererseits ist

$$(f - q, \varphi) = 0, \quad \forall \varphi \in S \quad \Leftrightarrow \quad (f - q, \varphi_i) = 0, \quad i = 1, \dots, n.$$

□

Die Orthogonalitätseigenschaft ermöglicht die Bestimmung der besten Approximation in Bezug auf eine beliebige Basis $\{\psi_1, \dots, \psi_n\}$ von S . Sei

$$p = \sum_{j=1}^n \gamma_j \psi_j$$

die eindeutige Darstellung der besten Approximation von $f \in C[a, b]$ bezüglich dieser Basis. Dann sind die Koeffizienten γ_j wegen der Orthogonalitätseigenschaft der besten Approximation Lösung des linearen Gleichungssystems

$$0 = \left(f - \sum_{j=1}^n \gamma_j \psi_j, \psi_i \right) = (f, \psi_i) - \sum_{j=1}^n \gamma_j (\psi_j, \psi_i).$$

Die Koeffizientenmatrix $A = (\psi_j, \psi_i)$ ist die *Gramsche Matrix* der Basis $\{\psi_1, \dots, \psi_n\}$; sie ist hermitisch und positiv definit,

$$(A\gamma, \gamma) = \sum_{i,j=1}^n a_{ij} \gamma_j \bar{\gamma}_i = \sum_{i,j=1}^n (\psi_j, \psi_i) \gamma_j \bar{\gamma}_i = \left(\sum_{j=1}^n \gamma_j \psi_j, \sum_{i=1}^n \bar{\gamma}_i \psi_i \right) = (p, p) = \|p\|^2 > 0$$

für jedes $p \neq 0$.

Betrachten wir als Beispiel für die Menge S die Polynome vom Grade kleiner gleich n auf dem Intervall $[-1, +1]$. Verwenden wir die Standardbasis des P_n , also $\{1, x, x^2, \dots, x^n\}$, so sind die Einträge der Gramschen Matrix

$$a_{ij} = \int_{-1}^1 x^j x^i dx = \begin{cases} 0 & \text{falls } i + j \text{ gerade} \\ \frac{2}{i + j + 1} & \text{falls } i + j \text{ ungerade.} \end{cases}$$

Die zugehörige Matrix ist zwar regulär, aber extrem schlecht konditioniert, für $n = 10$ berechnet MATLAB $\text{cond}(A) = 2.2 \cdot 10^{33}$. Wesentlich besser geeignet sind

die orthogonalen Legendre-Polynome $L_n(x)$, die wir im Zusammenhang mit der Gaußquadratur kennengelernt haben. Wegen

$$(L_n, L_n) = \frac{2}{2n+1}, \quad n = 0, 1, \dots$$

bilden die Polynome

$$\varphi_k(x) = \sqrt{\frac{2k-1}{2}} L_{k-1}(x), \quad k = 1, 2, \dots$$

ein orthogonales System. Die Berechnung der besten Approximation p von f gestaltet sich nun wesentlich einfacher

$$p(x) = \sum_{j=1}^n \int_{-1}^1 f(t) \varphi_j(t) dt \varphi_j(x).$$

Die Maximalabweichung

$$\|f - p\|_\infty = \max_{-1 \leq x \leq 1} |f(x) - p(x)|$$

kann jedoch (vor allem zu den Intervallenden hin) sehr groß werden. Um diesen Effekt zu unterdrücken, verwendet man das gewichtete Skalarprodukt

$$(f, g)_\omega = \int_{-1}^1 \omega(x) f(x) g(x) dx, \quad \omega(x) = \frac{1}{\sqrt{1-x^2}},$$

wodurch eine stärkere Bindung an den Intervallenden erreicht wird. Durch Orthogonalisierung von $\{1, x, \dots, x^{n-1}\}$ bezüglich des mit $1/\sqrt{1-x^2}$ gewichteten Skalarprodukts erhält man die Polynome

$$\varphi_1 = \sqrt{\frac{1}{\pi}}, \quad \varphi_k = \sqrt{\frac{2}{\pi}} T_{k-1}(x), \quad k = 2, \dots, n,$$

mit den *Tschebyscheff-Polynomen* T_k .

Theorem 5.1.3 Die *Tschebyscheff-Polynome* haben für $x \in [-1, 1]$ die Gestalt

$$T_k(x) = \cos(k \arccos(x)), \quad k = 0, 1, 2, \dots,$$

und genügen den Beziehungen

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_i(x) T_j(x) dx = \begin{cases} \pi & \text{für } i = j = 0 \\ \pi/2 & \text{für } i = j \neq 0 \\ 0 & \text{für } i \neq j. \end{cases}$$

Beweis. Wir zeigen zunächst die Rekursionsbeziehung für $g_k(x) := \cos(k \arccos(x))$. Wir haben unmittelbar $g_0(x) = 1$ und $g_1(x) = x$. Aus den Additionstheoremen leitet man

$$\cos(\alpha + \beta) + \cos(\alpha - \beta) = 2 \cos \alpha \cos \beta$$

oder mit $x = \alpha + \beta$ und $y = \alpha - \beta$

$$\cos x + \cos y = 2 \cos \frac{x+y}{2} \cos \frac{x-y}{2}$$

ab. Somit ist

$$\begin{aligned} g_{k+1}(x) + g_{k-1}(x) &= \cos((k+1) \arccos(x)) + \cos((k-1) \arccos(x)) \\ &= 2 \cos(k \arccos(x)) \cos(\arccos(x)) = 2xg_k(x). \end{aligned}$$

Damit ist g_k genau ein Polynom k -ten Grades und genügt der obigen Rekursionsbeziehung. Weiter gilt mit der Substitution $x = \cos t$

$$\begin{aligned} \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} g_j(x) g_i(x) dx &= \int_0^\pi \cos jt \cos it dt \\ &= \frac{1}{2} \int_0^\pi \cos(i+j)t + \cos(i-j)t dt \end{aligned}$$

woraus obige Orthogonalitätseigenschaft folgt. □

5.2 Tschebyscheff-Approximation

Im folgenden betrachten wir nur reell-wertige Funktionen. Im Unterschied zur Gauß-Approximation, die die beste Approximation im quadratischen Mittel bzw. im gewichteten quadratischen Mittel sucht, verwendet die Tschebyscheff-Approximation direkt die Maximumnorm

$$\|f\|_\infty = \max_{x \in [a,b]} |f(x)|.$$

Diese Norm wird nicht durch ein Skalarprodukt erzeugt, die Existenz einer besten Approximation kann daher nicht aus Theorem 5.1.1 gefolgert werden. Tatsächlich ist im allgemeinen die beste Approximation nicht einmal eindeutig bestimmt, wie das folgende Beispiel zeigt:

Beispiel: Seien $[a, b] = [0, 1]$ und $f(x) = 1$ auf $[0, 1]$. Wir wollen die Funktion f in der Maximumnorm bestmöglich durch ein Element der eindimensionalen Menge $S = \{p_\alpha(x) = \alpha x : \alpha \in \mathbb{R}\}$ approximieren. Wir haben $\|f - p_\alpha\|_\infty \geq 1$ für alle $p \in S$ und $\|f - p_\alpha\|_\infty = 1$ für $0 \leq \alpha \leq 2$. □

Theorem 5.2.1 Sei E ein normierter Vektorraum und $S \subset E$ ein endlich dimensionaler Teilraum. Dann gibt es zu jedem $f \in E$ eine beste Approximation $p \in S$:

$$\|f - p\| = \min_{q \in S} \|f - q\|.$$

Beweis. Ein Element $q_0 \in S$ mit $\|q_0\| > 2\|f\|$ kann keine beste Approximation sein, denn

$$\|f - q_0\| \geq \|q_0\| - \|f\| > \|f\| = \|f - 0\| \geq \inf_{q \in S} \|f - q\|.$$

Somit haben wir

$$\inf_{q \in S} \|f - q\| = \inf_{q \in S, \|q\| \leq 2\|f\|} \|f - q\|.$$

Nun ist die Abbildung $F : E \rightarrow \mathbf{R}$ mit $F(q) = \|f - q\|$ auf der kompakten Menge $A = \{q \in S : \|q\| \leq 2\|f\|\}$ stetig, denn

$$|F(q_1) - F(q_2)| = \left| \|f - q_1\| - \|f - q_2\| \right| \leq \|q_1 - q_2\|.$$

Nach dem Satz von Weierstraß nimmt F auf A das Infimum als Funktionswert an, es gibt also eine beste Approximation p von f . \square

Im Hinblick auf die Eindeutigkeit führen wir folgende Definition ein.

Definition: Wir nennen einen linearen normierten Raum E streng normiert, wenn in der Dreiecksungleichung

$$\|q_1 + q_2\| \leq \|q_1\| + \|q_2\|$$

das Gleichheitszeichen für $q_1 \neq 0$, $q_2 \neq 0$, nur im Fall $q_2 = \alpha q_1$ mit positivem α gilt. Beachte, dass der \mathbb{R}^3 versehen mit der euklidischen Norm streng normiert ist. Es ist sogar jeder unitäre Raum streng normiert. Zum Beweis nehmen wir an, dass

$$\|q_1 + q_2\| = \|q_1\| + \|q_2\|, \quad q_1 \neq 0, \quad q_2 \neq 0$$

gelte. Durch Quadrieren erhalten wir

$$\|q_1\|^2 + 2(q_1, q_2) + \|q_2\|^2 = (q_1 + q_2, q_1 + q_2) = \|q_1 + q_2\|^2 = \|q_1\|^2 + 2\|q_1\| \|q_2\| + \|q_2\|^2.$$

Also gibt es zwei Elemente $y_i = q_i/\|q_i\|$, $i = 1, 2$, der Länge 1, deren Skalarprodukt $(y_1, y_2) = 1$ ergibt. Wegen

$$\|y_1 - y_2\|^2 = (y_1 - y_2, y_1 - y_2) = \|y_1\|^2 - 2(y_1, y_2) + \|y_2\|^2 = 1 - 2 + 1 = 0$$

muss aber $y_1 = y_2$ bzw. $q_2 = (\|q_2\|/\|q_1\|) q_1$ sein. \square

Theorem 5.2.2 *Ist E streng normiert, gibt es höchstens ein Element $p \in S$ der besten Approximation von $f \in E$.*

