Prof. Dr. S. Sauter
Institut für Mathematik
Universität Zürich

Einführung in die Numerik

S. Sauter

Frühjahrsemester 2019

Version: 22. März 2019

Inhaltsverzeichnis

Als vorlesungsbegleitende Literatur empfehle ich die Bücher von Walter Gautschi (didaktisch sehr gut geschrieben) und Josef Stoer (sehr detailliert und über den Vorlesungsstoff auch hinausgehend).

Literatur

[1] W. Gautschi. *Numerical Analysis*. Birkhäuser, 1997.

[2] J. Stoer. *Numerische Mathematik*. Springer-Verlag, Heidelberg, 1989.

Diese Einführung in die Angewandte Mathematik behandelt die Entwicklung von Lösungsverfahren für mathematische Probleme für praktische Anwendungen. Sowohl numerische, algebraische und algorithmische Aspekte werden betrachtet.

Typischerweise gibt es mehrere Möglichkeiten, die Lösung einer mathematischen Fragestellung zu berechnen. Für die Bewertung der verschiedenen Lösungsalgorithmen spielen die folgenden Kriterien die wesentliche Rolle.

- Grösse der Problemklasse, für die das Lösungsverfahren anwendbar ist,
- Rechen- und Speicheraufwand in Abhängigkeit von der Genauigkeit (Effizienz),
- Implementierungsaufwand.

In der Regel ist es nicht möglich, die *exakte* Lösung eines mathematischen Problems zu berechnen, weil 1.) die Computerarithmetik mit Rundungsfehlern behaftet ist und 2.) im allgemeinen keine explizite endliche Darstellung der Lösung existiert und numerische Approximationen verwendet werden müssen. (Beispiel: Die Dezimaldarstellung der reellen Zahl $\sqrt{2}$ besitzt *unendliche* viele Stellen und kann daher auf dem Computer nicht explizit dargestellt werden.)

1 Computerarithmetik

Die Fragen, die in diesem Kapitel diskutiert werden, sind relevant in allen Bereichen der Mathematik, bei denen numerische Rechnungen auf dem Computer durchgeführt werden. Die Computerarithmetik unterscheidet sich von der üblichen mathematischen Arithmetik, weil auf Grund des beschränkten Rechenspeichers auf einem Computer nur eine *endliche* Menge von Zahlen zur Verfügung steht. Der Übergang von reellen Zahlen zu Zahlen auf dem Rechner nennt man Rundung. Die Kombination von elementaren arithmetischen Operationen $\circ \in \{+, -, \times, /\}$ auf dem Rechner führt typischerweise bei exakter Rechnung wieder auf reelle Zahlen, die auf dem Rechner nicht darstellbar sind und wiederum gerundet werden müssen. Das führt dazu, dass einige der Standardrechenregeln selbst für die elementaren arithmetischen Operationen nur eingeschränkt gelten.

1.1 Zahlendarstellung

In diesem Unterkapitel werden wir zunächst die reellen Zahlen einführen. Dies kann auf verschiedenste Weise gemacht werden, im Zusammenhang mit Computern hat sich allerdings die Darstellung im Dualsystem (Binärsystem) als vorteilhaft herausgestellt. Jede Zahl $x \in \mathbb{R}$ kann demnach in der Form

$$x = \pm(b_n 2^n + b_{n-1}2^{n-1} + \cdots + b_0 + b_{-1}2^{-1} + b_{-2}2^{-2} + \cdots) \tag{1.1}$$

geschrieben werden, wobei $n > 0$ eine natürliche Zahl ist und die b_i die sogenannten Binärkoeffizienten sind, für die

$$b_i = 0 \quad \text{oder} \quad b_i = 1$$

für alle i, gilt. Anstatt die Darstellung (1.1) werden wir im Folgenden die Kurzschreibweise

$$x = \pm(b_n b_{n-1} \cdots b_0 \,.\, b_{-1} b_{-2} \cdots)_2 \tag{1.2}$$

verwenden. Der Index 2 soll hier andeuten, dass die Zahl im Dualsystem vorliegt. Der Punkt in (1.2) wird Binärpunkt genannt und separiert den ganzzahligen Anteil links vom bruchzahligen Anteil rechts.

Aufgabe 1.1 *Zeigen Sie, dass die Darstellung (1.1) einer Zahl nicht eindeutig ist. Wie kann man Eindeutigkeit erreichen?*

Beispiel 1.2 *1.* $(10010.001)_2 = 1\cdot 2^4+0\cdot 2^3+0\cdot 2^2+1\cdot 2^1+0\cdot 2^0+0\cdot 2^{-1}+0\cdot 2^{-2}+1\cdot 2^{-3} = (18.125)_{10}$

2. $(0.01\bar{1})_2 = \sum_{k=2}^{\infty} 2^{-k} = -\frac{3}{2} + \sum_{k=0}^{\infty} 2^{-k} = -\frac{3}{2} + \frac{1}{1-1/2} = (0.5)_{10}$*, wobei hier verwendet wurde, dass*

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$$

falls $|r| < 1$ *(geometrische Reihe).*

3. $\frac{1}{5} = (0.2)_{10} = (.0011\overline{0011})_2$.

Das letzte Beispiel zeigt, dass eine reelle Zahl mit endlicher Dezimaldarstellung eine unendliche (nichttriviale) Binärdarstellung haben kann. Man kann deshalb nicht annehmen, dass eine Zahl mit endlicher Dezimaldarstellung exakt auf einem Binärcomputer darstellbar ist.

1.2 Zahlendarstellung auf Computern

Mit Hilfe der Binärdarstellung einer Zahl, lassen sich Gleitkommazahlen, so wie sie auf modernen Rechnern verwendet werden, einführen. Da auf Computern nur endlich viele Stellen einer Zahl gespeichert werden können, muss die maximale Anzahl an realisierbaren Stellen angegeben werden. Sei t die vom System erlaubte Anzahl von Stellen im bruchzahligen Teil der Zahl und s die zulässige Anzahl von Stellen im Exponenten. Wir bezeichnen die Menge der Gleitkommazahlen auf diesem Computer mit $\mathbb{R}(t,s)$, wobei

$$x \in \mathbb{R}(t,s) \quad \Longleftrightarrow \quad x = f\cdot 2^e \tag{1.3}$$

mit

$$f = \pm(.\, b_{-1}b_{-2}\cdots b_{-t})_2 \quad \text{und} \quad e = \pm(b_{s-1}b_{s-2}\cdots b_0\,.)_2.$$

Der bruchzahlige Teil f wird für gewöhnlich als Mantisse von x, die ganze Zahl e als Exponent von x bezeichnet. Die Zahl x in der Darstellung (1.3) heisst normalisiert, falls $b_{-1} = 1$. (Beachte: Die Zahl $x = 0$ besitzt keine normalisierte Darstellung).

Es ist offensichtlich, dass die Menge der Gleitkommazahlen $\mathbb{R}(t,s) \subset \mathbb{R}$ endlich ist. Ausserdem sind sie nicht gleichverteilt auf der reelen Achse.
Aus der Definition der (normalisierten) Gleitkommazahlen wird klar, dass die betragsmässig grösste und kleinste Zahl durch

$$\max_{x\in\mathbb{R}(t,s)} |x| = (1-2^{-t})2^{2^s-1}, \qquad \min_{x\in\mathbb{R}(t,s)} |x| = 2^{-2^s} \tag{1.4}$$

gegeben sind. Typische Werte für t und s sind $t = 23$ und $s = 7$. Damit ergibt sich für die betragsmässig grösste bzw. kleinste darstellbare Zahl $1.7\cdot 10^{38}$ bzw. $2.94\cdot 10^{-39}$. Jede Zahl, die

betragsmässig nicht in diesem Bereich liegt lässt sich auf einem solche Rechner nicht darstellen. Falls während einer Rechnung eine Zahl berechnet wird, die grösser als das Maximum in (1.4) ist, kommt es zu einem sogenannten *Überlauf*. Überläufe sind fatal und führen typischerweise zu einem sofortigen Abbruch der Berechnung. Falls während einer Rechnung eine Zahl produziert wird, die kleiner als das Minimum in (1.4) ist, kommt es zu einem sogenannten *Unterlauf*. Dies ist für gewöhnlich weniger folgenreich als ein Überlauf, da die entsprechende Zahl systemintern zu 0 gerundet wird. In bestimmten Fällen kann jedoch auch das zu falschen Resultaten führen.

1.3 Rundung

Wie oben beschrieben ist $\mathbb{R}(t,s)$ eine endliche, nicht gleichmässig verteilte Menge auf der reellen Achse. Da während einer Rechnung zwangsläufig auch Zahlen produziert werden, die nicht in dieser Menge liegen (sich also nicht in der Form (1.3) darstellen lassen), müssen solche Zahlen auf vom System darstellbare Zahlen abgebildet werden. Dies erfolgt durch Rundung. Um dies genauer zu erläutern, betrachten wir eine "exakte" reelle Zahl

$$x \in \mathbb{R}, \quad x = \pm \left(\sum_{k=1}^{\infty} b_{-k} 2^{-k} \right) 2^e.$$

Diese soll auf die Zahl

$$x^\star \in \mathbb{R}(t,s), \quad x^\star = \pm \left(\sum_{k=1}^{t} b^\star_{-k} 2^{-k} \right) 2^{e^\star}.$$

durch Rundung abgebildet werden. Eine Möglichkeit dies zu tun ist durch abschneiden:

$$x^\star = \mathrm{chop}(x), \quad \text{wobei} \quad e^\star = e, \quad b^\star_{-k} = b_{-k}$$

für $k = 1, 2, \ldots, t$. Der Rundungsfehler, der durch Abschneiden entsteht kann wie folgt abgeschätzt werden:

$$|x - x^\star| = |x - \mathrm{chop}(x)| = \left| \pm \sum_{k=t+1}^{\infty} b_{-k} 2^{-k} \right| 2^e \leq \sum_{k=t+1}^{\infty} 2^{-k} \cdot 2^e = 2^{-t} \cdot 2^e.$$

Die Grösse $|x - x^\star|$ ist der sogenannte *absolute Fehler*. Dieser Fehler hängt von der absoluten Grösse von x ab. Man geht deshalb oft zum sogenannten *relativen Fehler* $|x - x^\star|/|x|$, $x \neq 0$ über um diese Abhängigkeit zu eliminieren. Im Fall der Abschneideoperation kann dieser durch

$$\left| \frac{x - \mathrm{chop}(x)}{x} \right| \leq \frac{2^{-t} \cdot 2^e}{|\pm \left(\sum_{k=1}^{\infty} b_{-k} 2^{-k} \right) 2^e|} \leq \frac{2^{-t} \cdot 2^e}{|2^{-1} \cdot 2^e|} = 2 \cdot 2^{-t}$$

abgeschätzt werden.

Eine weitere Art des Rundens ist das symmetrische Runden. Dies entspricht dem gewöhnlich Auf- und Abrunden. Dies ist im Binärsystem aber einfacher, da es nur zwei Möglichkeiten gibt: Falls die $(t+1)$-ste Stelle eine 1 ist, wird aufgerundet, ansonsten abgerundet. Das symmetrische Runden lässt sich mit Hilfe der chop-Operation ausdrücken und lässt sich schreiben als:

$$x^\star = \mathrm{rd}(x), \quad \text{wobei} \quad \mathrm{rd}(x) = \mathrm{chop}\left(x + 2^{-(t+1)} \cdot 2^e \right).$$

Der relative Fehler für das symmetrische Runden erfüllt die etwas bessere Abschätzung

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq 2^{-t} \tag{1.5}$$

Die rechte Seite in (1.5) hängt vom verwendeten Computer ab und wird als relative Maschinengenauigkeit

$$\mathrm{eps} := 2^{-t}$$

bezeichnet. Sie bestimmt die Genauigkeit aller Gleitkommarechnungen auf dem zugehörigen Rechner. So ergibt sich beispielsweise für $t = 23$ ein Wert von $\mathrm{eps} = 1.19 \cdot 10^{-7}$, was 6 bis 7 signifikanten Dezimalstellen entspricht. In Matlab gilt beispielsweise $t = 53$, demnach erhält man für die relative Maschinengenauigkeit:

$$\mathrm{eps} = 2^{-53} \approx 1.11 \times 10^{-16},$$

d.h. 16 signifikante Stellen.
Im folgenden werden wir die Auswirkung von Rundungsfehlern auf elementare mathematische Operationen betrachten. Dazu ist es vorteilhaft mit Gleichheiten zu arbeiten und nicht mit Ungleichheiten. Wir schreiben deshalb

$$\mathrm{rd}(x) = x(1 + \varepsilon), \quad |\varepsilon| \leq \mathrm{eps}\,.$$

1.4 Rundungsfehler bei Gleitkommarechnungen

Alle vier Grundrechenarten, auf einem Computer ausgeführt, können als Resultat Zahlen produzieren, die nicht mehr vom Rechner darstellbar sind. Indem wir Über/Unterlauf vernachlässigen, sollten wir annehmen dass jede der Grundrechenarten $\circ \in \{+, -, \times, /\}$ ein korrekt gerundetes Resultat liefert. Für Fliesskommazahlen $x, y \in \mathbb{R}(t, s)$ bezeichnet $\mathrm{fl}(x \circ y)$ das Ergebnis auf dem Rechner und sollte

$$\mathrm{fl}(x \circ y) = x \circ y\,(1 + \varepsilon_{x \circ y})\,, \qquad |\varepsilon_{x \circ y}| \leq \mathrm{eps}$$

erfüllen. Im Folgenden werden wir den Rundungsfehlereinfluss bei den vier Grundrechenarten untersuchen.

Genauer nehmen wir an, dass die Aufgabe lautet, für zwei reelle Zahlen x, y die Operation $z = x \circ y$ zu berechnen. Wir nehmen weiter an, dass diese Operation auf dem Rechner wie folgt durchgeführt wird:

Input: $x, y \rightarrow$ Interne Darstellung $x^\star = x\,(1 + \varepsilon_x)$, $y^\star = y\,(1 + \varepsilon_y) \rightarrow$ Exakte Operation auf gerundeten Zahlen $z^\star := \mathrm{fl}(x \circ y) := x^\star \circ y^\star \rightarrow$ Output $z^\star$.

Da wir annehmen, dass $\varepsilon_x, \varepsilon_y$ sehr klein sind, vernachlässigen wir im folgenden quadratische Terme ε_x^2, ε_y^2, $\varepsilon_x \varepsilon_y$. Die Schreibweise

$$a \doteq b$$

bedeutet, dass $|a - b| \leq C\varepsilon^2$ gilt mit einer moderaten Konstanten C, die unabhängig von ε ist und $\varepsilon := \max\{|\varepsilon_x|, |\varepsilon_y|\}$.

Wie nennen die Operation $\circ$ gutartig, falls $z^\star \doteq z\,(1 + \varepsilon_{x \circ y})$ gilt mit $|\varepsilon_{x \circ y}| \leq C_1\,|\varepsilon_x| + C_2\,|\varepsilon_y|$ mit moderaten Konstanten C_1, C_2, die unabhängig von ε_x, ε_y sind.

a) *Multiplikation:* Mit den obigen Konventionen gilt

$$x^\star y^\star = xy\,(1+\varepsilon_x+\varepsilon_y+\varepsilon_x\varepsilon_y) \doteq xy\,(1+\varepsilon_x+\varepsilon_y)\,.$$

Der relative Fehler des Produkts ist dann durch $\varepsilon_{x\times y} = \varepsilon_x + \varepsilon_y$ gegeben. Das bedeutet, dass die Multiplikation gutartig ist.

b) *Division:* Sei $y \neq 0$ und ε_y hinreichend klein, z.B. $|\varepsilon_y| < 1/2$. Dann folgt mit $c = 1+\varepsilon_y$ die Beziehung

$$\frac{x^\star}{y^\star} = \frac{x}{y}\left(\frac{1+\varepsilon_x}{1+\varepsilon_y}\right) = \frac{x}{y}\,(1+\varepsilon_x)\left(1-\frac{\varepsilon_y}{c}\right) \doteq \frac{x}{y}\left(1+\varepsilon_x-c^{-1}\varepsilon_y\right).$$

Aus $|\varepsilon_y| < 1/2$ folgt $|c^{-1}| < 2$ so dass

$$\left|\varepsilon_{x/y}\right| \leq |\varepsilon_x| + 2\,|\varepsilon_y|$$

gilt. Daher ist die Division ebenfalls eine gutartige Operation.

c) *Addition und Subtraktion:* Durch Vorzeichenwechsel lässt sich der Fall der Subtraktion auf die Addition zurückführen. Es gilt

$$x^\star + y^\star = x\,(1+\varepsilon_x) + y\,(1+\varepsilon_y) = x+y+x\varepsilon_x+y\varepsilon_y = (x+y)\left(1+\frac{x\varepsilon_x+y\varepsilon_y}{x+y}\right),$$

wobei wir $x+y \neq 0$ annehmen. Daher gilt

$$\varepsilon_{x+y} = \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y.$$

Wie zuvor ist der Fehler eine Linearkombination aus den einzelnen relativen Fehlern, aber die Koeffizienten können betragsmässig beliebig gross werden. Falls x und y gleiches Vorzeichen haben sind beide Koeffizienten zwischen 0 und 1 und der Fehler verhält sich gutartig:

$$|\varepsilon_{x+y}| \leq |\varepsilon_x| + |\varepsilon_y|\,, \qquad xy \geq 0.$$

Die Addition zweier Zahlen mit gleichem Vorzeichen ist also eine gutartige Operation. Der einzig ungünstige Fall ist, wenn $x \approx -y$ gilt und $|x|$ gross ist. In diesem Fall spricht man von Auslöschung.

Bemerkung 1.3 *Die einzige rundungsfehlerkritische Situation tritt auf, falls zwei grosse und etwa gleichgrosse Zahlen subtrahiert werden. Dieses Phänomen heisst Auslöschung. In einigen Fällen kann das vermieden werden. Hier zeigt sich der fundamentale Unterschied zwischen idealisierter mathematischer und konkreter numerischer Berechnung.*

Beispiel 1.4 *Wir betrachten einen Computer, der 2 Dezimalstellen einer Zahl abspeichern kann und symmetrisch rundet. Die Identität* $(a-b)^2 = a^2 - 2ab + b^2$*, die zweifellos gilt wenn man in der Menge der reellen Zahlen rechnet, stimmt auf unserem Computer nicht mehr. Sei* $a = 1.8$ *und* $b = 1.7$ *dann gilt* $(a-b)^2 = 0.01$ *auf unserem Computer und mit exakter Arithmetik, jedoch gilt*

$$a^2 - 2ab + b^2 = 3.2 - 6.1 + 2.9 = 0$$

auf dem Computer. Hier tritt Auslöschung auf, da die Zahlen $a^2 + b^2$ *und* $2ab$ *intern auf eine Nachkommastelle gerundet werden und dann voneinander abgezogen werden.*

Beispiel 1.5 *Berechne $y = \sqrt{x+\delta} - \sqrt{x}$ mit $x > 0$ und sehr kleinem δ. Wie bereits analysiert können für grosses x und kleines δ ernsthafte Auslöschungseffekte auftreten. Dieses Problem lässt sich umformulieren. Die Darstellung*

$$y = \frac{\delta}{\sqrt{x+\delta} + \sqrt{x}}$$

ist mathematisch eine andere Darstellung des gleichen Ausdrucks. Hier treten keine Auslöschungseffekte auf.

Beispiel 1.5 zeigt, dass mathematisch-äquivalente Darstellungen eines Ausdrucks auf dem Computer unterschiedlich gutartig sein könnten, da die (ideale) mathematische Arithmetik nicht mehr gilt.

Übungsaufgabe 1.6 *Sei*

$$y := 333.75b^6 + a^2\left(11a^2b^2 - b^6 - 121b^4 - 2\right) + 5.5b^8 + \frac{a}{2b}.$$

a. *Schreiben Sie ein Programm, welches als Eingabe a und b benötigt und der Wert y ausgibt. Berechnen Sie y für $a = 77617$ und $b = 33096$. Verleichen Sie den Wert mit dem Ergebnis, welches ihr Taschenrechner liefert. Was stellen Sie fest und warum?*

b. *Auf wieviel Stellen genau müsste eine Arithmetik mindestens rechnen, um ein korrektes Resultat zu erreichen?*

2 Darstellung und Approximation von Funktionen

Konvention: In diesem Kapitel bezeichnet $I = [a, b] \subset \mathbb{R}$ immer ein reelles Intervall.

Beispiel 2.1 *Aus der Physik ist bekannt, dass ein physikalisches System bestrebt ist einen möglichst niedrigen Energiezustand anzunehmen. Dieser Zustand (beispielsweise die Geschwindigkeitsverteilung in einer Strömung) wird durch eine mathematische Funktion beschrieben, die als Minimum eines „Energiefunktionals" charakterisiert ist.*

Falls der physikalische Körper eindimensional modelliert werden kann (z.B. ein gerader Abschnitt eines Kanals ist), kann die Minimierungsaufgabe von folgender (mathematischer) Bauart sein: Sei $I = [0, 1]$. Finde eine differenzierbare Funktion $u : I \to \mathbb{R}$, welche das Funktional

$$J(u) = \int_0^1 \left(u'(x)^2 + u(x)^2\right) dx - 2\int_0^1 xu(x)\, dx$$

minimiert. Falls wir beispielsweise das Funktional J für die konstante Funktion $g = 1$ berechnen, ergibt sich

$$J(g) = \int_0^1 1dx - 2\int_0^1 xdx = 0$$

und für die lineare Funktion $h(x) = x$ ergibt sich

$$J(h) = \int_0^1 \left(1^2 + x^2\right) dx - 2\int_0^1 x^2 dx = 1 + \frac{1}{3} - 2 \times \frac{1}{3} = \frac{2}{3}.$$

Das heisst, dass für die beiden Funktion g und h die Ungleichung $J(g) < J(h)$ gilt. Die Aufgabe diejenige Funktion u zu konstruieren, für die das Funktional J den minimal-möglichen Wert annimmt, kann im allgemeinen nicht exakt gelöst werden und numerische Lösungsmethoden müssen zur Approximation eingesetzt werden.[1] *Das erfordert zunächst, effiziente Darstellungen von Funktionen zu finden, um danach mit diesen rechnen zu können.*

2.1 Darstellung von Funktionen

Funktionen $f : I \to \mathbb{R}$ sind charakterisiert durch die (überabzählbar) unendlich vielen Wertepaare

$$\{(x, f(x)) : x \in I\}.$$

Daher ist es nicht möglich, allgemeine Funktionen auf Computern darzustellen. Stattdessen beschränkt man sich auf Funktionen*systeme* beispielsweise auf die Menge aller Polynome von Maximalgrad n :

$$\mathbb{P}_n := \left\{a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \text{ mit reellen Koeffizienten } a_i, 0 \le i \le n\right\}.$$

Ein Polynom (kontinuierliche Funktion) $f \in \mathbb{P}_n$ ist dann eindeutig charakterisiert durch die Angabe der endlich vielen reellen Zahlen a_i, $0 \le i \le n$. Generell rechnet man auf dem Computer nicht mit Funktionen sondern mit *Koeffizienten*, welche Funktionen aus Funktionensysteme

[1] Im betrachteten Spezialfall lässt sich die Lösungs „ausnahmsweise" exakt berechnen und ist durch $u_{\min}(x) = x - \frac{\sinh\left(x-\frac{1}{2}\right)}{\cosh\frac{1}{2}}$ gegeben. Es gilt $J(u_{\min}) = -0.2575\ldots$.

charakterisieren. Da Funktionen im allgemeinen von unendlich vielen Werten abhängen, kann man nicht *alle* Funktion $f : I \to \mathbb{R}$ auf dem Computer durch endlich viele Koeffizienten charakterisieren. Man verwendet daher auf dem Computer Funktionensysteme (z.B. $\mathbb{P}_n$), um eine wesentlich grössere Menge von Funktionen zu *approximieren* und mit diesen Approximationen dann zu rechnen.

Polynome eignen sich in vielen Fällen sehr gut zur Approximation von Funktionen. Der Weierstrasssche Approximationssatz besagt, dass *jede* stetige Funktion auf I beliebig genau durch Polynome approximiert werden kann.

Wir werden uns daher im Folgenden mit einer effizienten Darstellungen von Polynomen beschäftigen.

Bemerkung 2.2 *Ein Polynom in der Darstellung*

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{2.1}$$

lässt sich in einem Punkt z mit $3n-1$ elementaren arithmetischen Operationen auswerten, wie der folgende Algorithmus zeigt.

$c_1 := z$; ***for*** $i = 2$ ***to*** n ***do*** $c_i = z \times c_{i-1}$;
$s := a_0$; ***for*** $i = 1$ ***to*** n ***do*** $s := s + a_i c_i$.

Die *Newtonsche Darstellung* mittels *dividierter Differenzen* ist für viele Anwendungen besonders geeignet. Eine dieser Anwendungen wollen wir zunächst beschreiben.

Sei $\Theta_n := \{x_0, x_1, \ldots, x_n\} \subset I$ eine Menge von $n+1$ verschiedenen Punkten, die alle im Intervall I enthalten sind und *Stützstellen* genannt werden. Die Werte einer Funktion seien in diesen Stützstellen gegeben:

$$f_i = f(x_i) \quad \forall 0 \leq i \leq n.$$

Ziel: Konstruiere ein Polynom p möglichst niedrigen Grades, welches die Funktion f in den Stützstellen *interpoliert*, d.h.

$$p(x_i) = f_i \quad \forall 0 \leq i \leq n \tag{2.2}$$

erfüllt.

Satz 2.3 *Seien (x_i, f_i), $0 \leq i \leq n$, eine gegebene Menge von Wertepaaren und die Stützstellen $x_i, 0 \leq i \leq n$, verschieden. Dann existiert genau ein Polynom $p \in \mathbb{P}_n$ mit der Eigenschaft (2.2).*

Beweis. Die Existenz dieses Polynoms wird in Satz 2.5 konstruktiv bewiesen.
Eindeutigkeit

Wir verwenden eine Folgerung aus dem Satz von Rolle. (Zur Erinnerung: Satz von Rolle: $f \in C^0([a,b])$ und $f \in C^1(]a,b[)$. Sei $f(a) = f(b)$. Dann existiert $\xi \in]a,b[$ mit $f'(\xi) = 0$.)

Seien also $p_1, p_2 \in \mathbb{P}_n$ zwei Polynome mit der Interpolationseigenschaft. Dann gilt für das Differenzpolynom $p = p_1 - p_2$

$$p \in \mathbb{P}_n \quad \text{und} \quad \forall 0 \leq i \leq n : p(x_i) = 0$$

Aus dem Satz von Rolle folgt, dass $p' \in \mathbb{P}_{n-1}$ n Nullstellen besitzt und induktiv, dass $p^{(n)} \in \mathbb{P}_0$ eine Nullstelle besitzt. Daher ist $p^{(n)} \equiv 0$ und $p^{(n-1)} \in \mathbb{P}_0$. Da $p^{(n-1)}$ zwei Nullstellen besitzt gilt $p^{(n-1)} \equiv 0$ und induktiv folgt $p \equiv 0$. ■

Das interpolierende Polynom hängt von der Stützstellenmenge Θ_n und der Funktion f ab, und wir schreiben $p(f, \Theta_n)$ statt p, um diese Abhängigkeit anzudeuten.

Wir stellen uns die Stützstellenmenge Θ_n sukzessive aufgebaut vor:

$$\Theta_0 := \{x_0\}, \quad \Theta_1 := \Theta_0 \cup \{x_1\}, \quad \Theta_2 := \Theta_1 \cup \{x_2\}, \ldots$$

und schreiben $p_j = p(f, \Theta_j)$ für das Polynom $p_j \in \mathbb{P}_j$, welches f in den Stützstellen $\Theta_j := \{x_k : 0 \leq k \leq j\}$ interpoliert. In vielen Anwendungen tritt nun die Aufgabe auf, *alle* Polynome p_j, $0 \leq j \leq n$, in einem Zwischenpunkt $z \in I$ auszuwerten.

Bemerkung 2.4 *Der Aufwand, alle Polynome p_j mit dem Algorithmus aus Bemerkung 2.2 in einem Punkt $z \in I$ auszuwerden, beträgt*

$$\sum_{j=1}^{n} (3j - 1) = \frac{n(3n+1)}{2} \tag{2.3}$$

elementare arithmetische Operationen. (Die Gleichheit in (2.3) folgt mit vollständiger Induktion.)

Die *Newtonsche Darstellung* erlaubt die Auswertung dieser Polynome in $O(n)$-Operationen.

Satz 2.5 *Für die Interpolationspolynome p_j, $0 \leq j \leq n$, gilt die Rekursion*

$$p_0(x) = b_0, \tag{2.4a}$$
$$p_n(x) = p_{n-1}(x) + b_n \omega_{n-1}(x), \qquad n = 1, 2, \ldots \tag{2.4b}$$

mit geeigneten Konstanten b_i, $0 \leq i \leq n$, und dem Stützstellenpolynom

$$\omega_{n-1}(x) = (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{2.5}$$

Beweis. Durch vollständige Induktion.

Anfang: $n = 0$:

Wir wählen $b_0 := f_0$. Dann ist $p_0 = f_0$ konstant und interpoliert offensichtlich den Wert (x_0, f_0).

Induktionsannahme: Aussage gelte für $0 \leq j \leq n - 1$.

Induktionsschluss: $n - 1 \to n$

Offensichtlich gilt $\omega_{n-1}(x_j) = 0$ für alle Stützstellen x_j, $0 \leq j \leq n - 1$, da ein Faktor in (2.5) dann gleich Null ist. Daraus folgt aus (2.4b) und der Induktionsannahme

$$p_n(x_j) = p_{n-1}(x_j) = f_j \quad \forall 0 \leq j \leq n - 1.$$

Wir müssen noch den Punkt x_n überprüfen. Einsetzen liefert[2]

$$p_n(x_n) = p_{n-1}(x_n) + b_n \omega_{n-1}(x_n) \stackrel{!}{=} f_n.$$

Da wegen (2.5) $\omega_{n-1}(x_n) \neq 0$, können wir dividieren und erhalten, dass für

$$b_n := \frac{f_n - p_{n-1}(x_n)}{\omega_{n-1}(x_n)} \tag{2.6}$$

[2] „$\stackrel{!}{=}$“ bedeutet: „soll gleich sein“ im Sinne einer Bedingung.

schliesslich auch $p_n(x_n) = f_n$ erfüllt ist. ■

Die Definition (2.6) zeigt, dass die Koeffizienten b_j von den Stützstellen x_i, $0 \leq i \leq j$, und der Funktion f abhängen. Wir drücken diese Abhängigkeit durch die Schreibweise aus

$$b_n = [x_0, x_1, \ldots, x_n] f, \qquad n = 0, 1, 2, \ldots$$

und nennen die rechte Seite die n-te *dividierte Differenz* von f bezüglich der Knotenpunkte $x_0, x_1, \ldots, x_n$. Für festes f ist dies eine Funktion von $n+1$ Variablen, den Stützstellen $x_i, 0 \leq i \leq n$.

Zentral für die Effizienz der Newtonschen Darstellung (2.4) ist die folgende Rekursionsformel für die Koeffizienten b_k

$$\begin{array}{ll} k = 0: & b_0 = [x_0] f := f_0, \\ k \geq 1: & b_k = [x_0, x_1, \ldots, x_k] f = \frac{[x_1, x_2, \ldots, x_k] f - [x_0, x_1, \ldots, x_{k-1}] f}{x_k - x_0}. \end{array} \tag{2.7}$$

Die k-te Differenz lässt sich also schreiben als Differenzenquotient der $(k-1)$-ten dividierten Differenzen.

Satz 2.6 *Sei $I = [a, b] \subset \mathbb{R}$ und eine Stützstellenmenge $\Theta_n = \{x_i : 0 \leq i \leq n\} \subset I$ gewählt. Für $f : C^0(I)$ besitzt $p_n(\Theta_n, f)$ die Darstellung*

$$p_n(\Theta_n, f) = \sum_{i=0}^{n} b_i \omega_{i-1},$$

wobei $\omega_{-1} \equiv 1$ und für $i \geq 1$, $\omega_{i-1} \in \mathbb{P}_i$ wieder das Stützstellenpolynom bezeichnet

$$\omega_{i-1}(x) := \textstyle\prod_{k=0}^{i-1} (x - x_k)$$

und die Koeffizienten rekursiv durch (2.7) gegeben sind.

Beweis. Wir setzen

$$\begin{aligned} r(x) &= p_{k-1}(f, \Theta_k')(x), \\ s(x) &= p_{k-1}(f, \Theta_{k-1})(x) \end{aligned}$$

mit $\Theta_k' = \Theta_k \setminus \{x_0\}$.