Beweis. Nehmen wir an, dass zwei verschiedene Elemente q_1, q_2 der besten Approximation von f in S existieren. Dann ist

$$\|f - q_1\| = \|f - q_2\| = m > 0,$$

da andernfalls $q_1 = q_2 = f$ wäre. Nun ist

$$\begin{aligned} m &\leq \left\| f - \frac{q_1 + q_2}{2} \right\| = \left\| \frac{1}{2}(f - q_1) + \frac{1}{2}(f - q_2) \right\| \\ &\leq \frac{1}{2}\|(f - q_1)\| + \frac{1}{2}\|(f - q_2)\| = m. \end{aligned}$$

Nun war E streng normiert, also muss $f - q_1 = \alpha(f - q_2)$ bzw. $(1 - \alpha)f = q_1 - \alpha q_2$ gelten. Wäre $\alpha \neq 1$, so ist f eine Linearkombination von Elementen aus S und $m = 0$ im Widerspruch zur Annahme. Somit ist $\alpha = 1$ und folglich $q_1 = q_2$. \square

Leider ist der lineare normierte Raum $C[a, b]$ versehen mit der Supremumsnorm nicht streng normiert. Hierzu betrachten wir $[a, b] = [0, 1]$ mit $q_1 = 1$ und $q_2 = x$. Es gilt nämlich

$$2 = \|q_1 + q_2\|_\infty = \|q_1\|_\infty + \|q_2\|_\infty = 1 + 1.$$

Die Eindeutigkeit der Tschebyscheff-Approximation wird durch die *Haarsche Bedingung* (H) an den Ansatzraum $S \subset C[a, b]$ mit $\dim S = n$ garantiert. Wir betrachten zunächst die Lösbarkeit der allgemeinen Interpolationsaufgabe

Finde $p \in S = \text{span}(\varphi_1, \dots, \varphi_n)$ mit $\dim S = n$, so dass zu paarweise verschiedenen Knoten $x_j, j = 1, \dots, n$, und Werten $y_j, j = 1, \dots, n$, die Interpolationsbedingung $p(x_j) = y_j, j = 1, \dots, n$, gilt.

Im Unterschied zur im Abschnitt 3.1 betrachteten Polynominterpolation hat die allgemeine Interpolationsaufgabe nicht notwendig eine Lösung. So verschwindet beispielsweise jede Linearkombination der Funktionen $\varphi_1(x) = \sin x$ und $\varphi_2(x) = \sin 2x$ an der Stelle $x = x_1 = 0$. Somit ist die Interpolationsaufgabe für alle $y_1 \neq 0$ unlösbar. Im allgemeinen Fall ist die Interpolationsaufgabe genau dann für beliebige Werte $y_j, j = 1, \dots, n$ lösbar, wenn die *Haar-Matrix* $H = (\varphi_i(x_j))$ nicht singular ist. Dies ist, wie wir oben gesehen haben, nicht notwendig für jeden Satz paarweise verschiedener Knoten x_1, \dots, x_n der Fall. Die lineare Unabhängigkeit der Funktionen $\varphi_1, \dots, \varphi_n$ sichert aber, dass es zumindest einen Satz paarweise verschiedener Knoten x_1, \dots, x_n gibt, für den $\det H \neq 0$ gilt. Im oben angegebenen Beispiel $S = \text{span}(\sin x, \sin 2x)$ könnte man beispielsweise $x_1 = \pi/4$ und $x_2 = \pi/2$ nehmen.

Theorem 5.2.3 (Haar) *Zu jeder gegebenen Funktion $f \in C[a, b]$ existiert genau dann ein eindeutig bestimmtes $p \in S$,*

(H) *wenn jedes Element $q \in S \setminus \{0\}$ des Unterraumes $S \subset C[a, b]$ mit $\dim S = n$ maximal $n - 1$ Nullstellen in $[a, b]$ hat. Das S erzeugende System $\varphi_1, \dots, \varphi_n$ heißt dann Tschebycheff-System.*

Beweis. Siehe z.B. I.S. Beresin, N.P. Shidkow: Numerische Methoden I, Deutscher Verlag der Wissenschaften, Berlin 1970. \square

So ist das System

$$S = \text{span}(\sin x, \sin 2x) \subset C\left[0, \frac{\pi}{3}\right], \quad \dim S = 2,$$

kein Tschebycheff-System, denn $q(x) = \sin x - \sin 2x$ verschwindet beispielsweise bei $x = 0$ und $x = \pi/3$.

Äquivalent zur Haarschen Bedingung ist die Forderung

(H') Für jeden Satz paarweise verschiedener Knotenwerte $a \leq x_1 < x_2 < \dots < x_n \leq b$ ist die Interpolationsaufgabe $p(x_i) = y_i, i = 1, \dots, n$ mit beliebigen Werten $y_1, \dots, y_n \in \mathbb{R}$ eindeutig durch ein $p \in S$ lösbar.

Beachte, dass (H') gleichbedeutend mit $\det H \neq 0$ für jeden Satz paarweise verschiedener Knotenwerte x_1, \dots, x_n ist.

Beweis der Äquivalenz. Wir zeigen zunächst $(H') \Rightarrow (H)$: Sei $q \in S \setminus \{0\}$ und nehmen wir indirekt an, q habe n Nullstellen $x_i, i = 1, \dots, n$. Die eindeutige Lösbarkeit der Interpolationsaufgabe $q(x_i) = 0$ liefert $q(x) = 0$ für alle $x \in [a, b]$ im Widerspruch zu $q \in S \setminus \{0\}$.

Umgekehrt gilt $\neg(H') \Rightarrow \neg(H)$: Wir haben also einen Satz paarweise verschiedener Knotenwerte x_1, \dots, x_n und Werte y_1, \dots, y_n , für den die Interpolationsaufgabe $q(x_i) = y_i, i = 1, \dots, n$ nicht eindeutig in S lösbar ist. Dann muss die Determinante der Haar-Matrix verschwinden und die Interpolationsaufgabe mit homogenen Werten $y_i = 0, i = 1, \dots, n$ ist nicht eindeutig lösbar. Neben $q(x) = 0$ gibt es also ein $q \in S \setminus \{0\}$ mit $q(x_i) = 0$, also gibt es ein Element $q \in S \setminus \{0\}$, das mindestens n Nullstellen besitzt. \square

Beispiele: Die Polynomräume $S = P_{n-1}$ erfüllen die Haarsche Bedingung auf jedem Intervall. Im Fall $0 \in [a, b]$ und $S = \text{span}(x, x^2, \dots, x^n)$ ist die Bedingung jedoch nicht erfüllt. \square

Theorem 5.2.4 (Alternantensatz von Tschebyscheff) *Für den Teilraum $S \subset C[a, b]$ mit $\dim S = n$ sei die Haarsche Bedingung erfüllt. Dann ist die*

Tschebyscheff-Approximation p einer Funktion $f \in C[a, b]$ durch folgende Eigenschaft charakterisiert. Es gibt mindestens $n + 1$ Stellen $a \leq x_1 < x_2 < \dots < x_{n+1} \leq b$ (Alternante genannt), so dass für die Fehlerfunktion $e(x) = f(x) - p(x)$ gilt:

$$|e(x_i)| = \|e\|_\infty, \quad e(x_i) = -e(x_{i+1}); \quad i = 1, \dots, n$$

Insbesondere ist die beste Approximation eindeutig bestimmt.

Beweis. Siehe etwa G. Maß Vorlesungen über Numerische Mathematik II, Akademie-Verlag Berlin 1988. Wir wollen hier nur die Eindeutigkeit der besten Approximation für den Spezialfall $S = P_{n-1}$ zeigen. Seien q_i , $i = 1, 2$, zwei beste Approximationen mit den Fehlerfunktionen $e_i = f - q_i$. Für $\lambda \in (0, 1)$ ist dann $\|\lambda e_2\|_\infty < \|e_1\|_\infty$. Aufgrund der Existenz einer Alternante schneidet der Graph von e_1 den von λe_2 mindestens n -mal. Jede der Funktionen $\varphi_\lambda(x) = e_1(x) - \lambda e_2(x)$ hat somit mindestens n (paarweise verschiedene) Nullstellen. Gehen wir zur Grenze $\lambda \rightarrow 1$ über, so muss $\varphi_1(x) = e_1(x) - e_2(x) = q_2(x) - q_1(x) \in P_{n-1}$ mindestens n (ihrer Vielfachheit entsprechend oft gezählte) Nullstellen haben, dies bedeutet zwangsläufig $q_2 = q_1$. \square

Zur Anwendung der Tschebyscheff-Approximation betrachten wir nun das Problem der "optimalen" Wahl der Stützstellen bei der Lagrange-Interpolation. Für das Lagrange-Interpolationspolynom $p_n \in P_n$ einer Funktion $f \in C^{n+1}[a, b]$ zu den Stützstellen $a \leq x_0 < x_1 < \dots < x_n \leq b$ gilt die Fehlerdarstellung

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^n (x - x_j).$$

Die Stützstellen sollen nun so gewählt werden, dass das Maximum

$$\max_{x \in [a, b]} \left| \prod_{j=0}^n (x - x_j) \right|$$

minimal wird. Aus der Darstellung

$$\prod_{j=0}^n (x - x_j) = x^{n+1} - p, \quad p \in P_n$$

sieht man, dass diese Aufgabe äquivalent dazu ist, die Tschebyscheff-Approximation von $f(x) = x^{n+1}$ bezüglich des Raumes $S = P_n$ zu bestimmen. Nach dem Alternantensatz ist der Betrag der Fehlerfunktion $e = x^{n+1} - p$ an mindestens $n + 2$ Stellen im Intervall $[a, b]$ gleich dem Minimum.

Theorem 5.2.5 *Auf dem Intervall $[a, b] = [-1, 1]$ ist die Tschebyscheff-Approximation $p \in P_n$ zu $f(x) = x^{n+1}$ gegeben durch*

$$p(x) = x^{n+1} - 2^{-n} T_{n+1}(x)$$

mit dem $(n + 1)$ -ten Tschebyscheff-Polynom

$$T_{n+1}(x) = \cos[(n + 1) \arccos(x)].$$

Die Nullstellen

$$x_k = \cos\left(\frac{\pi}{2} \cdot \frac{2k + 1}{n + 1}\right), \quad k = 0, \dots, n$$

von T_{n+1} sind die "optimalen" Stützstellen der Lagrange-Interpolation auf $[-1, 1]$.

Bemerkung. Die "optimale" Wahl der Stützstellen als Nullstellen des Tschebyscheff-Polynoms T_{n+1} bedeutet nicht, dass das zugeordnete Lagrangesche Interpolationspolynom, ein Polynom bester Approximation in der Maximumnorm ist, denn in der Fehlerdarstellung

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{2^n(n + 1)!} T_{n+1}(x), \quad x \in [-1, 1]$$

ist ξ_x noch von x abhängig. □

Beweis von Theorem 5.2.5. Das Polynom $T_{n+1} \in P_{n+1}$ hat $n + 1$ Nullstellen

$$x_k = \cos\left(\frac{\pi}{2} \cdot \frac{2k + 1}{n + 1}\right), \quad k = 0, \dots, n.$$

Aus der rekursiven Beziehung

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

folgt, dass $T_{n+1}(x) = 2^n x^{n+1} + q(x)$ mit einem geeigneten $q \in P_n$ ist. Damit haben wir die Darstellung

$$2^{-n} T_{n+1}(x) = \prod_{k=0}^n (x - x_k),$$

aus der insbesondere

$$\max_{-1 \leq x \leq 1} \prod_{k=0}^n |x - x_k| = \max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - x_k) \right| = 2^{-n} \max_{-1 \leq x \leq 1} |T_{n+1}(x)| = 2^{-n}$$

folgt. Nun nimmt $T_{n+1}(x) = \cos[(n + 1) \arccos(x)]$ im Intervall $[-1, +1]$ genau $n + 2$ mal einen Extremwert an, abwechselnd ± 1 . Diese $n + 2$ Extremalstellen

$$x_k^* = \cos \frac{k\pi}{n + 1}, \quad k = 0, \dots, n + 1,$$

bilden eine Alternante für die Approximation $p(x) = x^{n+1} - 2^{-n} T_{n+1}(x) \in P_n$ zu x^{n+1} . Nach dem Alternantensatz ist somit p die eindeutig bestimmte beste

Approximation zu x^{n+1} .

Es bleibt noch die Optimalitätseigenschaft

$$\max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - x_k) \right| = \min_{-1 \leq \zeta_0 < \zeta_1 < \dots < \zeta_n \leq 1} \max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - \zeta_k) \right|$$

zu zeigen. Mit den Nullstellen x_k von $T_{n+1}(x)$ gilt

$$\begin{aligned} \max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - x_k) \right| &= 2^{-n} \max_{-1 \leq x \leq 1} |T_{n+1}(x)| \\ &= \max_{-1 \leq x \leq 1} \left| x^{n+1} - [x^{n+1} - 2^{-n} T_{n+1}(x)] \right| = \min_{q \in P_n} \max_{-1 \leq x \leq 1} |x^{n+1} - q(x)|, \end{aligned}$$

also

$$\begin{aligned} \max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - x_k) \right| &= \min_{q \in P_n} \max_{-1 \leq x \leq 1} |x^{n+1} - q(x)| \\ &= \min_{-1 \leq \zeta_0 < \zeta_1 < \dots < \zeta_n \leq 1} \max_{-1 \leq x \leq 1} \left| \prod_{k=0}^n (x - \zeta_k) \right|. \end{aligned}$$

und damit die Optimalitätseigenschaft. \square

Kapitel 6

Nichtlineare Gleichungen

6.1 Nullstellen reeller Funktionen

Sei $f : [a, b] \rightarrow \mathbb{R}$ eine auf dem Intervall $[a, b]$ stetige Funktion. Das einfachste Verfahren zur Bestimmung von Nullstellen von f beruht auf dem Zwischenwertsatz für stetige Funktionen: *Gibt es ein Teilintervall $[a_0, b_0] \subset [a, b]$ mit $f(a_0)f(b_0) < 0$, so hat f in (a_0, b_0) mindestens eine Nullstelle.*

Das **Intervallschachtelungs-Verfahren** erzeugt ausgehend von einem Intervall $[a_0, b_0]$ mit $f(a_0)f(b_0) < 0$ eine Folge von Intervallen $[a_i, b_i]$, $i = 0, 1, \dots$ die mindestens eine Nullstelle von f besitzen durch die Vorschrift

$$\begin{aligned} x_i = \frac{1}{2}(a_i + b_i), \quad f(x_i) = 0 &\Rightarrow \text{STOP} \\ f(a_i)f(x_i) < 0 &\Rightarrow a_{i+1} = a_i, \quad b_{i+1} = x_i \\ f(a_i)f(x_i) > 0 &\Rightarrow a_{i+1} = x_i, \quad b_{i+1} = b_i. \end{aligned}$$

Im Fall, dass der Algorithmus nicht nach endlich vielen Schritten mit einer Nullstelle abbricht, gilt $a_i \leq a_{i+1} \leq b_{i+1} \leq b_i$ und

$$|b_{i+1} - a_{i+1}| = \frac{1}{2}|b_i - a_i| = 2^{-(i+1)}|b_0 - a_0|.$$

Die monotonen nach oben bzw. nach unten beschränkten Zahlenfolgen $(a_n)_{n \in \mathbb{N}}$ und $(b_n)_{n \in \mathbb{N}}$ konvergieren gegen ein $x_0 \in \mathbb{R}$, das wegen

$$f(x_0)^2 = \lim_{n \rightarrow \infty} f(a_n)f(b_n) \leq 0$$

eine Nullstelle von f ist. Hat man also ein Ausgangsintervall $[a_0, b_0] \subset [a, b]$ mit $f(a_0)f(b_0) < 0$ gefunden, so liefert das Verfahren für stetige Funktionen immer eine Nullstelle. Nachteil des Verfahrens ist die sehr langsame Konvergenz.

Ist die gegebene Funktion auf $[a, b]$ stetig differenzierbar, so kann diese Zusatzinformation zur effizienteren Berechnung einer Nullstelle verwendet werden. Das klassische **Newton-Verfahren** basiert auf Approximation der Funktion in Umgebung einer gesuchten Nullstelle durch das Taylorpolynom ersten Grades (graphisch durch die Tangente). Sei also x_n eine Näherung für die gesuchte Nullstelle x_0 . Die Gleichung $f(x) = 0$ wird nun ersetzt durch

$$f(x) \approx f(x_n) + f'(x_n)(x - x_n) = 0$$

mit der Lösung

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 1, 2, \dots$$

Die Iteration ist durchführbar, wenn die Ableitungswerte $f'(x_n)$ nicht zu klein werden.

Theorem 6.1.1 *Die Funktion $f \in C^2[a, b]$ habe im Innern von $[a, b]$ eine Nullstelle x_0 und es gelte*

$$m := \min_{a \leq x \leq b} |f'(x)| > 0, \quad M := \max_{a \leq x \leq b} |f''(x)| < \infty.$$

Sei der Radius ρ der abgeschlossenen Kugel $\overline{B}(x_0, \rho)$ klein genug gewählt, so dass

$$\overline{B}(x_0, \rho) \subset [a, b] \quad \text{und} \quad q := \frac{M}{2m}\rho < 1.$$

Dann ist für jeden Startwert $x_1 \in \overline{B}(x_0, \rho)$ die Newton-Iteration durchführbar und die Folge $(x_n)_{n \in \mathbb{N}}$ der Newton-Iterierten konvergiert gegen die Nullstelle x_0 . Darüber hinaus gelten die a-priori Fehlerabschätzung

$$|x_{n+1} - x_0| \leq \frac{2m}{M} q^{2^n} \quad n \in \mathbb{N}$$

sowie die a-posteriori Fehlerabschätzung

$$|x_{n+1} - x_0| \leq \frac{1}{m} |f(x_{n+1})| \leq \frac{M}{m} |x_{n+1} - x_n|^2.$$

Beweis. Wir beginnen mit einigen Vorbereitungen. Für $x, y \in [a, b]$ mit $x \neq y$ folgt aus dem Mittelwertsatz der Differentialrechnung mit einer Zwischenstelle $\xi \in (a, b)$

$$\left| \frac{f(x) - f(y)}{x - y} \right| = |f'(\xi)| \geq m > 0$$

und somit

$$|x - y| \leq \frac{1}{m} |f(x) - f(y)|. \quad (6.1.1)$$

Hieraus folgt insbesondere, dass die Nullstelle $x_0 \in [a, b]$ eindeutig bestimmt ist. Wir betrachten nun die Abbildung $P : \overline{B}(x_0, \rho) \rightarrow \mathbb{R}$, gegeben durch

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

mit deren Hilfe das Newton-Verfahren in der Form

$$x_{n+1} = P(x_n) \quad n = 1, 2, \dots$$

beschrieben werden kann. Mit Hilfe der Taylorschen Formel zur Entwicklung von $f(x_0)$ an der Entwicklungsstelle x folgt

$$0 = f(x_0) = f(x) + f'(x)(x_0 - x) + \frac{f''(\xi)}{2}(x_0 - x)^2,$$

wobei $\xi \in (a, b)$ wieder eine geeignete Zwischenstelle bezeichnet. Für $x \in \overline{B}(x_0, \rho)$ folgt

$$\begin{aligned} P(x) - x_0 &= x - \frac{f(x)}{f'(x)} - x_0 = -\frac{1}{f'(x)} \{f(x) + f'(x)(x_0 - x)\} \\ &= \frac{1}{f'(x)} \frac{f''(\xi)}{2}(x_0 - x)^2 \\ |P(x) - x_0| &\leq \frac{M}{2m} |x - x_0|^2 \leq \left(\frac{M}{2m}\rho\right) \rho < \rho, \end{aligned} \quad (6.1.2)$$

also liegt $P(x)$ in der Kugel $\overline{B}(x_0, \rho)$, vorausgesetzt dass bereits x in dieser Kugel lag, d.h. $x_n \in \overline{B}(x_0, \rho)$ für alle $n \in \mathbb{N}$. Setzt man $\rho_n = |x_n - x_0|M/(2m)$, so erhält man aus (6.1.2) für alle $n \in \mathbb{N}$

$$\rho_{n+1} = \frac{M}{2m} |x_{n+1} - x_0| = \frac{M}{2m} |P(x_{n+1}) - x_0| \leq \left(\frac{M}{2m} |x_n - x_0|\right)^2 = \rho_n^2.$$

Somit haben wir

$$\rho_{n+1} \leq (\rho_1)^{2^n} \quad \Rightarrow \quad |x_{n+1} - x_0| \leq \frac{2m}{M} (\rho_1)^{2^n},$$

woraus wegen $\rho_1 = |x_1 - x_0|M/(2m) \leq \rho M/(2m) = q < 1$ die Konvergenz von x_n gegen die Nullstelle x_0 folgt. Die letzte Abschätzung beinhaltet zugleich die im Satz behauptete a-priori Fehlerabschätzung. Zum Beweis der a-posteriori Fehlerabschätzung nutzt man die Taylorentwicklung von $f(x_{n+1})$ an der Entwicklungsstelle x_n :

$$f(x_{n+1}) = f(x_n) + f'(x_n)(x_{n+1} - x_n) + \frac{f''(\xi)}{2}(x_{n+1} - x_n)^2,$$

woraus unter Beachtung von $f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0$ und (6.1.1)

$$|x_{n+1} - x_n| \leq \frac{1}{m} |f(x_{n+1}) - f(x_n)| \leq \frac{M}{2m} |x_{n+1} - x_n|^2$$

folgt. □

Für zweimal stetig differenzierbare Funktionen f gibt es zu jeder einfachen Nullstelle x_0 , d.h. $f(x_0) = 0$ und $f'(x_0) \neq 0$ eine abgeschlossenen Kugel $\overline{B}(x_0, \rho)$, so dass die Voraussetzungen des Satzes erfüllt sind. Das Problem beim Newton-Verfahren ist die Bestimmung eines im *Einzugsbereich der Nullstelle* liegenden Startwertes x_1 .