Zwischenbehauptung:
Es gilt:

$$p_k(f, \Theta_k)(x) = r(x) + \underbrace{\frac{x - x_k}{x_k - x_0} \underbrace{(r(x) - s(x))}_{=: d^{\mathrm{I}}(x)}}_{d^{\mathrm{II}}(x)}. \tag{2.8}$$

Beweis der Zwischenbehauptung:

Offensichtlich gilt $d^{\mathrm{I}}(x_j) = 0$ für alle $1 \leq j \leq k-1$. Zusätzlich verschwindet d^{II} auch für $x = x_k$. Daraus folgt $p_k(f, \Theta_k)(x_j) = r(x_j) = f_j$ für alle $1 \leq j \leq k$. Für x_0 rechnet man schliesslich nach

$$p_k(f, \Theta_k)(x_0) = r(x_0) - d^{\mathrm{I}}(x_0) = s(x_0) = f_0.$$

Wegen $p_k \in \mathbb{P}_k$, folgt daher, dass p_k das (eindeutig bestimmte) Interpolationspolynom bezüglich Θ_k ist, also (2.8) gilt. Damit ist die Hilfsbehauptung bewiesen.

Wir haben gezeigt, dass das Interpolationspolynom eindeutig durch die Werte in den Stützstellen festgelegt ist. Die Funktionen r und s lassen sich mittels der dividierten Differenzen gemäss

$$r = ([x_1, \ldots, x_k]\, f)\, x^{k-1} + \tilde{r}_{k-1}$$
$$s = ([x_0, \ldots, x_{k-1}]\, f)\, x^{k-1} + \tilde{s}_{k-1}$$

darstellen mit $\tilde{r}_{k-1}, \tilde{s}_{k-1} \in \mathbb{P}_{k-2}$. Damit besitzt p_k aus (2.8) die Darstellung

$$p_k(x) = \frac{x}{x_k - x_0} \left(([x_1, \ldots, x_k]\, f) - ([x_0, \ldots, x_{k-1}]\, f) \right) x^{k-1} + \tilde{p}_{k-1}$$

mit $\tilde{p}_{k-1} \in \mathbb{P}_{k-1}$. Andererseits ist der führende Koeffizient in p_k durch

$$[x_0, x_1, \ldots, x_k]\, f$$

gegeben und durch Koeffizientenvergleich ergibt sich (2.7). ■

Die Formel (2.7) kann verwendet werden, um das Schema der *dividierten Differenzen* zu erzeugen.

x	f			
x_0	f_0			
x_1	f_1	$[x_0, x_1]\, f$		
x_2	f_2	$[x_1, x_2]\, f$	$[x_0, x_1, x_2]\, f$	
x_3	f_3	$[x_2, x_3]\, f$	$[x_1, x_2, x_3]\, f$	$[x_0, x_1, x_2, x_3]\, f$
$\vdots$	$\vdots$	$\ldots$	$\ldots$	$\ldots$

Man beachte, dass gemäss (2.7) jeder Eintrag in obigem Schema aus dem linken Nachbar in der gleichen Zeile und dem linken Nachbar aus der darüberliegenden Zeile zu berechnen ist. Die dividierten Differenzen b_n, welche in Newton's Formel (2.4a) auftreten, sind genau die ersten $n+1$ Diagonaleinträge im Differenzenschema.

Das Hinzufügen eines zusätzlichen Wertepaares (x_{n+1}, f_{n+1}) erfordert lediglich das Erzeugen einer weiteren Zeile von links nach rechts in obigem Schema.

Beispiel 2.7 *Ziel ist es, die Wertepaare* $(0,3)$, $(1,4)$, $(2,7)$, $(4,19)$ *durch ein kubische Polynom (Polynom vom Grad 3) zu interpolieren. Das Berechnungsschema besitzt in diesem Fall die folgende Form:*

x	f			
0	3			
1	4	$(4-3)\,/\,(1-0) = 1$		
2	7	$(7-4)\,/\,(2-1) = 3$	$(3-1)\,/\,(2-0) = 1$	
4	19	$(19-7)\,/\,(4-2) = 6$	$(6-3)\,/\,(4-1) = 1$	$(1-1)\,/\,(4-0) = 0$

(2.9)

Die Koeffizienten sind daher durch $b_0 = 3$, $b_1 = 1$, $b_2 = 1$, $b_3 = 0$ *gegeben und das interpolierende Polynom durch*

$$p_3(f, \Theta_3)(x) = 3 + 1 \times (x-0) + 1 \times (x-0)(x-1) + 0 \times (x-0)(x-1)(x-2) = 3 + x^2.$$

Bemerkung 2.8 *Der Aufwand pro Kästchen im Schema (2.9) (rechts von den ersten beiden Spalten) beträgt* 3 *arithmetische Operationen. Insgesamt beträgt die Anzahl der Kästchen rechts von den ersten beiden Spalten*

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

und somit der Gesamtaufwand zur Berechnung der Koeffizienten b_j, $0 \leq j \leq n$, $3\frac{n(n+1)}{2}$ *arithmetische Operationen. Sind diese berechnet und gespeichert lassen sich alle Polynome* p_j, $0 \leq j \leq n$, *an einer Zwischenstelle mit dem folgenden Algorithmus auswerten*

$$\begin{aligned} &c_0 := z - x_0; \quad \textbf{\textit{for}}\ j = 1\ \textbf{\textit{to}}\ n-1\ \textbf{\textit{do}}\ c_j = (z - x_j)\, c_{j-1}; \\ &s_0 := f_0; \quad \textbf{\textit{for}}\ j = 1\ \textbf{\textit{to}}\ n\ \textbf{\textit{do}}\ s_j = s_{j-1} + b_j c_{j-1}. \end{aligned}$$

Die berechneten Zahlen s_j, $0 \leq j \leq n$, *sind dann genau die Auswertungen von* p_j *im Punkt* z.

Der Aufwand zur Auswertung aller Polynome beträgt daher lediglich $4n$ *Operationen.*

Man beachte, dass ***alle*** *Polynome mittels der Koeffizienten* b_j, $0 \leq j \leq n$, *abgespeichert werden können (Aufwand* $n+1$ *reelle Zahlen) im Gegensatz zur Speicherung der Polynome in der Darstellung (2.1), bei der für jedes Polynom* p_j *insgesamt* $j+1$ *reelle Koeffizienten abzuspeichern sind. Der Speicheraufwand würde dann quadratisch mit* n *wachsen:*

$$\sum_{j=0}^{n} (j+1) = \frac{(n+2)\,(n+1)}{2}.$$

2.2 Fehlerabschätzungen

In diesem Abschnitt werden wir uns damit beschäftigen, wie die Genauigkeit der Polynominterpolation von der Funktion f, der Wahl der Stützstellen und deren Anzahl abhängt.

Zunächst wird ein Mass für die Genauigkeit einer Funktionsapproximation eingeführt. Sei dazu wieder $I = [a, b] \subset \mathbb{R}$ mit $a < b$ ein reelles Intervall und $f : I \to \mathbb{R}$ eine Funktion.[3]

Definition 2.9 *Die Menge aller stetigen Funktionen auf* I *ist durch*

$$C^0(I) := \{f : I \to \mathbb{R} \mid f \text{ ist stetig in jedem Punkt } x \in I\}$$

gegeben. Für $k \in \mathbb{N}$ *ist die Menge aller* k*-mal stetig differenzierbaren Funktionen auf* I *gegeben durch*

$$C^k(I) := \left\{f : I \to \mathbb{R} \mid \forall 0 \leq i \leq k : f^{(i)} \text{ existiert und ist stetig}\right\}.$$

Hierbei bezeichnet $f^{(i)}$ *die* i*-te Ableitung von* f.

Die Mengen $C^k(I)$ sind Vektorräume und es lassen sich *Normen* darauf definieren. Die „Maximumsnorm“ auf $C^0(I)$ ist definiert für eine Funktion $f \in C^0(I)$ durch[4]

$$\|f\|_{\max} := \max_{a \leq x \leq b} |f(x)|.$$

[3] Wir setzen voraus, dass die Begriffe „Stetigkeit“ und „Differenzierbarkeit“ einer reellen Funktion bekannt sind.

[4] Aus der Analysis ist bekannt, dass das Maximum einer stetigen Funktion auf einem beschränkten, abgeschlossenen Intervall I immer existiert.

Man prüft leicht nach, dass die Normeigenschaften erfüllt sind:

$$\begin{array}{lll} \forall f \in C^0(I): & \|f\|_{\max} \geq 0 & \text{und} \quad \|f\|_{\max} = 0 \iff f = 0, \\ \forall f \in C^0(I), \forall \alpha \in \mathbb{R}: & \|\alpha f\|_{\max} = |\alpha| \, \|x\|_{\max} & \\ \forall f, g \in C^0(I): & \|f+g\|_{\max} \leq \|f\|_{\max} + \|g\|_{\max} & \text{(Dreiecksungleichung).} \end{array}$$

Beispiel 2.10

a. Sei $I = [-1, 2]$ *und* $f(x) = x^2$. *Dann gilt* $\|f\|_{\max} = 4$.

b. Sei $I = [0, 10]$ *und* $f(x) = \mathrm{e}^x$. *Dann gilt* $\|f\|_{\max} = \mathrm{e}^{10} \approx 2.2 \times 10^4$.

Mit Hilfe der Maximumsnorm lässt sich der „Abstand“ zweier Funktionen messen. Sei f gegeben und p eine Approximation von f. Dann gilt

$$\|f - p\|_{\max} = \max_{a \leq x \leq b} |f(x) - p(x)|.$$

Insbesondere folgt aus den Normeigenschaften, dass $\|f - p\|_{\max} = 0$ gilt, genau dann wenn $f = p$ ist.

Wir kommen nun zur Abschätzung des Interpolationsfehlers.

Bezeichnungen: $f \in C^0(I)$ ist eine gegebene Funktion. Die Menge der gegebenen Stützstellen wird wieder mit $\Theta_n := \{x_i : 0 \leq i \leq n\}$ bezeichnet. Das Interpolationspolynom, welches f in den Stützstellen Θ_n interpoliert, wird mit p_n bezeichnet. Dieses hängt von f und Θ_n ab.

Für die Fehlerdarstellung benötigen wir das „Stützstellenpolynom“ ω_n, welches nur von Θ_n abhängt und gegeben ist durch

$$\omega_n(x) := (x - x_0)(x - x_1) \cdots (x - x_n).$$

Es ist ein Polynom vom Grad $n + 1$ und besitzt Nullstellen genau in den Stützstellen.

Satz 2.11 *Sei* $f \in C^{n+1}(I)$ *mit* $I = [a, b] \subset \mathbb{R}$ *und* $\Theta_n \subset I$. *Dann gilt die Fehlerdarstellung*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_n(x), \qquad x \in [a, b] \tag{2.10}$$

an einer geeigneten Zwischenstelle $\xi \in [a, b]$, *die im allgemeinen von* x, *den Stützstellen und von* f *abhängt.*

Beweis. Für $x \in \Theta_n$ verschwinden beide Seiten der Gleichung und die Behauptung gilt.

Sei $x \notin \Theta_n$ und $p = p(f, \Theta_n)$. Wir definieren eine Hilfsfunktion:

$$F_x(t) = f(t) - p(t) - \frac{f(x) - p(x)}{\omega_n(x)} \omega_n(t).$$

(Man beachte, dass der Term $\frac{f(x)-p(x)}{\omega_n(x)} \omega_n(t)$ den Fehler interpoliert.) Offensichtlich gilt $F_x \in C^{n+1}[a, b]$ und

$$\begin{aligned} F_x(t) &= 0, \qquad \forall t \in \Theta_n, \\ F_x(x) &= 0. \end{aligned}$$

Daraus folgt, dass F_x $n+2$ verschiedene Nullstellen in $[a,b]$ besitzt. Indem wir den Satz von Rolle anwenden folgt:

F_x' hat mindestens $n+1$ verschiedene Nullstellen,

F_x'' hat mindestens n verschiedene Nullstellen

und induktiv:

$F_x^{(n+1)}$ hat mindestens eine Nullstelle und –wegen der vorausgesetzten Glattheit von f– ist $F_x^{(n+1)}$ stetig. Eine Nullstelle von $F_x^{(n+1)}$ sei mit $\xi = \xi(x)$ bezeichnet. Indem F_x $(n+1)$-mal differenziert wird und $t = \xi$ gesetzt wird, erhält man

$$F_x^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{\omega_n(x)}(n+1)!.$$

(Hier wurde benutzt, dass $\omega_n(t) = t^{n+1} + p_n$ gilt mit einem Polynom $p_n \in \mathbb{P}_n$.) Indem diese Gleichung nach dem Fehler $f - p$ aufgelöst wird, erhält man die behauptete Fehlerdarstellung. ■

Korollar 2.12 *Der Beweis des vorigen Satzes hat gezeigt, dass die Zwischenstelle ξ im kleinsten Intervall liegt, welches die Interpolationspunkte Θ_n enthält.*

Da ξ in (2.10) in $[a,b]$ enthalten ist, gilt

$$\left|f^{(n+1)}(\xi)\right| \leq \max_{a\leq x\leq b}\left|f^{(n+1)}(x)\right| = \left\|f^{(n+1)}\right\|_{\max} =: C_n \tag{2.11}$$

und für alle $x \in [a,b]$

$$|\omega_n(x)| \leq \max_{a\leq x\leq b}|\omega_n(x)| = \|\omega_n\|_{\max} =: M(\Theta_n). \tag{2.12}$$

Korollar 2.13 *Aus Satz 2.11 folgt*

$$\|f - p_n\|_{\max} \leq \frac{C_n}{(n+1)!}M(\Theta_n). \tag{2.13}$$

Bemerkung 2.14

a. *Die Grösse $M(\Theta_n)$ hängt nur von der Wahl der Stützstellen ab und nicht von der Funktion f.*

b. *Umgekehrt hängt C_n nicht von der Wahl der Stützstellen ab, sondern lediglich von deren Anzahl und von der Funktion f.*

c. *Aus Korollar 2.13 folgt: Eine Folge von Polynominterpolationen p_n, $n \in \mathbb{N}$, konvergiert gegen f bezüglich der Maximumsnorm, falls*

$$\frac{C_n}{(n+1)!}M(\Theta_n) \overset{n\to\infty}{\to} 0$$

gilt. Man beachte, dass diese Bedingung lediglich hinreichend, aber nicht unbedingt notwendig ist.

Beispiel 2.15 *Wir betrachten die lineare Interpolation mit zwei Stützstellen* $x_0 = a$ *und* $x_1 = b$*, d.h.*

$$p_1(x) = f(a) + (x-a)\frac{f(b)-f(a)}{b-a}.$$

Es gilt die Fehlerdarstellung

$$f(x) - p_1(x) = (x-a)(x-b)\frac{f''(\xi)}{2}, \qquad \text{für ein } \xi \in [a,b]$$

und die Fehlerabschätzung

$$\|f - p_1\|_{\max} \leq \frac{C_1}{8}(b-a)^2.$$

(Hier wurde benutzt, dass für $x \in [a,b]$ *gilt* $|\omega_1(x)| = (x-a)(b-x)$ *und dieser Ausdruck maximal wird für* $x = (a+b)/2$*. Der Maximalwert ist* $\frac{1}{4}(b-a)^2$*.)*

Bemerkung 2.16 *Für* $x \in I$ *gilt* $|x - x_i| \leq (b-a)$*. Daraus folgt*

$$M(\Theta_n) \leq (b-a)^{n+1},$$

und das ergibt mit (2.13) die Abschätzung

$$\|f - p_n\|_{\max} \leq C_n \frac{(b-a)^{n+1}}{(n+1)!}. \tag{2.14}$$

Die Abschätzung der Konstante C_n ist in vielen Fällen nicht einfach. Um die Konvergenz für $n \to \infty$ betrachten zu können, folgt aus der Definition (2.11) sofort, dass wir $f \in C^\infty(I)$ fordern müssen (d.h. $f \in C^k(I)$ für alle $k \in \mathbb{N}$). Das folgende Beispiel zeigt jedoch, dass diese Bedingung nicht *hinreichend* ist für die Konvergenz.

Beispiel 2.17 *Wir betrachten die Funktion* $f(x) = \frac{1}{x+1}$ *in einem Intervall* $[0,b]$ *für ein* $b > 0$*. Dann gilt*

$$f^{(n)}(x) = (-1)^n \frac{n!}{(x+1)^{n+1}}$$

und für die Konstante C_n

$$C_n = \left\|f^{(n+1)}\right\|_{\max} = \max_{0\leq x\leq b}\left|f^{(n+1)}(x)\right| = \max_{0\leq x\leq b}\left|(-1)^{n+1}\frac{(n+1)!}{(x+1)^{n+2}}\right| = (n+1)! \quad .$$

Eingesetzt in (2.14) ergibt sich die Interpolationsfehlerabschätzung (beachte $a = 0$*)*

$$\|f - p_n\|_{\max} \leq b^{n+1}.$$

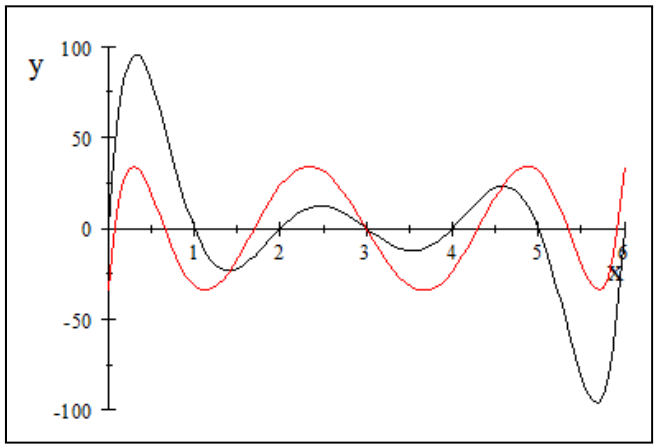
Die Interpolation konvergiert sehr schnell für $0 < b < 1$ *aber divergiert, falls* $b > 1$ *zu gross ist.*

Die Wahl der Stützstellen fliesst lediglich in die Konstante $M(\Theta_n)$ ein. Das folgende Beispiel zeigt das charakteristische Verhalten des Stützstellenpolynoms für äquidistante Stützstellen.

Beispiel 2.18 (Interpolation vom Grad n) *Sei $I = [a,b]$, $n \in \mathbb{N}$ und $f \in C^{n+1}(I)$. Äquidistante Gitterpunkte auf dem Intervall I sind durch $x_i = a + ih$ mit $h := (b-a)/n$ für $i = 0, 1, \ldots, n$ gegeben. Das Stützstellenpolynom besitzt die Form*

$$\omega_n(x) = (x-a)(x-a-h)(x-a-2h)\cdots(x-b).$$

Dies ist eine hochoszillierende Funktion (siehe Abbildung); die maximalen Funktionswerte und Steigungen sind an den Intervallenden.



Das Maximum des Stuetzstellenpolynoms (rot) fuer die Cebysev-Stuetzstellen ist deutlicher kleiner verglichen zum Maximum des Stuetzstellenpolynoms (schwarz) fuer aequidistante Stuetzstellen.

Die Frage, ob durch eine geschicktere Wahl der $n+1$ Stützstellen die Konstante $M(\Theta_n)$ verkleinert werden kann, lässt sich positiv beantworten. Die Čebyšev-Stützstellen (vgl. Übungsaufgabe) sind durch

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

gegeben. Die zugehörige Stützstellenmenge $\Theta_n := \{x_i : 0 \le i \le n\}$ minimiert die Konstante $M(\Theta_n)$ in (2.13).

2.3 Adaptive Verfeinerung

In diesem Abschnitt werden wir uns mit der folgenden Problemstellung beschäftigen.

Sei $I = [a,b] \subset \mathbb{R}$ ein Intervall. Eine Funktion $f : I \to \mathbb{R}$ beschreibt ein physikalisches Verhalten und kann in Punkten $x \in [a,b]$ gemessen werden. Die Messungen seien "teuer", und daher ist es das Ziel, die Funktion f rekursiv mit Polynomen $p_n \in \mathbb{P}_n$, $n \in \mathbb{N}_0$, zu approximieren und die Stützstellen "möglichst problemangepasst" zu wählen.

Im Gegensatz zu den vorigen Abschnitten bezeichnet Θ_ℓ nun die Stützstellenmenge der Stufe ℓ, die aber aus mehr als $\ell+1$ Stützstellen bestehen darf: $\Theta_\ell := \{x_i : 0 \le i \le n_\ell\}$ mit streng monoton wachsenden Zahlen n_ℓ.

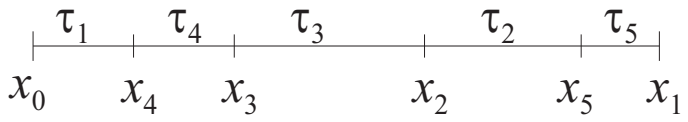


Abbildung 1: Schrittweise Unterteilung des Intervalls $[a, b]$ mit $x_0 = a$ und $x_1 = b$. Man beachte, dass die Gitterpunkte und die Teilintervalle τ_i im allgemeinen nicht angeordnet numeriert sind.

Die verwendeten Stützstellenmengen Θ_ℓ, $\ell \in \mathbb{N}_0$, werden schrittweise konstruiert. Lediglich Θ_0 und $\Theta_1 = \Theta_0 \cup \Theta_0^+$ müssen vordefiniert werden. Die anderen Mengen seien durch die Rekursion

$$\text{für} \quad \ell = 0, 1, \ldots : \qquad \Theta_{\ell+1} = \Theta_\ell \cup \Theta_\ell^+$$

gegeben. Die Wahl der jeweils neuen Stützstellen, d.h., Θ_ℓ^+, geschieht "adaptiv" mit Hilfe eines "a-posteriori Fehlerindikators", der rekursiv die Kenntnis der bereits berechneten Polynome $p_{\ell-1}, p_{\ell-2}$ verwendet und im Folgenden hergeleitet wird.

Der Einfachheit halber wählen wir immer $x_0 = a$ und $x_1 = b$. Die Stützstellenmenge Θ_ℓ, $\ell \geq 1$, zerlegt dann das Intervall $[a, b]$ in (abgeschlossene) Teilintervalle, die wir im Gitter

$$\mathcal{G}_\ell := \{\tau_{i,\ell} : 1 \leq i \leq n_\ell\}$$

zusammenfassen (siehe Abb. 1). Für zwei beliebige, verschiedene Teilintervalle $\tau_{i,\ell}, \tau_{j,\ell} \in \mathcal{G}_\ell$ ist der Schnitt $\tau_{i,\ell} \cap \tau_{j,\ell}$ dann entweder leer oder eine gemeinsame Stützstelle.

Wahl der neuen Stützstellen:

Seien für $n \geq 2$ die Stützstellenmengen Θ_ℓ und zugehörige Gitter $\mathcal{G}_\ell$ für $1 \leq \ell \leq n$ bereits erzeugt. Bezeichne p_ℓ das Interpolationspolynom zum Gitter Θ_ℓ. Wir berechnen für $\tau \in \mathcal{G}_n$ die Differenzen

$$d_\tau := |p_n(M_\tau) - p_{n-1}(M_\tau)| \tag{2.15}$$

in den Mittelpunkten M_τ der Intervalle $\tau \in \mathcal{G}_n$ und bestimmen

$$d_{\max} = \max_{\tau \in \mathcal{G}_n} d_\tau.$$

Für einen festgewählten Steuerungsparameter $\alpha \in [0, 1]$ definieren wir die neuen Gitterpunkte gemäss

$$\Theta_n^+ := \{M_\tau : \tau \in \mathcal{G}_n \wedge d_\tau \geq \alpha d_{\max}\}.$$

Dann ist die neue Stützstellenmenge durch $\Theta_{n+1} = \Theta_n \cup \Theta_n^+$ gegeben und das zugehörige Gitter wird mit $\mathcal{G}_{n+1}$ bezeichnet.

Bemerkung 2.19 *Die Polynome p_j werden am günstigsten in der Newtonsche Darstellung aus dem vorigen Abschnitt gespeichert. Im Schritt $\ell \to \ell + 1$ müssen dann lediglich die dividierten Differenzen b_j, $j = n_\ell + 1, \ldots, n_{\ell+1}$ berechnet werden (vgl. (2.7)), d.h. $n_{\ell+1} - n_\ell$ neue Zeilen im Schema zur Berechnung der dividierten Differenzen erzeugt werden. Die Auswertung der Differenzen in (2.15) erfolgt dann effizient mit dem Algorithmus aus Bemerkung 2.8.*

Bemerkung 2.20 *Als Stoppkriterium für das Hinzufügen von Gitterpunkten kann eine maximale Anzahl von Stützstellen vorgegeben werden. Alternativ kann eine Toleranz $\varepsilon > 0$ festgelegt werden und der Algorithmus abgebrochen werden, falls $d_{k_\star} \leq \varepsilon$ gilt.*

2.4 Approximation durch Splinefunktionen

Wesentlich flexibler als globale Polynomapproximationen sind Approximationen mit Splinefunktionen. Aus dem Weierstrassschen Approximationssatz folgt, dass sich stetige Funktionen auf abgeschlossenen Intervallen durch Polynome hinreichend hohen Grades beliebig genau approximieren lassen. Für alle quantitativen Fehlerabschätzungen wurde in den vorigen Abschnitten immer vorausgesetzt, dass die Funktion hinreichend glatt, sogar teilweise unendlich oft differenzierbar sein musste. Falls eine Funktion lediglich endliche Glattheit besitzt, $f \in C^k(I)$ aber $f \notin C^{k+1}(\Gamma)$, so lässt sich Konvergenz durch Reduzierung der Intervalllänge beweisen. Sei dazu $I = [a, b]$ wieder ein kompaktes Intervall und Θ eine Menge von Gitterpunkten:

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b,$$

die eine Intervallpartitionierung

$$\begin{aligned} \mathcal{G} &= \{\tau_i : 1 \leq i \leq n\}, \\ \tau_i &= [x_{i-1}, x_i] \end{aligned}$$

induzieren. Auf den einzelnen Intervallen τ_i wird nun mit Polynomen niedrigen (festen) Grades approximiert. Die Konvergenz erhält man durch Verkleinerung der *Schrittweite*

$$h := \max_{1 \leq i \leq n} h_i \quad \text{mit} \quad h_i := x_i - x_{i-1}.$$

Wir definieren den Raum aller Splinefunktionen zur Glattheit $k \in \{-1, 0, 1, 2, \ldots\}$ und lokalem Polynomgrad $m \in \mathbb{N}_0$:[5]

$$\mathcal{S}_{\mathcal{G}}^{k,m} := \left\{ u \in C^k(I) \mid \forall \tau \in \mathcal{G} : u \mid_{\tau} \in \mathbb{P}_m \right\}$$

Die *globale* Glattheitsanforderung ist offensichtlich im Innern der Intervalle immer erfüllt. Eine wesentliche Bedingung stellt sie nur in den Stützstellen dar. Für den Fall $k \geq m$ lässt sich zeigen, dass dann $\mathcal{S}_{\mathcal{G}}^{k,m} = \mathbb{P}_m$ gilt. Da wir die Splinefunktionen als Alternative zur Approximation mit *globalen* Polynomen betrachten wollen, setzen wir im folgenden immer

$$k < m$$

voraus.

2.4.1 Stückweise lineare Interpolation

Die wesentlichen Eigenschaften der Spline-Interpolation lassen sich am besten für den Fall $k = 0$, $m = 1$, also der linearen Spline-Interpolation, erklären. Hierbei stellt sich die Aufgabe, $u_1 \in \mathcal{S}_{\mathcal{G}}^{0,1}$ zu finden, so dass für eine gegebene stetige Funktion $f \in C^0(I)$

$$\forall x \in \Theta : \quad f(x) = u_1(x)$$

[5] Mit $C^{-1}(I)$ wird der Raum aller Funktionen von I nach $\mathbb{R}$ bezeichnet, d.h., diese Funktionen sind im Allgemeinen unstetig.

gilt. Die Lösung kann einfach *explizit* angegeben werden. Auf dem Intervall $\tau_i = [x_{i-1}, x_i]$ gilt

$$u_1(x) = f_{i-1} + (x - x_{i-1})[x_{i-1}, x_i] f$$

Zur Fehleranalyse können wir die bereits erzielten Ergebnisse der linearen Spline-Interpolation heranziehen. Für den Fehler im Intervall τ_i gilt die Darstellung:

$$f(x) - u_1[f](x) = (x - x_{i-1})(x - x_i)\frac{f''(\xi_x)}{2}, \qquad x \in \tau_i, \xi_x \in \tau_i$$

Für $f \in C^2(I)$ ergibt sich

$$|f(x) - u_1[f](x)| \leq \frac{h_i^2}{8} \|f''\|_{\max,\tau_i}, \qquad \forall x \in \tau_i. \tag{2.16}$$

Diese Fehlerabschätzung zeigt, dass der Fehler beliebig klein (gleichmässig in I) wird für $h \to 0$. Andererseits wird das Problem für kleiner werdende Schrittweite h zunehmend aufwendiger, da immer mehr Daten (Polynomabschnitte) gespeichert werden müssen. Um die stückweise lineare Interpolation zu konstruieren, verwenden wir die *Hut*-Basis für $\mathcal{S}_{\mathcal{G}}^{0,1}$: Für einen Knotenpunkt $x_i \in \Theta$ setzen wir für $0 \leq i \leq N$

$$b_i(x) := \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{falls } i \geq 1 \text{ und } x \in \tau_i, \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{falls } i < N \text{ und } x \in \tau_{i+1}, \\ 0 & \text{sonst.} \end{cases} \tag{2.17}$$

Dann ist durch $(b_i)_{i=0}^N$ eine Basis von $\mathcal{S}_{\mathcal{G}}^{0,1}$ gegeben. Die Interpolation einer stetigen Funktion $f \in C^0(I)$ ist dann durch

$$u_1[f](x) = \sum_{i=0}^{N} f(x_i) b_i(x)$$

gegeben.

2.4.2 Konvergenzgeschwindigkeit.

Numerische Verfahren hängen typischerweise von einem oder mehreren Kontrollparametern ab, die ein Mass für den Berechnungsaufwand darstellen. Je höher der Aufwand ist, desto kleiner sollte der Fehler sein. Für die globale Polynominterpolation war das beispielsweise der maximale Polynomgrad n oder für die Approximation in $\mathcal{S}_{\mathcal{G}}^{0,1}$ der Anzahl N der Teilintervalle bzw. die maximale Schrittweite h. In beiden Fällen haben wir Fehlerschranken hergeleitet, die für wachsende Parameter n oder N unter geeigneten Voraussetzungen an die Daten (Eingabefunktion f, Intervall, Stützstellen) gegen Null konvergieren. Der Begriff der Konvergenzgeschwindigkeit wird in der Numerik für verschiedene Anwendungen etwas unterschiedlich definiert. Für die Approximation von Funktionen wird typischerweise zwischen exponentieller und algebraischer Konvergenz unterschieden.

Definition 2.21 *Sei n der Aufwandsparameter der numerischen Approximationsmethode und ε_n die obere Fehlerschranke, welche von der (Glattheit der) zu approximierenden Funktion f,*

dem Intervall I, den Stützstellen und deren Anzahl n abhängt. Falls Zahlen $C_0 > 0$ und $\rho \in (0,1)$ existieren, so dass

$$\varepsilon_n \leq C_0 \rho^n \quad \forall n \tag{2.18}$$

gilt, sprechen wir von exponentieller Konvergenz *charakterisiert durch die Parameter C_0 und ρ.*

Falls Konstanten $C_1, s > 0$ existieren, so dass

$$\varepsilon_n \leq C_1 n^{-s} \quad \forall n$$

gilt, sprechen wir von algebraischer Konvergenz *charakterisiert durch die Parameter C_1 und s.*

Bemerkung 2.22

1. *Man beachte, dass Definition 2.21 lediglich eine Aussage über die Konvergenzgeschwindigkeit der Fehlerschranke macht. Numerische Experimente werden dann benötigt, ob die Schärfe der Abschätzung für Beispiele zu testen oder ob der Fehler sich für betrachtete Beispiele eventuell deutlich besser verhält.*

2. *Um numerisch ein exponentielles Konvergenzverhalten zu ermitteln, trägt man am einfachsten den Fehler e_n in* logarithmischer Skala *auf; plottet also $(n, \log e_n)$ für den tatsächlichen Fehler e_n. Anwendung des Logarithmus auf die rechte Seite von (2.18) ergibt die lineare Funktion $\log C_0 + (\log \rho)\, n$. Falls also die Abbildung der Paare $(n, \log e_n)$ gut durch eine (fallende) Gerade gefittet werden kann, ist das ein starkes Indiz für exponentielle Konvergenz. Aus der Steigung m dieser Gerade kann man dann mittels $\log \rho = m$ die Konstante ρ und aus dem y-Achsenabschnitt y_0 die Konstante C_0 mittels $\log C_0 = y_0$ ermitteln.*

3. *Analog wie bei (2.) lässt sich algebraische Konvergenz ermitteln: Man trägt nun den Fehler in einem* log-log-Plot *auf, d.h., plottet die Wertepaare $(\log n, \log e_n)$. Falls diese wieder gut durch eine Gerade gefittet werden (mit Steigung m und y-Achsenabschnitt y_0), legt das algebraische Konvergenz nahe. Die Konstanten C_1 und s lassen sich gemäss $s = -m$ und $\log C_1 = y_0$ numerisch bestimmen.*

2.4.3 Kleinste-Quadrate-Approximation mit linearen Splines

Die Interpolation eignet sich, um stetige Funktionen mit Polynomen oder Splinefunktionen zu approximieren. Falls die Funktion f unstetig oder sogar unbeschränkt ist (Beispiel: $f :]0,1[\to \mathbb{R}$, $f(x) = \log x$), ist die Interpolation nicht zu empfehlen oder ist nicht definiert. In diesem Abschnitt behandeln wir eine sehr allgemeine Approximationsmethode – die kleinste-Quadrate-Approximation – die zwar aufwändiger als die Interpolationsmethode ist, aber für eine wesentlich grössere Funktionenklasse anwendbar ist. Sei $\Omega =]a,b[$ für $a < b$ und

$$L^2(\Omega) := \left\{ f : \Omega \to \mathbb{R} \mid \int_a^b |f(x)|^2 \, dx < \infty \right\}.$$

Beispielsweise ist die unbeschränkte Funktion $f(x) := x^{-1/3}$ in $L^2(]0,1[)$ enthalten:

$$\int_0^1 \left(x^{-1/3}\right)^2 dx = 3.$$

Da die Funktion f für $x \to 0$ gegen $+\infty$ strebt, ist die Maximumnorm nicht definiert. Als Mass für die Grösse einer Funktion verwenden wir die L^2-Norm:

$$f \in L^2(\Omega): \qquad \|f\|_{L^2(\Omega)} := \sqrt{\int_a^b |f(x)|^2 \, dx}.$$

Unser Ziel ist es, die beste Approximation $f_{\text{opt}} \in \mathcal{S}_{\mathcal{G}}^{0,1}$von f bezüglich der L^2-Norm zu bestimmen, d.h.,

$$\|f - f_{\text{opt}}\|_{L^2(\Omega)} = \min_{u \in \mathcal{S}_{\mathcal{G}}^{0,1}} \|f - u\|_{L^2(\Omega)} \tag{2.19}$$

zu berechnen. Da das Kleinste-Quadrate-Verfahren für eine viel grössere Klasse von Problemen anwendbar ist, wollen wir es hier abstrakt erklären. Wir benötigen dazu nur zwei Voraussetzungen:

- V ist ein (möglicherweise unendlichdimensionaler) Funktionenraum über dem Körper $\mathbb{R}$ mit einem Skalarprodukt $(\cdot,\cdot)$ und Norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$.
- $S \subset V$ ist ein endlichdimensionaler Untervektorraum, für den eine Basis b_i, $0 \le i \le n$, gewählt sei.