Beispiel: Newton-Verfahren zur Wurzelberechnung

Die Quadratwurzel einer Zahl $a > 0$ ist die Nullstelle der Funktion $f(x) = x^2 - a$. Die Newton-Iteration ist gegeben durch

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right).$$

Wir untersuchen die Konvergenz der Folge und stellen zunächst fest, dass für einen beliebigen positiven Startwert $x_1 > 0$ die Positivität aller Folgenglieder x_{n+1} folgt. Die Ungleichung zum arithmetischen und geometrischen Mittel ergibt

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right) \geq \sqrt{a}$$

wobei das Gleichheitszeichen nur für $x_n = \sqrt{a}$ gilt. Startet man also mit einem positiven $x_1 \neq \sqrt{a}$, so ist $x_n > \sqrt{a}$ für alle $n \geq 2$. Eine derartige Folge $(x_n)_{n \geq 2}$ fällt streng monoton, denn

$$x_{n+1} - x_n = \frac{1}{2x_n} (a - x_n^2) < 0.$$

Als streng fallende Folge, die nach unten beschränkt ist, konvergiert $(x_n)_{n \geq 2}$ gegen g , wobei g der Gleichung

$$g = \frac{1}{2} \left(g + \frac{a}{g} \right) \quad \Rightarrow \quad g = \sqrt{a}$$

genügt. Für hinreichend große n ist x_n im Einzugsbereichs der Nullstelle $x_0 = \sqrt{a}$. Im konkreten Fall haben wir

$$x_{n+1} - \sqrt{a} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right) - \sqrt{a} = \frac{1}{2x_n} (x_n^2 + a - 2\sqrt{a}x_n) = \frac{1}{2x_n} (x_n - \sqrt{a})^2,$$

$$|x_{n+1} - \sqrt{a}| \leq \frac{1}{2\sqrt{a}} |x_n - \sqrt{a}|^2 \quad n \geq 2.$$

Quadratische Konvergenz liegt bereits ab dem zweiten Folgenglied vor. \square

Wir betrachten nun den Fall, dass eine mehrfache Nullstelle mit dem Newton-Verfahren berechnet werden soll. Wir beschränken uns auf den Fall einer zweifachen Nullstelle x_0 , d.h. $f(x_0) = f'(x_0) = 0$ und $f''(x_0) \neq 0$. Wir betrachten eine Umgebung von x_0 in der $|f''(x)| \geq m > 0$ gilt. Für die Newton-Iteration gilt nach dem Mittelwertsatz

$$x_{n+1} = x_n - \frac{f(x_n) - f(x_0)}{f'(x_n) - f'(x_0)} = x_n - \frac{f'(\xi_n)}{f''(\eta_n)}$$

für geeignete Zwischenpunkte ξ_n, η_n . Der Quotient $f(x_n)/f'(x_n)$ bleibt also für x_n gegen x_0 wohldefiniert. Hinsichtlich der quadratischen Konvergenz ist aber zu beachten, dass wegen (Nachweis durch partielle Integration!)

$$\begin{aligned} f(x) &= (x - x_0)^2 \int_0^1 (1-t) f''(x_0 + t(x - x_0)) dt = (x - x_0)^2 Q(x, x_0) \\ f'(x) &= 2(x - x_0) Q(x, x_0) + (x - x_0)^2 Q'(x, x_0) \end{aligned}$$

der Quotient $f(x)/f'(x)$ sich verhält wie

$$\begin{aligned} \frac{f(x)}{f'(x)} &= \frac{(x - x_0)^2 Q(x, x_0)}{2(x - x_0) Q(x, x_0) + (x - x_0)^2 Q'(x, x_0)} \\ &= \frac{(x - x_0)}{2} - (x - x_0)^2 \left\{ \frac{Q'(x, x_0)}{2Q(x, x_0) + (x - x_0) Q'(x, x_0)} \right\}, \end{aligned}$$

wobei

$$\lim_{x \rightarrow x_0} Q(x, x_0) = \frac{f''(x_0)}{2}, \quad \lim_{x \rightarrow x_0} Q'(x, x_0) = \frac{f'''(x_0)}{6}.$$

Quadratische Konvergenz kann durch die modifizierte Iteration

$$x_{n+1} = x_n - \frac{2f(x_n)}{f'(x_n)}, \quad n \geq 1,$$

erreicht werden. In der Tat gilt für die modifizierte Iteration

$$\begin{aligned} x_{n+1} - x_0 &= x_n - x_0 - \frac{2f(x_n)}{f'(x_n)} \\ &= (x - x_0)^2 \left\{ \frac{Q'(x, x_0)}{2Q(x, x_0) + (x - x_0) Q'(x, x_0)} \right\} \\ |x_{n+1} - x_0| &\leq C |x_n - x_0|^2. \end{aligned}$$

Das Newton-Verfahren benötigt in jedem Iterationsschritt die Auswertung der ersten Ableitung $f'(x_n)$, was bei komplizierten (möglicherweise nur implizit

durch ein Rechenprogramm definierter) Funktionen viel Aufwand erfordert. In derartigen Fällen geht man zum *vereinfachten Newton-Verfahren*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(s)}, \quad n \geq 1$$

mit einem fest gewählten Punkt s über. Das Verfahren ist ein Spezialfall des allgemeineren *Fixpunktverfahrens*, $\sigma \neq 0$,

$$x_{n+1} = x_n + \sigma f(x_n), \quad n \geq 1$$

zur Berechnung einer Nullstelle von f (oder äquivalent eines Fixpunktes der Abbildung $P(x) = x + \sigma f(x)$). Konvergenzaussagen folgen aus dem Banachschen Fixpunktsatz.

6.2 Konvergenzverhalten iterativer Verfahren

Das Newton-Verfahren zur Bestimmung einer einfachen Nullstelle besitzt lokal die Eigenschaft

$$|x_{n+1} - x_0| \leq C|x_n - x_0|^2.$$

Man nennt es deshalb *quadratisch* oder *von zweiter Ordnung* konvergent. Allgemein konvergiert ein Iterationsverfahren zur Berechnung einer Größe x_0 mit der *Ordnung* p , wenn es eine Konstante $C > 0$ gibt mit

$$|x_{n+1} - x_0| \leq C|x_n - x_0|^p.$$

Im Falle $p > 1$ impliziert die Abschätzung die Konvergenz des Verfahrens für Startwerte x_1 , die genügend nahe an x_0 liegen (analoge Argumentationskette, wie beim Newton-Verfahren!). Im Falle *linearer Konvergenz*, d.h. $p = 1$, heißt die bestmögliche Konstante *lineare Konvergenzrate* und wir haben Konvergenz im Falle $C < 1$:

$$|x_{n+1} - x_0| \leq C|x_n - x_0| \leq \dots \leq C^n|x_1 - x_0| \rightarrow 0 \quad (n \rightarrow \infty).$$

Gilt die Abschätzung

$$|x_{n+1} - x_0| \leq C_n|x_n - x_0|$$

mit einer Nullfolge $c_n \rightarrow 0$ ($n \rightarrow \infty$), so nennt man das Verfahren *superlinear konvergent*. Bei Fixpunktiterationen $x_{n+1} = P(x_n)$ mit stetig differenzierbarer Abbildung P gilt

$$\left| \frac{x_{n+1} - x_0}{x_n - x_0} \right| = \left| \frac{P(x_n) - P(x_0)}{x_n - x_0} \right| \rightarrow |P'(x_0)| \quad (n \rightarrow \infty),$$

die lineare Konvergenzrate ist asymptotisch also gerade $|P'(x_0)|$, für $|P'(x_0)| = 0$ liegt mindestens superlineare Konvergenz vor.

Theorem 6.2.1 Die Funktion P sei in Umgebung des Fixpunktes x_0 p -mal stetig differenzierbar mit $p \geq 2$. Die Fixpunktiteration $x_{n+1} = P(x_n)$ hat genau dann die Ordnung p , wenn

$$P'(x_0) = P''(x_0) = \dots = P^{(p-1)}(x_0) = 0 \quad \text{und} \quad P^{(p)}(x_0) \neq 0.$$

Beweis. Anwendung der Taylorschen Formel. □

In Anwendung des Satzes betrachten wir das Newton-Verfahren zur Bestimmung einer einfachen Nullstelle. Mit $P(x) = x - f(x)/f'(x)$ folgt im allgemeinen

$$P'(x_0) = 1 - \frac{f'(x_0)^2 - f(x_0)f''(x_0)}{f'(x_0)^2} = 0 \quad \text{und} \quad P''(x_0) = \frac{f''(x_0)}{f'(x_0)} \neq 0,$$

es ist also von zweiter Ordnung.

6.3 Interpolationsverfahren

Ziel dieser Klasse von Verfahren ist die Nullstellenbestimmung ohne Auswertung von Ableitungen, die jedoch effizienter als das Intervallschachtelungs-Verfahren bzw. die einfache sukzessive Approximation sind.