Aufgabe: Berechne die Kleinste-Quadrate-Approximation $f_{\text{opt}} \in S$ für eine gegebene Funktion $f \in V$ bezüglich der Norm $\|\cdot\|$.

1. **Algebraisierung.**

   (a) Die Minimierungsfunktion f_{opt} von (2.19) lässt sich äquivalent durch

   $$\|f - f_{\text{opt}}\|^2 = \min_{u \in S} F^2(u) \quad \text{mit} \quad F^2(u) := \|f - u\|^2 \tag{2.20}$$

   charakterisieren.

   (b) Das Quadrat der Norm lässt sich mit Hilfe des Skalarprodukts ausmultiplizieren

   $$\|f - u\|^2 = (f - u, f - u) = \|f\|^2 - 2(f, u) + (u, u).$$

   (c) Jede Funktion $u \in S$ besitzt eine eindeutige Basisdarstellung

   $$u = \sum_{i=0}^{n} u_i b_i,$$

   und das Ziel ist nun, die Koeffizienten u_i, $0 \le i \le n$, zu berechnen, so dass die zugehörige Funktion u das Funktional F^2 (und damit auch F) minimiert.

   (d) Die Linearität des Skalarprodukts ergibt

   $$(u, v) = \sum_{i,j=0}^{n} u_i v_j (b_i, b_j).$$

   Wir definieren die Matrix

   $$\mathbf{M} := ((b_i, b_j))_{i,j=0}^{n} \tag{2.21}$$

und den Vektor

$$\mathbf{r} := ((f, b_i))_{i=0}^n .$$

Dann lässt sich das Minimierungsproblem für S überführen in ein Minimierungsproblem über $\mathbb{R}^{n+1}$

$$\|f - f_{\text{opt}}\|^2 = \min_{(u_i)_{i=0}^n \in \mathbb{R}^{n+1}} \tilde{F}^2(\mathbf{u})$$

mit

$$\tilde{F}^2(\mathbf{u}) = \tilde{F}^2(u_0, \ldots, u_n) := \|f\|^2 - 2\sum_{j=0}^N r_j u_j + \sum_{j,k=0}^N u_j u_k M_{j,k}. \tag{2.22}$$

2. **Minimierung**

   (a) Ziel ist es nun, einen Vektor $\mathbf{u}^{\text{opt}} = \left(u_i^{\text{opt}}\right)_{i=0}^n \in \mathbb{R}^{n+1}$ zu bestimmen, der

   $$\tilde{F}^2\left(\mathbf{u}^{\text{opt}}\right) = \min_{(u_i)_{i=0}^n \in \mathbb{R}^{n+1}} \tilde{F}^2(u_0, u_2, \ldots, u_n)$$

   erfüllt.

   (b) Die Funktion $\tilde{F}$ hängt von $(n+1)$-Variablen ab. Das Minimum wird angenommen, falls für alle Ableitungen $\partial\tilde{F}/\partial u_i = 0$ gilt. Das ist äquivalent[6] zur Lösung des linearen Gleichungssystems

   $$\mathbf{M}\mathbf{u}_{\text{opt}} = \mathbf{r}. \tag{2.23}$$

   (c) Die Lösung des Ausgangsproblems ist dann durch

   $$f_{\text{opt}} = \sum_{i=0}^N u_i^{\text{opt}} b_i$$

   gegeben.

Anwendung auf (2.19)

Für diese Anwendung gilt: $V := C^0(I)$, $S = S_{\mathcal{G}}^{0,1}$, $(b_i)_{i=0}^n$ ist die Hut-Basis, $(\cdot,\cdot) = (\cdot,\cdot)_{L^2(\Omega)}$ und $\|\cdot\| = \|\cdot\|_{L^2(\Omega)}$.

Um die Matrixeinträge $M_{i,j}$ zu berechnen, starten wir mit einigen Vorüberlegungen. Der Träger einer Funktion $g \in C^0(\Omega)$ ist die Menge

$$\operatorname{tr} g = \overline{\{x : g(x) \neq 0\}} \cap \Omega.$$

[6] Aus der Definition von $\tilde{F}$ folgt mit dem Kronecker-Delta $\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$ und den Beziehungen $\partial u_j/\partial u_i = \delta_{i,j}$ und der Produktregel $\frac{\partial(u_j u_k)}{\partial u_i} = \delta_{i,j}u_k + \delta_{i,k}u_j$, dass $\partial\tilde{F}^2/\partial u_i = 0$ äquivalent ist zu

$$\partial\tilde{F}^2/\partial u_i = -2\sum_{j=0}^N r_j\delta_{i,j} + \sum_{j,k=0}^N \delta_{i,j}u_k M_{j,k} + \sum_{j,k=0}^N \delta_{i,k}u_j M_{j,k} = -2r_i + \sum_{k=0}^N M_{i,k}u_k + \sum_{j=0}^N u_j M_{j,i}$$
$$= \left(-2\mathbf{r} + \mathbf{M}\mathbf{u} + \mathbf{M}^\intercal\mathbf{u}\right)_i \stackrel{!}{=} 0.$$

Aus der Definition (2.21) und der Symmetrie des Skalarprodukts ergibt sich $\mathbf{M}^\intercal = \mathbf{M}$, so dass wir (2.23) erhalten.

Offensichtlich gilt:

$$\int_\Omega g(x)\,dx = \int_{\operatorname{tr} g} g(x)\,dx.$$

Für die Basisfunktion b_i gilt:

$$\operatorname{tr} b_i = \overline{\tau_i} \cup \overline{\tau_{i+1}} \tag{2.24}$$

mit entsprechenden Modifikationen für $i \in \{0, N\}$. Daraus folgt

$$(b_i, b_j)_{L^2(\Omega)} = \int_a^b b_i(x)\, b_j(x)\,dx = \int_{\operatorname{tr} b_i \cap \operatorname{tr} b_j} b_i(x)\, b_j(x)\,dx.$$

Wegen (2.24) gilt für $|i-j| \geq 2$

$$(b_i, b_j)_{L^2(\Omega)} = 0.$$

Daher ist $\mathbf{M}$ tridiagonal. Einfache Rechnung ergibt für

$$\mathbf{M} = \left((b_i, b_j)_{L^2(\Omega)}\right)_{i,j=0}^{N} = \frac{1}{6}\begin{bmatrix} 2h_1 & h_1 & 0 & \dots & & 0 \\ h_1 & 2(h_1+h_2) & h_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & \ddots & & & & 0 \\ & & & h_{N-1} & 2(h_{N-1}+h_N) & h_N \\ 0 & \dots & & 0 & h_N & 2h_N \end{bmatrix}.$$

Wir haben damit das Interpolationsproblem auf die Frage zurückgeführt, ein lineares Gleichungssystem zu lösen.

Fehlerabschätzungen.

Wir wollen an dieser Stelle noch auf die Konvergenz-Eigenschaften eingehen und zwei Fälle betrachten.

1. $V = C^0(I)$, $S = \mathbb{P}_n$ und $(\cdot,\cdot)_{L^2(I)}$. Die beste Approximation einer Funktion $f \in V$ bezüglich der $L^2(I)$-Norm wird mit $f_n \in \mathbb{P}_n$ bezeichnet. Dann gilt für den Fehler auf Grund der Monotonie des Integrals

   $$\|f - f_n\|_{L^2(I)}^2 = \min_{u\in\mathbb{P}_n}\left(\int_a^b (f-u)^2\right) \leq \min_{u\in\mathbb{P}_n}\left(\|f-u\|_{\max}^2 \int_a^b 1\right)$$

   und damit

   $$\|f - f_n\|_{L^2(I)} \leq \sqrt{b-a}\, \min_{u\in\mathbb{P}_n} \|f-u\|_{\max}.$$

   Der Weierstrasssche Approximationssatz besagt, dass stetige Funktionen auf einem kompakten Intervall (beschränkt und abgeschlossen) beliebig gut mit Polynomen approximiert werden können, genauer:

   $$\min_{u\in\mathbb{P}_n} \|f-u\|_{\max} \overset{n\to\infty}{\to} 0$$

   gilt. Damit haben wir gezeigt, dass durch die kleinste-Quadrate-Approximation eine Folge von Polynomen konstruiert werden kann, die für beliebiges, gegebenes $f \in C^0(I)$ konvergiert.

2. Wir betrachten $V = C^0(I)$, $S = S_{\mathcal{G}}^{0,1}$ und $(\cdot,\cdot)_{L^2(I)}$. Dann erhalten wir analog wie zuvor eine Folge von Funktionen $f_n \in S_{\mathcal{G}}^{0,1}$ mit

$$\|f - f_n\|_{L^2(I)} \leq \sqrt{b-a} \min_{u \in S_{\mathcal{G}}^{0,1}} \|f - u\|_{\max}. \tag{2.25}$$

Wählen wir nun zur Abschätzung der rechten Seite in (2.25) die Interpolierende $p[f, \Theta_n] \in S_{\mathcal{G}}^{0,1}$, erhalten wir mit der Voraussetzungen $f \in C^2(I)$ aus (2.16)

$$\|f - f_n\|_{L^2(I)} \leq \sqrt{b-a} \frac{h^2}{8} \|f''\|_{\max,I}.$$

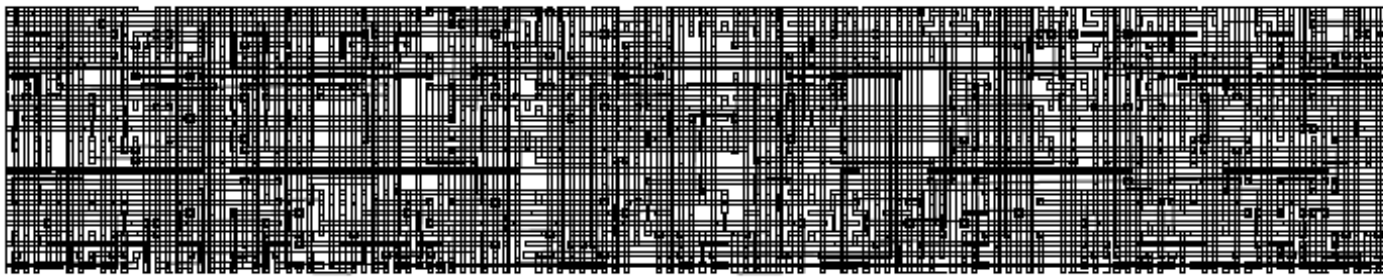


Abbildung 2: Schaltkreis eines Routing Channels (Ausschnitt).

3 Lineare Gleichungssysteme – das QR-Verfahren

In diesem Abschnitt werden wir die Lösung linearer Gleichungssysteme (LGS) behandeln. Dieses Problem tritt in vielen unterschiedlichen Bereichen der Mathematik und in vielen praktischen Anwendungen auf, und wir beginnen mit einem einleitenden Beispiel.

Beispiel 3.1 *Ziel sei es, das elektrische Feld in einem komplizierten Schaltkreis zu berechnen. Der Schaltkreis sei gegeben durch eine Menge von Verzweigungspunkten (Knotenpunkte)* $\Theta = \{x_i : 1 \leq i \leq n\}$ *und einer Menge von Verbindungskanten (elektrische Leiter)* $\mathcal{E} \subset \Theta \times \Theta$*, beispielsweise bedeutet die Notation* $e = (x_i, x_j) \in \mathcal{E}$*, dass* e *die Knotenpunkte* x_i *und* x_j *als Endpunkte besitzt. Aus Symmetriegründen setzen wir voraus, dass mit jedem* $e = (x_i, x_j)$ *auch die "umgekehrte" Kante* $\tilde{e} = (x_j, x_i)$ *in* $\mathcal{E}$ *enthalten ist.*

Auf Computerchips können derartige Schaltkreise äusserst komplex sein, wie die Abbildung 2 zeigt.

Auf einem Leiterstück $e = (x_i, x_j) \in \mathcal{E}$ *der Länge* $h_e := \|x_i - x_j\|$ *beträgt dann die elektrische Energie*

$$a_e \frac{(u(x_i) - u(x_j))^2}{h_e},$$

wobei $a_e > 0$ *einen gegebenen Leitfähigkeitskoeffizienten für die Kante* e *bezeichnet. Die Gesamtenergie des Systems bei einem angelegten Feld* $\mathbf{f} = (f(x_i))_{i=1}^n$ *beträgt dann für einen Spannungsvektor* $\mathbf{v} = (v(x_i))_{i=1}^n$ *folglich*

$$J(\mathbf{v}) := \frac{1}{2} \sum_{e=(x_i,x_j)\in\mathcal{E}} \frac{a_e}{h_e} (v(x_i) - v(x_j))^2 - 2 \sum_{x_i \in \Theta} f(x_i)\, v(x_i).$$

Der Zustand, den das elektrische System einnimmt, ist derjenige, welcher die Energie $J(\cdot)$ *minimiert. Mathematisch ist dieser Zustand charakterisiert durch die Gleichung*

$$\nabla J(\mathbf{u}) = 0$$

Der Gradient[7] *wird hierbei auf die Variablen* $u(x_i)$ *angewendet. Durch Berechnung dieser Ableitungen erhält man schliesslich, dass der Spannungsvektor für das elektrische System die Lösung des linearen Gleichungssystems*

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

[7] Der Gradient wirkt auf eine Funktion $f(x, y, z...)$, die von mehreren Variablen abhängt und bezeichnet den Vektor der partiellen Ableitungen $\nabla f = \left(\frac{\partial}{\partial x} f, \frac{\partial}{\partial y} f, \frac{\partial}{\partial z} f, \ldots\right)^\intercal$.

ist, wobei

$$\mathbf{A}_{i,j} := \begin{cases} \displaystyle\sum_{\substack{x_k \in \Theta: \\ (x_i,x_k)\in\mathcal{E}}} \frac{a_{(x_i,x_k)}}{h_{(x_i,x_k)}} & \textit{falls } i = j, \\ -\dfrac{a_{(x_i,x_j)}}{h_{(x_i,x_j)}} & \textit{falls } (x_i, x_j) \in \mathcal{E}, \\ 0 & \textit{andernfalls.} \end{cases}$$

Sind einige der Werte von $\mathbf{u}$ *durch Messungen bekannt, können diese Variablen bereits eingesetzt werden und die entsprechenden Zeilen und Spalten aus der Matrix gestrichen werden.*

Dies ist ein praktisches Beispiel, für das die robuste und effiziente Lösung grosser linearer Gleichungssysteme eine wesentliche Rolle spielt. Generell unterscheidet man zwei Typen von Verfahren zur Lösung linearer Gleichungssysteme.

- *Direkte* Lösungsmethoden lassen sich zur Auflösung aller linearer Gleichungssysteme mit regulären Matrizen einsetzen. Vernachlässigt man den Rundungsfehler liefern diese Verfahren nach **endlich** vielen Schritten die exakte Lösung.
- *Iterative* Lösungsverfahren konstruieren aus einem Startvektor eine Folge von Vektoren, die gegen die exakte Lösung des linearen Gleichungssystems konvergieren. Im allgemeinen konvergiert ein iteratives Verfahren nicht für alle regulären Gleichungssysteme. Die *Konvergenzanalyse* dient u.a. zur Bestimmung von hinreichenden Kriterien an das LGS, damit die Konvergenz gesichert ist.

In diesem Kapitel werden wir uns mit direkten Lösungsverfahren beschäftigen. Im zweiten Teil der Vorlesung Numerik I werden iterative Verfahren zur Lösung von LGS behandelt. Die wesentlichen Vor- und Nachteile dieser Verfahren sind im folgenden kurz zusammengestellt. n bezeichnet immer die Dimension des (quadratischen) Gleichungssystems.

direkte Lösungsverfahren	iterative Verfahren
anwendbar auf alle regulären Matrizen	Konvergenz für kleinere Klassen von Matrizen
Aufwand $O(n^3)$	Aufwand der schnellsten Verfahren (Mehrgitterverfahren) beträgt $O(n)$.

Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$ eine reguläre, quadratische Matrix der Dimension n mit (im Allgemeinen) komplexen Einträgen $a_{i,j} \in \mathbb{C}$. Wir betrachten die Aufgabe, für gegebene rechte Seite $\mathbf{b} \in \mathbb{C}^n$ einen Vektor $\mathbf{x} \in \mathbb{C}^n$ zu finden mit

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{3.1}$$

Die Dimension der Matrix $\mathbf{A}$ sei gross, d.h. 5-10000 Unbekannte.

Für einige wenige Matrixtypen ist die Lösung von (3.1) sehr einfach.

Dreiecksmatrizen: Sei $\mathbf{A} = (a_{i,j})_{i,j=1}^n \in \mathbb{C}^{n\times n}$ eine linke untere Dreiecksmatrix, d.h.

$$\mathbf{A} = \begin{bmatrix} \star & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ & & & 0 \\ \star & \dots & & \star \end{bmatrix} \quad \text{d.h.} \quad a_{i,j} = 0 \quad \forall 1 \le i < j \le n.$$

Bemerkung 3.2 *Die Voraussetzung „***A** *ist regulär“ impliziert für Dreiecksmatrizen, dass die die Diagonalelemente $a_{i,i}$ von Null verschieden sind. (Beachte:* $\det \mathbf{A} = a_{1,1} \times a_{2,2} \times \ldots \times a_{n,n}$.*)*

Die Lösung eines regulären LGS mit linker unterer Dreiecksmatrix geschieht durch einfaches Rückwärtseinsetzen.

procedure solve_lower_triangular_system$(\mathbf{A}, \mathbf{b}, \mathbf{x})$;
begin
  for $i := 1$ **to** n **do begin**
    $s := b_i$;
    for $j := 1$ **to** $i - 1$ **do** $s := s - a_{i,j} x_j$;
    $x_i := s / a_{i,i}$
  end;
end;

Wie man leicht abzählt, beträgt die erforderliche Anzahl arithmetischer Operationen

$$\sum_{i=1}^{n} \left(1 + \sum_{j=1}^{i-1} 2 \right) = n^2.$$

Analog lassen sich reguläre LGS mit rechten oberen Dreiecksmatrizen einfach lösen.

Unitäre Matrizen

Konvention: $\mathbb{C}$ bezeichnet die Menge der komplexen Zahlen. Jede komplexe Zahl $w \in \mathbb{C}$ lässt sich eindeutig zerlegen gemäss $w = u + \mathrm{i}\, v$ mit reellen Zahlen $u, v \in \mathbb{R}$ und der imaginären Einheit $\mathrm{i} = \sqrt{-1}$. Die Zahl $u = \mathrm{Re}\,(w)$ bezeichnet den *Realteil* und $v = \mathrm{Im}\,(w)$ den *Imaginärteil* von w. Für eine komplexe Zahl $w = u + \mathrm{i}\, v \in \mathbb{C}$ ist die *komplexe Konjugation* durch $\overline{w} := u - \mathrm{i}\, v$ definiert. Für Vektoren $\mathbf{u} = (u_k)_{k=1}^n \in \mathbb{C}^n$ und $\mathbf{v} = (v_k)_{k=1}^n \in \mathbb{C}^n$ ist das Euklidsche Skalarprodukt durch

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=1}^{n} u_k \overline{v_k}$$

gegeben und die Euklidsche Norm (Länge) durch $\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}$. Die Spaltenvektoren einer quadratischen Matrix $\mathbf{Q} \in \mathbb{C}^{n \times n}$ werden mit $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ bezeichnet.

Definition 3.3 *Die Matrix* $\mathbf{Q} \in \mathbb{C}^{n \times n}$ *heisst* ***unitär,*** *falls deren Spaltenvektoren* $\mathbf{q}_i$, $1 \le i \le n$, *eine Orthonormalbasis in* $\mathbb{C}^n$ *bilden:*

$$\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \begin{cases} 1 & i = j, \\ 0 & \textit{sonst.} \end{cases}$$

Die ***komplex transponierte*** *Matrix einer Matrix* $\mathbf{A} = (a_{i,j})_{i,j=1}^n$ *ist durch*

$$\mathbf{A}^{\mathrm{H}} = (\overline{a_{j,i}})_{i,j=1}^n$$

gegeben.

Bemerkung 3.4
(a) Für eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n \times n}$ *gilt* $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{H}}$.
(b) Für beliebiges $\mathbf{x} \in \mathbb{C}^n$ *gilt* $\|\mathbf{Q}\mathbf{x}\| = \|\mathbf{x}\|$.

Beweis. Für $\mathbf{Q} = (q_{i,j})_{i,j=1}^n$ gilt $\mathbf{Q}^{\mathrm{H}} = (\overline{q_{j,i}})_{i,j=1}^n$. Wir bezeichnen die Spaltenvektoren in $\mathbf{Q}$ wieder mit $\mathbf{q}_1$, $\mathbf{q}_2$, ...,$\mathbf{q}_n$. Dann gilt

$$\left(\mathbf{Q}^{\mathrm{H}}\mathbf{Q}\right)_{i,j} = \sum_{m=1}^{n} (\overline{q_{m,i}})\, q_{m,j} = \langle \mathbf{q}_j, \mathbf{q}_i \rangle = \begin{cases} 1 & i = j, \\ 0 & \text{sonst.} \end{cases}$$

Daher ist $\mathbf{Q}^{\mathrm{H}}\mathbf{Q}$ die Einheitsmatrix, und daraus folgt $\mathbf{Q}^{\mathrm{H}} = \mathbf{Q}^{-1}$.

Für beliebiges $\mathbf{x} = (x_i)_{i=1}^n \in \mathbb{C}^n$ gilt

$$\begin{aligned} \|\mathbf{Q}\mathbf{x}\|^2 = \langle \mathbf{Q}\mathbf{x}, \mathbf{Q}\mathbf{x} \rangle &= \left\langle \sum_{i=1}^{n} x_i \mathbf{q}_i, \sum_{j=1}^{n} x_j \mathbf{q}_j \right\rangle = \sum_{i,j=1}^{n} x_i \overline{x_j} \langle \mathbf{q}_i, \mathbf{q}_j \rangle \\ &= \sum_{i=1}^{n} |x_i|^2 = \|\mathbf{x}\|^2 . \end{aligned}$$

Durch Wurzelziehen erhalten wir die Behauptung (b). ■

Aus Bemerkung 3.4 folgt, dass sich lineare Gleichungssysteme mit unitären Matrizen sehr einfach auflösen lassen. Aus

$$\mathbf{Q}\mathbf{x} = \mathbf{b} \quad \text{folgt} \quad \mathbf{x} = \mathbf{Q}^{\mathrm{H}}\mathbf{b},$$

und die Lösung ist durch eine einfache Matrix-Vektor-Multiplikation gegeben.

```
procedure solve_unitary_system(Q, b, x);
begin
    for i := 1 to n do begin
        x_i := 0;
        for j := 1 to n do x_i := x_i + conj(q_{j,i}) b_j
    end;
end;
```

Der Aufwand beträgt offensichtlich $2n^2$ arithmetische Operationen.

Das bekannteste direkte Lösungsverfahren ist die Gauss-Elimination. Hierbei werden eine linke untere Dreiecksmatrix $\mathbf{L}$ und eine rechte obere Dreiecksmatrix $\mathbf{R}$ konstruiert, so dass die folgende Faktorisierung der Originalmatrix

$$\mathbf{A} = \mathbf{L}\mathbf{R}$$

gilt.[8] Dann lässt sich ein (beliebiges) reguläres Gleichungssystem der Form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{d.h.} \quad \mathbf{L}\mathbf{R}\mathbf{x} = \mathbf{b}$$

in zwei Schritten lösen: Berechne $\mathbf{y}$ in $\mathbf{L}\mathbf{y} = \mathbf{b}$ mit Hilfe von **procedure solve_lower_triangular_system** und anschliessend $\mathbf{x}$ in $\mathbf{R}\mathbf{x} = \mathbf{y}$ mit Hilfe der analogen Prozedur für rechte obere Dreiecksmatrizen. Das Verfahren besitzt jedoch zwei entscheidende Nachteile: (a) Das Verfahren kann numerisch instabil sein, d.h. Rundungs- und Eingabefehler können kritisch

[8] Im allgemeinen ist dies nur nach geeigneten Zeilen- bzw. Spaltenvertauschungen bei $\mathbf{A}$ möglich. Wir verzichten hier auf die Details, da diese Schwierigkeiten beim QR-Verfahren entfällt.

verstärkt werden, durch den Gauss-Algorithmus. (b) Die Implementierung der Spalten und Zeilenvertauschungen kompliziert den Algorithmus.

In diesem Kapitel werden wir einen Eliminationsalgorithmus vorstellen, bei dem beide Nachteile beseitigt sind.

Generell setzen wir immer voraus, dass $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{A}$ regulär, gilt.

Das Ziel des QR-Verfahrens ist es, sukzessive eine unitäre Matrix $\mathbf{U}$ und eine obere Dreiecksmatrix $\mathbf{R}$ zu konstruieren, so dass

$$\mathbf{UA} = \mathbf{R} \tag{3.2}$$

gilt. Das Gleichungssystem

$$\mathbf{U}^{\mathrm{H}}\mathbf{Rx} = \mathbf{Ax} = \mathbf{b}$$

ist daher in zwei Stufen zu lösen:

$$\text{Berechne } \mathbf{y} := \mathbf{Ub} \quad \text{und löse danach } \mathbf{Rx} = \mathbf{y}.$$

Die unitäre Matrix $\mathbf{U}$ in (3.2) wird spaltenweise aufgebaut. Wir betrachten die Aufgabe etwas allgemeiner: Sei $\mathbf{B} \in \mathbb{C}^{m\times m}$ eine quadratische Matrix, und $\mathbf{b}_1$ bezeichnet die erste Spalte von $\mathbf{B}$. Wie suchen eine unitäre Matrix $\mathbf{P} = \mathbf{P}(\mathbf{B})$, so dass für die Multiplikation mit $\mathbf{B}$ gilt

$$\mathbf{PB} = \begin{bmatrix} k & \star & \dots & \star \\ 0 & \star & \ddots & \vdots \\ \vdots & \vdots & & \\ 0 & \star & \dots & \star \end{bmatrix}.$$

Falls die erste Spalte $\mathbf{b}_1$ der Matrix $\mathbf{B}$ bereits ein Vielfaches des ersten Einheitsvektors ist, d.h., $\mathbf{b}_1 = k\mathbf{e}_1$, setzen wir $\mathbf{P} = \mathbf{I}$ also gleich der Einheitsmatrix. Im Folgenden wird der (typischere) Fall betrachtet, dass $\mathbf{b}_1$ kein Vielfaches von $\mathbf{e}_1$ ist.

Wir setzen[9]

$$\mathbf{P} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^{\mathrm{H}} \tag{3.3}$$

mit

$$\mathbf{w} := \mathbf{w}(\mathbf{b}_1) := \frac{\mathbf{b}_1 - k\mathbf{e}_1}{\|\mathbf{b}_1 - k\mathbf{e}_1\|}. \tag{3.4a}$$

Wählt man

$$k := k(\mathbf{b}_1) := \begin{cases} -\frac{b_{1,1}}{|b_{1,1}|}\|\mathbf{b}_1\| & b_{1,1} \neq 0, \\ \|\mathbf{b}_1\| & b_{1,1} = 0. \end{cases} \tag{3.4b}$$

rechnet man nach (vgl. Lemma 3.5), dass

$$\mathbf{Pb}_1 = k\mathbf{e}_1$$

gilt, d.h., die erste Spalte von $\mathbf{B}$ eliminiert ist.

[9] Das *dyadische Product* zweier Vektoren $\mathbf{a} = (a_i)_{i=1}^n, \mathbf{c} = (c_i)_{i=1}^n \in \mathbb{C}^n$ ist eine $n\times n$-Matrix, deren Einträge durch $\left(\mathbf{ac}^{\mathrm{H}}\right)_{i,j} := a_i\overline{c_j}$, $1 \le i,j \le n$, definiert sind.

Lemma 3.5 *Sei* $\mathbf{b}_1 \notin \operatorname{span}\{\mathbf{e}_1\}$. *Mit der Wahl (3.4) von* $\mathbf{w}$ *und* k *gilt*

$$\mathbf{P}\mathbf{b}_1 = k\mathbf{e}_1.$$

Die Matrix $\mathbf{P}$ *ist unitär.*

Beweis. Zunächst rechnet man nach, dass

$$\begin{aligned}\|\mathbf{b}_1 - k\mathbf{e}_1\|^2 &= \|\mathbf{b}_1\|^2 - b_{1,1}\overline{k} - k\overline{b_{1,1}} + |k|^2 \\ &\overset{\text{Einsetzen der Definition von } k}{=} 2\|\mathbf{b}_1\|\left(|b_{1,1}| + \|\mathbf{b}_1\|\right)\end{aligned}$$

gilt. Daraus folgt

$$\mathbf{P}\mathbf{b}_1 = \left(\mathbf{I} - 2\mathbf{w}\mathbf{w}^{\mathrm{H}}\right)\mathbf{b}_1 = k\mathbf{e}_1 + (\mathbf{b}_1 - k\mathbf{e}_1)\underbrace{\left(1 - 2\frac{\langle \mathbf{b}_1, \mathbf{b}_1 - k\mathbf{e}_1\rangle}{\|\mathbf{b}_1 - k\mathbf{e}_1\|^2}\right)}_{=:\delta}.$$

Die Behauptung ergibt sich, falls wir $\delta = 0$ zeigen. Wir verwenden

$$\begin{aligned}\delta &= 1 - 2\frac{\langle \mathbf{b}_1, \mathbf{b}_1 - k\mathbf{e}_1\rangle}{\|\mathbf{b}_1 - k\mathbf{e}_1\|^2} = 1 - 2\frac{\langle \mathbf{b}_1 - k\mathbf{e}_1, \mathbf{b}_1 - k\mathbf{e}_1\rangle}{\|\mathbf{b}_1 - k\mathbf{e}_1\|^2} + 2\frac{\langle -k\mathbf{e}_1, \mathbf{b}_1 - k\mathbf{e}_1\rangle}{\|\mathbf{b}_1 - k\mathbf{e}_1\|^2} \\ &= -1 - 2k\frac{\overline{b_{1,1}} - \overline{k}}{\|\mathbf{b}_1 - k\mathbf{e}_1\|^2} \overset{\text{Einsetzen von } k}{=} 0.\end{aligned}$$

Die Unitarität von $\mathbf{P}$ folgt wegen $\|\mathbf{w}\| = 1$ gemäss

$$\mathbf{P}_1^{\mathrm{H}}\mathbf{P}_1 = \left(\mathbf{I} - 2\mathbf{w}\mathbf{w}^{\mathrm{H}}\right)\left(\mathbf{I} - 2\mathbf{w}\mathbf{w}^{\mathrm{H}}\right) = \mathbf{I} - 4\mathbf{w}\mathbf{w}^{\mathrm{H}} + 4\mathbf{w}\underbrace{\left(\mathbf{w}^{\mathrm{H}}\mathbf{w}\right)}_{=1}\mathbf{w}^{\mathrm{H}} = \mathbf{I}.$$

■

Damit ist gezeigt, dass sich $\mathbf{P}$ in der Form

$$\mathbf{P} = \mathbf{I} - \beta\mathbf{u}\mathbf{u}^{\mathrm{H}}$$

schreiben lässt mit

$$\beta := \beta(\mathbf{b}_1) := \frac{1}{\|\mathbf{b}_1\|\left(|b_{1,1}| + \|\mathbf{b}_1\|\right)} \tag{3.5a}$$

und

$$\mathbf{u} := \mathbf{u}(\mathbf{b}_1) := \mathbf{b}_1 - k\mathbf{e}_1 = \begin{pmatrix} b_{1,1} + \sigma_1\|\mathbf{b}_1\| \\ b_{2,1} \\ \vdots \\ b_{n,1} \end{pmatrix} \quad \text{wobei} \quad \sigma_1 := \begin{cases} b_{1,1}/|b_{1,1}| & \text{für } b_{1,1} \neq 0, \\ -1 & \text{für } b_{1,1} = 0 \end{cases}. \tag{3.5b}$$

Die Anwendung von $\mathbf{P}(\mathbf{A})$ auf $\mathbf{A}^{(0)} := \mathbf{A}$ liefert also ein System $\mathbf{A}^{(1)}$ mit erstem Spaltenvektor $k\mathbf{e}_1$. Nach $j-1$ Schritten erhält man eine Matrix der Bauart

$$\mathbf{A}^{(j-1)} = \begin{bmatrix} \mathbf{D} & \mathbf{B} \\ \mathbf{0} & \tilde{\mathbf{A}}^{(j-1)} \end{bmatrix}$$

mit einer $(j-1)\times(j-1)$ oberen Dreiecksmatrix $\mathbf{D}$. Man setzt dann für den nächsten Schritt

$$\mathbf{P}_j = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}\left(\tilde{\mathbf{A}}^{(j-1)}\right) \end{bmatrix}.$$

Anwendung von $\mathbf{P}_j$ auf $\mathbf{A}^{(j-1)}$ lässt die ersten $(j-1)$ Zeilen unverändert und auch den linken unteren Nullblock. Die ersten Zeile von $\tilde{\mathbf{A}}^{(j-1)}$ ist danach eliminiert. Nach $(n-1)$ Schritten erhält man eine obere Dreiecksmatrix:

$$\mathbf{R} := \mathbf{A}^{(n-1)}.$$

Die Eliminationsmatrix $\mathbf{P}_j$ ist durch $(n-j+1)$ wesentliche Komponenten von $\mathbf{u}_j$ charakterisiert. In $\tilde{\mathbf{A}}^{(j-1)}$ werden aber nur $(n-j)$ Plätze unterhalb des Diagonalelements $\tilde{A}_{j,j}$ frei, so dass man typischerweise das Diagonalelement von $\tilde{\mathbf{A}}^{(j-1)}$ in einem separaten Vektor $\mathbf{d}$ abspeichert und alle wesentlichen Komponenten von $\mathbf{u}$ in die frei werdenden Plätze.