Die **Sekantenmethode** berechnet ausgehend von einem Wertepaar (x_{n-1}, x_n) die neue Iterierte als Nullstelle der Geraden durch die Punkte $(x_{n-1}, f(x_{n-1}))$ und $(x_n, f(x_n))$. Dies führt zur Iteration:

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad n \geq 1.$$

Bei diesem Verfahren handelt es sich um ein 2-Schrittverfahren, da zur Berechnung die beiden vorangegangenen Iterierten benötigt (und damit auch gespeichert) werden. Ähnlich wie für das Newton-Verfahren haben wir auch eine lokale Konvergenzaussage für die Sekantenmethode. Hierzu benötigt man die durch

$$\gamma_0 = \gamma_1 = 1, \quad \gamma_{n+1} = \gamma_n - \gamma_{n-1}, \quad n \geq 1$$

definierten *Fibonacci-Zahlen*.

Theorem 6.3.1 Die Funktion $f \in C^2[a, b]$ habe im Innern des Intervalls $[a, b]$ eine Nullstelle x_0 und es seien

$$m := \min_{x \in [a, b]} |f'(x)| > 0, \quad M := \max_{x \in [a, b]} |f''(x)| < \infty.$$

Sei ferner $\rho > 0$ klein genug gewählt, so dass $q := \rho M / (2m) < 1$ gilt und die abgeschlossene Kugel $\overline{B}(x_0, \rho)$ Teilmenge von $[a, b]$ ist. Dann sind für jedes Paar

von Startwerten $x_1, x_2 \in \overline{B}(x_0, \rho)$ mit $x_1 \neq x_2$ die Iterierten $x_n \in \overline{B}(x_0, \rho)$ der Sekantenmethode wohl definiert und konvergieren gegen die Nullstelle x_0 . Dabei gelten die a-priori Fehlerabschätzung

$$|x_{n+1} - x_0| \leq \frac{2m}{M} q^{\gamma^n}, \quad n \geq 1,$$

und die a-posteriori Fehlerabschätzung

$$|x_{n+1} - x_0| \leq \frac{1}{m} |f(x_{n+1})| \leq \frac{M}{2m} |x_{n+1} - x_n| |x_{n+1} - x_{n-1}|, \quad n \geq 1.$$

Beweis. Wie im Beweis zum Newton-Verfahren haben wir zunächst als Folgerung aus dem Mittelwertsatz für zwei Punkte $x, y \in [a, b]$ mit $x \neq y$

$$|x - y| \leq \frac{1}{m} |f(x) - f(y)|,$$

woraus die Eindeutigkeit der Nullstelle im Intervall $[a, b]$ folgt. Weiter ist

$$\frac{f(x) - f(y)}{x - y} = \int_0^1 f'(x + t(y - x)) dt.$$

Mit einem dritten Punkt $z \in [a, b]$, $z \neq x$, ergibt sich

$$\frac{f(x) - f(y)}{x - y} - \frac{f(x) - f(z)}{x - z} = \int_0^1 (f'(x + t(y - x)) - f'(x + t(z - x))) dt.$$

Wegen

$$f'(x + t(y - x)) - f'(x + t(z - x)) = (y - z) \int_0^t f''(x + t(y - x) + s(z - y)) ds$$

haben wir für $t \in (0, 1)$

$$|f'(x + t(y - x)) - f'(x + t(z - x))| \leq M|y - z|t,$$

woraus

$$\left| \frac{f(x) - f(y)}{x - y} - \frac{f(x) - f(z)}{x - z} \right| \leq \frac{M}{2} |y - z|$$

folgt. Für Punkte $x, y \in \overline{B}(x_0, \rho)$ mit $x \neq y$, $x \neq x_0$, $y \neq x_0$ definieren wir

$$P(x, y) := x - f(x) \frac{x - y}{f(x) - f(y)}.$$

Dann gilt

$$\begin{aligned}
P(x, y) - x_0 &= x - x_0 - f(x) \frac{x - y}{f(x) - f(y)} \\
&= \frac{x - y}{f(x) - f(y)} \left\{ (x - x_0) \frac{f(x) - f(y)}{x - y} - f(x) + f(x_0) \right\} \\
|P(x, y) - x_0| &\leq \frac{1}{m} |x - x_0| \left| \frac{f(x) - f(y)}{x - y} - \frac{f(x) - f(x_0)}{x - x_0} \right| \\
&\leq \frac{M}{2m} |x - x_0| |y - x_0| \leq \frac{M}{2m} \rho^2 < \rho.
\end{aligned}$$

Die Iterierten der Sekantenmethode x_n bleiben also in der Menge $\overline{B}(x_0, \rho)$ und es gilt die Abschätzung

$$|x_{n+1} - x_0| \leq \frac{M}{2m} |x_n - x_0| |x_{n-1} - x_0|.$$

Setzt man $\rho_n := |x_n - x_0| M / (2m)$, so folgt

$$\rho_{n+1} \leq \rho_n \rho_{n-1}, \quad n \geq 2,$$

d.h. mit $\rho_1 \leq q$, $\rho_2 \leq q$ gilt $\rho_{n+1} \leq q^{\gamma_n}$. Wegen $\gamma_n \rightarrow \infty$ ($n \rightarrow \infty$) und $q < 1$ konvergiert damit die Folge und wir haben die a-priori Abschätzung

$$|x_{n+1} - x_0| = \frac{2m}{M} \rho_{n+1} \leq \frac{2m}{M} q^{\gamma_n} \rightarrow 0 \quad (n \rightarrow \infty).$$

Zum Nachweis der a-posteriori Fehlerabschätzung setzen wir oben $x = x_n$, $y = x_{n+1}$ und $z = x_{n-1}$ und erhalten

$$\begin{aligned}
|x_{n+1} - x_0| &\leq \frac{1}{m} |f(x_{n+1}) - f(x_0)| = \frac{1}{m} |f(x_{n+1})| \\
&\leq \frac{1}{m} \left| f(x_n) + (x_{n+1} - x_n) \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n} \right| \\
&\leq \frac{1}{m} |x_{n+1} - x_n| \left| \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n} - \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \right| \\
&\leq \frac{M}{2m} |x_{n+1} - x_n| |x_n - x_{n-1}|
\end{aligned}$$

durch Ausnutzung der Bildungsvorschrift für x_{n+1} . □

Um die Konvergenzgeschwindigkeit der Sekantenmethode zu beurteilen, benötigen wir Informationen über das Anwachsen der Fibonacci-Zahlen γ_n für $n \rightarrow \infty$. Diese genügen der homogenen Differenzgleichung

$$\gamma_{n+2} - \gamma_{n+1} - \gamma_n = 0, \quad n \geq 0.$$

Zur Lösung machen wir den Ansatz $\gamma_n = \lambda^n$ und erhalten

$$\lambda^n(\lambda^2 - \lambda - 1) = 0$$

zur Bestimmung von λ . Mit den Wurzeln $\lambda_{1,2} = (1 \pm \sqrt{5})/2$ der quadratischen Gleichung $\lambda^2 - \lambda - 1 = 0$ lautet die allgemeine Lösung der Differenzengleichung

$$\gamma_n = c_1 \lambda_1^n + c_2 \lambda_2^n.$$

Die unbekanntenen Koeffizienten bestimmen wir aus den Forderungen $\gamma_0 = \gamma_1 = 1$. Die Lösung des zugeordneten Gleichungssystems und Einsetzen in die allgemeine Darstellung der Lösung ergibt

$$\gamma_n = \frac{1}{\sqrt{5}} (\lambda_1^{n+1} - \lambda_2^{n+1}) = \frac{\lambda_1^{n+1}}{\sqrt{5}} \left(1 - \left(\frac{\lambda_2}{\lambda_1} \right)^{n+1} \right) \sim \frac{\lambda_1^{n+1}}{\sqrt{5}} \quad (n \rightarrow \infty).$$

Die Sekantenmethode konvergiert also mindestens so schnell wie ein 1-Schrittverfahren der Ordnung $p = 1.6$. Fasst man jedoch zwei Schritte des Sekantenverfahrens zu einem Makro-Schritt zusammen, so erhält man wegen

$$|x_{2n+1} - x_0| \leq \frac{2m}{M} q^{2\gamma_n}, \quad \gamma_{2n} \sim \frac{1}{\sqrt{5}} \lambda_1^{2n} = \frac{1}{\sqrt{5}} (\lambda_1 + 1)^n$$

ein Verfahren der Ordnung $p \geq 2.6$. Da ein Schritt im Newton-Verfahren zwei Funktionsauswertungen erfordert und zwei Schritte des Sekantenverfahrens ebenso, ist das Sekantenverfahren asymptotisch bei gleichem Aufwand schneller. Dem theoretischen Vorteil steht aber die Gefahr von Auslöschungseffekten im Sekantenschritt bei monotoner Konvergenz von $f(x_n) \rightarrow 0$ gegenüber. Man stabilisiert die Methode deshalb auch mit der Intervallschachtelungsidee zur *regula falsi*:

Seien $a_n < b_n$ derart bestimmt, dass $f(a_n)f(b_n) < 0$, d.h. f hat eine Nullstelle $x_0 \in (a_n, b_n)$. Beim Sekantenschritt

$$x_n = a_n - f(a_n) \frac{a_n - b_n}{f(a_n) - f(b_n)}$$

tritt nun keine Auslöschung im Term $f(a_n) - f(b_n)$ auf, solange die Intervalllänge $b_n - a_n \gg \text{eps}$. Das neue Intervall $[a_{n+1}, b_{n+1}]$ bestimmt sich wie beim Intervallschachtelungsverfahren durch die Vorschrift:

$$\begin{aligned} f(x_n) = 0 &\Rightarrow \text{STOP} \\ f(a_n)f(x_n) < 0 &\Rightarrow a_{n+1} = a_n, \quad b_{n+1} = x_n \\ f(a_n)f(x_n) > 0 &\Rightarrow a_{n+1} = x_n, \quad b_{n+1} = b_n. \end{aligned}$$

Die höhere Stabilität der regula falsi gegenüber dem Sekantenverfahren wird durch eine geringere Konvergenzgeschwindigkeit erkauft.

Die Methode der *quadratischen Interpolation* ist eine Weiterentwicklung der regula falsi. Die neue Iterierte wird als Nullstelle des quadratischen Interpolationspolynoms zu den Knoten a_n , b_n und $(a_n + b_n)/2$ bestimmt. Man kann zeigen, dass es im Fall $f(a_n)f(b_n) < 0$ genau eine Nullstelle x_n innerhalb von (a_n, b_n) gibt, mit der dann wie beim Intervallschachtelungsverfahren weitergearbeitet wird.

6.4 Newton-Verfahren im \mathbb{R}^d

Wir betrachten das Newton-Verfahren zur Lösung nichtlinearer Gleichungssysteme der Form $f(x) = 0$ mit stetig differenzierbarer Abbildung $f : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}^d$. Formal erhalten wir das Newton-Verfahren durch Approximation der Gleichung $f(x) = 0$ durch das lineare Taylor-Polynom an einer bereits bekannten Näherungsstelle x_n :

$$f(x) \approx f(x_n) + f'(x_n)(x - x_n) = 0.$$

Damit lautet die Iterationsvorschrift

$$x_{n+1} = x_n - (f'(x_n))^{-1} f(x_n), \quad n \geq 1,$$

mit der Jacobi-Matrix $f'(x_n)$ von f an der Stelle x_n . Wir können mit dem Banachschen Fixpunktsatz lokale Konvergenz für Lösungen x_0 zeigen, für die die Jacobi-Matrix nicht verschwindet. Bei der Anwendung hat man vor allem mit zwei Problemen zu tun:

- (a) hoher Aufwand pro Iterationsschritt
(Lösung eines linearen Gleichungssystems) und
- (b) Wahl geeigneter Startwerte.

Zur Überwindung dieser Probleme verwendet man gegebenenfalls das *vereinfachte Newton-Verfahren*

$$x_{n+1} = x_n - (f'(s))^{-1} f(x_n), \quad n \geq 1,$$

mit einem geeigneten $s \in \mathbb{R}^d$, zum Beispiel $s = x_1$. In diesem Fall haben die in jedem Schritt zu lösenden Gleichungssysteme alle die gleiche Koeffizientenmatrix $f'(s)$ und mit Hilfe einer einmal berechneten *LR-Zerlegung* kann der Aufwand beträchtlich reduziert werden. Zur Vergrößerung des Konvergenzbereiches bietet sich das *gedämpfte Newton-Verfahren* an:

$$x_{n+1} = x_n - \omega_n (f'(x_n))^{-1} f(x_n), \quad n \geq 1,$$

bei dem $\omega_n \in (0, 1]$ zu Beginn zunächst klein gesetzt und im Verlaufe der Rechnung nach einer geeigneten Dämpfungsstrategie schrittweise bis auf $\omega_n = 1$ erhöht wird.

Kapitel 7

Lineare Gleichungssysteme II

In der Praxis kommen häufig dünn besetzte Bandmatrizen hoher Dimension $n \gg 10^6$ (z.B. bei der Diskretisierung von Differentialgleichungen) vor. Für derart große Gleichungssysteme ist das Gaußsche Eliminationsverfahren wenig geeignet, da durch das fill in bei der LR -Zerlegung ein sehr hoher Speicherplatzbedarf entsteht, der sogar die Größe des Kernspeichers des Rechners übertreffen kann. Die im folgenden betrachteten *iterativen Verfahren* benötigen zur näherungsweise Lösung des Gleichungssystems $Ax = b$ nicht viel mehr Speicherplatz, als zur Speicherung von A erforderlich ist. Außerdem benötigt eine Iteration deutlich weniger als $n^3/3$ arithmetische Operationen.

7.1 Einzelschritt- und Gesamtschrittverfahren

Wir betrachten das Gleichungssystem $Ax = b$, das ausgeschrieben unter Hervorhebung des Diagonalelementes

$$a_{ii}x_i + \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j = b_i, \quad i = 1, \dots, n$$

lautet. Im Falle $a_{ii} \neq 0$ erhalten wir

$$x_i = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j \right\}, \quad i = 1, \dots, n.$$

Das *Gesamtschritt-* oder *Jacobi-Verfahren* erzeugt nun aus einer Iterierten $x^{(k)}$ eine neue Iterierte $x^{(k+1)}$ durch die Vorschrift

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j^{(k)} \right\}, \quad i = 1, \dots, n.$$

Zum Zeitpunkt der Berechnung von $x_i^{(k+1)}$ sind die neuen Komponenten $x_l^{(k+1)}$ mit $l < i$ berechnet und könnten bereits in die Rechnung eingehen. Diese Idee führt zum *Einzelschritt-* oder *Gauß-Seidel-Verfahren*

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j < i}}^n a_{ij} x_j^{(k+1)} - \sum_{\substack{j=1 \\ j > i}}^n a_{ij} x_j^{(k)} \right\}, \quad i = 1, \dots, n.$$

Beide Iterationen $x^{(k)} \mapsto x^{(k+1)}$ können als Fixpunktiteration zur Lösung des Gleichungssystems $Ax = b$ aufgefaßt werden. Zur kompakten Schreibweise zerlegen wir A in $A = D + L + R$ mit

$$D = \begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{bmatrix} \quad L = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

Das Jacobi-Verfahren kann damit in der Form

$$x^{(k+1)} = D^{-1} \{ b - (L + R)x^{(k)} \} = -D^{-1}(L + R)x^{(k)} + D^{-1}b$$

geschrieben werden. Entsprechend gilt für das Gauß-Seidel-Verfahren

$$x^{(k+1)} = D^{-1} \{ b - Lx^{(k+1)} - Rx^{(k)} \} = -(D + L)^{-1}Rx^{(k)} + (D + L)^{-1}b.$$

Beide Verfahren lassen sich also in der Form

$$x^{(k+1)} = Bx^{(k)} + c$$

mit den *Iterationsmatrizen* $B = -D^{-1}(L + R)$ und $B = -(D + L)^{-1}R$ schreiben. Konvergiert die Folge, so konvergiert sie gegen einen Fixpunkt der Abbildung $P(x) = Bx + c$. Beim Jacobi- und Gauß-Seidel-Verfahren sind per Konstruktion Fixpunkte von P Lösungen des Gleichungssystems. Zur Konstruktion allgemeiner iterativer Verfahren dieses Typs wählt man eine einfach zu invertierende Matrix C und iteriert ausgehend von

$$Ax = b \quad \Leftrightarrow \quad Cx = Cx - Ax + b \quad \Leftrightarrow \quad x = x + C^{-1}(b - Ax)$$

in der Form

$$x^{(k+1)} = (I - C^{-1}A)x^{(k)} + C^{-1}b.$$

Für den Fall $C = I$ heißt die Iteration auch *Richardson-Verfahren*. Nach dem Banachschen Fixpunktsatz ist ein hinreichendes Kriterium für die Konvergenz dieser Iteration, dass

$$\|B\| = \|I - C^{-1}A\| < 1$$

für irgendeine Matrixnorm $\|\cdot\|$ auf $\mathbb{R}^{n \times n}$. Die Gültigkeit dieser Beziehung ist bei gegebener Iterationsmatrix B aber von der gewählten Matrixnorm abhängig. Zur Charakterisierung verwendet man daher besser den *Spektralradius der Iterationsmatrix* B

$$\text{spr}(B) = \max\{|\lambda| : \lambda \text{ Eigenwert von } B\}.$$

Aus

$$|\lambda|\|x\| = \|\lambda x\| = \|Bx\| \leq \|B\| \|x\|, \quad x \neq 0 \text{ Eigenvektor zu } \lambda,$$

folgt, dass für jede natürliche Matrixnorm $\|\cdot\|$

$$\text{spr}(B) \leq \|B\|,$$

für symmetrische Matrizen gilt sogar

$$\text{spr}(B) = \|B\|_2 = \sup_{x \in \mathbb{R}^{n \times n}} \frac{\|Bx\|_2}{\|x\|_2}.$$

Im allgemeinen ist jedoch der Spektralradius keine Norm, wie das Beispiel

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \neq 0 \quad \Rightarrow \quad \text{spr}(A) = 0$$

zeigt. Es gilt aber das folgende Resultat.

Theorem 7.1.1 *Für jede Matrix $B \in \mathbb{R}^{n \times n}$ gibt es zu jedem $\varepsilon > 0$ eine natürliche Matrixnorm, so dass*

$$\text{spr}(B) \leq \|B\|_\varepsilon \leq \text{spr}(B) + \varepsilon.$$

Beweis. Siehe z.B. J. Stoer, R. Bulirsch: Numerische Mathematik 2. Springer-Verlag 1990. \square

Theorem 7.1.2 *Die Matrizen $A \in \mathbb{R}^{n \times n}$ und $C \in \mathbb{R}^{n \times n}$ seien regulär. Die durch*

$$x^{(k+1)} = (I - C^{-1}A)x^{(k)} + C^{-1}b$$

erzeugten Iterierten $x^{(k)} \in \mathbb{R}^n$ konvergieren genau dann für jeden Startwert $x^{(1)} \in \mathbb{R}^n$ gegen die Lösung $x_0 \in \mathbb{R}^n$ des Gleichungssystems $Ax = b$, wenn für den Spektralradius der Iterationsmatrix $B = I - C^{-1}A$ die Beziehung $\text{spr}(B) < 1$ gilt.

Beweis. Wir bezeichnen die Fehlervektoren durch $e^{(k)} := x^{(k)} - x_0$. Da die Lösung $x_0 \in \mathbb{R}^n$ Fixpunkt von P mit $P(x) = Bx + c$ ist, folgt für die Folge der Fehlervektoren

$$e^{(k+1)} = x^{(k+1)} - x_0 = Bx^{(k)} + c - Bx_0 - c = B(x^{(k)} - x_0) = Be^{(k)}.$$

Damit haben wir $e^{(k+1)} = B^k e^{(1)}$. Im Fall $\text{spr}(B) < 1$ gibt es ein $\varepsilon > 0$ mit $\text{spr}(B) + \varepsilon < 1$ und eine natürliche Matrixnorm $\|\cdot\|_\varepsilon$, so dass

$$\|e^{(k+1)}\|_\varepsilon = \|B^k e^{(1)}\|_\varepsilon \leq \|B^k\|_\varepsilon \|e^{(1)}\|_\varepsilon \leq \|B\|_\varepsilon^k \|e^{(1)}\|_\varepsilon \rightarrow 0.$$

Da alle Normen im \mathbb{R}^n äquivalent sind, konvergiert die Folge $x^{(k)}$ gegen x_0 .