In konstruktiver Weise haben wir damit gezeigt, dass sich jede reguläre Matrix $\mathbf{A}\in\mathbb{C}^{n\times n}$ in das Produkt einer unitären Matrix und einer oberen Dreiecksmatrix zerlegen lässt. Bezeichnet man das Produkt der Eliminationsmatrizen mit $\mathbf{U}=\mathbf{P}_{n-1}\mathbf{P}_{n-2}\cdots\mathbf{P}_1$, dann ist $\mathbf{I}$ wieder unitär (aber nicht notwendigerweise hermitesch), und es gilt

$$\mathbf{UA}=\mathbf{R} \quad \text{bzw.} \quad \mathbf{A}=\mathbf{QR} \quad \text{mit} \quad \mathbf{Q}=\mathbf{U}^{\mathrm{H}}.$$

Schliesslich wollen wir den Aufwand für das QR-Verfahren abschätzen.

Satz 3.6 *Der Aufwand, die QR-Zerlegung einer Matrix $\mathbf{A}\in\mathbb{C}^{n\times n}$ zu berechnen, beträgt $\left(\frac{4}{3}n^3+p\left(n\right)\right)$ arithmetische Operationen, wobei p ein Polynom vom Maximalgrad 2 bezeichnet.*

Beweis. Der Algorithmus zur Berechnung von $\mathbf{P}_1$ gliedert sich in die folgenden Schritte:

1. Berechnung von $\|\mathbf{a}_1\|$: Anzahl Operationen

$$\underbrace{(n-1)}_{\text{Additionen}} + \underbrace{n}_{\text{Multiplikationen}} + \underbrace{1}_{\text{Quadratwurzel}} = 2n.$$

2. Berechnung von k: Aufwand: $O\left(1\right)$ Operationen (unabhängig von n)

3. Berechnung von $\mathbf{u}$ und β (vgl. (3.5)): Aufwand: $O\left(1\right)$ Operationen (unabhängig von n)

4. Auswertung von $\mathbf{P}_1\mathbf{A}=\mathbf{A}-\beta\mathbf{u}\mathbf{u}^{\mathrm{H}}\mathbf{A}$. Die Berechnung erfolgt in 2 Schritten

   (a) Berechnung von $\mathbf{v}^{\mathrm{H}}:=\beta\mathbf{u}^{\mathrm{H}}\mathbf{A}$. Dazu sind $n-1$ Skalarprodukte der Länge n auszuwerten. Gesamtaufwand zur Berechnung von $\mathbf{v}^{\mathrm{H}}:=\left(n-1\right)2n$ Operationen.

   (b) Berechnung von $\mathbf{A}-\mathbf{u}\mathbf{v}^{\mathrm{H}}$. Die Berechnung erfordert $n\left(n-1\right)$ Multiplikationen und genauso viele Subtraktionen.

Der Gesamtaufwand aus Schritt (a)-(d) beträgt

$$2n + O(1) + 2n(n-1) + 2n(n-1) = 4n^2 + \tilde{p}(n)$$

Operationen, wobei $\tilde{p}$ ein Polynom vom Maximalgrad 1 bezeichnet.

Dieser Eliminationsschritt muss für die kleiner werdenden, nicht-eliminierten unteren Blockmatrizen durchgeführt werden und beträgt

$$\sum_{i=1}^{n-1} 4(n-i+1)^2 + \tilde{p}(n-i+1) = \frac{4}{3}n^3 + p(n)$$

Operationen, wobei p ein Polynom vom Maximalgrad 2 bezeichnet. ■

Bemerkung 3.7 *Der Aufwand, das faktorisierte Gleichungssystem* $\mathbf{QRx} = \mathbf{b}$ *nach* $\mathbf{x}$ *aufzulösen, beträgt* $3n^2$ *arithmetische Operationen.*

4 Eine weitere Anwendung des QR-Verfahrens: Lineare Ausgleichungsrechnung

In diesem Kapitel werden wir die Lösung des folgenden Problems betrachten. Sei f eine physikalische Grösse z.B. Spannung, Temperatur etc, die von verschiedenen Parametern abhängt z.B. den Ortskoordinaten, der Zeit, etc. Mathematisch formuliert, betrachten wir eine Abbildung $f : \mathbb{R}^k \to \mathbb{R}$. Wir nehmen an, dass die Funktion f nicht kontinuierlich sondern lediglich gemessene Wertepaare vorliegen

$$\left(\mathbf{y}^{(i)}, b^{(i)}\right) \quad \text{für } 1 \leq i \leq n.$$

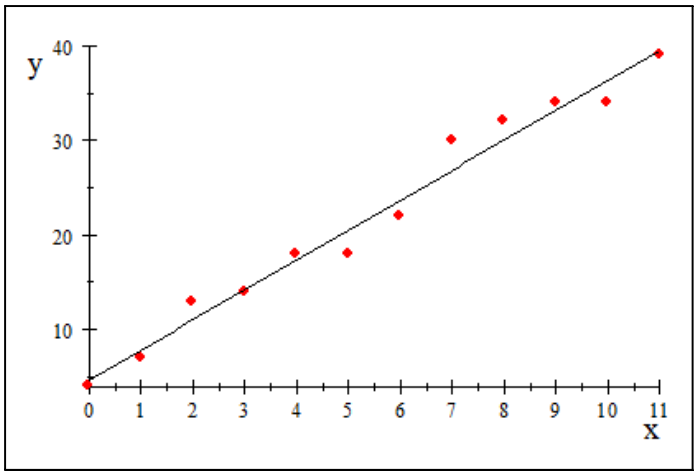
Hierbei bezeichnet $b^{(i)} \in \mathbb{R}$ den Messwert an Parameterpunkt $\mathbf{y}^{(i)} \in \mathbb{R}^k$, der mit $f\left(\mathbf{y}^{(i)}\right)$ bis auf Messfehler übereinstimmt.

Ziel ist es nun, eine lineare (genauer: affine) Abbildung

$$\tilde{f}(\mathbf{y}) = a_0 + \sum_{j=1}^{k} a_j y_j \tag{4.1}$$

zu finden, welche „möglichst gut“ die Wertepaare $\left(\mathbf{y}^{(i)}, b^{(i)}\right)$ approximiert. Gesucht ist hier der Koeffizientenvektor $\mathbf{a} = (a_j)_{j=0}^{k}$, welcher die lineare Abbildung festlegt. Die folgende Ab-

bildung illustriert ein charakteristisches Beispiel.



Linearer Fit von Messdaten.

Um den Begriff „möglichst gut“ mathematisch zu präzisieren, führen wir ein geeignetes Mass ein. Zunächst stellen wir fest, dass im idealen Fall

$$\tilde{f}\left(\mathbf{y}^{(i)}\right) = b^{(i)} \quad \forall 1 \leq i \leq n \tag{4.2}$$

gilt und alle Wertepaare exakt von der Funktion $\tilde{f}$ „getroffen“ werden. Setzt man den Ansatz (4.1) in (4.2) ein, ergibt sich

$$a_0 + \sum_{j=1}^{k} a_j y_j^{(i)} = b^{(i)} \quad \forall 1 \leq i \leq n. \tag{4.3}$$

Wir setzen $\mathbf{Y}^{(j)} := \left(y_j^{(i)}\right)_{i=1}^{n}$ für $1 \leq j \leq k$ und $\mathbf{b} := \left(b^{(i)}\right)_{i=1}^{n} \in \mathbb{R}^n$ und definieren die Matrix $\mathbf{A} \in \mathbb{R}^{n\times(k+1)}$ durch die Spaltenvektoren

$$\mathbf{A} = \left[\mathbf{1} \mid \mathbf{Y}^{(1)} \mid \mathbf{Y}^{(2)} \mid \mathbf{Y}^{(3)} \mid \ldots \mid \mathbf{Y}^{(k)}\right] \quad \text{mit} \quad \mathbf{1} = (1, 1, \ldots, 1) \ .$$

Damit lässt sich (4.3) auch kompakt schreiben gemäss

$$\mathbf{Aa} - \mathbf{b} = \mathbf{0}.$$

Für $n > (k+1)$ ist dieses lineare Gleichungssystem (LGS) überbestimmt, d.h., es existieren mehr Gleichungen als Unbekannte und das LGS besitzt im allgemeinen keine Lösung. In der linearen Ausgleichsrechnung besteht die Aufgabe darin, einen Vektor $\mathbf{a} = (a_i)_{i=0}^{k}$ zu finden, so dass $\mathbf{Aa} - \mathbf{b}$ „im quadratischen Mittel“ möglichst klein ist. Die Aufgabe lautet: Finde $\mathbf{a} \in \mathbb{R}^{k+1}$ so dass die Abbildung

$$F : \mathbb{R}^{k+1} \to \mathbb{R}, \qquad F(\mathbf{x}) := \sum_{i=1}^{n} (\mathbf{Ax} - \mathbf{b})_i^2$$

an der Stelle $\mathbf{x} = \mathbf{a}$ minimal ist:

$$F(\mathbf{a}) \leq F(\mathbf{x}) \quad \forall \mathbf{x} = (x_i)_{i=0}^{k} \in \mathbb{R}^{k+1}.$$

Aus der Analysis ist bekannt, dass eine notwendige Bedingung für eine Minimalstelle durch

$$\forall 0 \le i \le k: \quad \frac{\partial}{\partial a_i} F(\mathbf{a}) = 0$$

gegeben ist. Diese *Normalengleichungen* lauten explizit:

$$\forall 0 \le i \le k: \quad 2\sum_{j=1}^{n} A_{j,i} \left((\mathbf{Aa})_j - b_j \right) = 2\sum_{j=1}^{n} A_{j,i} \left(\sum_{\ell=1}^{k} A_{j,\ell} a_\ell - b_j \right) = 0.$$

Der Wert $\mathbf{a}$ ist dann charakterisiert durch

$$\mathbf{A}^{\cdot} \mathbf{Aa} = \mathbf{A}^{\cdot} \mathbf{b}. \tag{4.4}$$

Dies ist ein lineares Gleichungssystem mit quadratischer Matrix $\mathbf{A}^{\cdot} \mathbf{A} \in \mathbb{R}^{k \times k}$ und rechter Seite $\mathbf{A}^{\cdot} \mathbf{b} \in \mathbb{R}^k$.

Satz 4.1

a. *Das lineare Ausgleichsproblem: Suche $\mathbf{a} \in \mathbb{R}^{k+1}$, so dass*

$$\|\mathbf{Aa} - \mathbf{b}\|^2 = \min_{\mathbf{y} \in \mathbb{R}^{k+1}} \|\mathbf{Ay} - \mathbf{b}\|^2 =: \min_{\mathbf{y} \in \mathbb{R}^{k+1}} F(\mathbf{y}) \tag{4.5}$$

gilt, besitzt mindestens eine Lösung.

b. *Die Bilder verschiedener Lösungen $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{k+1}$ sind gleich: $\mathbf{Aa}_1 = \mathbf{Aa}_2$. Daher ist das Residuum $\mathbf{r} = \mathbf{b} - \mathbf{Aa}_1 = \mathbf{b} - \mathbf{Aa}_2$ eindeutig und genügt der Beziehung*

$$\mathbf{A}^{\cdot} \mathbf{r} = \mathbf{0}.$$

c. *Das Problem (4.4) ist äquivalent zum linearen Ausgleichsproblem (4.5).*

Beweis. Teil 1: Für beliebiges $\mathbf{b} \in \mathbb{R}^n$ besitzt die Normalengleichung eine Lösung. Das Bild von $\mathbb{R}^{k+1}$ unter $\mathbf{A}$ bezeichnen wir mit

$$\operatorname{Im} \mathbf{A} := \mathbf{A}\mathbb{R}^{k+1} := \left\{ \mathbf{Ax} : \mathbf{x} \in \mathbb{R}^{k+1} \right\}.$$

Damit lässt sich $\mathbb{R}^n$ als direkte Summe in der Form schreiben:

$$\mathbb{R}^n = \operatorname{Im} \mathbf{A} \oplus (\operatorname{Im} \mathbf{A})^{\perp}$$

mit dem orthogonalen Komplement

$$\begin{aligned} (\operatorname{Im} \mathbf{A})^{\perp} &= \{ \mathbf{v} \in \mathbb{R}^n \mid \forall \mathbf{w} \in \operatorname{Im} \mathbf{A}: \quad \langle \mathbf{v}, \mathbf{w} \rangle = 0 \} \\ &= \left\{ \mathbf{v} \in \mathbb{R}^n \mid \forall \mathbf{x} \in \mathbb{R}^{k+1}: \quad \langle \mathbf{A}^{\cdot} \mathbf{v}, \mathbf{x} \rangle = 0 \right\} \\ &= \{ \mathbf{v} \in \mathbb{R}^n \mid \mathbf{A}^{\cdot} \mathbf{v} = 0 \}. \end{aligned} \tag{4.6}$$

Jeder Vektor $\mathbf{b} \in \mathbb{R}^n$ besitzt eine eindeutige Zerlegung:

$$\mathbf{b} = \mathbf{b}^{\parallel} + \mathbf{b}^{\perp} \qquad \text{mit} \quad \mathbf{b}^{\parallel} \in \operatorname{Im} \mathbf{A} \quad \text{und} \quad \mathbf{b}^{\perp} \in (\operatorname{Im} \mathbf{A})^{\perp}.$$

Wegen $\mathbf{b}^{\|} \in \operatorname{Im} \mathbf{A}$ existiert (mindestens) ein $\mathbf{x}^{\|} \in \mathbb{R}^{k+1}$ mit $\mathbf{b}^{\|} = \mathbf{A}\mathbf{x}^{\|}$. Wegen (4.6) gilt $\mathbf{A}^{\top}\mathbf{b}^{\perp} = \mathbf{0}$ und daher

$$\mathbf{A}^{\top}\mathbf{A}\mathbf{x}^{\|} = \mathbf{A}^{\top}\mathbf{b}^{\|} = \mathbf{A}^{\top}\mathbf{b}^{\|} + \mathbf{A}^{\top}\mathbf{b}^{\perp} = \mathbf{A}^{\top}\mathbf{b}.$$

Teil 2: Zwei Lösungen $\mathbf{x}_1$, $\mathbf{x}_2$ der Normalengleichung erfüllen $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2$.
Sei $\mathbf{x}$ eine Lösung der Normalengleichung (4.4). Dann gilt

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \underbrace{(\mathbf{b} - \mathbf{A}\mathbf{x})}_{=\mathbf{r}}.$$

Offensichtlich gilt $\mathbf{A}\mathbf{x} \in \operatorname{Im} \mathbf{A}$ und für alle $\mathbf{w} \in \mathbb{R}^{k+1}$ die Gleichheit $\langle \mathbf{r}, \mathbf{A}\mathbf{w} \rangle = \langle \mathbf{b} - \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{w} \rangle = \langle \mathbf{A}^{\top}\mathbf{b} - \mathbf{A}^{\top}\mathbf{A}\mathbf{x}, \mathbf{w} \rangle = 0$, also $\mathbf{r} \in (\operatorname{Im} \mathbf{A})^{\perp}$. Da die Aufspaltung in eine direkte Summe eindeutig ist, muss $\mathbf{A}\mathbf{x}$ für alle Lösungen übereinstimmen mit $\mathbf{b}^{\|}$.

Teil 3: Jede Lösung $\mathbf{x}$ der Normalengleichung minimiert das Funktional $F(\cdot)$ in (4.5).
Für jedes $\mathbf{w} \in \mathbb{R}^{k+1}$ gilt

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{w}\|^2 &= \|\mathbf{b} - \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{w}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 + 2\langle \mathbf{r}, \mathbf{A}(\mathbf{x} - \mathbf{w}) \rangle + \|\mathbf{A}(\mathbf{x} - \mathbf{w})\|^2 \\ &= \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 + \|\mathbf{A}(\mathbf{x} - \mathbf{w})\|^2 \geq \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2, \end{aligned}$$

und daher minimiert $\mathbf{x}$ das Funktional $F(\cdot)$.

Teil 4: Die Umkehrung von Teil 3 folgt aus der Herleitung der Normalengleichung (4.4).
■

Der einfachste Fall liegt vor, wenn die Matrix $\mathbf{A}$ vollen Spaltenrang besitzt, also die Spalten von $\mathbf{A}$ linear unabhängig sind. Dann gilt

$$\mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}.$$

Die Matrix $\mathbf{A}^{\top}\mathbf{A}$ der Normalengleichung ist dann regulär, da die Symmetrie von $\mathbf{A}^{\top}\mathbf{A}$ offensichtlich ist und für alle $\mathbf{x} \neq \mathbf{0}$ gilt $\mathbf{A}\mathbf{x} \neq 0$ und

$$\langle \mathbf{x}, \mathbf{A}^{\top}\mathbf{A}\mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \|\mathbf{A}\mathbf{x}\|^2 > 0.$$

Die Minimallösung kann dann mit Hilfe des QR-Verfahren, angewendet auf die Normalengleichung (4.4), gelöst werden.

	Zürich 1901-1960	Zürich 1961-1990	Säntis 1961-1990	Jungfraujoch 1961-1990
Jan	−1.0	−0.5	−7.6	−13.6
Feb	0.2	0.9	−8.0	−14.2
März	3.9	4.2	−7.0	−13.1
April	7.7	7.9	−4.6	−10.8
Mai	12.1	12.2	−0.5	−6.6
Juni	15.0	15.4	2.4	−3.7
Juli	16.7	17.7	4.9	−1.2
Aug	16.0	16.8	4.9	−1.2
Sept	12.9	13.9	3.4	−2.6
Okt	7.8	9.2	1.0	−5.2
Nov	3.0	3.9	−4.2	−10.4
Dez	0.0	0.6	−6.4	−12.3

Tabelle 1: Gemittelte Klimatabelle

5 Schnelle Fouriertransformation (FFT)

In vielen Anwendungen der Statistik liegen Messdaten vor, aus denen empirische Gesetze herauszulesen sind. Gerade in der Signalverarbeitung oder bei der Ausbreitung akustischer oder elektromagnetischer Wellen (Hören, Handy-Empfang) sind die beschreibenden Funktionen eine (verrauschte) Überlagerung (Linearkombination) weniger signifikankter Wellen der Bauart $e^{i\lambda_n x}$. Die Fouriertransformation erlaubt es, aus gemessenen Signalen die wesentlichen Schwingungsanteile herauszufiltern. Diese reduzierte Information kann dann sehr effizient übermittelt werden und wird dann vom Empfänger mittels inverser Fouriertransformation zurückübersetzt.

Es sei hier betont, dass das effiziente Herausfiltern der wesentlichen Information aus komplizierten Signalen und deren schnelle Übertragung eine immense Bedeutung für viele Bereiche unseres Alltags besitzt.

5.1 Einleitendes Beispiel (Teil I)

In der Klimatologie interessiert man sich unter anderem für Temperaturmittelwerde über gewisse Zeitintervalle an verschiedenen Orten. Für einen festgewählten Ort, hängt dieser Wert noch vom Mittelungszeitraum ab. Je länger diese Epoche ist, desto weniger sollten die Mittelwerte um einen Gleichgewichtszustand schwanken. Berechnen wir an einem Ort beispielsweise die Temperatur gemittelt über einen festen Monat, so sind die Schwankungen dieses Wert über die vergangenen hundert Jahre relativ klein. Dennoch kann man daraus die durchschnittliche Erwärmung der Erde ablesen. Eine Mittelung über längere Zeiträume „verschmiert“ immer stärker die Detailinformation. So ist beispielsweise bekannt, dass der durchschnittliche Temperaturanstieg in den Sommermonaten Juli/August signifikant grösser ist als in den Wintermonaten Januar/Februar. Diese Information geht natürlich verloren, wenn über das ganze Jahr gemittelt wird.

Im folgenden Experiment betrachten wir langjährige Monatsmittel der Temperatur an verschiedenen Orten und für verschiedene zeitliche Mittel.

Die Fourier-Analyse wird uns erlauben, diese Daten miteinander zu vergleichen und physi-

kalisch zu interpretieren. Im folgenden Abschnitt werden wir die diskrete Fouriertransformation (DFT) einführen und danach die DFT zur Interpretation dieses Beispiels verwenden.

5.2 Diskrete Fouriertransformation

Konvention: Mit $\mathbb{R}$ wird die Menge der reellen und mit $\mathbb{C}$ die Menge der komplexen Zahlen bezeichnet. Die Komponenten eines Vektor $\mathbf{a} \in \mathbb{C}^n$ bzw. $\mathbf{a} \in \mathbb{R}^n$ werden mit a_i, $0 \leq i \leq n-1$, bezeichnet, und wir verwenden die Kurzschreibweise $\mathbf{a} = (a_i)_{i=0}^{n-1}$.

Jede komplexe Zahl $w \in \mathbb{C}$ besitzt eine eindeutige Darstellung $w = u + \mathrm{i}\, v$ mit $u, v \in \mathbb{R}$ und $\mathrm{i} = \sqrt{-1}$ der imaginären Einheit; u wird der Real- und v der Imaginärteil von w genannt:

$$u = \operatorname{Re} w, \qquad v = \operatorname{Im} w.$$

Die „komplexe Konjugation“ einer komplexen Zahl $w = u + \mathrm{i}\, v$ wird mit $\overline{w}$ bezeichnet und durch

$$\overline{w} := u - \mathrm{i}\, v$$

definiert.

Wir betrachten eine stetige, periodische Funktion $f : [a, b] \to \mathbb{C}$ (im folgenden „Signal“ genannt), die in regelmässigen Abständen abgetastet wird. Dies ergibt den Vektor

$$\mathbf{f} = (f_r)_{r=0}^{N-1} = (f(a + r\Delta t))_{r=0}^{N-1}$$

mit $\Delta t = (b-a)/N$. In der Praxis ist von der Funktion f nur der Vektor $\mathbf{f}$ und die Periodenlänge $(b-a)$ bekannt und das Ziel ist, eine Funktion mit Hilfe der Wertepaare $(a + r\Delta t, f_r)_{r=0}^{N-1}$ zu bestimmen, welche f approximiert. Der Ansatz ist durch Exponentialfunktionen gegeben

$$\tilde{f}(t) = \sum_{m=0}^{N-1} \hat{f}_m \, \mathrm{e}^{\mathrm{i}\, \frac{2\pi}{b-a} m t}.$$

Die Interpolationsbedingungen

$$f_r = \tilde{f}(r\Delta t) \qquad \text{für } 0 \leq r \leq N-1$$

führen auf die Bestimmungsgleichungen

$$f_r = \sum_{m=0}^{N-1} \hat{f}_m \, \mathrm{e}^{\mathrm{i}\, \frac{2\pi}{b-a} m r \Delta t}. \tag{5.1}$$

Für $m \in \{0, 1, 2, \ldots N-1\}$ definieren wir die Vektoren $\mathbf{w}_m \in \mathbb{C}^N$ durch $\mathbf{w}_m := \left(\mathrm{e}^{\mathrm{i}\, m \frac{2\pi}{b-a} r \Delta t}\right)_{r=0}^{N-1}$. Diese erfüllen die folgenden Orthogonalitätsrelationen bezüglich des Euklidischen Skalarprodukt[10]

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=0}^{N-1} u_k \overline{v_k} \qquad \forall \mathbf{u} = (u_k)_{k=0}^{N-1} \in \mathbb{C}^N, \quad \forall \mathbf{v} = (v_k)_{k=0}^{N-1} \in \mathbb{C}^N.$$

[10] Das Zeichen „$\forall$“ steht für „für alle“.

Satz 5.1 *Für $m, k \in \{0, 1, 2, \ldots N-1\}$ gilt*

$$\langle \mathbf{w}_m, \mathbf{w}_k \rangle = \begin{cases} N & m = k \\ 0 & m \neq k \end{cases}$$

Beweis. Es gilt

$$\langle \mathbf{w}_m, \mathbf{w}_k \rangle = \sum_{r=0}^{N-1} \mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a} mr\Delta t}\, \mathrm{e}^{-\mathrm{i}\,\frac{2\pi}{b-a} kr\Delta t} = \sum_{r=0}^{N-1} \alpha^r$$

mit $\alpha = \mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)\Delta t}$. Offensichtlich gilt $|\alpha| = 1$.

1.Fall: $\alpha = 1$, d.h., $m = k$. Dann gilt

$$\langle \mathbf{w}_m, \mathbf{w}_k \rangle = N.$$

2. Fall: $\alpha \neq 1$, d.h., $m \neq k$. Dann gilt mit $\Delta t = (b-a)/N$

$$\sum_{r=0}^{N-1} \alpha^r = \frac{1-\alpha^N}{1-\alpha} = \frac{1-\mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)\Delta t N}}{1-\mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)\Delta t}} = \frac{1-\mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)(b-a)}}{1-\mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)\Delta t}} = \frac{1-\left(\mathrm{e}^{2\pi\,\mathrm{i}}\right)^{(m-k)}}{1-\mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}(m-k)\Delta t}} = 0.$$

■

Die Interpolationsgleichung (5.1) lässt sich in der kompakten Form schreiben

$$\mathbf{f} = \mathbf{W}\hat{\mathbf{f}} \tag{5.2}$$

mit

$$\mathbf{f} = (f_i)_{i=0}^{N-1}, \qquad \hat{\mathbf{f}} = \left(\hat{f}_i\right)_{i=0}^{N-1}.$$

$\mathbf{W}$ bezeichnet die Matrix, welche die Vektoren $\mathbf{w}_m$, $0 \leq m \leq N-1$, als Spaltenvektoren besitzt. Man beachte, dass $\mathbf{W}$ symmetrisch aber nicht hermitesch ist. Aus den Orthogonalitätsrelationen folgt $\mathbf{W}^{\mathrm{H}}\mathbf{W} = N\mathbf{I}$ und daher aus (5.2) $\hat{\mathbf{f}} = \frac{1}{N}\mathbf{W}^{\mathrm{H}}\mathbf{f}$, d.h.,

$$\hat{f}_s = \frac{1}{N}\langle \mathbf{f}, \mathbf{w}_s \rangle = \frac{1}{N}\sum_{k=0}^{N-1} f_k\, \mathrm{e}^{-\mathrm{i}\,s\frac{2\pi}{b-a}k\Delta t} \tag{5.3}$$

Definition 5.2 *Für einen beliebigen Vektor $\mathbf{f} = (f_i)_{i=0}^{N-1} \in \mathbb{R}^N$ ist die diskrete Fouriertransformation (DFT) zur Periodenlänge $(b-a)$ durch*

$$\hat{f}_s = \frac{1}{N}\sum_{k=0}^{N-1} f_k\, \mathrm{e}^{-\mathrm{i}\,s\frac{2\pi}{b-a}k\Delta t} \tag{5.4}$$

gegeben.

Lemma 5.3 *Die inverse Fourier-Transformation ist durch*

$$f_s = \sum_{k=0}^{N-1} \hat{f}_k\, \mathrm{e}^{\mathrm{i}\,s\frac{2\pi}{b-a}k\Delta t}$$

gegeben.

Beweis. Es gilt

$$\sum_{k=0}^{N-1} \hat{f}_k \, \mathrm{e}^{\mathrm{i}\, s \frac{2\pi}{b-a} k\Delta t} = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} f_\ell \, \mathrm{e}^{-\mathrm{i}\, k \frac{2\pi}{b-a} \ell \Delta t} \, \mathrm{e}^{\mathrm{i}\, s \frac{2\pi}{b-a} k\Delta t} = \frac{1}{N} \sum_{\ell=0}^{N-1} f_\ell \left(\sum_{k=0}^{N-1} \mathrm{e}^{-\mathrm{i}\, k \frac{2\pi}{b-a} \ell \Delta t} \, \mathrm{e}^{\mathrm{i}\, s \frac{2\pi}{b-a} k\Delta t} \right).$$

Man beachte, dass der Klammerausdruck gleich dem Skalarprodukt $\langle \mathbf{w}_s, \mathbf{w}_\ell \rangle$ ist. Mit der Orthogonalitätsrelation aus Satz 5.1 erhalten wir daher das Gewünschte

$$\frac{1}{N} \sum_{\ell=0}^{N-1} f_\ell \, \langle \mathbf{w}_s, \mathbf{w}_\ell \rangle = f_s.$$

■

Bemerkung 5.4 *Man beachte, dass die Matrix* $\mathbf{W}$ *vollbesetzt ist und die naive Anwendung von (5.4) einer Matrix-Vektor-Multiplikation entstricht mit einem Aufwand von* $2N^2$ *arithmetischen Operationen.*

Lemma 5.5 *Die gegebenen Daten seien reell* $\mathbf{f} = (f_r)_{r=0}^{N-1} \in \mathbb{R}^N$.

a. Dann gilt für $1 \le s \le N-1$

$$\hat{f}_s = \overline{\hat{f}_{N-s}}.$$

Für gerades N *und* $s = N/2$ *gilt insbesondere* $\hat{f}_s \in \mathbb{R}$.

b. Dann interpoliert das trigonometrische Polynom

$$T(t) := \hat{A}_0 + \sum_{m=1}^{\lfloor \frac{N-1}{2} \rfloor} \left(\hat{A}_m \cos \frac{2\pi m t}{b-a} + \hat{B}_m \sin \frac{2\pi m t}{b-a} \right) + \begin{cases} \hat{A}_{\frac{N}{2}} \cos \frac{\pi N}{b-a} t & N \text{ gerade} \\ 0 & \text{sonst} \end{cases} \tag{5.5}$$

die Wertepaare $(r\Delta t, f_r)$, $r = 0, 1, \ldots, N-1$. *Die Koeffizienten sind hierbei durch* $\hat{A}_0 := \hat{f}_0$ *und für* $1 \le m \le \lfloor \frac{N-1}{2} \rfloor$

$$\hat{A}_m := \left(2 \operatorname{Re} \hat{f}_m \right), \quad \hat{B}_m := -\left(2 \operatorname{Im} \hat{f}_m \right) \tag{5.6}$$

gegeben. Für gerades N *wird* $\hat{A}_{\frac{N}{2}} := \hat{f}_{\frac{N}{2}}$ *gesetzt.*

Beweis. @ a: Die Rechenregeln für komplexe Konjugation ergeben

$$\overline{\hat{f}_{N-s}} = \frac{1}{N} \sum_{k=0}^{N-1} f_k \, \mathrm{e}^{\mathrm{i}(N-s) \frac{2\pi}{b-a} k\Delta t} = \frac{1}{N} \sum_{k=0}^{N-1} f_k \underbrace{\mathrm{e}^{\mathrm{i}\, 2\pi}}_{=1} \mathrm{e}^{-\mathrm{i}\, s \frac{2\pi}{b-a} k\Delta t} = \frac{1}{N} \sum_{k=0}^{N-1} f_k \, \mathrm{e}^{-\mathrm{i}\, s \frac{2\pi}{b-a} k\Delta t} = \hat{f}_s.$$

@ b: Sei $t = r\Delta t$ für ein $r \in \{0, 1, \ldots, N-1\}$. Wir verifizieren die Interpolationsbedingung. Für $1 \le m < N/2$ gilt

$$\begin{aligned} \hat{f}_m \, \mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} m t} + \hat{f}_{N-m} \, \mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} (N-m) t} &= \left(\operatorname{Re} \hat{f}_m \right) \left(\mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} m t} + \mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} (N-m) t} \right) \\ &\quad + \left(\mathrm{i} \operatorname{Im} \hat{f}_m \right) \left(\mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} m t} - \mathrm{e}^{\mathrm{i} \frac{2\pi}{b-a} (N-m) t} \right) \\ &= 2 \left(\operatorname{Re} \hat{f}_m \right) \cos \frac{2\pi m t}{b-a} - 2 \left(\operatorname{Im} \hat{f}_m \right) \sin \frac{2\pi m t}{b-a}. \end{aligned}$$

	Zürich 1901-1960	Zürich 1961-1990	Säntis 1961-1990	Jungfraujoch 1961-1990
$\hat{A}_0$	7.85833	8.51667	−1.80833	−7.90833
$\hat{A}_1$	−8.90393	−8.9645	−6.39046	−6.37602
$\hat{A}_2$	−0.025	−0.025	0.25	0.308333
$\hat{A}_3$	0.0666667	−0.0333333	0.266667	0.316667
$\hat{A}_4$	−0.0583333	0.0583333	0.233333	0.208333
$\hat{A}_5$	−0.0127363	−0.10217	−0.126208	−0.140642
$\hat{A}_6$	0.075	0.05	−0.025	−0.00833333
$\hat{B}_1$	−0.0688996	−0.510406	−2.24206	−2.26706
$\hat{B}_2$	0.418579	0.534049	0.288675	0.274241
$\hat{B}_3$	−0.116667	0.0333333	0.25	0.2
$\hat{B}_4$	−0.0721688	−0.0433013	−0.0288675	−0.101036
$\hat{B}_5$	−0.0977671	−0.106261	−0.307938	−0.332938

Tabelle 2: Fourierkoeffizienten der gemittelten Klimatabelle

Daraus folgt

$$\tilde{f}(t) = \sum_{m=0}^{N-1} \hat{f}_m \, \mathrm{e}^{\mathrm{i}\frac{2\pi}{b-a}mt} = \hat{A}_0 + \sum_{m=1}^{\lfloor\frac{N-1}{2}\rfloor} \left(\hat{A}_m \cos\frac{2\pi mt}{b-a} + \hat{B}_m \sin\frac{2\pi mt}{b-a} \right) + \gamma_N(t),$$

wobei der Koeffizient $\gamma_N(t)$ für gerades N durch

$$\gamma_N(t) := \hat{f}_{\frac{N}{2}} \left(\cos\frac{\pi N}{b-a}t + \mathrm{i}\sin\frac{\pi N}{b-a}t \right)$$

definiert ist und andernfalls $\gamma_N(t) = 0$ gesetzt wird. Da der Imaginärteil von $\gamma_N(t)$ in allen Stützstellen $r\Delta t$ verschwindet, interpoliert das trigonometrische Polynom $T(t)$ aus (5.5) die Wertepaare $(r\Delta t, f_r)$ für $0 \le r \le N-1$. ■

5.3 Einleitendes Beispiel (Teil II)

Wir wollen nun die Daten aus Tabelle 1 mit Hilfe der Fourier-Transformation analysieren und beginnen mit dem Verhalten des zeitlichen Verlaufs. Die Periodenlänge beträgt ein Jahr und die Zeitschritte einen Monat, d.h., $b-a = 1$ und $\Delta t = 1/12$. Damit gilt

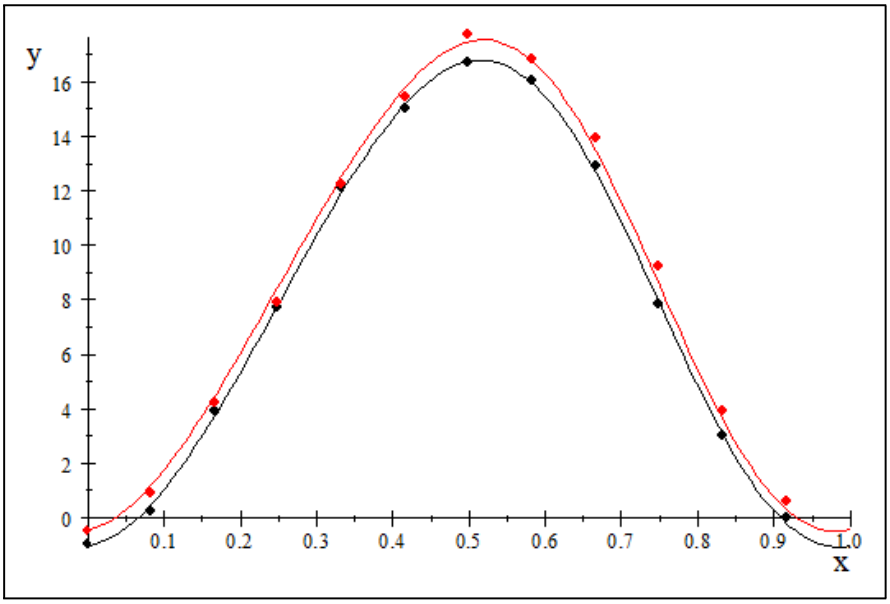
$$\mathbf{w}_m := \left(\mathrm{e}^{\frac{\mathrm{i}\,m2\pi r}{12}} \right)_{r=0}^{11}.$$

Die Matrix $\mathbf{W}$ ist eine 12×12 Matrix. Da die gegebenen Daten reell sind, verwenden wir ein trigonometrisches Polynom zur Interpolation. Die Kombination von (5.4) und Lemma 5.5 liefert nach expliziter Rechnung die Koeffizienten. Aus der Koeffiziententabelle liest man leicht ab, dass beispielsweise der Temperaturverlauf in Zürich zwischen 1901 und 1960 gut durch die Funktion

$$\tilde{f}(t) = 7.85833 - 8.90393\cos 2\pi t + 0.418579\sin 4\pi t$$

und für der Temperaturverlauf in Zürich zwischen 1961-1990 durch die Funktion

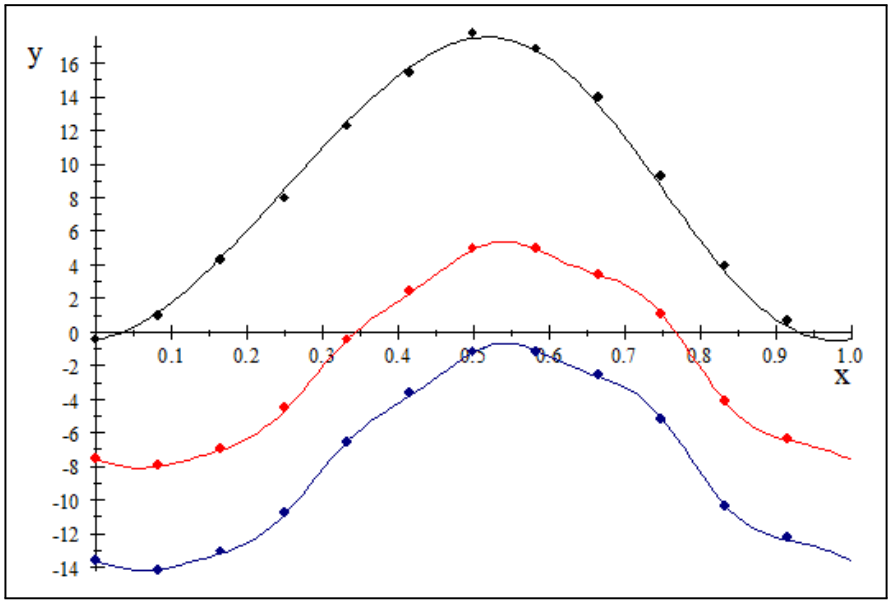
$$\tilde{g}(t) = 8.51667 - 8.9645\cos 2\pi t + 0.534049\sin 4\pi t$$



Trigonometrische Approximation des Temperaturverlaufs und gemessene Datenpunkte. Schwarz: Zeitraum 1901-1960. Rot: Zeitraum 1961-1990.

wiedergegeben wird (5.3) und anstelle von 12 Datenpunkten lediglich 3 Koeffizienten zur Wiedergabe des Verlaufs benötigt werden.

Die folgende Grafik vergleicht die Temperaturverläufe von Zürich (1961-1990) mit den entsprechenden Daten für den Säntis und das Jungfraujoch.



Trigonometrische Approximation des Temperaturverlaufs und gemessene Datenpunkte. Zeitraum 961-1990. Schwarz: Zuerich. Rot: Saentis. Blau: Jungfraujoch.

Zur physikalischen Interpretation der Fourierkoeffizienten:

Der Koeffizient $\hat{a}_0$ entspricht dem Jahresmittel. Es fällt auf, dass die Fourierkoeffizienten von je zwei Messreihen ausser $\hat{a}_0$ sehr gut übereinstimmen. Der mittlere Temperaturgang entspricht fast einer reinen Grundschwingung mit der Periode von einem Jahr. Im Vergleich zu den Messwerten in Zürich sind bei den beiden Bergstationen merkliche Anteile von höheren Frequenzen vorhanden und die Amplituden der Grundschwingung wesentlich geringer. Wir versuchen im folgenden die Beobachtungen physikalisch zu deuten: Die Landmassen im Mittelland erwärmen sich hauptsächlich durch die Sonneneinstrahlung und bestimmen die Monatsmittel der Lufttemperatur wesentlich. Vermutlich wird der mittlere Temperaturverlauf in Zürich wesentlich von der Sonneneinstrahlung und der Tageslänge diktiert. Gegenüber dem Sonnenstand (Maximum am 21.6.) hinkt das Temperatursignal hinterher. Die Verzögerung ist noch deutlicher erkennbar in den Grundschwingungen, die aus den Daten der Bergstationen abgeleitet wurde.

5.4 Die schnelle Fouriertransformation

Für das betrachtete Beispiel mit zwölf Datenpunkten spielt die Komplexität der Fouriertransformation keine Rolle. In der Praxis treten jedoch häufig Probleme auf, wo eine riesige Menge an Daten gemessen wurde und die charakteristischen Eigenschaften herausgefiltert werden sollen. Der Erfolg der Fouriertransformation zur Analyse und Kompression von Daten beruht wesentlich auf der Tatsache, dass ein schneller Algorithmus zur Fouriertransformation existiert.

In Bemerkung 5.4 haben wir festgestellt, dass eine "naive" Anwendung der Definition einen Aufwand von $O\left(N^2\right)$ arithmetischen Operationen benötigt. Dies würde dazu führen, dass sehr grosse Datenmengen auch mit modernsten Computern nicht analysierbar wären.

In diesem Abschnitt werden wir die schnelle Fouriertransformation behandeln, welche die Transformation mit einem Aufwand von $O\left(N \log N\right)$ berechnet. Dies ist ein Beispiel, wie das Verwenden hierarchischer, baumartiger Algorithmen die Komplexität einer Berechnung drastisch reduzieren kann.

Wir betrachten die Aufgabe für gegebene periodische Funktion $f : [0, 2\pi] \to \mathbb{C}$ und zugehörigen Koeffizienten $(f_k)_{k=0}^{N-1} \in \mathbb{C}^N$ die Fourierkoeffizienten

$$\hat{f}_j = \frac{1}{N} \sum_{k=0}^{N-1} f_k \omega_n^{jk}, \quad j = 0, 1, 2, \ldots, N-1 \tag{5.7}$$

zu berechnen, wobei wir

$$\omega_n := \mathrm{e}^{-2\pi \mathrm{i}/2^n}$$

gesetzt und $N = 2^n$ für eine Ganzzahl $n > 0$ angenommen haben. (Das Verfahren von Cooley und Tukey lässt sich am einfachsten erklären für den Fall $N = 2^n$, ist aber verallgemeinerbar für beliebiges N.)

Für die "Stufen" $m = 0, 1, \ldots, n$, setzen wir $N_m := 2^m$, $R_m := 2^{n-m}$ und betonen, dass die Stufe $m = n$ der globalen N-punktigen trigonometrischen Interpolation (5.7) entspricht. Das zugehörige Interpolationspolynom bezeichnen wir mit $p_0^{(n)}$. Die Herleitung des Algorithmus basiert auf der Interpolationseigenschaft:

$$p_0^{(n)}\left(t_k\right) = f_k, \quad k = 0, 1, \ldots, N-1$$

mit

$$t_k = 2\pi k/N,$$

des trigonometrischen Polynoms

$$p_0^{(n)}(t) := \sum_{k=0}^{N_n-1} \hat{f}_{0,k}^{(n)} \, \mathrm{e}^{\mathrm{i}\,kt}$$

mit den Fourier-Koeffizienten $\hat{f}_{0,k}^{(n)}$ aus (5.7)[11]. Ferner seien

$$p_0^{(n-1)}(t) := \sum_{k=0}^{N_{n-1}-1} \hat{f}_{0,k}^{(n-1)} \, \mathrm{e}^{\mathrm{i}\,kt},$$
$$p_1^{(n-1)}(t) := \sum_{k=0}^{N_{n-1}-1} \hat{f}_{1,k}^{(n-1)} \, \mathrm{e}^{\mathrm{i}\,kt}$$

diejenigen trigonometrischen Polynome der halben Ordnung N_{n-1} :

$$p_0^{(n-1)}(t_{2\ell}) = f_{2\ell}, \quad p_1(t_{2\ell}) = f_{2\ell+1}, \quad \text{für } \ell = 0, 1, \ldots, N_{m-1} - 1.$$

Dann interpoliert $p_0^{(n-1)}(t)$ an allen *geradzahlig* indizierten Stützstellen (t_{2k}, f_{2k}) und $q(t) := p_1^{(n-1)}\left(t - \frac{2\pi}{N_m}\right) = p_1^{(n-1)}\left(t - \frac{\pi}{N_{m-1}}\right)$ an allen Stützpunkten (t_{2k+1}, f_{2k+1}) mit *ungeradem* Index. Wegen

$$\mathrm{e}^{\mathrm{i}\,t_k N_{n-1}} = \mathrm{e}^{\mathrm{i}\,\frac{2\pi}{N_n} k N_{n-1}} = \mathrm{e}^{\mathrm{i}\,\pi k} = \begin{cases} 1 & \text{für gerades } k, \\ -1 & \text{für ungerades } k \end{cases}$$

interpoliert daher das trigonometrische Polynom

$$\tilde{p}_0^{(n)}(t) := \left(\frac{1 + \mathrm{e}^{\mathrm{i}\,N_{n-1}t}}{2}\right) p_0^{(n-1)}(t) + \left(\frac{1 - \mathrm{e}^{\mathrm{i}\,N_{n-1}t}}{2}\right) p_1^{(n-1)}\left(t - \frac{\pi}{N_{n-1}}\right) \tag{5.8}$$

an allen Stützstellen: $\tilde{p}_0^{(n)}(t_k) = f_k$, $k = 0, 1, \ldots, N_n - 1$. Da es offensichtlich ein Polynom $(N_n - 1)$-ten Grades ist, stimmt $\tilde{p}_0^{(n)}$ mit $p_0^{(n)}$ überein. Wir haben so die Bestimmung von $p_0^{(n)}$ (d.h. der Koeffizienten $\hat{f}_{0,k}^{(n)}$) auf die Bestimmung zweier anderer Polynome halben Grades zurückgeführt. Dieses Vorgehen kann man natürlich wiederholen. Man bekommt so ein n-stufiges Rekursionsverfahren: Allgemein hat man auf jeder Stufe $m = 0, 1, \ldots, n$ nun R_m trigonometrische Polynome vom Grad $N_m - 1$ der Form

$$p_r^{(m)}(t) = \sum_{k=0}^{N_m-1} \hat{f}_{r,k}^{(m)} \, \mathrm{e}^{\mathrm{i}\,kt}, \quad r = 0, 1, \ldots, R_m - 1 \tag{5.9}$$

mit den Interpolationseigenschaften

$$p_r^{(m)}(t_{R_m\ell}) = f_{R_m\ell+r}, \quad \ell = 0, 1, \ldots, N_m - 1, \quad r = 0, 1, \ldots, R_m - 1. \tag{5.10}$$

Diese Polynome genügen analog zu (5.8) für $m = 1, 2, \ldots, n$ den Rekursionsformeln

$$2p_r^{(m)}(t) = \left(1 + \mathrm{e}^{\mathrm{i}\,N_{m-1}t}\right) p_r^{(m-1)}(t) + \left(1 - \mathrm{e}^{\mathrm{i}\,N_{m-1}t}\right) p_{R+r}^{(m-1)}\left(t - \frac{\pi}{N_{m-1}}\right), \quad r = 0, 1, \ldots, R_m - 1. \tag{5.11}$$

[11] Die Bedeutung der zusätzlichen Indizes (n) und 0 wird aus der folgenden Herleitung klar.

Die Polynome $p_r^{(0)}(t) = \hat{f}_{r,0}^{(0)}$, $r = 0, 1, \ldots, N-1$, können sofort angegeben werden: (5.10) ergibt

$$\hat{f}_{r,0}^{(0)} = f_r, \quad r = 0, 1, \ldots, N-1. \tag{5.12}$$

Für die Koeffizienten $\hat{f}_{r,k}^{(m)}$ der übrigen Polynome $p_r^{(m)}(t)$ erhält man durch Koeffizientenvergleich aus (5.11) die Rekursionsformeln:

$$\left.\begin{aligned} 2\hat{f}_{r,k}^{(m)} &= \hat{f}_{r,k}^{(m-1)} + \hat{f}_{R_m+r,k}^{(m-1)}\omega_m^k \\ 2\hat{f}_{r,N_{m-1}+k}^{(m)} &= \hat{f}_{r,k}^{(m-1)} - \hat{f}_{R_m+r,k}^{(m-1)}\omega_m^k \end{aligned}\right\} \text{ für } \left\{\begin{array}{l} m = 1, 2, \ldots, n \\ r = 0, 1, \ldots, R_m - 1; \\ k = 0, 1, \ldots, N_{m-1} - 1. \end{array}\right. \tag{5.13}$$

Ausgehend von der Startwerten (5.12) liefern diese Formeln für $m = 1, 2, \ldots, n$ schliesslich die gesuchten Koeffizienten $\hat{f}_k = \hat{f}_{0,k}^{(n)}$, $k = 0, 1, \ldots, N-1$. Zur praktischen Realisierung dieses Verfahrens benötigt man wiederum nur einen Vektor $\tilde{\mathbf{f}} = \left(\tilde{f}_i\right)_{i=0}^{N-1}$ zur Speicherung der Koeffizienten $\hat{f}_{r,k}^{(m)}$, wenn man nach der Auswertung von (5.13) $\hat{f}_{r,k}^{(m-1)}$ und $\hat{f}_{R_m+r,k}^{(m-1)}$ durch $\hat{f}_{r,k}^{(m)}$ bzw. $\hat{f}_{r,N_{m-1}+k}^{(m)}$ überschreibt. Die folgende Permutationsabbildung wird Bitumkehrung genannt und ist für eine Ganzzahl z mit Binärdarstellung

$$z = \alpha_0 + \alpha_1 2 + \ldots + \alpha_{n-1} 2^{n-1}, \quad a_j \in \{0, 1\}$$

durch

$$\rho(z) := \alpha_{n-1} + \alpha_{n-2} 2 + \ldots + \alpha_0 2^{n-1}$$

definiert. Wir permutieren den *Startvektor* $\tilde{\mathbf{f}}$ entsprechend $\rho(r)$

$$\tilde{f}_{\rho(r)} = f_r, \quad r = 0, 1, \ldots, N-1.$$

Das folgende Programm der schneller Fourier-Transformation liefert dann schliesslich

$$\tilde{f}_k = N\hat{f}_k, \quad k = 0, 1, \ldots N-1$$

in der natürlichen Anordnung. (Der Faktor $N = 2^n$ erklärt sich daraus, dass im folgenden Programm der Faktor 2 in (5.13) weggelassen wurde.

```
for m := 1 to n do
begin
  for k := 0 to 2^(m-1) - 1 do
  begin
    e := ω_m^k;
    for r := 0 to 2^n - 1 step 2^m do
    begin
                    u := f̃_{r+k}          v := e × f̃_{r+k+2^(m-1)}
                    f̃_{r+k} := u + v      f̃_{r+k+2^(m-1)} := u - v;
    end
  end
end;
```

5.5 Weitere Anwendungen: Die schwingende Saite und die gedämpfte Schwingung eines Massepunktes

Wir betrachten eine Seite (beispielsweise eine Geigenseite) der Länge $a > 0$, die in den Punkten 0 und a eingespannt ist. Die Seite kann durch Zupfen oder Streichen zum Schwingen angeregt werden, und wir bezeichnet die Auslenkung zum Zeitpunkt $t \geq 0$ im Punkt $x \in (0, a)$ mit $u(x, t)$. Um diese Auslenkung zu berechnen, benötigen wir das zugehörige physikalische Bewegunsgesetz, welches aus den Newtonschen Kraftgesetzen hergeleitet werden kann,

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2} \qquad x \in (0, a)\,,\, t \in [0, T]\,, \tag{5.14}$$

wobei T einen festen Endzeitpunkt und α eine positive Materialkonstante bezeichnet. Diese Gleichung enthält Ableitungen nach mehreren Variablen (t, x) und wird daher *partielle Differentialgleichung* genannt. Um die Gleichung zu lösen, verwenden wir einen *Separationsansatz*

$$u(x, t) = v(x)\, w(t)\,,$$

welcher für partiellen Differentialgleichungen geeignet ist, bei denen die Variablen in einem Rechtecksgebiet $(0, a) \times (0, T)$ liegen und bei denen die Gleichung (5.14) linear in der Funktion u ist (also keine Potenzen u^2, $\left(\frac{\partial u}{\partial x}\right)^3$, etc. enthält) und homogen ist (also keine Funktion f als weiteren additiven Term in (5.14) enthält). Einsetzen des Separationsansatzes in (5.14) ergibt

$$v(x)\, \ddot{w}(t) = \alpha^2 v''(x)\, w(t)\,. \tag{5.15}$$

Hier bezeichnet der doppelte Punkt auf einer Funktion die zweite Zeitableitung und der Doppelstrich die zweite Ableitung nach dem Ort x. Falls $w(t) \neq 0$ und $v(x) \neq 0$ gilt, lässt sich (5.15) auch in der Form

$$\frac{\ddot{w}(t)}{w(t)} = \alpha^2 \frac{v''(x)}{v(x)}$$

schreiben. Die linke Seite dieser Gleichung hängt nur von t ab und die rechte nur von x. Gleichheit für alle x und t kann daher nur gelten, falls die jeweiligen Quotienten konstant sind:

$$\frac{v''(x)}{v(x)} = -\lambda \qquad \text{und} \qquad \frac{\ddot{w}(t)}{w(t)} = -\alpha^2 \lambda.$$

Das sind zwei *gewöhnliche* Differentialgleichungen (d.h. Differentialgleichungen für Funktionen mit einer Variablen)

$$v''(x) = -\lambda v(x) \quad x \in (0, a) \tag{5.16a}$$
$$\ddot{w}(t) = -\alpha^2 \lambda w(t) \quad t \in [0, T]\,. \tag{5.16b}$$

Die Vorgabe, dass die Saite in den Punkten 0 und a eingespannt ist, spiegelt sich in den *Randbedingungen* wieder:

$$u(0, t) = v(0)\, w(t) = 0 \quad \text{und} \quad u(a, t) = v(a)\, w(t) = 0 \qquad \forall t \in [0, T]\,. \tag{5.17}$$

Da wir uns nicht für die Nulllösung ($u(x, t) = 0$ für alle x, t) interessieren, folgt dass $t \in [0, T]$ existiert mit $w(t) \neq 0$. Daher ist (5.17) äquivalent zu

$$v(0) = 0 \quad \text{und} \quad v(a) = 0, \tag{5.16c}$$

und das sind die Randbedingungen für die Gleichung (5.16a). Gleichung (5.16a) mit (5.16c) hat nicht für jedes $\lambda \in \mathbb{R}$ eine nichttriviale Lösung (d.h. eine Lösung, die verschieden von der Nulllösung ist). Mit der Produktregel für Integrale folgt für $v \neq 0$ (hier bezeichnet 0 die Nullfunktion auf $(0, a)$)

$$\lambda \int_0^a v^2 = -\int_0^a vv'' = -\left. vv' \right|_0^a + \int_0^a (v')^2 = \int_0^a (v')^2 > 0,$$

da die Funktion v nicht gleichzeitig konstant, Nullrandwerte haben und von der Nullfunktion verschieden sein kann. In der Theorie der gewöhnlichen Differentialgleichung wird bewiesen, dass sich jede Lösung von (5.16a) als Linearkombination *zweier* linear unabhängiger Lösungen von (5.16a) schreiben lässt. Einfache Rechnung zeigt, dass die Funktionen $v_1(x) = \cos\left(\sqrt{\lambda}x\right)$ und $v_2(x) = \sin\left(\sqrt{\lambda}x\right)$ zwei linear unabhängige Lösungen von (5.16a) sind und die allgemeine Lösung die Form

$$v(x) = A\cos\left(\sqrt{\lambda}x\right) + B\sin\left(\sqrt{\lambda}x\right)$$

besitzt. Einsetzen in die Randbedingungen liefert die Bedingungen

$$A = 0 \qquad \text{und} \qquad A\cos\sqrt{\lambda}a + B\sin\sqrt{\lambda}a = 0,$$

d.h. $A = 0$ und

$$B\sin\sqrt{\lambda}a = 0. \tag{5.18}$$

Die Wahl $B = 0$ würde wiederum auf die Nulllösung führen, und wir betrachten daher den Fall $B \neq 0$. Dann folgt aus (5.18)

$$\sqrt{\lambda}a = k\pi \qquad \forall k = 1, 2, \ldots$$

Wir haben damit die Lösungen von (5.16a), (5.16c) bestimmt:

$$v(x) = B\sin\left(\frac{k\pi}{a}x\right).$$

Wir müssen daher die zweite Gleichung (5.16b) nur für diese Werte $\lambda_k = (k\pi/a)^2$ lösen. Die Überlegungen für die Funktion v lassen sich übertragen auf die Funktion w, und wir erhalten

$$w(t) = C\cos\alpha\sqrt{\lambda_k}t + D\sin\alpha\sqrt{\lambda_k}t.$$

Insgesamt haben wir gezeigt, dass für $k \in \mathbb{N}_{\geq 1}$ die Funktionen

$$u_k(x,t) := \sin\left(\sqrt{\lambda_k}x\right)\left(\tilde{C}\cos\left(\alpha\sqrt{\lambda_k}t\right) + \tilde{D}\sin\left(\alpha\sqrt{\lambda_k}t\right)\right) \quad \text{mit} \quad \lambda_k = (k\pi/a)^2$$

die Gleichung (5.14) mit Randbedingungen (5.17) lösen. Vom praktischen Standpunkt ist diese Lösung noch nicht ganz befriedigend. Interessanter ist die Frage, wie die eingespannte Saite schwingt, wenn deren Ausgangslage und die durch Zupfen oder Streichen mitgegebene Anfangsgeschwindigkeit vorgegeben ist:

$$u(x,0) = g(x) \quad \text{und} \quad \frac{\partial u}{\partial t}(x,0) = h(x) \quad \forall x \in (0,a).$$

Um dieses Problem zu lösen, stellen wir zunächst fest, dass jede Linearkombination aus Funktionen u_k die Gleichung (5.14) mit Randbedingungen (5.17) löst. Das führt zum Ansatz

$$u(x,t) = \sum_{k=1}^{\infty} u_k(x,t) = \sum_{k=1}^{\infty} \sin\left(\sqrt{\lambda_k}x\right)\left(C_k \cos\left(\alpha\sqrt{\lambda_k}t\right) + D_k \sin\left(\alpha\sqrt{\lambda_k}t\right)\right). \qquad (5.19)$$

Dieser Ansatz macht nur Sinn, falls die unbekannten Koeffizienten C_k, D_k so gewählt werden können, dass die Reihe konvergiert und gliedweise zweimal differenzierbar sind. Die Anfangsbedingungen ergeben dann

$$\sum_{k=1}^{\infty} \sin\left(\sqrt{\lambda_k}x\right) C_k = g(x) \quad \text{und} \quad \alpha \sum_{k=1}^{\infty} D_k \sqrt{\lambda_k} \sin\left(\sqrt{\lambda_k}x\right) = h(x) \quad \forall x \in (0,a). \qquad (5.20)$$

Mit diesem mathematischen Zugang sind wir also auf das Problem gestossen, gegebene „physikalische“ Funktionen in trigonometrische Reihen zu entwickeln und durch Koeffizientenvergleich die gesuchten Konstanten C_k, D_k abzulesen. Wegen der Periodizität der Sinus- und Kosinusfunktion beschränken wir uns auf periodische Anfangsbedingungen, d.h., $g(0) = g(a) = 0$ und $h(0) = h(a) = 0$ („$= 0$“ wegen der Einspannbedingung) . Der Ansatz (5.19) enthält unendlich viele Summanden, so dass selbst bei bekannten Koeffizienten C_k, D_k die Funktion $u(x,t)$ im Allgemeinen nicht in endlicher Zeit ausgewertet werden kann. Indem die Summe im Ansatz (5.19) durch eine endliche Summe ersetzt wird:

$$u^m(x,t) = \sum_{k=1}^{m} \sin\left(\sqrt{\lambda_k}x\right)\left(C_k \cos\left(\alpha\sqrt{\lambda_k}t\right) + D_k \sin\left(\alpha\sqrt{\lambda_k}t\right)\right)$$

können jedoch die Gleichheiten in (5.20) im Allgemeinen nicht gelten. Die Idee ist nun die Funktion g in geeigneten Punkten zu interpolieren. Der Einfachheit halber nehmen wir für das folgende $a = \pi$ an, so dass $\sqrt{\lambda_k} = k$ gilt. Wir zerlegen das Intervall $[0,\pi]$ in Gitterpunkte $x_k = 2\pi\frac{k}{N}$ für $k = 0,1,2,\ldots N-1$ und bestimmen die Fourierkoeffizienten $(\hat{g}_k)_{k=0}^{N-1}$ im trigonometrischen Polynom

$$p(x) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{g}_k \, \mathrm{e}^{\mathrm{i}\,\frac{2\pi}{b-a}kx},$$

so dass die Interpolationsbedingung

$$p(x_k) = g_k \quad \forall 0 \le k \le N-1$$

erfüllt ist. Dann werden die Koeffizienten im Sinus-Kosinus-Polynom (vgl. (5.5)) gemäss (5.6) bestimmt. Die Koeffizienten zu den Kosinustermen müssen (näherungsweise) verschwinden und werden daher weggelassen. Durch diesen *Ansatz* erhält man also eine numerische Approximation der Lösung der schwingenden Saite.

Ein ähnliches Problem, welches sich durch einen Fourierreihenansatz näherungsweise lösen lässt, wird durch eine gedämpfte Schwingung unter dem Einfluss periodischer Zwangskräfte beschrieben. Wir stellen uns einen metallenen Massepunkt an einer Feder aufgehängt vor, der periodischen Kräften ausgesetzt ist – beispielsweise in einem periodischen äusseren magnetischem Feld schwingt – und wollen dessen Auslenkung $u(t)$ zum Zeitpunkt $t \in [0,T]$ zu bestimmen (damit stellt $\dot{u}(t)$ die Geschwindigkeit des Massepunktes zum Zeitpunkt t dar und $\ddot{u}(t)$

die Beschleunigung.) Die Bewegung eines solchen elastisch angebundenen und Reibungskräften unterliegenden Punktes mit der Masse m wird durch die gewöhnliche Differentialgleichung

$$m\ddot{u} = -k^2 u - r\dot{u} + K(t)$$

beschrieben, wobei $m > 0$ die Masse, $k > 0$ eine Materialkonstante für die Feder, $r > 0$ den inneren Reibungskoeffizienten und $K(t)$ die periodische äussere Kraft darstellt. Mit den Abkürzungen $\rho = r/(2m)$, $\omega_0 = k/\sqrt{m}$ lässt sich diese Differentialgleichung auch in der Form schreiben

$$\ddot{u} + 2\rho\dot{u} + \omega_0^2 u = \frac{1}{m} K(t).$$

Die Periodizität von K wird durch die Bedingung $K(T) = K(0)$ wiedergespiegelt. Der Einfachheit halber nehmen wir an, dass die Periodenlänge $T = 2\pi$ erfüllt. Wir betrachten hier lediglich den Fall einer geraden Zwangskraft[12] (vgl. (5.5))

$$\frac{1}{m} K(t) := T_\infty(t) := \sum_{k=0}^{\infty} a_k \cos(kt). \tag{5.21}$$

Man beachte, dass die endliche Reihe $T_N(t)$ dann eine Approximation von $\frac{1}{m}K$ darstellt, deren Koeffizienten wieder durch die Interpolationsbedingung mittels schneller Fouriertransformation gewonnen werden können. Die Idee ist nun Lösungen der Gleichung

$$\ddot{u}_k + 2\rho\dot{u}_k + \omega_0^2 u_k = \cos(kt)$$

für alle $k \in \mathbb{N}$ zu bestimmen, so dass die exakte Lösung durch die Linearkombination

$$u(t) = \sum_{k=0}^{\infty} a_k u_k(t) \tag{5.22}$$

gegeben ist.