Sei umgekehrt die Iteration konvergent für jeden Startwert. Wir setzen als Startwert $x^{(1)} = w + x_0$, wobei $w \in \mathbb{R}^n \setminus \{0\}$ Eigenvektor zum betragsmäßig größten Eigenwert λ von B ist. Aus

$$\lambda^k w = B^k w = B^k(x^{(1)} - x_0) = e^{(k+1)} \rightarrow 0 \quad (k \rightarrow \infty)$$

folgt notwendig $|\lambda| < 1$ für jeden Eigenwert von B , d.h. $\text{spr}(B) < 1$. \square

Bei der Konstruktion von Iterationsverfahren, d.h. der Wahl einer geeigneten Matrix C , sind zwei Ziele zu berücksichtigen:

- (a) $\text{spr}(I - C^{-1}A)$ soll möglichst klein sein und
- (b) die Gleichungssysteme $Cx^{(k+1)} = (C - A)x^{(k)} + b$ sollen möglichst leicht und mit wenig zusätzlichem Speicherplatzbedarf lösbar sein.

Ein Extremfall ist $C = A$ mit $\text{spr}(I - C^{-1}A) = 0$, aber die Lösung der Gleichungssysteme unter (b) entspricht der Lösung des Ausgangssystems. Wir werden also gewisse Kompromisse eingehen müssen. Beim Jacobi- bzw. Gauß-Seidel-Verfahren lauten die Iterationsmatrizen $J = -D^{-1}(L + R)$ bzw. $H = -(D + L)^{-1}R$, es sind also nur Dreieckssysteme zu lösen. Hinsichtlich der Konvergenz gilt folgendes Resultat.

Theorem 7.1.3 *Die Matrix $A \in \mathbb{R}^{n \times n}$ sei strikt diagonal dominant, d.h.*

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

Dann ist $\text{spr} J < 1$ und $\text{spr} H < 1$, d.h. das Jacobi- und das Gauß-Seidel-Verfahren konvergieren.

Beweis. Sei v Eigenvektor zum Eigenwert λ der Iterationsmatrix J des Jacobi-Verfahrens, d.h.

$$\lambda v = Jv = -D^{-1}(L + R)v.$$

In der Maximumnorm $\|\cdot\|_\infty$ folgt für $\|v\|_\infty = 1$

$$\begin{aligned} |\lambda| &= \|\lambda v\|_\infty \leq \|D^{-1}(L + R)\|_\infty \|v\|_\infty = \|D^{-1}(L + R)\|_\infty \\ &= \max_{1 \leq i \leq n} \left\{ \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \right\} < 1. \end{aligned}$$

Ähnlich erhalten wir für den Eigenvektor w zum Eigenwert μ der Iterationsmatrix H des Gauß-Seidel-Verfahrens

$$\mu w = Hw = -(D + L)^{-1}Rw \quad \Leftrightarrow \quad \mu w = -D^{-1}(\mu L + R)w,$$

woraus für $\|w\|_\infty = 1$

$$\begin{aligned} |\mu| &= \|\mu w\|_\infty \leq \|D^{-1}(\mu L + R)\|_\infty \|w\|_\infty = \|D^{-1}(\mu L + R)\|_\infty \\ &= \max_{1 \leq i \leq n} \left\{ \frac{1}{|a_{ii}|} \left[\sum_{\substack{j=1 \\ j < i}}^n |\mu| |a_{ij}| + \sum_{\substack{j=1 \\ j > i}}^n |a_{ij}| \right] \right\} \end{aligned}$$

folgt. Angenommen $1 \leq |\mu|$, dann folgt der Widerspruch

$$|\mu| \leq |\mu| \|D^{-1}(L + R)\|_\infty < |\mu|.$$

Also kann nur $|\mu| < 1$ gelten. □

In der Praxis vorkommende große Matrizen besitzen leider Spektralradien nahe bei 1 und konvergieren daher viel zu langsam. Man versucht daher die Konvergenz durch Einführung von *Relaxationsparametern* zu beschleunigen. Beim *SOR-Verfahren* (Successive-OverRelaxation method) berechnet man im k -ten Iterationsschritt ausgehend vom Gauß-Seidel-Zwischenwert

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j < i}}^n a_{ij} x_j^{(k+1)} - \sum_{\substack{j=1 \\ j > i}}^n a_{ij} x_j^{(k)} \right\}, \quad i = 1, \dots, n$$

den nächsten Wert $x_i^{(k+1)}$ als Linearkombination

$$x_i^{(k+1)} = \omega \tilde{x}_i^{(k+1)} + (1 - \omega) x_i^{(k)}.$$

mit einem positiven Relaxationsparameter ω . Im Falle $\omega = 1$ erhalten wir gerade das Gauß-Seidel-Verfahren. Ist $\omega \in (0, 1)$ spricht man von Unterrelaxation, für $\omega > 1$ von Überrelaxation. Die Iterationsmatrix H_ω des Relaxationsverfahrens erhält man aus der Beziehung

$$x^{(k+1)} = \omega D^{-1} \{ b - Lx^{(k+1)} - Rx^{(k)} \} + (1 - \omega)x^{(k)}$$

in der Darstellung

$$H_\omega = (D + \omega L)^{-1} [(1 - \omega)D - \omega R].$$

Der folgende Satz zeigt, dass das SOR-Verfahren höchstens für Werte um $\omega = 1$ konvergieren kann.

Theorem 7.1.4 *Notwendig für die Konvergenz des SOR-Verfahrens ist die Bedingung $0 < \omega < 2$.*

Beweis. Seien $\lambda_1, \dots, \lambda_n$ die Eigenwerte der Iterationsmatrix

$$H_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega R].$$

Die Iterationsmatrix ist das Produkt zweier Dreiecksmatrizen, daher gilt

$$\begin{aligned} \det H_\omega &= \det [(D + \omega L)^{-1}] \cdot \det [(1 - \omega)D - \omega R] \\ &= \prod_{i=1}^n \frac{1}{a_{ii}} \cdot \prod_{j=1}^n (1 - \omega)a_{jj} = (1 - \omega)^n. \end{aligned}$$

Andererseits ist die Determinante einer Matrix das Produkt ihrer Eigenwerte, woraus

$$|1 - \omega|^n = |(1 - \omega)^n| = \left| \prod_{i=1}^n \lambda_i \right| = \prod_{i=1}^n |\lambda_i| \leq (\text{spr}(H_\omega))^n$$

folgt. Nun ist $|1 - \omega| < 1$ äquivalent zu $0 < \omega < 2$. □

Theorem 7.1.5 *Für eine symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n \times n}$ gilt*

$$\text{spr}(H_\omega) < 1 \quad \text{für } 0 < \omega < 2.$$

Beweis. Wegen der Symmetrie von A haben wir $R = L^T$, d.h.

$$A = L + D + L^T.$$

Sei λ Eigenwert von H_ω mit zugehörigem Eigenvektor $v \in \mathbb{R}^n$. Es gilt

$$H_\omega v = \lambda v \quad \Leftrightarrow \quad [(1 - \omega)D - \omega L^T] v = \lambda(D + \omega L)v$$

und damit auch

$$\omega(D + L^T)v = (1 - \lambda)Dv - \lambda\omega Lv.$$

Hieraus berechnen wir

$$\omega Av = \omega(D + L^T)v + \omega Lv = (1 - \lambda)Dv + (1 - \lambda)\omega Lv$$

und

$$\begin{aligned} \lambda\omega Av &= \lambda\omega(D + L^T)v + \lambda\omega Lv \\ &= \lambda\omega(D + L^T)v + (1 - \lambda)Dv - \omega(D + L^T)v \\ &= (\lambda - 1)\omega(D + L^T)v + (1 - \lambda)Dv \\ &= (1 - \lambda)(1 - \omega)Dv - (1 - \lambda)\omega L^T v. \end{aligned}$$

Multiplikation beider Gleichungen mit v^T und Addition ergibt

$$\begin{aligned}(1 + \lambda)\omega v^T A v &= (1 - \lambda)(2 - \omega)v^T D v + \omega(1 - \lambda)(v^T L v - v^T L^T v) \\ &= (1 - \lambda)(2 - \omega)v^T D v,\end{aligned}$$

woraus wegen der positiven Definitheit von A und D folgt, dass $\lambda \neq \pm 1$ und damit

$$\mu := \frac{1 + \lambda}{1 - \lambda} = \frac{2 - \omega}{\omega} \frac{v^T D v}{v^T A v} > 0.$$

Nach λ umgestellt, sehen wir $|\lambda| = |(\mu - 1)/(\mu + 1)| < 1$. \square

Die Frage nach optimalen Relaxationsparametern ist nicht leicht zu beantworten. Wir betrachten hier als Beispiel nur das Richardson-Verfahren

$$\tilde{x}^{(k+1)} = x^{(k)} + (b - Ax^{(k)}), \quad x^{(k+1)} = \omega \tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}$$

mit der Iterationsmatrix $B = I - \omega A$. Wir setzen A symmetrisch und positiv definit voraus. Dann ist auch die Iterationsmatrix symmetrisch. Ist λ Eigenwert von A mit zugeordnetem Eigenvektor $v \in \mathbb{R}^n$, so ist $1 - \omega\lambda$ Eigenwert von $I - \omega A$, denn

$$(I - \omega A)v = v - \omega\lambda v = (1 - \omega\lambda)v.$$

Für $\omega \neq 0$ gilt auch die Umkehrung. Der Spektralradius der Iterationsmatrix wird damit

$$\text{spr}(I - \omega A) = \max(|1 - \omega\lambda_{\min}(A)|, |1 - \omega\lambda_{\max}(A)|).$$

Minimierung von $\text{spr}(I - \omega A)$ über ω führt auf $\omega_{\text{opt}} = 2/(\lambda_{\min}(A) + \lambda_{\max}(A))$.

7.2 Abstiegsverfahren

Wir betrachten in diesem Abschnitt ausschließlich symmetrische, positiv definite Matrizen $A \in \mathbb{R}^{n \times n}$. Dann sind alle Eigenwerte von A positiv, d.h.

$$0 < \lambda_{\min} \leq \lambda_i \leq \lambda_{\max},$$

und durch

$$\|x\|_A := \sqrt{(Ax, x)}, \quad x \in \mathbb{R}^n$$

wird eine Norm definiert, die wir A -Norm nennen wollen.