Bemerkung 5.6 *Man beachte, dass die Konvergenz von Funktionenreihen wie beispielsweise (5.21) vom Abklingen der Koeffizienten a_k abhängt. Offensichtlich gilt*

$$|a_k \cos(kt)| \leq |a_k| \quad \forall t,$$

so dass aus dem Majorantenkriterium die Konvergenz von (5.21) für jedes t aus der absoluten Konvergenz der Reihe $\sum a_k$ gefolgert werden kann.

Die detaillierte Konvergenztheorie von unendlichen Fourierreihen würde den Rahmen dieser Vorlesung sprengen. Wir verwenden diese lediglich als formale Reihen, die für Berechnungen immer durch eine endliche Summe ersetzt werden.

Für $k = 0$ gilt

$$\ddot{u}_0 + 2\rho\dot{u}_0 + \omega_0^2 u_0 = 1$$

und wir sehen, dass die konstante Funktion

$$u_0 = \omega_0^{-2}$$

[12] Der Fall einer ungeraden Zwangskraft lässt sich analog behandeln und daraus dann die Lösung für den allgemeinen Fall gewinnen.

die Gleichung löst. Für $k \geq 1$ müssen wir die Gleichung

$$\ddot{u}_k + 2\rho\dot{u}_k + \omega_0^2 u_k = \cos(kt) \tag{5.23}$$

lösen. Der Ansatz

$$u_k = A_k \cos(kt) + B_k \sin(kt),$$

eingesetzt in (5.23), ergibt daher die Bestimmungsgleichung

$$\begin{aligned} -A_k k^2 \cos(kt) - B_k k^2 \sin(kt) + 2\rho\left(-A_k k \sin(kt) + B_k k \cos(kt)\right) \\ + \omega_0^2\left(A_k \cos(kt) + B_k \sin(kt)\right) = \cos(kt). \end{aligned}$$

Sortieren nach Kosinus- und Sinustermen liefert

$$\left(-A_k k^2 + 2\rho B_k k + \omega_0^2 A_k - 1\right)\cos(kt) + \left(-B_k k^2 - 2\rho A_k k + \omega_0^2 B_k\right)\sin(kt) = 0.$$

Für A_k, B_k erhalten wir also folgendes lineares Gleichungssystem

$$\begin{bmatrix} -k^2 + \omega_0^2 & 2\rho k \\ -2\rho k & -k^2 + \omega_0^2 \end{bmatrix} \begin{pmatrix} A_k \\ B_k \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

mit der exakten Lösung

$$A_k = \frac{\omega_0^2 - k^2}{\left(\omega_0^2 - k^2\right)^2 + 4k^2\rho^2} \quad \text{und} \quad B_k = \frac{2k\rho}{\left(\omega_0^2 - k^2\right)^2 + 4k^2\rho^2}.$$

Damit haben wir gezeigt, dass die Lösung der gedämpften Schwingungsgleichung unter periodischen Zwangskräften (formal) die Entwicklung

$$u(t) = \sum_{k=0}^{\infty} \frac{a_k}{\left(\omega_0^2 - k^2\right)^2 + 4k^2\rho^2}\left(\left(\omega_0^2 - k^2\right)\cos(kt) + 2k\rho\sin(kt)\right)$$

besitzt. Indem die Summe bei einem Index $N-1$ abgebrochen wird, erhält man eine numerische Näherung der exakten Lösung.

Zusammenfassend erhält man auch für dieses Anwendungsbeispiel durch trigonometrische Approximation der rechten Seite $\frac{1}{m}K$ mit Hilfe der schnellen Fouriertransformation zunächst die Koeffizienten $(a_k)_{k=0}^{N-1}$ und durch

$$u^N(t) = \sum_{k=0}^{N-1} \frac{a_k}{\left(\omega_0^2 - k^2\right)^2 + 4k^2\rho^2}\left(\left(\omega_0^2 - k^2\right)\cos(kt) + 2k\rho\sin(kt)\right)$$

eine Approximation u^N der exakten Lösung u.

6 Numerische Integration und Differentiation

In diesem Kapitel werden wir die numerische Approximation von Integralen und Ableitungen behandeln. Wir werden sehen, dass die meisten dieser Näherungsverfahren auf dem Prinzip basieren:

> "Ersetze den Integranden bzw. die Funktion durch eine *Approximation* (vgl. Kap. 2) und *approximiere* das Integral bzw. die Ableitung der Funktion durch das exakte Integral bzw. die exakte Ableitung der *Approximation*."

6.1 Numerische Integration

Beispiel 6.1 *Probleme in der Akustik –genauer der Schallabstrahlung– werden häufig mit Integralgleichungen modelliert. Die Diskretisierung erfolgt mit der Randelementmethode und der rechenintensivste Schritt ist das Aufstellen und Lösen eines linearen Gleichungssystem mit einer sehr grossen, vollbesetzten Systemmatrix* $\mathbf{A}$*, d.h.,* $n = \dim \mathbf{A} \sim 10^4 - 10^5$*. Die Einträge dieser Matrix sind durch Integrale gegeben:*

$$A_{i,j} = \int_{\Delta_i} \int_{\Delta_j} \frac{e^{\mathrm{i}\,\kappa\|x-y\|}}{4\pi\,\|x-y\|} ds_x ds_y \qquad 1 \le i,j \le n.$$

Dabei bezeichnen Δ_i*,* Δ_j *im Allgemeinen gekrümmte Dreiecke im Raum. Diese Integrale lassen sich nicht exakt berechnen. Durch geeignete Variablentransformationen lässt sich dieses Problem auf die Aufgabe zurückführen, das parameterabhängige Integral*

$$\int_0^1 \frac{e^{ik\sqrt{r^2+\delta^2}}}{\sqrt{r^2+\delta^2}} r dr$$

für alle Werte von δ *mit einer vorgegebenen Genauigkeit effizient zu approximieren. Die Bedeutung der Effizienz hängt mit den* n^2 *Aufrufen dieses numerische Quadraturverfahren im Computerprogramm zusammen.*

6.1.1 Newton-Cotes-Formeln

Die Newton-Cotes-Formeln basieren auf der Approximation des Integranden durch Interpolation. Wir beginnen mit den in der Praxis am häufigsten verwendeten Methoden: *Trapezregel* und *Simpsonregel*.

In diesem Abschnitt bezeichnet I immer ein reelles Intervall $I = [a,b] \subset \mathbb{R}$ und $f \in C^0(I)$. Wir verwenden eine äquidistante Zerlegung des Intervalls I. Sei dazu $n \in \mathbb{N}$ und $h := (b-a)/n$ die *Schrittweite*. Die Stützstellen werden mit $x_i = a + hi$, $0 \le i \le n$, bezeichnet. Dann gilt

$$\int_I f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx. \tag{6.1}$$

Summierte Trapezregel:

Die lineare Interpolation der Funktion f im Intervall (x_{i-1}, x_i) ist durch

$$p_{1,i}[f](x) = \frac{x_i - x}{h} f(x_{i-1}) + \frac{x - x_{i-1}}{h} f(x_i) \qquad \forall x \in [x_{i-1}, x_i]$$

gegeben. Ersetzen von $f|_{[x_{i-1},x_i]}$ durch $p_i[f]$ in (6.1) ergibt die Approximation

$$\mathcal{I}(f) := \int_I f(x)\,dx \approx \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} p_{1,i}[f](x)\,dx = \sum_{i=1}^{n} \frac{h}{2}\{f(x_{i-1}) + f(x_i)\} = h\sum_{i=0}^{n} \alpha_i^T f(x_i) =: Q_T^n(f)$$

mit

$$\alpha_i^T := \begin{cases} 1 & 1 \le i \le n-1, \\ \frac{1}{2} & i = 0, n. \end{cases}$$

Für den Fehler dieser Approximation verwenden wir die Fehlerdarstellung aus (2.10)

$$R_{1,i}[f](x) = (x - x_i)(x - x_{i-1})\frac{f''(\xi_x)}{2} \quad \text{für ein } \xi_x \in (x_{i-1}, x_i)$$

und erhalten eine lokale Abschätzung des Quadraturfehlers durch

$$\left| \int_{x_{i-1}}^{x_i} R_{1,i}[f](x) \right| \le \frac{\|f''\|_{\max,[x_{i-1},x_i]}}{2} \int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1}) = \|f''\|_{\max,[x_{i-1},x_i]} \frac{h^3}{12}.$$

Für den globalen Fehler erhalten wir

$$|\mathcal{I}(f) - Q_T^n(f)| \le \sum_{i=1}^{n} \|f''\|_{\max,[x_{i-1},x_i]} \frac{h^3}{12} = \frac{h^3 n}{12} \|f''\|_{\max,I} = \frac{b-a}{12} h^2 \|f''\|_{\max,I}.$$

Satz 6.2 *Sei $f \in C^2(I)$. Dann konvergiert die summierte Trapezregel mit äquidistanten Stützstellen mit der Rate h^2 :*

$$|\mathcal{I}(f) - Q_T^n(f)| \le \frac{b-a}{12} h^2 \|f''\|_{\max,I}$$

mit $h = (b-a)/n$.

Summierte Simpsonregel:

Um eine Formel höherer Ordnung zu konstruieren, wird die Funktion f nun lokal quadratisch interpoliert. Dazu sind drei Stützstellen erforderlich. Für $x_i, 1 \le i \le n-1$ wählen wir dazu die Punkte x_{i-1}, x_i, x_{i+1} und bezeichnen die zugehörige quadratische Interpolation der Funktion f mit

$$p_{2,i}[f](x) = f_{i-1} + (x - x_{i-1})[x_{i-1}, x_i]f + (x - x_{i-1})(x - x_i)[x_{i-1}, x_i, x_{i+1}]f$$

Integration über $[x_{i-1}, x_{i+1}]$ liefert

$$\int_{x_{i-1}}^{x_{i+1}} p_{2,i}[f](x)\,dx = f_{i-1}(2h) + (2h^2)\frac{f_i - f_{i-1}}{x_i - x_{i-1}} + \left(\frac{2}{3}h^3\right)\frac{\frac{f_{i+1}-f_i}{x_{i+1}-x_i} - \frac{f_i - f_{i-1}}{x_i - x_{i-1}}}{x_{i+1} - x_{i-1}} = \frac{h}{3}\{f_{i-1} + 4f_i + f_{i+1}\}.$$

Um zu einer summierten Formel zu gelangen, nehmen wir an, dass n gerade ist und erhalten

$$\mathcal{I}(f) = \sum_{k=1}^{n/2} \int_{x_{2k-2}}^{x_{2k}} f(x)\,dx \approx \frac{h}{3}\sum_{k=1}^{n/2}\{f_{2k-2} + 4f_{2k-1} + f_{2k}\} =: Q_S^n(f).$$

Unter Berücksichtigung mehrfach auftretender Funktionswerte ergibt sich

$$Q_S^n(f) = \frac{h}{3}\{f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \ldots + 4f_{n-1} + f_n\} = \frac{h}{3}\sum_{i=0}^{n} \alpha_i^S f_i$$

mit

$$\alpha_i^S = \begin{cases} 4 & i \text{ ungerade,} \\ 2 & i \text{ gerade und } i \notin \{0, n\}, \\ 1 & i \in \{0, n\}. \end{cases}$$

Satz 6.3 *Sei $f \in C^4(I)$ und die Simpsonregel bezüglich n, n gerade, äquidistanten Stützstellen definiert. Dann gilt die Fehlerabschätzung:*

$$|\mathcal{I}(f) - Q_S^n(f)| \leq \frac{h^4}{180}(b-a)\left\|f^{(4)}\right\|_{\max,I}.$$

Beweis.

a) Fehlerabschätzung auf dem Einheitsintervall:

Wir betrachten zunächst das Einheitsintervall $I = [-1, 1]$ und wenden die Simpsonregel bezüglich der drei Punkte $x_0 = -1$, $x_1 = 0$ und $x_2 = 1$ an. Sei $\tilde{p} \in \mathbb{P}_3$ das Polynom, welches durch die Bedingungen

$$\begin{aligned} \tilde{p}(x_i) &= f(x_i) \quad 0 \leq i \leq 2 \\ \tilde{p}'(0) &= f'(0) \end{aligned}$$

eindeutig charakterisiert ist. Die Lösung erhält man am einfachsten mit der Verallgemeinerung der Newtonschen dividierten Differenzen auf die *Hermite-Interpolation*, bei der Ableitungen in den Stützstellen vorgegeben werden können:

$$\tilde{p}(x) = f_{-1} + (f_0 - f_{-1})(x+1) + (f_0' - f_0 + f_{-1})\,x(x+1) + (f_1 - 2f_0' - f_{-1})\frac{x^2(x+1)}{2}.$$

Die Integration über $[-1, 1]$ liefert

$$\int_{-1}^{1} \tilde{p}(x)\,dx = \frac{1}{3}(f_{-1} + 4f_0 + f_1).$$

Daraus folgt für die Simpsonregel Q_S^2 (beachte $h = 1$) die Fehlerdarstellung

$$\int_{-1}^{1} f(x)\,dx - Q_S^2(f) = \int_{-1}^{1} (f - \tilde{p})(x)\,dx.$$

Die Fehlerdarstellung (Satz 2.11) des Interpolationsfehlers liefert

$$|(f - \tilde{p})(x)| \leq \frac{|(x+1)\,x^2\,(x-1)|}{4!}\left\|f^{(4)}\right\|_{\max,[-1,1]}.$$

Integration über $[-1, 1]$ ergibt schliesslich die Fehlerabschätzung

$$\left|\int_{-1}^{1} f(x)\,dx - Q_S^2(f)\right| \leq \frac{1}{90}\left\|f^{(4)}\right\|_{\max,[-1,1]}. \tag{6.2}$$

Lokale Fehlerabschätzungen auf dem Intervall $[x_{i-1}, x_{i+1}]$

Die Fehlerabschätzung (6.2) übertragen wir nun auf das Intervall $[x_{i-1}, x_{i+1}]$ behelfs der Transformation

$$\chi_i : [-1,1] \to [x_{i-1}, x_{i+1}], \qquad \chi_i(t) = \frac{1-t}{2} x_{i-1} + \frac{t+1}{2} x_{i+1}.$$

Die transformierte Funktion wird mit $\hat{f} = f \circ \chi_i$ bezeichnet. Damit gilt

$$\begin{aligned}\int_{x_{i-1}}^{x_{i+1}} f(x)\,dx - Q_S^2(f) &= h \int_{-1}^{1} \hat{f}(t)\,dt - \frac{h}{3}\left(f_{i-1} + 4f_i + f_{i+1}\right) \\ &= h\left\{\int_{-1}^{1} \hat{f}(t)\,dt - \frac{1}{3}\left(\hat{f}(-1) + 4\hat{f}(0) + \hat{f}(1)\right)\right\}.\end{aligned}$$

Für die geschweifte Klammer können wir die Abschätzung (6.2) verwenden und erhalten

$$\left|\int_{x_{i-1}}^{x_{i+1}} f(x)\,dx - Q_S^2(f)\right| \le \frac{h}{90}\left\|\hat{f}^{(4)}\right\|_{\max,[-1,1]}.$$

Wir müssen die Norm auf der rechten Seite zurücktransformieren auf das Intervall $[x_{i-1}, x_{i+1}]$. Die Kettenregel liefert wegen der Linearität von χ_i die Beziehung

$$\hat{f}^{(4)} \circ \chi_i^{-1}(x) = f^{(4)}(x)\left(\frac{d\chi_i}{dt}\right)^4 = f^{(4)}(x)\,h^4.$$

Daraus folgt

$$\begin{aligned}\left|\int_{x_{i-1}}^{x_{i+1}} f(x)\,dx - Q_S^2(f)\right| &\le \frac{h}{90}\left\|\hat{f}^{(4)}\right\|_{\max,[-1,1]} = \frac{h}{90}\max_{x\in[x_{i-1},x_{i+1}]}\left|\hat{f}^{(4)} \circ \chi_i^{-1}(x)\right| \\ &= \frac{h^5}{90}\max_{x\in[x_{i-1},x_{i+1}]}\left|f^{(4)}(x)\right| = \frac{h^5}{90}\left\|f^{(4)}\right\|_{\max,[x_{i-1},x_{i+1}]}.\end{aligned}$$

Globale Fehlerabschätzung für den Quadraturfehler:

Summation über alle Teilintervalle ergibt die globale Fehlerabschätzung

$$|\mathcal{I}(f) - Q_S^n(f)| \le \frac{h^5}{90}\sum_{k=1}^{n/2}\left\|f^{(4)}\right\|_{\max,[x_{2k-2},x_{2k}]} \le \left\|f^{(4)}\right\|_{\max,I}\frac{h^5}{90} \times \frac{n}{2} = \frac{b-a}{180}h^4\left\|f^{(4)}\right\|_{\max,I}.$$

■

Bemerkung 6.4 *Man beachte, dass –hinreichende Glattheit der Funktion f vorausgesetzt– die summierte Simpsonregel um 2 Ordnungen schneller als die Trapezregel konvergiert bei gleichem Aufwand.*

Nach diesen einführenden Beispielen werden wir nun systematisch Quadraturformeln beliebig hoher Ordnung konstruieren.

Ziel: Approximiere das Integral über $I = [a, b]$ durch eine *Quadraturformel* der Form:

$$\mathcal{I}(f) := \int_a^b f(x)\,\omega(x)\,dx = \underbrace{\sum_{k=0}^{n} w_k f(x_k)}_{=:Q^n(f)} + E_n(f), \tag{6.3}$$

wobei $\omega : I \to \mathbb{R}$ eine **positive** Gewichtsfunktion bezeichnet. Der Zugang besteht darin, die Koeffizienten w_k und Stützstellen x_k so zu bestimmen, dass Polynome möglichst hohen Grades exakt integriert werden. Wir setzen dazu voraus, dass

$$\int_a^b \omega(x)\,x^k dx$$

für alle $k \in \mathbb{N}_0$ existiert.

Definition 6.5 *Eine Quadraturformel der Form (6.3) besitzt den Exaktheitsgrad k, falls*

$$\forall p \in \mathbb{P}_k : \quad E_n(p) = 0$$

gilt und mindestens ein $p \in \mathbb{P}_{k+1}$ existiert mit $E_n(p) \neq 0$.

Sei $f \in C^0(I)$. Für vorgegebene Stützstellenmenge $\Theta_n = \{x_i : 0 \leq i \leq n\}$ ist das Interpolationspolynom $p_n(f, \Theta_n)$ eindeutig bestimmt und besitzt die Lagrange-Darstellung

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

mit der Lagrangebasis

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

Für die Quadraturformel lässt sich daher der Ansatz

$$Q^n(f) = \int_a^b p_n(f, \Theta_n)(x)\,\omega(x)\,dx = \sum_{i=0}^{n} f(x_i) \int_a^b \ell_i(x)\,\omega(x)\,dx$$

verwenden. Diese Formel besitzt die Bauart wie in (6.3) mit den Quadraturgewichten

$$w_i := \int_a^b \ell_i(x)\,\omega(x)\,dx \qquad 0 \leq i \leq n. \tag{6.4}$$

Da die Interpolation $p_n(q, \Theta_n)$ eines Polynoms $q \in \mathbb{P}_n$ eindeutig ist, gilt $p_n(q, \Theta_n) = q$ und die Quadraturformel ist exakt für alle Polynome vom Maximalgrad n. Diese Überlegung führt auf die folgende Definition.

Definition 6.6 *Eine Quadraturformel ist interpolatorisch, falls der Exaktheitsgrad k die Beziehung*

$$k \geq n$$

erfüllt.

Bemerkung 6.7 *Für festgelegte Stützstellenmenge Θ_n sind die Gewichte einer interpolatorischen Quadraturformel eindeutig durch (6.4) festgelegt. Derartige Quadraturformeln nennt man* ***Newton-Cotes-Formeln****.*

Der allgemeine Rahmen für Fehlerabschätzungen benötigt den Begriff der Stabilität.

Definition 6.8 *Seien X, Y normierte Vektorräume. Eine lineare Abbildung $f : X \to Y$ heisst beschränkt, falls eine Konstante $C > 0$ existiert mit*

$$\forall x \in X : \quad \|f(x)\|_Y \leq C \|x\|_X .$$

Bemerkung 6.9 *Sei $I = [a, b] \subset \mathbb{R}$.*

a. *Die Abbildung $\mathcal{I} : C^0(I) \to \mathbb{R}$ aus (6.3) ist beschränkt mit $C = \int_a^b \omega(t)\, dt$.*

b. *Sei $Q^n : C^0(I) \to \mathbb{R}$ eine Quadraturformel mit Exaktheitsgrad $k \geq 0$ der Form (6.3). Dann ist Q^n beschränkt mit*

$$C = C_{Q,n} \int_I \omega(t)\, dt \quad \text{und} \quad C_{Q,n} := \frac{\sum_{i=0}^n |w_i|}{\sum_{i=0}^n w_i}. \tag{6.5}$$

Beweis. Teil a der Bemerkung folgt aus der Monotonie des Integrals. Für Teil b verwenden wir

$$|Q^n(f)| = \left| \sum_{i=0}^n w_i f(x_i) \right| \leq \max_{1 \leq k \leq n} |f(x_k)| \sum_{i=0}^n |w_i| \leq \|f\|_{\max,I} \left(\sum_{i=0}^n w_i \right) \left(\frac{\sum_{i=0}^n |w_i|}{\sum_{i=0}^n w_i} \right).$$

Da für den Exaktheitsgrad der Quadratur $k \geq 0$ gilt, folgt

$$\sum_{i=0}^n w_i = \int_I \omega(t)\, dt$$

und damit die Behauptung. ■

Die Konstante $C_{Q,n}$ in (6.5) wird *Stabilitätskonstante* der Quadraturformel genannt.

Satz 6.10 *Sei $I = [a, b] \subset \mathbb{R}$ ein Intervall. Die Quadraturformel Q^n besitze den Exaktheitsgrad $k \geq 0$. Dann gilt*

$$\forall f \in C^0(I) : \quad |\mathcal{I}(f) - Q^n(f)| \leq \left(\int_a^b \omega(t)\, dt \right) (1 + C_{Q,n}) \inf_{q \in \mathbb{P}_k} \|f - q\|_{\max,I} . \tag{6.6}$$

Beweis. Für $f \in C^0(I)$ bezeichnen wir den Quadraturfehler mit $E_n(f) = \mathcal{I}(f) - Q^n(f)$. Aus dem Exaktheitsgrad der Quadraturformel folgt

$$E_n(q) = 0 \qquad \forall q \in \mathbb{P}_k.$$

Offensichtlich ist $E_n : C^0(I) \to \mathbb{R}$ linear und beschränkt:

$$|E_n(f)| \leq |\mathcal{I}(f)| + |Q^n(f)| \leq \left(\int_a^b \omega(t)\, dt \right) (1 + C_{Q,n}) \|f\|_{\max,I}.$$

Zusammen ergibt sich für beliebiges $q \in \mathbb{P}_k$

$$|E_n(f)| = |E_n(f-q)| \leq \left(\int_a^b \omega(t)\, dt \right) (1 + C_{Q,n}) \|f-q\|_{\max,I}.$$

Da q beliebig war, können wir zum Infimum auf der rechten Seite übergehen und erhalten die Behauptung. ■

Bemerkung 6.11

a. *Man beachte, dass die Zahl der Quadraturpunkte und der Exaktheitsgrad einer Quadraturformel nicht unabhängig voneinander sind. (Genauer werden wir für alle interpolatorischen Formeln $n \leq k \leq 2n+1$ zeigen.) Der Weierstrasssche Approximationssatz besagt zwar, dass das Infimum in (6.6) gegen Null konvergiert aber nicht, dass die Konstante $C_{Q,n}$ beschränkt bleibt. Man kann zeigen, dass für äquidistante Gitterpunkte und interpolatorische Quadraturverfahren die Konstante $C_{Q,n}$ für $n \to \infty$ gegen Unendlich strebt und somit die Konvergenz des Quadraturfehlers durch Vergrössern des Exaktheitsgrades nicht gesichert ist. Die folgende Tabelle zeigt das Anwachsen der Stabilitätskonstanten $C_{Q,n}$ für $I = [0,1]$ und $\omega \equiv 1$*

n	2	3	4	5	6	7	8	9	10	11	...
$C_{Q,n}$	1	1	1	1	1	1	1.45	1	3.06	1.58	...

...	12	13	14	15	16	17	18	19	20
...	7.53	3.24	20.34	8.34	58.45	22.21	175.46	63.24	544.17

b. *Formeln hoher Ordnung konvergieren im allgemeinen nur dann schneller als Formeln niedriger Ordnung, falls die Funktionen hinreichend glatt sind.*

6.1.2 Gauss-Quadratur

In diesem Abschnitt werden wir die Frage behandeln, ob man durch eine geschickte Wahl der Stützstellen, einen Exaktheitsgrad $k > n$ erhalten kann.

Für $n+1$ Stützstellen $\Theta_n = \{x_i : 0 \leq i \leq n\} \subset I$ definieren wir das Stützstellenpolynom

$$\pi_n(x) = \prod_{i=0}^n (x - x_i).$$

Satz 6.12 *Sei $d \in \mathbb{N}_0$ gegeben. Eine Quadraturformel des Typs (6.3) besitzt den Exaktheitsgrad $k = n + 1 + d$ genau dann, wenn beide untenstehenden Bedingungen erfüllt sind:*

a. *Die Formel (6.3) ist interpolatorisch.*

b. *Das Stützstellenpolynom π_n erfüllt*

$$\int_a^b \pi_n(x)\, p(x)\, \omega(x)\, dx = 0 \qquad \forall p \in \mathbb{P}_d.$$

Bemerkung 6.13

1. *Bedingung (b) ist als Bedingung an die Stützstellen $x_i, 0 \leq i \leq n$, zu verstehen.*

2. *Sei $\omega > 0$ und $\omega \in C^0(]a,b[)$ eine Gewichtsfunktion, die $\int_a^b \omega(x)\, dx < \infty$ erfüllt. Dann ist durch*

$$(u,v)_\omega = \int_a^b u(x)\, v(x)\, \omega(x)\, dx$$

ein Skalarprodukt auf $C^0(I)$ gegeben. Zwei Funktionen u, v sind orthogonal bzgl. $(\cdot,\cdot)_\omega$, falls $(u,v)_\omega = 0$ gilt. Bedingung (b) besagt daher, dass das Stützstellenpolynom orthogonal zu allen Polynomen vom Grad d bzgl. des $(\cdot,\cdot)_\omega$-Skalarprodukts ist. Daraus folgt $\pi_n \notin \mathbb{P}_d$, woraus $d \leq n$ folgt. Der grösste Wert von d ist daher durch $d = n$ gegeben. Aus Satz 6.12 ergibt sich der maximale Exaktheitsgrad von $k = 2n+1$ für eine Quadraturformel basierend auf $n+1$ Stützstellen. Die in diesem Sinne optimalen Quadraturformeln werden Gausssche Quadraturformeln bezeichnet.

Beweis von Satz 6.12:

Teil 1: Exaktheitsgrad $k = n+1+d$ impliziert die Bedingungen (a), (b).

Der Exaktheitsgrad der Formel ist $k = n+1+d \geq n$ und daher ist die Formel interpolatorisch (vgl. Definition 6.6).

Um Bedingung (b) zu zeigen, wählen wir ein beliebiges $p \in \mathbb{P}_d$ und betrachten das Produkt $\pi_n p \in \mathbb{P}_{n+1+d}$. Es gilt

$$\int_a^b \pi_n(x)\, p(x)\, \omega(x)\, dx = \sum_{i=0}^n w_i \pi_n(x_i)\, p(x_i).$$

Da π_n aber genau Nullstellen in x_i besitzt, verschwindet die rechte Seite.

Teil 2: Aus Bedingungen (a), (b) folgt der Exaktheitsgrad $k = n+1+d$.

Wir wählen $p \in \mathbb{P}_{n+1+d}$ und zeigen, dass der Fehler in (6.3) verschwindet. Wir verwenden dazu Polynomdivision mit Rest, wählen also $q \in \mathbb{P}_d$ und $r \in \mathbb{P}_n$ mit

$$p = q\pi_n + r.$$

Daraus folgt

$$\int_a^b p(x)\, \omega(x)\, dx = \int_a^b q(x)\, \pi_n(x)\, \omega(x)\, dx + \int_a^b r(x)\, \omega(x)\, dx.$$

Das erste Integral auf der rechten Seite verschwindet wegen der Orthogonalitätseigenschaft von π_n. Für das zweite verwenden wir die Bedingung (a) und erhalten

$$\int_a^b r(x)\, \omega(x)\, dx = \sum_{i=0}^n w_i r(x_i) = \sum_{i=0}^n w_i \left(p(x_i) - q(x_i)\, \pi_n(x_i)\right) = \sum_{i=0}^n w_i p(x_i).$$

Wiederum haben wir ausgenützt, dass π_n in den Knotenpunkten verschwindet. Zusammen haben wir gezeigt, dass

$$\int_a^b p(x)\,\omega(x)\,dx = \sum_{i=0}^{n} w_i p(x_i)$$

gilt und der Fehler daher verschwindet. ■

Beispiel 6.14 *Wir betrachten Integrale der Bauart*

$$\mathcal{I}(f) = \int_0^1 \omega(x)\,f(x)\,dx \quad \textit{mit der Gewichtsfunktion } \omega(x) = x^{-1/2}.$$

Die Zwei-Punkt-Newton-Cotes-Formel bzw. Zwei-Punkt-Gauss-Formel sind durch

$$Q_{NC}^{(2)}(f) = w_0^{NC} f(0) + w_1^{NC} f(1)$$
$$Q_G^{(2)}(f) = w_0^G f(x_0) + w_1^G f(x_1)$$

gegeben. Dabei gilt für die Newton-Cotes-Formel

$$w_0^{NC} = \int_0^1 \omega(x)(1-x)\,dx = \int_0^1 x^{-1/2}(1-x)\,dx = \frac{4}{3},$$
$$w_1^{NC} = \int_0^1 \omega(x)\,x dx = \int_0^1 x^{1/2} dx = \frac{2}{3}.$$

Die Konstruktion der Gauss-Formel startet mit der Konstruktion des Knotenpolynoms. Wir verwenden den Ansatz

$$\pi_1(x) = (x-x_0)(x-x_1) = x^2 + bx + a$$

Die Orthognalitätsrelationen lauten

$$\int_0^1 \omega(x)\,\pi_1(x)\,dx = \int_0^1 x^{-1/2}\left(x^2+bx+a\right)dx = 2a + \frac{2}{3}b + \frac{2}{5} \stackrel{!}{=} 0,$$
$$\int_0^1 \omega(x)\,\pi_1(x)\,x dx = \int_0^1 x^{-1/2}\left(x^2+bx+a\right)x dx = \frac{2}{3}a + \frac{2}{5}b + \frac{2}{7} \stackrel{!}{=} 0.$$

Dies ist ein lineares Gleichungssystem für die Koeffizienten a, b, und wir erhalten

$$a = 3/35 \quad \textit{und} \quad b = -6/7.$$

Die Nullstellen des Knotenpolynoms $\pi_1(x) = x^2 - 6/7x + 3/35$ sind durch $x_{0,1} = \frac{3}{7} \pm \frac{2}{35}\sqrt{30}$ gegeben. Die Gewichte können wir wie zuvor berechnen

$$w_0^G = \int_0^1 x^{-1/2}(x-x_1)/(x_0-x_1)\,dx = 1 + \frac{1}{3}\sqrt{\frac{5}{6}},$$
$$w_1^G = \int_0^1 x^{-1/2}(x-x_0)/(x_1-x_0)\,dx = 1 - \frac{1}{3}\sqrt{\frac{5}{6}}.$$

Als Beispiel verwenden wir die Funktion $f(x) = \cos\frac{\pi x}{2}$. *Der exakte Integralwert lautet:*

$$\int_0^1 x^{-1/2} \cos\frac{\pi x}{2} dx = 1.559786...$$

Die Newton-Cotes-Formel bzw. Gauss-Formel liefert

$$\begin{array}{ll} Q_{NC}^2(f) = 1.33333..... & E_{NC}^2(f) = 0.226... \\ Q_G^2(f) = 1.557589... & E_G^2(f) = 0.00220... \end{array}$$

Bemerkung 6.15 *Die Gewichtsfunktion erfülle* $\omega > 0$, $\omega \in C^0(]a,b[)$ *und* $\int_a^b \omega < \infty$.

a. *Die Stützstellen als Nullstellen von Orthogonalpolynomen sind alle reell, einfach und enthalten im offenen Intervall* (a,b).

b. *Die Gewichte sind alle positiv.*

c. *Die Gauss-Formeln konvergieren für alle stetigen Funktionen.*

Beweis. Zu a: Mit Hilfe der Gewichtsfunktion ω lässt sich das Skalarprodukt $(u,v)_\omega = \int_a^b \omega(x)\, u(x)\, v(x)\, dx$ definieren.

Wir zeigen: Sei $(p_n)_{n\in\mathbb{N}_0}$ ein Orthogonalsystem bzgl. dieses Skalarprodukts. Dann sind die Nullstellen x_i, $1 \le i \le n$, reell und einfach. Sie liegen im offenen Intervall (a,b).

Seien $a < x_1 < x_2 < \ldots < x_\ell < b$ diejenigen Nullstellen von p_n im Intervall (a,b), in denen p_n das Vorzeichen wechselt, d.h. die Vielfachheit dieser Nullstellen ist ungerade. Wir zeigen $\ell = n$ durch Widerspruch: Sei $\ell < n$. Dann hat das Polynom

$$q(x) = \prod_{j=1}^{\ell} (x - x_j) \in \mathbb{P}_\ell$$

den Grad $\ell < n$, so dass $(p_n, q)_\omega = 0$ gilt. Andererseits wurde q so konstruiert, dass $\omega(\cdot)\, q(\cdot)\, p_n(\cdot)$ sein Vorzeichen nicht ändert auf (a,b). Daraus folgt

$$(p_n, q)_\omega = \int_a^b \omega(x)\, p_n(x)\, q(x)\, dx \neq 0.$$

Dies ist ein Widerspruch und daraus folgt $\ell = n$. Damit sind alle Nullstellen reell, einfach und liegen im offenen Intervall (a,b).

Zu b: Seien ℓ_i die Lagrange-Basisfunktionen vom Grad n. Dann gilt (wegen des Exaktheitsgrades $2n+1$ der $n+1-$Punkt-Gauss-Formeln)

$$0 < \int_a^b \ell_i^2(x)\, \omega(x)\, dx = \sum_{j=0}^{n} w_j \ell_i^2(x_j) = w_i \qquad 0 \le i \le n.$$

Zu c: Wir verwenden die Fehlerabschätzung aus Satz 6.10. Da die Gewichte der Gauss-Quadratur positiv sind, ist die Konstante $C_{Q,n}$ in (6.6) gleich 1, und wir erhalten:

$$|\mathcal{I}(f) - Q^n(f)| \le 2\left(\int_a^b \omega(t)\, dt\right) \inf_{q\in\mathbb{P}_{2n+1}} \|f - q\|_{\max,I}.$$

Wegen des Weierstrassschen Approximationssatzes konvergiert die rechte Seite für $n \to \infty$ gegen Null. ■

Im Folgenden wenden wir uns der Berechnung der Stützstellen für die Gauss-Quadratur zu.

Zunächst muss ein Orthogonalsystem von Polynomen bezüglich des Skalarprodukts

$$(u, v)_\omega = \int_a^b \omega(x)\, u(x)\, v(x)\, dx \tag{6.7}$$

erzeugt werden. Das Gram-Schmidtsche Orthogonalisierungsverfahren ist aufwendig und numerisch instabil. Wir werden hier eine elegante Vorgehensweise vorstellen, welche die Stützstellen als Eigenwerte von Tridiagonalmatrizen ermittelt.

Sei ein Skalarprodukt durch (6.7) mit $\omega > 0$, $\omega \in C^0(]a, b[)$, $\int_a^b \omega < \infty$, gegeben. Polynome p_k vom Grade k werden durch die folgende 3-Term-Rekursion definiert

$$\begin{gathered} p_0(x) \equiv 1, \\ p_{k+1}(x) = (x - \alpha_k)\, p_k(x) - \beta_k p_{k-1}(x) \quad k = 0, 1, 2, \ldots \end{gathered}$$

Formal wird $p_{-1} \equiv 0$ gesetzt. Die Koeffizienten sind gegeben durch

$$\begin{aligned} \alpha_k &= \frac{(xp_k, p_k)_\omega}{(p_k, p_k)_\omega} \qquad k = 0, 1, 2, \ldots, \\ \beta_k &= \frac{(p_k, p_k)_\omega}{(p_{k-1}, p_{k-1})_\omega} \quad k = 1, 2, 3, \ldots, \\ \beta_0 &= \textstyle\int_a^b \omega. \end{aligned}$$

Man trägt nun diese Koeffizienten wie folgt in eine Tridiagonalmatrix ein

$$\mathbf{J} := \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & 0 & \cdots & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & \ddots & & & \vdots \\ 0 & \sqrt{\beta_2} & \ddots & \ddots & & & \\ \vdots & \ddots & \ddots & & & & \\ & & & & \alpha_{n-3} & \sqrt{\beta_{n-2}} & 0 \\ & & & & \sqrt{\beta_{n-2}} & \alpha_{n-2} & \sqrt{\beta_{n-1}} \\ 0 & \cdots & & & 0 & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix}.$$

Satz 6.16 *Die Stützstellen der n-Punkt-Gauss-Quadratur sind die Eigenwerte der Gleichung*

$$\mathbf{J}\mathbf{v}_i = x_i \mathbf{v}_i \qquad 1 \le i \le n$$

mit der Normierungsbedingung $\|\mathbf{v}_i\| = 1$.

Beweis. Wir zeigen $\det(x\mathbf{I} - \mathbf{J}) = p_n(x)$ durch Induktion über n. Um die Dimension $n := \dim \mathbf{J}$ anzudeuten, schreiben wir $\mathbf{J}_n$ statt $\mathbf{J}$.

- $n = 1$.

  Dann gilt $\det(x\mathbf{I} - \mathbf{J}_0) = x - \alpha_0 = p_1(x)$.

- Die Aussage sei bewiesen für $\ell = 0, 1, \ldots, n-1$.
- $n-1 \to n$:

  Für das charakteristische Polynom gilt die Rekursion

  $$\det(x\mathbf{I} - \mathbf{J}_n) = (x - \alpha_{n-1}) \det(x\mathbf{I} - \mathbf{J}_{n-1}) - \beta_{n-1} \det(x\mathbf{I} - \mathbf{J}_{n-2}).$$

  Diese Rekursion ist identisch mit der Rekursion der Orthogonalpolynome und daher stimmen die Polynome überein.

■

Bemerkung 6.17 *Wir werden später sehen, dass die Eigenwerte von Tridiagonalmatrizen sehr effizient und stabil berechnet werden können.*

6.1.3 Fehlerabschätzungen linearer Funktionale nach Peano

Integrale und Quadraturformeln können als lineare Funktionale auf der Menge $C^0([a,b])$ aufgefasst werden. In diesem Abschnitt werden wir einen allgemeinen Rahmen für Fehlerabschätzungen für lineare Funktionale herleiten, der auf Peano zurückgeht.

Der Quadraturfehler $E : C^0([a,b]) \to \mathbb{R}$ besitzt die abstrakte Bauart

$$E(f) = L(f) - \sum_{i=0}^{n} L_i(f),$$

wobei $L_i : C^0([a,b]) \to \mathbb{R}$ im Fall von Quadraturverfahren die (mit w_i gewichtete) Auswertung einer Funktion in der Stützstelle x_i bezeichnet und Lf die Integration über $[a,b]$. Für die Abschätzung des Fehlerfunktionals spielt der Exaktheitsgrad eine wesentliche Rolle, also das grösste $k \in \mathbb{N}_0$, für das

$$\forall p \in \mathbb{P}_k : \quad E(p) = 0$$

gilt. Der Exaktheitsgrad kann für Fehlerabschätzungen nur dann ausgenützt werden, falls die Funktion f hinreichend glatt ist, genauer $f \in C^{k+1}([a,b])$ erfüllt. In diesem Fall lässt sich f in eine Taylorreihe entwickeln:

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(a)}{i!} (x-a)^i + (R_k f)(x), \tag{6.8}$$

wobei das Restglied $R_k f$ die folgende Integraldarstellung besitzt

$$(R_k f)(x) = \frac{1}{k!} \int_a^x (x-t)^k f^{(k+1)}(t)\, dt = \frac{1}{k!} \int_a^b (x-t)_+^k f^{(k+1)}(t)\, dt,$$
$$(x-t)_+ := \max\{x-t, 0\}.$$

Wir wenden nun das Fehlerfunktional auf beide Seiten von (6.8) an und erhalten wegen des Exaktheitsgrades und der Linearität von E

$$Ef = ER_k f = E\frac{1}{k!} \int_a^b (x-t)_+^k f^{(k+1)}(t)\, dt = \frac{1}{k!} \int_a^b f^{(k+1)}(t)\, E_x\left((x-t)_+^k\right) dt. \tag{6.9}$$

(Abhängig von der konkreten Definition von E muss nachgeprüft werden, ob das Vertauschen des Fehlerfunktionals und der Integration erlaubt ist. In den meisten Fällen ist diese Voraussetzung unproblematisch.) Der Index x in E_x deutet an, dass das Funktional auf die $x-$Variable in $(x-t)_+^k$ bezogen ist. Die rechte Seite in (6.9) motiviert die Definition des *Peano-Kerns*

$$K_k(t) := \frac{1}{k!} E_x\left((x-t)_+^k\right),$$

der nicht von f, sondern lediglich von L und $(L_i)_{i=0}^n$ abhängt. Die *Peano-Darstellung* des Fehlers ist damit durch

$$Ef = \int_a^b f^{(k+1)}(t)\, K_k(t)\, dt \tag{6.10}$$

gegeben. Das Funktional E wird *definit* genannt, falls der Peano-Kern sein Vorzeichen in (a,b) nicht wechselt. Für derartige Funktionale lässt sich mit dem Mittelwertsatz die Darstellung weiter vereinfachen:

$$Ef = f^{(k+1)}(\tau) \int_a^b K_k(t)\, dt \quad \text{für ein } \tau \in (a,b).$$

Die direkte Berechnung des Integrals $\alpha := \int_a^b K_k(t)\, dt$ ist in vielen Fällen etwas aufwendig. Eine elegante Methode, α zu berechnen, geht von der Wahl $f(t) = t^{k+1}/(k+1)!$ aus. Offensichtlich gilt $f^{(k+1)}(\tau) = 1$ und wir erhalten

$$\int_a^b K_k(t)\, dt = E\left(t^{k+1}/(k+1)!\right).$$

Beispiel 6.18 *Das Integral*

$$\mathcal{I}f := \int_0^1 \omega(x)\, f(x)\, dx \quad mit \quad \omega(x) = \sqrt{x}$$

soll approximiert werden. Wir nehmen an, dass von der Funktion f die Funktionale $f(0)$ und $\int_0^1 f(x)\, dx$ bekannt sind und der Ansatz

$$Qf = \alpha_0 f(0) + \alpha_1 \int_0^1 f(x)\, dx$$

verwendet wird. Die Bedingungen $Qp = \mathcal{I}p$ für alle $p \in \mathbb{P}_1$ führt auf

$$Qf = -\frac{2}{15} f(0) + \frac{4}{5} \int_0^1 f(x)\, dx.$$

Der Fehler besitzt also die Darstellung

$$Ef = \int_0^1 \sqrt{x} f(x)\, dx + \frac{2}{15} f(0) - \frac{4}{5} \int_0^1 f(x)\, dx$$

und erfüllt $Ep = 0$ für alle $p \in \mathbb{P}_1$. Für $f \in C^2([0,1])$ ergibt sich

$$Ef = \int_0^1 K_1(t)\, f''(t)\, dt \quad mit \quad K_1(t) = E_x(x-t)_+.$$

Wir berechnen nun K_1 gemäss

$$\begin{aligned} K_1(t) &= \int_0^1 \sqrt{x}\,(x-t)_+\,dx + \frac{2}{15}(0-t)_+ - \frac{4}{5}\int_0^1 (x-t)_+\,dx \\ &= \int_t^1 \sqrt{x}\,(x-t)\,dx + \frac{2}{15}\times 0 - \frac{4}{5}\int_t^1 (x-t)\,dx = \frac{2}{15}t - \frac{2}{5}t^2 + \frac{4}{15}t^{\frac{5}{2}}. \end{aligned}$$

Der Kern K_1 ist positiv in $(0,1)$. Für die Grösse α erhalten wir

$$\alpha = E\left(\frac{t^2}{2}\right) = \int_0^1 \sqrt{t}\frac{t^2}{2}dt + \frac{2}{15}\times 0 - \frac{4}{5}\int_0^1 \frac{t^2}{2}dt = \frac{1}{105}.$$

(Alternativ hätte man auch $\int_0^1 K_1(t)\,dt = 1/105$ verwenden können.) Damit haben wir

$$Ef = \frac{1}{105}f''(\tau) \qquad \tau \in (0,1)$$

gezeigt.

6.2 Numerische Differentiation

In diesem Abschnitt behandeln wir die Approximation der *ersten* Ableitung einer Funktion. Höhere Ableitungen lassen sich approximieren durch Hintereinanderausführung der hier vorgestellten Methoden.

Ziel: Für eine gegebene differenzierbare Funktion f soll die Ableitung in einem Punkt x_0 durch geeignete Kombination der Werte von f in x_0 und nahegelegenen Punkten $x_1, \ldots, x_n$ approximiert werden.

Generell werden wir in diesem Abschnitt voraussetzen, dass $I = [a,b] \in \mathbb{R}$ gilt und die Stützstellenmenge $\Theta_n = \{x_i : 0 \le i \le n\}$ aus $n+1$ verschiedenen Punkten besteht, die alle in I enthalten sind. Für die Funktion f setzen wir die Stetigkeit $f \in C^0(I)$ und Differenzierbarkeit in x_0 voraus.

6.2.1 Eine allgemeine Formel zur Approximation der Ableitung einer Funktion

Ähnlich wie für die interpolatorische Integration besteht ein Zugang zur Approximation der Ableitung einer Funktion darin, die Funktion f zunächst zu interpolieren (allgemeiner: approximieren) und dann die Approximation zu differenzieren.

Die Polynominterpolation der Funktion $f \in C^0(I)$ in den Stützstellen Θ_n wird mit $p_n(f)$ bezeichnet und besitzt die Darstellung mit Hilfe der Newtonschen dividierten Differenzen

$$\begin{aligned} p_n(f) = f_0 &+ (x-x_0)\,[x_0,x_1]\,f + (x-x_0)(x-x_1)\,[x_0,x_1,x_2]\,f + \ldots \\ &+ (x-x_0)(x-x_1)\cdots(x-x_{n-1})\,[x_0,x_1,\ldots,x_n]\,f. \end{aligned}$$

Der Fehler $r_n = f - p_n(f)$ ist durch

$$r_n(x) = \prod_{i=0}^{n}(x-x_i)\,\frac{f^{(n+1)}(\xi_x)}{(n+1)!} \tag{6.11}$$

gegeben, wobei hier $f \in C^{n+1}(I)$ vorausgesetzt wurde. Daraus ergibt sich die Approximation der ersten Ableitung von f gemäss

$$f'(x_0) = p_n'(x_0) + e_n. \tag{6.12}$$

Der Fehler ist durch

$$e_n = r_n'(x_0) = \prod_{i=1}^{n} (x_0 - x_i) \frac{f^{(n+1)}(\xi_{x_0})}{(n+1)!} \tag{6.13}$$

gegeben.

Wichtig: Die Formel (6.13) gilt nur unter der Voraussetzung $f \in C^{n+2}(I)$, da die Ableitung von (6.11) (zunächst in einem algemeinen Punkt $x \in I$) auch den additiven Term

$$\prod_{i=0}^{n} (x - x_i) \frac{d}{dx} \frac{f^{(n+1)}(\xi_x)}{(n+1)!} = \prod_{i=0}^{n} (x - x_i) \frac{f^{(n+2)}(\xi_x)}{(n+1)!} \frac{d\xi_x}{dx}$$

enthält und *danach* $x = x_0$ eingesetzt wird.

Die Länge des kleinsten Intervalls, welches die Stützstellenmenge Θ_n enthält, wird mit H bezeichnet. Aus (6.13) ergibt sich

$$|e_n| \leq M_n H^n \tag{6.14}$$

mit $M_n := \left\| f^{(n+1)} \right\|_{\max} / (n+1)!$.

Satz 6.19 *Sei $f \in C^0(I)$ und die Stützstellenmenge $\Theta_n \subset I$ gegeben. Dann ist eine Approximation von $f'(x_0)$ durch*

$$p_n'(f)(x_0) = [x_0, x_1] f + (x_0 - x_1)[x_0, x_1, x_2] f + (x_0 - x_1) \cdots (x_0 - x_{n-1})[x_0, x_1, \ldots, x_n] f.$$

gegeben.

Für $f \in C^{n+2}(I)$ gilt die Fehlerabschätzung

$$|f'(x_0) - p_n'(f)(x_0)| \leq M_n H^n$$

mit M_n und H wie in (6.14).

Bemerkung 6.20 *Die Fehlerabschätzung aus Satz 6.19 enthält zwei Parameter, n und H, die verwendet werden können, um eine vorgegebene Genauigkeit zu erreichen.*

a. *Erhöhung der Interpolationsordnung n. In diesem Fall muss wie im Abschnitt über die (globale) Polynominterpolation vorausgesetzt werden, dass f unendlich oft differenzierbar ist. Dies genügt im Allgemeinen jedoch nicht – genauer folgt die Konvergenz (ähnlich wie für Taylorreihen), falls $\frac{\left\| f^{(n+1)} \right\|_{\max}}{(n+1)!} H^n \overset{n \to \infty}{\to} 0$ gilt.*

b. *Verkleinerung der Intervallgrösse, $H \to 0$ und n fest. In diesem Fall genügt es, $f \in C^{n+2}(I)$ für ein festes $n \in \mathbb{N}$ vorauszusetzen. Die Konvergenz ist dann für $H \to 0$ algebraisch:*

$$|f'(x_0) - p_n'(f)(x_0)| \leq M_n H^n.$$

Beispiel 6.21

1. $n = 1$, $x_1 = x_0 + h$. *Dann gilt*

$$p_1'(f)(x_0) = [x_0, x_1] f = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f_1 - f_0}{h}. \tag{6.15}$$

Für den Fehler

$$e_1 = f'(x_0) - p_1'(f)(x_0) = f'(x_0) - \frac{f_1 - f_0}{h}$$

gilt unter der Voraussetzung $f \in C^3(I)$ *die Fehlerdarstellung*

$$e_1 = (x_0 - x_1)\frac{f''(\xi)}{2} = -h\frac{f''(\xi)}{2} \quad \text{für ein } \xi \in (x_0, x_1).$$

Die Approximation (6.15) wird Vorwärtsdifferenz genannt.

2. $n = 2$, $x_{-1} = x_0 - h$, $x_1 = x_0 + h$. *Dann gilt*

$$p_2'(f)(x_0) = [x_{-1}, x_0] f + (x_0 - x_{-1})[x_{-1}, x_0, x_1] f.$$

Das Differenzenschema zur Berechnung der dividierten Differenzen ist hierbei wie folgt gegeben

x_0	f_0		
x_{-1}	f_{-1}	$\frac{f_{-1}-f_0}{-h}$	
x_1	f_1	$\frac{f_1-f_{-1}}{2h}$	$\frac{\frac{f_1-f_{-1}}{2h} - \frac{f_{-1}-f_0}{-h}}{h} = \frac{f_{-1} - 2f_0 + f_1}{2h^2}$

Daher gilt

$$p_2'(f)(x_0) = \frac{f_{-1} - f_0}{-h} + \underbrace{(x_0 - x_{-1})}_{h}\frac{f_{-1} - 2f_0 + f_1}{2h^2} = \frac{f_1 - f_{-1}}{2h}. \tag{6.16}$$

Für $f \in C^4(I)$ *gilt die Fehlerdarstellung*

$$f'(x_0) = \frac{f_1 - f_{-1}}{2h} + e_2 \quad \textit{mit} \quad e_2 = (x_0 - x_{-1})(x_0 - x_1)\frac{f'''(\xi)}{3!} = -h^2\frac{f'''(\xi)}{6}.$$

Die Approximation (6.16) wird symmetrische Differenz genannt. Man beachte, dass die symmetrische Differenz bei gleichem Aufwand wie für die Vorwärtsdifferenz eine höhere Genauigkeit liefert, falls f *hinreichend glatt ist. Symmetrische Differenzen lassen sich definieren, falls* f *auf beiden Seiten vom Punkt* x_0 *definiert ist. Ist* f *lediglich auf einer Seite von* x_0 *definiert, müssen unsymmetrische Formeln verwendet werden.*

3. $n = 2$, $x_1 = x_0 + h$, $x_2 = x_0 + 2h$.

Das Differenzenschema ist wie folgt gegeben

x_0	f_0		
x_1	f_1	$\frac{f_1-f_0}{h}$	
x_2	f_2	$\frac{f_2-f_1}{h}$	$\frac{\frac{f_2-f_1}{h} - \frac{f_1-f_0}{h}}{2h} = \frac{f_2 - 2f_1 + f_0}{2h^2}$

und die Approximation der Ableitung lautet

$$p_2'(f)(x) = \frac{f_1 - f_0}{h} - h\frac{f_2 - 2f_1 + f_0}{2h^2} = \frac{4f_1 - 3f_0 - f_2}{2h}.$$

Für den Fehler erhalten wir

$$e_2 = 2h^2 \frac{f'''(\xi)}{6}.$$

Die Konvergenzordnung dieser einseitigen Differenz ist zwar gleich wie für die symmetrische Differenz, jedoch erfordert deren Berechnung die Auswertung von f in drei Stützstellen.

6.2.2 Numerische Differentiation mit gestörten Daten

Die numerische Differentiation tritt praktisch nie als idealisiertes mathematisches Problem auf. Die Daten sind meist durch Messfehler oder durch vorangegangene numerische Approximationen gestört. Den Einfluss derartiger Störungen werden wir in diesem Abschnitt behandeln. Wir betrachten die Situation, dass eine glatte Funktion f nur in *gestörter* Form f_S vorliegt, d.h.

$$f_S = f + \Delta f.$$

Von der Störung Δf und damit auch von der gestörten Funktion f_S können wir keine Glattheit erwarten; setzen jedoch voraus, dass f_S punktweise ausgewertet werden kann, also stetig ist. Damit lässt sich beispielsweise die symmetrische Differenz zur Approximation der Ableitung verwenden:

$$f'(x_0) \approx \frac{f_S(x_1) - f_S(x_{-1})}{2h} = \underbrace{\frac{f(x_1) - f(x_{-1})}{2h} + \frac{\Delta f(x_1) - \Delta f(x_{-1})}{2h}}_{=:\widetilde{f'}(x_0)}.$$

Analog lässt sich der Fehler in zwei Summanden zerlegen:

$$\varepsilon_1 := f'(x_0) - \frac{f(x_1) - f(x_{-1})}{2h} \quad \text{und} \quad \varepsilon_2 := \frac{\Delta f(x_1) - \Delta f(x_{-1})}{2h},$$

so dass $f'(x_0) - \tilde{f}'(x_0) = \varepsilon_1 + \varepsilon_2$ gilt. Der erste Teil des Fehlers liefert die gewünschte quadratische Konvergenzordnung. Setzen wir $\delta_1 = \Delta f(x_1)$ und $\delta_{-1} = \Delta f(x_{-1})$, so entsteht im ungünstigsten Fall eine Fehlerverstärkung im zweiten Term der Form

$$|\varepsilon_2| \leq \frac{|\delta_1| + |\delta_2|}{2h}. \tag{6.17}$$

Bemerkung 6.22 *Um eine Abschätzung der Form $|\varepsilon_2| \leq Ch^2$ sicherzustellen, muss die Genauigkeit in der Auswertung der Funktion die Grössenordnung $O(h^3)$ betragen.*

In gewissen Fällen lässt sich diese Genauigkeit aber nicht erreichen, da die Daten/Funktion f *gegeben* sind und damit die Genauigkeiten $|\delta_1|$, $|\delta_2|$ positiv sind und nicht von h abhängen. Offensichtlich **divergiert** dann die rechte Seite in (6.17) für $h \to 0$.

Bemerkung 6.23 *Falls die Daten $f(x_1)$, $f(x_{-1})$ mit einem **festen** Fehler behaftet sind, divergiert im allgemeinen der Fehler $f'(x_0) - \widetilde{f'}(x_0)$ für $h \to 0$.*

Das Ziel in dieser Situation ist es daher, eine (positive) Schrittweite h zu finden, so dass der Fehler in der Ableitung minimal ist. Die Fehlerabschätzung für die symmetrische Approximation der ersten Ableitung mit gestörten Daten lautet

$$\left| f'(x_0) - \widetilde{f}'(x_0) \right| \leq \frac{\delta}{h} + C_f h^2 =: E(h),$$

wobei C_f von der dritten Ableitung von f abhängt und wir $\|\Delta f\|_{\max} \leq \delta$ annehmen.

Diskussion der Funktion $E(h)$ zeigt, dass das Minimum für

$$h_{\text{opt}} := C_1 \delta^{1/3} \text{ mit } C_1 = (2C_f)^{-1/3}$$

angenommen wird und beträgt $E(h_{\text{opt}}) = C_2 \delta^{2/3}$ mit $C_2 := \frac{3}{2}(2C_f)^{1/3}$. Das bedeutet, dass selbst bei einer optimalen Wahl der Schrittweite h eine **Fehlerverstärkung** stattfindet: Der Eingabefehler $O(\delta)$ in den Funktionswerten wird zu einem Fehler $O\left(\delta^{2/3}\right)$ für die Approximation der Ableitung verstärkt.

7 Nichtlineare Gleichungen

In diesem Kapitel behandeln wir nichtlineare Gleichungen der abstrakten Form

$$f(x) = 0.$$

Im einfachsten Fall ist $f : \mathbb{R} \to \mathbb{R}$ eine skalare Gleichung -allgemein gilt $f : \mathbb{R}^d \to \mathbb{R}^d$. Falls alle Komponenten von f linear von x abhängen, werden die Gleichungen als lineare, algebraische Gleichungen bezeichnet. Falls mindestens eine Gleichung nichtlinear von x abhängt, sprechen wir von nichtlinearen Gleichungen.

Da selbst einfachste, nichtlineare algebraische Gleichungen keine rationalen Lösungen besitzen, lassen sich die Lösungen auf einem Rechner nicht exakt berechnen. Typischerweise werden Folgen konstruiert $x_0, x_1, \ldots, x_n, \ldots$ mit Hilfe einer *Iterationsvorschrift* $x_{i+1} = \Phi(x_i)$, die gegen x konvergieren. Man bricht diese Iteration dann nach endlich vielen Schritten ab und benötigt Fehlerabschätzungen, um die Genauigkeit der Näherungslösungen abzuschätzen. Wesentlich für die *Effizienz* des Verfahrens ist die Konvergenzgeschwindigkeit, genauer die Konvergenzordnung.

Definition 7.1 *Die Konvergenz $x_i \to x$ heisst (mindestens) linear, falls*

$$|x_i - x| \leq \varepsilon_i \tag{7.1}$$

gilt mit einer Folge ε_i, die

$$\lim_{i \to \infty} \frac{\varepsilon_{i+1}}{\varepsilon_i} = c \tag{7.2}$$

für eine $0 < c < 1$ erfüllt.

Man beachte, dass die Konvergenzordnung mit Hilfe der *Schranken* ε_i definiert wurde. Der tatsächliche Fehler könnte wesentlich schneller bzw. sogar nicht-monoton konvergieren.

Definition 7.2 *Die Konvergenz $x_i \to x$ heisst (mindestens) von Ordnung $p > 1$, falls anstelle von (7.2) die Bedingung*

$$\lim_{i \to \infty} \frac{\varepsilon_{i+1}}{\varepsilon_i^p} = c \tag{7.3a}$$

mit einem $c > 0$ erfüllt ist.

Analog lässt sich die Konvergenzordnung für vektorwertige Funktionen definieren, wobei der Absolutbetrag in (7.1) durch geeignete Normen ersetzt werden muss.

7.1 Bisektion und Sturmsche Ketten

7.1.1 Bisektion

Die einfachste Möglichkeit, die Nullstellen einer nichtlinearen Funktion zu bestimmen, ist die Bisektion. Man geht dabei von einem endlichen Startintervall aus, welches mindestens eine Nullstelle enthält und unterteilt dieses durch Halbierung in zwei Teilintervalle. Durch Betrachtung von Vorzeichenwechsel lässt sich dann immer ein Intervall bestimmen, welches die Nullstelle enthält und der Prozess kann fortgesetzt werden.

Algorithmisch lautet das Verfahren der Bisektion wie folgt. Sei $[a, b]$ ein Intervall, welches die Nullstelle enthält. Annahme:

$$f \in C^0([a, b]), \qquad f(a) < 0, \qquad f(b) > 0. \tag{7.4}$$

Algorithmus 7.3 *Es gelte (7.4). Setze* $a_1 := a$ *und* $b_1 := b$;
for $n := 1, 2, 3, \ldots$ ***do begin***
$x_n := \frac{1}{2}(a_n + b_n)$;
if $f(x_n) < 0$ ***then begin*** $a_{n+1} = x_n$; $b_{n+1} := b_n$;***end***
else begin $a_{n+1} = a_n$; $b_{n+1} = x_n$;***end***
end;

Nach n Schritten dieses Prozesses ist die Nullstelle sicherlich im Intervall (a_n, b_n) enthalten. Daraus folgt, dass die exakte Nullstelle vom Mittelpunkt x_n von (a_n, b_n) lediglich um $(b_n - a_n)/2$ abweichen kann. Genauer gilt:

$$|x_n - x| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}.$$

Das führt zu einer Fehlerabschätzung mit $\varepsilon_n = (b-a)/2^n$. Weiter gilt $\varepsilon_{n+1}/\varepsilon_n = 1/2$. Das bedeutet, dass die Bisektionsmethode mindestens linear konvergiert. Falls nun die Nullstelle bis zu einer Genauigkeit von $\varepsilon > 0$ berechnet werden soll, lässt sich die Zahl der Iterationsschritte a-priori bestimmen. Sei n die kleinste Ganzzahl, die

$$\frac{b-a}{2^n} < \varepsilon$$

erfüllt. Auflösen ergibt

$$n = \left\lceil \log_2 \frac{b-a}{\varepsilon} \right\rceil,$$

wobei $\lceil \cdot \rceil$ „Aufrunden“ bedeutet.

Zusammenfassend lässt sich sagen, dass die Bisektionsmethode eine sehr robuste und einfach zu implementierende Methode ist, um Nullstellen einer Funktion zu berechnen. Die Glattheitsvoraussetzungen an die Funktion sind sehr gering. Auf der anderen Seite ist die Konvergenzgeschwindigkeit sehr niedrig und die Methode daher für komplizierte Anwendungen zu teuer. Konzeptionell ist sie auf skalare Funktionen beschränkt, die lediglich von einer Variablen abhängen.

7.1.2 Sturmsche Ketten

In sehr vielen Fällen besitzen nichtlineare Funktionen mehrere Nullstellen. In einigen Anwendungen ist man jedoch lediglich an den niedrigsten n Nullstellen interessiert. In diesem Abschnitt betrachten wir die Aufgabe, eine oder mehrere spezielle Nullstellen einer Funktion zu berechnen. Für orthogonale Polynome oder das charakteristische Polynom einer Matrix ist a-priori bekannt, wieviele Nullstellen insgesamt existieren. Unter Verwendung dieser Zusatzinformation lassen sich die Nullstellen mit Hilfe *Sturmscher Ketten* berechnen.

Wir sind im Zusammenhang mit der Gaussschen Quadraturmethode auf das Problem gestossen, die Nullstellen von Polynomen zu bestimmen. Diese Polynome besitzen gewisse Orthogonalitätseigenschaften. Genauer haben wir gezeigt, dass deren Nullstellen Eigenwerte symmetrischer Tridiagonalmatrizen sind. Sei dazu ein Skalarprodukt mit Gewichtsfunktion $\omega : [a, b] \to \mathbb{R}$ mit $\omega(t) > 0$ für alle $t \in [a, b]$ gegeben:

$$(u, v)_\omega = \int_a^b \omega(t)\, u(t)\, v(t)\, dt.$$

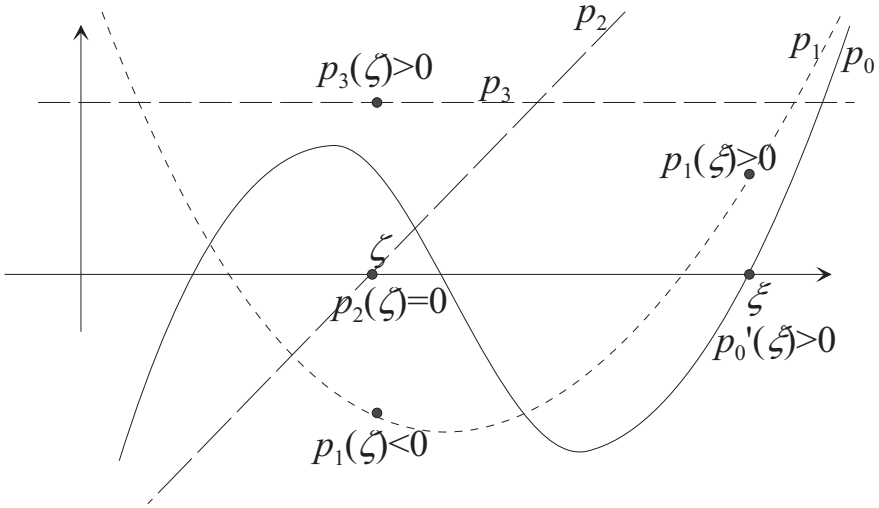


Abbildung 3: Sturmsche Kette, bestehend aus vier Polynomen.