Aus der Analysis wissen wir, dass die notwendige Bedingung für ein lokales Minimum einer skalaren Funktion mehrerer Veränderlicher $Q : \mathbb{R}^n \rightarrow \mathbb{R}$ (oder eines Funktionals) das Verschwinden der ersten partiellen Ableitungen ist. Wir

suchen nun ein Funktional $Q : \mathbb{R}^n \rightarrow \mathbb{R}$, so dass die notwendige Bedingung für Minima gerade der Aufgabe $Ax = b$ entspricht. Sei

$$Q(y) := \frac{1}{2}(Ay, y) - (b, y).$$

Theorem 7.2.1 *Sei A s.p.d. Dann ist x Lösung von $Ax = b$ genau dann, wenn*

$$Q(x) = \min_{y \in \mathbb{R}^n} Q(y).$$

Ferner ist das Minimum eindeutig bestimmt, also $Q(x) < Q(y) \quad \forall y \neq x$.

Beweis. Notwendig für ein Minimum von Q in $x \in \mathbb{R}^n$ ist das Verschwinden der Richtungsableitung in einer beliebigen Richtung $r \in \mathbb{R}^n$, $\|r\| = 1$, d.h.

$$\frac{\partial Q}{\partial r} = \lim_{t \rightarrow 0} \frac{Q(x + tr) - Q(x)}{t} = \frac{1}{2} \lim_{t \rightarrow 0} [(Ax, r) + (Ar, x) - 2(b, r) + t(Ar, r)] = 0.$$

Mit der Symmetrie der Matrix A folgt hieraus $(Ax - b, r) = 0$ für jede Richtung, also ist jede Minimalstelle von Q Lösung des Gleichungssystems. Sei nun x die Lösung des Gleichungssystems $Ax = b$. Dann gilt aufgrund der Symmetrie von A

$$\begin{aligned} \|y - x\|_A^2 &= (A(y - x), y - x) = (Ay, y) - (Ax, y) - (Ay, x) + (Ax, x) \\ &= (Ay, y) - 2(Ax, y) - (Ax, x) + 2(Ax, x) \\ &= (Ay, y) - 2(b, y) - \{(Ax, x) - 2(b, x)\} \\ &= 2Q(y) - 2Q(x). \end{aligned}$$

Somit ist $Q(y) > Q(x)$ für alle $y \neq x$. □

Die Richtungsableitung von Q (siehe Beweis oben) wird maximal für den Gradienten

$$\text{grad } Q(y) = Ay - b.$$

Sei $r^{(k)}$ eine zunächst beliebige Richtung. Wir wollen in Richtung von $r^{(k)}$ fortschreiten und suchen den bestmöglichen Abstieg, also $x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)}$, so dass (*line search*)

$$Q(x^{(k+1)}) = \min_{\alpha \in \mathbb{R}} Q(x^{(k)} + \alpha r^{(k)}).$$

Für das optimale α ergibt sich

$$\alpha_k = - \frac{(Ax^{(k)} - b, r^{(k)})}{(Ar^{(k)}, r^{(k)})},$$

das berechenbar ist, solange $r^{(k)} \neq 0$. Die Idee des Gradienten-Verfahrens besteht nun darin, als Richtung den steilsten Abstieg im Punkt $x^{(k)}$ zu wählen, also

$$r^{(k)} = - \text{grad } Q(x^{(k)}) = -(Ax^{(k)} - b),$$

und in dieser Richtung bestmöglich abzustiegen. Das **Gradienten-Verfahren** lautet demnach:

$$\begin{aligned}
 \text{Start:} \quad & x^{(1)}, \quad g^{(1)} := Ax^{(1)} - b \\
 \mathbf{k} \rightarrow \mathbf{k} + 1 : \quad & w^{(k)} = Ag^{(k)} \\
 & \alpha_k = -\frac{(g^{(k)}, g^{(k)})}{(w^{(k)}, g^{(k)})} \\
 & x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} \\
 & g^{(k+1)} = g^{(k)} - \alpha_k w^{(k)}
 \end{aligned}$$

Zur Durchführung des Verfahrens benötigt man also die drei Vektoren x , g und w . Man beachte, dass für $g^{(k)} \neq 0$ der Nenner in der Berechnung von α_k wegen

$$(w^{(k)}, g^{(k)}) = (Ag^{(k)}, g^{(k)}) > 0$$

nicht verschwindet. Hinsichtlich der Konvergenz des Verfahrens gilt

Theorem 7.2.2 Sei A s.p.d. und $\kappa := \text{cond}_2(A) = \lambda_{\max}/\lambda_{\min}$. Dann gilt

$$\|x^{(k+1)} - x\|_A \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x^{(1)} - x\|_A \quad k \geq 1,$$

d.h. das Gradientenverfahren konvergiert für beliebigen Startwert $x^{(1)}$ gegen die Lösung x des Gleichungssystems $Ax = b$.

Ist κ nahe bei Eins, die Eigenwerte also nahe beieinander, konvergiert das Gradientenverfahren ziemlich schnell. Für schlecht konditionierte Matrizen konvergiert im allgemeinen das Gradientenverfahren sehr langsam.

Das Verfahren der konjugierten Gradienten (conjugate gradient method) oder kurz **CG-Verfahren** von Hestenes und Stiefel (1952) zielt auf eine Verbesserung des Gradientenverfahrens durch Modifikation der Abstiegsrichtungen. Hierzu überlegen wir uns zunächst, wann ein Punkt x optimal bezüglich einer Suchrichtung d ist, d.h. wann

$$Q(x) = \min_{\alpha \in \mathbb{R}} Q(x + \alpha d).$$

Anwendung der Extremwerttheorie ergibt, dass x optimal bezüglich der Suchrichtung d genau dann ist, wenn $(g, d) = 0$ mit $g = Ax - b$ gilt. Wir sagen, dass x optimal bezüglich des Unterraums $B_k := \text{span}(d^{(1)}, \dots, d^{(k)})$ ist, wenn

$$Q(x) = \min_{y \in x + B_k} Q(y) = \min_{d \in B_k} Q(x + d).$$

Obige Argumentation zeigt, dass dies äquivalent zu

$$(g, d^{(i)}) = 0 \quad i = 1, \dots, k, \quad g = Ax - b,$$

ist. Sei x optimal bezüglich B_k . Wir suchen eine Richtung d , so dass $x' = x + d$ optimal bezüglich B_k bleibt. Die Beziehung

$$0 = (A(x + d) - b, d^{(i)}) = (Ax - b, d^{(i)}) + (Ad, d^{(i)}) = (Ad, d^{(i)})$$

zeigt, dass notwendig und hinreichend ist, dass d konjugiert (oder A -orthogonal) zu allen $d^{(i)}$ sein muss, also

$$(d, d^{(i)})_A = (Ad, d^{(i)}) = 0 \quad i = 1, \dots, k.$$

Die Idee des CG-Verfahrens besteht nun darin, anstelle des steilsten Abstiegs

$$-g^{(k+1)} = b - Ax^{(k+1)}$$

eine neue Richtung

$$d^{(k+1)} \perp_A B_k = \text{span} (d^{(1)}, \dots, d^{(k)})$$

zu wählen. Dazu setzen wir mit unbekanntem Koeffizienten $\beta_j^{(k+1)}$

$$d^{(k+1)} = -g^{(k+1)} + \sum_{j=1}^k \beta_j^{(k+1)} d^{(j)}$$

an und erhalten aus der Forderung der A -Orthogonalität von $d^{(k+1)}$ zu $d^{(i)}$

$$0 = (d^{(k+1)}, d^{(i)})_A = -(g^{(k+1)}, d^{(i)})_A + \beta_i^{(k+1)} (d^{(i)}, d^{(i)})_A$$

beziehungsweise

$$\beta_i^{(k+1)} = \frac{(g^{(k+1)}, Ad^{(i)})}{(d^{(i)}, Ad^{(i)})} \quad i = 1, \dots, k.$$

Es stellt sich heraus, dass bei dieser Wahl

$$B_k = \text{span} (d^{(1)}, \dots, d^{(k)}) = K_k(d^{(1)}; A) := \text{span} (d^{(1)}, Ad^{(1)}, \dots, A^{k-1}d^{(1)}),$$

wobei K_k Krylov-Raum genannt wird. Ferner gilt

$$\beta_i^{(k+1)} = 0 \quad i < k, \quad \beta_k^{(k+1)} \neq 0 \quad \text{falls } g^{(k+1)} \neq 0.$$

Wie beim Gradienten-Verfahren erhält man aus der neuen Suchrichtung $d^{(k)}$ die neue Iterierte $x^{(k+1)}$ sowie $g^{(k+1)} = Ax^{(k+1)} - b$. Die neue Iterierte ist dann optimal bezüglich $d^{(k+1)}$ und B_k , d.h. bezüglich B_{k+1} :

$$Q(x^{(k+1)}) = \min_{d \in B_{k+1}} Q(x^{(k+1)} + d) = \min_{y \in x^{(1)} + K_{k+1}} Q(y).$$

Das **CG-Verfahren** lautet damit:

$$\begin{array}{lll}
 \text{Start:} & x^{(1)}, & g^{(1)} = Ax^{(1)} - b, & d^{(1)} = -g^{(1)} \\
 \mathbf{k} \rightarrow \mathbf{k} + 1 : & & w^{(k)} = Ad^{(k)} & \\
 & & \gamma_k = (g^{(k)}, g^{(k)}) & \alpha_k = \gamma_k / (d^{(k)}, w^{(k)}) \\
 & & x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} & g^{(k+1)} = g^{(k)} + \alpha_k w^{(k)} \\
 & & \beta_k = (g^{(k+1)}, g^{(k+1)}) / \gamma_k & d^{(k+1)} = -g^{(k)} + \beta_k d^{(k)}
 \end{array}$$

Man braucht also 4 Vektoren x , g , d und w . Da die Richtungen $d^{(i)}$ paarweise A -orthogonal sind, sind sie linear unabhängig (solange das Verfahren (nicht abbricht); damit gilt spätestens nach n Schritten

$$\text{span}(d^{(1)}, \dots, d^{(n)}) = \mathbb{R}^n.$$

und bei exakter Arithmetik bricht das CG-Verfahren spätestens nach n Schritten mit der exakten Lösung ab. Hinsichtlich der Konvergenz gilt

Theorem 7.2.3 *Sei A s.p.d. und $\kappa := \text{cond}_2(A) = \lambda_{\max}/\lambda_{\min}$. Dann gilt*

$$\|x^{(k+1)} - x\|_A \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x^{(1)} - x\|_A \quad k \geq 1,$$

d.h. das CG-Verfahren konvergiert für beliebigen Startwert $x^{(1)}$ gegen die Lösung x des Gleichungssystems $Ax = b$.

Beweis. Deuffhard, Hohmann: Numerische Mathematik I, 1993

□