Zur Berechnung der Stützstellen und Gewichte für die Gaussquadratur benötigen wir ein Orthogonalsystem $(\pi_i)_{i\in\mathbb{N}_0}$ mit den Eigenschaften

$$\pi_i \in \mathbb{P}_i, \qquad \pi_i = t^i + \sum_{k=0}^{i-1} \alpha_k t^k, \tag{7.5}$$
$$(\pi_i, \pi_j)_\omega = 0 \qquad \forall i \neq j.$$

Wir haben gesehen, dass ein solches System mit Hilfe einer Dreitermrekursion berechnet werden kann

$$\pi_0(t) \equiv 1, \qquad \pi_1 = t - \alpha_0$$
$$\pi_{k+1}(t) = (t - \alpha_k)\,\pi_k(t) - \beta_k \pi_{k-1}(t), \qquad k = 1, 2, \ldots,$$

wobei alle β_k positive Zahlen sind. Im vorigen Kapitel wurde gezeigt, dass die Nullstellen dieser Polynome Eigenwerte symmetrischer reeller Tridiagonalmatrizen sind und damit alle reell sind. Weiter wurde bewiesen, dass alle Nullstellen einfach sind und im Intervall (a, b) liegen. Des weiteren gilt die folgende **Separationseigenschaft:**

$$\text{Die Nullstellen von } \pi_{i-1} \text{ trennen die Nullstellen von } \pi_i. \tag{7.6}$$

Die Methode von Sturm basiert auf folgendem Prinzip: Aus den Polynomen π_i lässt sich die Kette

$$\pi_d(x), \pi_{d-1}(x), \ldots, \pi_0(x) \tag{7.7}$$

bilden. Man zeigt zunächst, dass diese Kette eine Sturmsche Kette bildet (vgl. Abb. 3).

Definition 7.4 *Eine Folge reeller Polynome*[13]

$$p_0, p_1, \ldots, p_m$$

heisst Sturmsche Kette, falls die Eigenschaften 1-4 erfüllt sind.

[13] Achtung: In der untenstehenden Kette ist die Numerierung umgekehrt.

1. *p_0 besitzt nur einfache Nullstellen.*

2. *Für alle ξ mit $p_0(\xi) = 0$ gilt* $\operatorname{sgn} p_1(\xi) = \operatorname{sgn} p_0'(\xi)$

3. *Für $i = 1, 2, \ldots, m-1$ und alle ξ mit $p_i(\xi) = 0$ gilt*

$$p_{i+1}(\xi)\, p_{i-1}(\xi) < 0.$$

4. *Das letzte Polynom ändert sein Vorzeichen nicht.*

Lemma 7.5 *Die Kette (7.7) bildet eine Sturmsche Kette.*

Beweis. Im Zusammenhang mit Gaussschen Quadraturverfahren wurde bereits gezeigt: "Alle π_i besitzen einfache, reelle Nullstellen". Damit ist (1) gezeigt.

Zu (3): Wir betrachten ein $1 \leq k < m$. Sei $\pi_k(t) = 0$. Die 3-Term-Rekursion der π_k liefert:

$$\pi_{k+1}(t) = (t - \alpha_k)\, \pi_k(t) - \beta_k \pi_{k-1}(t) = -\beta_k \pi_{k-1}(t).$$

Wir hatten bereits $\beta_k > 0$ gezeigt. Daher folgt aus $\pi_k(t) = 0$ und der Separationseigenschaft (7.6) der Nullstellen: $\pi_{k-1}(t), \pi_{k+1}(t) \neq 0$ die Eigenschaft 3.).

Zu (4): Offensichtlich ändert $p_m \equiv \pi_0 \equiv 1$ das Vorzeichen nicht.

Zu (2): Wegen der Separationseigenschaft der Nullstellen folgt aus $p_0(\xi) = \pi_m(\xi) = 0$ die Eigenschaft $\pi_{m-1}(\xi) \neq 0$. Es gilt $\pi_k(x) = x^k$+Terme niedrigerer Ordnung. Setzen wir

$$s_k := \operatorname{sgn} \pi_k(-\infty)$$

gilt $s_k = -s_{k-1}$. Für die kleinste Nullstelle ξ_m von π_m gilt daher $\operatorname{sgn} \pi_m'(\xi_m) = -s_m$. Da π_{m-1} bei ξ_m auf Grund der Separationseigenschaft noch keinen Vorzeichenwechsel durchläuft, gilt $\operatorname{sgn} \pi_{m-1}(\xi_m) = s_{m-1}$. Daraus folgt:

$$\operatorname{sgn} \pi_{m-1}(\xi_m) = s_{m-1} = -s_m = \operatorname{sgn} \pi_m'(\xi_m).$$

Induktiv nach rechts zu den nächst grösseren Nullstellen von π_m fortschreitend, ergibt sich die Behauptung. ■

Mit $\sigma(x)$ bezeichnen wir die Zahl der Vorzeichenwechsel in dieser Kette. Falls $\pi_k(x) = 0$ gilt, wird das *nicht* als Vorzeichenwechsel gezählt. Genauer werden alle Glieder der Kette mit $\pi_k(x) = 0$ aus der Kette entfernt und dann die Vorzeichenwechsel gezählt. Dann gilt: Die Zahl der Nullstellen von π_d in $(a, b]$ ist gleich $\sigma(a) - \sigma(b)$.

Satz 7.6 *Die Anzahl der reellen Nullstellen von π_d im Intervall $(a, b]$ ist gleich $\sigma(a) - \sigma(b)$, wobei $\sigma(x)$ die Anzahl der Vorzeichenwechsel in der Kette*

$$\pi_d(x), \pi_{d-1}(x), \ldots, \pi_0(x)$$

bezeichnet.

Beweis. Wir untersuchen, wie sich die Zahl der Vorzeichenwechsel $\sigma(a)$ sich mit wachsendem a ändert. Solange für alle Polynome $\pi_k(a) \neq 0$ gilt, ändert sich die Zahl der Vorzeichenwechsel nicht. Betrachten wir nun ein a mit $\pi_k(a) = 0$. Wir unterscheiden zwei Fälle:

1. $\pi_k(a) = 0,\ \ k \neq d$

2. $\pi_k(a) = 0,\ k = d.$

Im ersten Fall gilt wegen Eigenschaft (3) der Sturmschen Ketten:

$$\pi_{k+1}(a), \pi_{k-1}(a) \neq 0.$$

Wir betrachten eine Umgebung von a und wählen h hinreichend klein. Dann trifft eine der folgenden Situationen zu: (Beachte, dass bereits festgestellt wurde, dass alle Nullstellen von π_k einfach sind).

	$a-h$	a	$a+h$
$k-1$	$-$	$-$	$-$
k	$-$	0	$+$
$k+1$	$+$	$+$	$+$

	$a-h$	a	$a+h$
$k-1$	$+$	$+$	$+$
k	$-$	0	$+$
$k+1$	$-$	$-$	$-$

	$a-h$	a	$a+h$
$k-1$	$-$	$-$	$-$
k	$+$	0	$-$
$k+1$	$+$	$+$	$+$

	$a-h$	a	$a+h$
$k-1$	$+$	$+$	$+$
k	$+$	0	$-$
$k+1$	$-$	$-$	$-$

Offensichtlich gilt, dass sich $\sigma(a)$ beim Durchgang durch eine Nullstellen nicht ändert. Wir betrachten den zweiten Fall. Dann trifft eine der folgenden Situationen zu (wg. Separationseigenschaft und Eigenschaft 2 (Steigung $\pi_d(a)$ positiv, dann $\operatorname{sgn}\pi_{d-1}$ positiv und umgekehrt) aus der Definition von Sturmschen Ketten)

	$a-h$	a	$a+h$
d	$-$	0	$+$
$d-1$	$+$	$+$	$+$

	$a-h$	a	$a+h$
d	$+$	0	$-$
$d-1$	$-$	$-$	$-$

(7.8)

Daraus folgt, dass sich die Vorzeichenanzahl genau beim Durchgang durch eine Nullstelle von π_d um eins verändert (reduziert). Offensichtlich gilt $\sigma(-\infty) = d$ wegen (7.5). Seien die Nullstellen von π_d gemäss

$$\xi_d < \xi_{d-1} < \ldots < \xi_1 \tag{7.9}$$

angeordnet. O.B.d.A. sei $s_d = \operatorname{sgn}(\pi_d(-\infty)) = 1$. Dann trifft der rechte Teil von (7.8) zu und es gilt: $\sigma(\xi_d) = d-1$ und rekursiv $\sigma(\xi_i) = i-1$. Daraus folgt (vgl. Abb. 4)), dass die Anzahl der Nullstellen von π_d in $(-\infty, a]$ gleich $d - \sigma(a)$ ist und in $(-\infty, b]$ gleich $d - \sigma(b)$. Daraus folgt, dass die Anzahl der Nullstellen in $(a, b]$ gleich $\sigma(a) - \sigma(b)$ ist. ■

Mit Hilfe von Sturmsche Ketten lässt sich also einfach bestimmen, wieviele Nullstellen von π_d in einem Intervall liegen. Seien die Nullstellen von π_d gemäss (7.9) angeordnet. Man startet mit einem Intervall (beispielsweise, dem Intervall, bezüglich dessen die Orthogonalpolynome definiert sind), welches sicher alle Nullstelle enthält. Ziel sei es, die Nullstelle ξ_i zu berechnen. Dann halbiert man das aktuelle Intervall mit Hilfe der Bisektion. Die Entscheidung, welches Teilintervall man für die weitere Unterteilung verwendet, wird mit Hilfe der Sturmsche Kette bestimmt. Algorithmisch lautet das Verfahren zur Bestimmung der Nullstelle ξ_i:

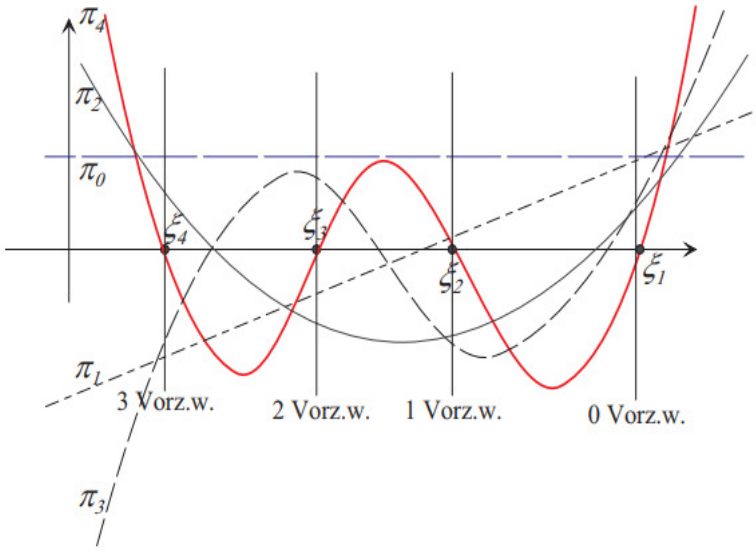


Abbildung 4: Die Zahl der Vorzeichenwechsel in Sturmschen Ketten reduziert sich um 1 genau beim Durchgang durch eine Nullstelle von π_d von links nach rechts gezählt.

Für $j = 0, 1, 2, \ldots$ berechne:

$$\mu_j := (a_j + b_j)/2,$$
$$a_{j+1} := \begin{cases} a_j & \text{falls } \sigma(\mu_j) \leq i-1 \\ \mu_j & \text{falls } \sigma(\mu_j) > i-1 \end{cases}$$
$$b_{j+1} := \begin{cases} \mu_j & \text{falls } \sigma(\mu_j) \leq i-1 \\ b_j & \text{falls } \sigma(\mu_j) > i-1 \end{cases}$$

Die Vorzeichenwechsel lassen sich einfach mit Hilfe der Rekursion der Polynome π_k abzählen. Es gilt dann stets: $\xi_i \in [a_{j+1}, b_{j+1}]$. Die Konvergenzrate ist wie bei Bisektion linear. Das Verfahren ist sehr genau und man kann einzelne Nullstellen berechnen, ohne die anderen zu berechnen. Der Nachteil ist, dass die Konvergenzgeschwindigkeit nur linear, also langsam ist.

7.2 Das Newton-Verfahren

Die Newtonsche Methode zur Nullstellenbestimmung von Funktionen basiert auf Linearisierung. Wir nehmen an, dass die Funktion f in der Gleichung

$$f(x) = 0 \tag{7.10}$$

differenzierbar ist. Sei $x^{(i)}$ eine alte Iterierte. Die Idee ist nun, die Funktion f in $x^{(i)}$ in eine Taylorreihe zu entwickeln und nur Terme bis zur ersten Ordnung zu berücksichtigen

$$f(x) \approx f\left(x^{(i)}\right) + f'\left(x^{(i)}\right)\left(x - x^{(i)}\right). \tag{7.11}$$

Falls f affin ist in einer Umgebung von $x^{(i)}$, ist diese Formel exakt -allgemein gilt, dass (7.11) in einer hinreichend kleinen Umgebung $x \in \mathcal{U}\left(x^{(i)}\right)$ eine gute Approximation von f darstellen sollte. Die Gleichung (7.10) wird nun durch die affine Gleichung:

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

ersetzt. Auflösen nach x ergibt die neue Iterierte:

$$x^{(i+1)} = x^{(i)} - \frac{f\left(x^{(i)}\right)}{f'\left(x^{(i)}\right)}, \qquad i = 0, 1, 2, \ldots. \tag{7.12}$$

Um diese Rekursion zu initialisieren, benötigen wir noch einen geeigneten Startwert $x^{(0)}$. Die Bestimmung eines geeigneten Startwerts ist für das Newton-Verfahren ein delikates Problem, da die Konvergenz nur in einer *hinreichend kleinen Umgebung* einer Lösung $x_\star$ von (7.10) gesichert ist (vgl. Bemerkung 7.16).

Mit Hilfe der Interpretation als lokale, lineare Approximation von f, kann die Newtonsche Methode direkt auch auf Systeme von Gleichungen angewendet werden. Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ eine differenzierbare Funktion (Annahme: Die Jacobi-Matrix $\mathbf{J}(x)$ und ihre Inverse existieren in jedem Punkt x). Dann gilt in einem Punkt $x^{(i)}$ und $x \in \mathcal{U}\left(x^{(i)}\right)$:

$$f(x) \approx f\left(x^{(i)}\right) + \mathbf{J}\left(x^{(i)}\right)\left(x - x^{(i)}\right).$$

Ersetzen der (dieses Mal vektorwertigen) Gleichung (7.10) durch die linearisierte Gleichung liefert:

$$0 \overset{!}{=} f\left(x^{(i)}\right) + \mathbf{J}\left(x^{(i)}\right)\left(x - x^{(i)}\right).$$

Um nach x aufzulösen, muss ein Gleichungssystem mit Matrix $\mathbf{J}\left(x^{(i)}\right)$ gelöst werden. Als Lösung erhalten wir die neue Iterierte:

$$x = x^{(i)} - \mathbf{J}^{-1}\left(x^{(i)}\right) f\left(x^{(i)}\right) \tag{7.13}$$

und damit ein Iterationsverfahren zur Nullstellenbestimmung nichtlinearer, vektorwertiger Gleichungen.

Zur Illustration des Newton-Verfahrens betrachten wir einige Beispiele.

Berechnung der Wurzel einer Zahl $a > 0$. Betrachte die Funktion

$$f(x) = x^2 - a.$$

Ziel ist die Gleichung $f(x) = 0$ zu lösen. Gleichung (7.12) liefert die Iterationsvorschrift:

$$x^{(i+1)} = x^{(i)} - \frac{\left(x^{(i)}\right)^2 - a}{2x^{(i)}} = \frac{1}{2}\left(x^{(i)} + \frac{a}{x^{(i)}}\right), \qquad i = 0, 1, 2 \ldots. \tag{7.14}$$

Die folgende Tabelle zeigt die schnelle Konvergenz des Newton-Verfahrens verglichen zur Bisektion für die Wahl $a = 2$. Als Startwert wurde für das Newton-Verfahren $x_0 = 1$ gewählt und als Startintervall für die Bisektion $[0, 1]$.

Anzahl Iteration	x^i_{Newton}	$\text{Fehler}^i_{\text{Newton}}$	$x^i_{\text{Bisektion}}$	$\text{Fehler}^i_{\text{Bisektion}}$
1	1.5	8.58×10^{-2}	1	0.41
2	$1.41\bar{6}$	2.45×10^{-3}	1.5	0.086
3	1.41421568	2.12×10^{-6}	1.25	0.16
4	1.4142135623747	1.59×10^{-12}	1.375	0.039
5	1.41421356237309504880168	8.99×10^{-25}	1.4375	0.023
6	1.41421356237309504880168872421	2.86×10^{-49}	1.40625	0.008

Übungsaufgabe 7.7 *Beweisen Sie, dass die Iteration (7.14) für jeden positiven Startwert $x^{(0)}$ gegen die positive Wurzel von $a > 0$ konvergiert.*

Beispiel 7.8 *Wir betrachten die Funktion $f : (-\pi/2, \pi/2) \to \mathbb{R}$:*

$$f(x) := \sin x.$$

Bekanntlich liegt im betrachteten Intervall genau eine Nullstelle $x = 0$. Das Newton-Verfahren liefert die Iterationsvorschrift:

$$x^{(i+1)} = x^{(i)} - \tan x^{(i)}.$$

Wählen wir nun den Startwert so, dass $x^{(0)} - \tan x^{(0)} = -x^{(0)}$ gilt, (nämlich $x^{(0)} = -1.16556118520721...$) so folgt, $x^{(1)} = -x^{(0)}$ und $x^{(2)} = x^{(0)}$. Offensichtlich konvergiert die Methode nicht und die beiden alternierenden Werte haben nichts mit der exakten Lösung zu tun. Wir werden zeigen, dass das Newton-Verfahren in einer hinreichend kleinen Umgebung von 0 gegen die exakte Lösung $x^{(0)}$ konvergiert.

Ein weitere charakteristische Eigenschaft des Newton-Verfahrens ist die sehr hohe (quadratische) Konvergenz in der *Nähe* der exakten Lösung. Die *lokale* Konvergenz des Newtonschen Verfahrens wird unter geeigneten Voraussetzungen in folgendem Satz gezeigt.

Satz 7.9 *Sei $x_\star$ eine einfache Nullstelle der Gleichung $f(x) = 0$ und $I_\varepsilon := \{x \in \mathbb{R} : |x - x_\star| \le \varepsilon\}$. Sei $f \in C^2[I_\varepsilon]$. Definiere*

$$M(\varepsilon) := \max_{s,t \in I_\varepsilon} \left| \frac{f''(s)}{2f'(t)} \right|.$$

Für hinreichend kleines $\varepsilon > 0$ – genauer:

$$2\varepsilon M(\varepsilon) < 1, \tag{7.15}$$

ist für jeden Startwert $x^{(0)} \in I_\varepsilon$ das Newton-Verfahren wohldefiniert und konvergiert quadratisch gegen die einzige Nullstelle $x_\star \in I_\varepsilon$.

Beweis. Wir erinnern an die Definition der quadratischen Konvergenz:

$$\left|x^{(i)} - x_\star\right| \le \varepsilon_i \qquad \text{und} \qquad \frac{\varepsilon_{i+1}}{\varepsilon_i^2} \le c.$$

1. Schritt. Wir zeigen: $x_\star$ ist die einzige Nullstelle von f in I_ε.
Verwende die Taylorsche Formel um die Nullstelle $x_\star$

$$f(x) = f(x_\star) + f'(x_\star)(x - x_\star) + \frac{f''(\xi)}{2}(x - x_\star)^2$$

mit einem Zwischenwert $\xi \in [x_\star, x]$. Falls also $x \in I_\varepsilon$ gilt, liegt auch $\xi \in I_\varepsilon$. Wir zeigen, dass $x \neq x_\star$ keine Nullstelle sein kann. Aus $x \neq x_\star$ und der Voraussetzung "$x_\star$ ist einfache Nullstelle" folgt:

$$f(x) = f'(x_\star)(x - x_\star)\left\{1 + \frac{f''(\xi)}{2f'(x_\star)}(x - x_\star)\right\}.$$

Wegen $x \neq x_\star$, der Annahme, dass $x_\star$ eine einfache Nullstelle ist und

$$\left|\frac{f''(\xi)}{2f'(x_\star)}(x - x_\star)\right| \le \varepsilon M(\varepsilon) \le 2\varepsilon M(\varepsilon) < 1 \Rightarrow \{\ldots\} \neq 0$$

folgt $f(x) \neq 0$.

2. Schritt: Wir zeigen: Alle Iterierten liegen in I_ε und aufeinanderfolgende Itererierte (falls $x^{(i)} \neq x_\star$ gilt) sind voneinander verschieden (andernfalls, würde die Iteration an einer falschen Stelle stehen bleiben).

Beweis durch Induktion: Sei $x^{(i)} \in I_\varepsilon$ und $x^{(i)} \neq x^{(i-1)}$. (Dies ist für $i = 1$ gesichert:

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

und $f\left(x^{(0)}\right) \neq 0$ für $x^{(0)} \neq x_\star$). Da Newtonsche dividierte Differenzen an einer Zwischenstelle (bis auf einen Faktor) mit der Ableitung von f gleicher Ordnung übereinstimmen, folgt

$$\left[x^{(i)}, x_\star\right] f = f'(\xi_1), \qquad \left[x^{(i)}, \xi_1\right] f' = f''(\xi_2)$$

für geeignete $\xi_1 \in \left[x_\star, x^{(i)}\right]$ und $\xi_2 \in \left[\xi_1, x^{(i)}\right] \subset \left[x_\star, x^{(i)}\right]$. Daraus folgt mit der Definition des Newton-Verfahrens und unter Verwendung von $x^{(i)} \neq x_\star$:

$$\begin{aligned} x^{(i+1)} - x_\star &= x^{(i)} - x_\star - \frac{f\left(x^{(i)}\right)}{f'\left(x^{(i)}\right)} = \left(x^{(i)} - x_\star\right)\left(1 - \frac{f\left(x^{(i)}\right) - f(x_\star)}{f'\left(x^{(i)}\right)\left(x^{(i)} - x_\star\right)}\right) \\ &= \left(x^{(i)} - x_\star\right)\left(1 - \frac{f'(\xi_1)}{f'\left(x^{(i)}\right)}\right) = \left(x^{(i)} - x_\star\right)\left(\frac{f'\left(x^{(i)}\right) - f'(\xi_1)}{f'\left(x^{(i)}\right)}\right) \\ &= \left(x^{(i)} - x_\star\right)\left(x^{(i)} - \xi_1\right) 2\frac{f''(\xi_2)}{2f'\left(x^{(i)}\right)}. \end{aligned} \tag{7.16}$$

Daraus folgt mit der Annahme über den Quotienten aus erster und zweiter Ableitung die Abschätzung

$$\left|x^{(i+1)} - x_\star\right| \leq \varepsilon\left(2\varepsilon M(\varepsilon)\right) < \varepsilon.$$

Daraus folgt $x^{(i+1)} \in I_\varepsilon$. Die Definition der Netwon-Iteration impliziert weiter, dass $x^{(i+1)} = x^{(i)}$ lediglich für $x^{(i)} = x_\star$ gelten kann.

3. Schritt: Beweis der Konvergenzordnung.

Indem wir nochmals (7.16) verwenden, erhalten wir unter Verwendung von $\left|x^{(i)} - \xi_1\right| \leq \left|x^{(i)} - x_\star\right|$:

$$\left|x^{(i+1)} - x_\star\right| \leq \left|x^{(i)} - x_\star\right|^2 2M(\varepsilon), \qquad i = 1, 2, 3, \ldots.$$

Daraus folgt für den Fehler $e_i = x^{(i)} - x_\star$.

$$\frac{e_{i+1}}{e_i^2} \leq 2M(\varepsilon)$$

und die Iteration konvergiert quadratisch. ■

Für praktische Anwendungen ist Satz 7.9 ungeeignet, da die Kenntnis der Nullstelle $x_\star$ in die Definition des Intervalls I_ε einfliesst. Häufiger ist man mit der Situation konfrontiert, dass ein Intervall $I_\delta(a) := [a - \delta, a + \delta]$ – beispielsweise durch das Bisektionsverfahren – bestimmt werden kann, welches die Nullstelle $x_\star$ sicher enthält. Das Ziel ist dann, δ so zu bestimmen, dass das Newtonverfahren mit Startwerten in I_δ sicher konvergiert.

Korollar 7.10 *Sei $I_\delta(a)$ wie zuvor definiert und δ erfülle*

$$4\delta \max_{s,t\in I_{3\delta}(a)} \left|\frac{f''(s)}{2f'(t)}\right| < 1.$$

Des weiteren sei x_* die einzige Nullstelle in $I_{3\delta}(a)$. Dann konvergiert das Newtonverfahren für jeden Startwert in $I_\delta(a)$.

Beweis. Wir müssen zeigen, dass für jedes $x \in I_\delta(a)$ die Bedingung (7.15) erfüllt ist. Sei $x_\star \in I_\delta(a)$ die Nullstelle und I_ε wie in Satz 7.9 definiert. Dann gilt

$$I_\delta(a) \subset I_\varepsilon \subset I_{3\delta}(a) \quad \text{mit } \varepsilon = 2\delta. \tag{7.17}$$

Die Bedingung

$$4\delta \max_{s,t\in I_{3\delta}(a)} \left|\frac{f''(s)}{2f'(t)}\right| < 1$$

impliziert dann $2\varepsilon M(\varepsilon) < 1$ und das Newton-Verfahren konvergiert gemäss Satz 7.9 für jeden Startwert in I_ε. Die linke Inklusion in (7.17) impliziert dann die Behauptung. ■

Wir wollen nun das Newton-Verfahren anwenden, um Nullstellen von Polynomen zu berechnen.

Die Idee hierbei ist, die Nullstellen nacheinander zu berechnen und jeweils per Polynomdivision aus dem Polynom herauszudividieren. Wir beginnen zunächst damit, ein Polynom effizient durch einen Linearfaktor zu dividieren.

Notation:

$$f(x) = x^d + \sum_{i=0}^{d-1} a_i x^i, \tag{7.18}$$

wobei der führende Koeffizienten o.B.d.A. gleich 1 gewählt wurde, da wir lediglich an den Nullstellen des Polynoms interessiert sind.

Wir betrachten zunächst das Teilproblem, das Polynom f aus (7.18) in einem Parameterpunkt t auszuwerten. Die Rekursion basiert auf der Polynomdivision mit Rest und verwenden den Ansatz

$$f(x) = (x-t)\left(x^{d-1} + \sum_{i=0}^{d-2} b_i x^i\right) + b_{-1}. \tag{7.19}$$

Offensichtlich gilt $f(t) = b_{-1}$, und wir leiten eine Rekursion zur Berechnung von b_{-1} her. Ausmultiplizieren der rechten Seite in (7.19) liefert

$$x^d + (-t + b_{d-2})\, x^{d-1} + (-tb_{d-2} + b_{d-3})\, x^{d-2} + \ldots (-tb_1 + b_0)\, x + (-tb_0 + b_{-1})$$

und Koeffizientenvergleich:

$$\begin{aligned} b_{d-2} &= a_{d-1} + t, \qquad & b_{d-3} &= a_{d-2} + tb_{d-2} \\ b_k &= a_{k+1} + tb_{k+1}, \qquad & k &= d-3, d-4, \ldots -1. \end{aligned} \tag{7.20}$$

Man prüft leicht nach, dass b_k ein Polynom in t vom Grad $d-k-1$ ist.

Die Rekursion (7.20) stellt eine effiziente Möglichkeit dar, ein Polynom in einem Punkt t auszuwerten und wird Horner-Schema genannt. Der Aufwand besteht in der Ausführung

von d Multiplikationen und d Additionen, was weniger ist, als die direkte Auswertung eines Polynoms vom Grad d und darüber hinaus noch bessere numerische Stabilitätseigenschaften besitzt.

Falls nun t eine Nullstelle von f ist, wird durch das Horner-Schema (7.20) eine Vorschrift definiert, um die Koeffizienten des um einen Grad reduzierten Polynoms $f(\cdot)/(\cdot - t)$ zu berechnen.

Die Idee, das Newton-Verfahren zur Berechnung der Nullstellen eines Polynoms zu verwenden, besteht nun in einer Iteration der folgenden beiden Schritte:

1. Berechne eine Nullstelle t von f nach dem Newton-Schema
2. Dividiere das Polynom durch den Linearfaktor $(x - t)$ zu dieser Nullstelle mit Hilfe des Horner-Schemas und ersetze das vorige Polynom durch das neue, um einen Grad reduzierte Polynom.

Im Newtonverfahren entsteht die Aufgabe die Funktion f und ihre Ableitung auszuwerten. Diese Auswertung sollte ebenfalls mit Hilfe des Horner-Schemas durchgeführt werden. Da die Ableitung ebenfalls ein Polynom ist, lässt sich diese effizient mit dem Horner-Schema ausgewerten. Beide Auswertungen lassen sich miteinander verbinden.

Übungsaufgabe 7.11 *Entwickeln Sie einen Algorithmus zur Berechnung der Nullstellen von Polynomen mit dem Newtonverfahren unter Verwendung der Polynomdivision. Verwenden Sie das Horner-Schema zur kombinierten Auswertung des Polynoms und dessen Ableitung.*

Bemerkung 7.12 *Da die iterative Anwendung des oben beschriebenen Doppelschritts zu einer Fehlerverstärkung führen kann (insbesondere, wenn der Grad des Polynom hoch ist), kann die Genauigkeit durch Nachiteration verbessert werden. Die berechneten, approximativen Nullstellen werden als Startwerte verwendet, um mit Hilfe des Newtonverfahrens bessere Näherungen dieser Nullstellen zu erhalten.*

Das Newton-Verfahren gehört zur Klasse der Fixpunktiterationen. Diese werden im folgenden systematisch diskutiert.

7.3 Fixpunkt-Iterationen

Viele nichtlineare Probleme lassen sich in natürlicher Weise als Fixpunktgleichungen schreiben

$$x = \varphi(x).$$

Jede Zahl x, welche diese Gleichung erfüllt, nennt man Fixpunkt. Beispielsweise definiert die Iterationsvorschrift des Newton-Verfahrens die Funktion

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

und die Lösung $x_\star$ der Fixpunktgleichung $x = \varphi(x)$ ist auch kontinuierliche Lösung $f(x_\star) = 0$, vorausgesetzt $f'(x_\star) \neq 0$. Die Iterationsvorschrift für das Newton-Verfahren lautet dann:

$$x^{(i+1)} = \varphi\left(x^{(i)}\right).$$

Falls diese Iteration konvergiert, ist der Grenzwert ein Fixpunkt, vorausgesetzt φ ist stetig. (Es gilt nämlich:

$$\alpha := \lim_{i\to\infty} x^{(i)} = \lim_{i\to\infty} \varphi\left(x^{(i)}\right) \overset{\varphi \text{ stetig}}{=} \varphi\left(\lim_{i\to\infty} x^{(i)}\right) = \varphi(\alpha)$$

und α ist daher ein Fixpunkt).

Für eine konvergente Fixpunktiteration lässt sich die Konvergenzordnung sehr einfach bestimmen. Sei $x_\star = \lim x^{(i)}$. Wir nehmen an, dass φ in einer Umgebung des Fixpunkts hinreichend häufig differenzierbar ist. Wir definieren eine Ganzzahl p durch

$$\begin{aligned} \varphi^{(m)}(x_\star) &= 0, \qquad 1 \le m \le p-1, \\ \varphi^{(p)}(x_\star) &\neq 0. \end{aligned} \tag{7.21}$$

Dann gilt mit Hilfe einer Taylorentwicklung um den Fixpunkt:

$$\begin{aligned} \varphi\left(x^{(i)}\right) &= \varphi(x_\star) + \sum_{m=1}^{p-1} \frac{1}{m!}\varphi^{(m)}(x_\star)\left(x^{(i)} - x_\star\right)^m + \frac{\left(x^{(i)} - x_\star\right)^p}{p!}\varphi^{(p)}\left(\xi^{(i)}\right) \\ &= \varphi(x_\star) + \frac{\left(x^{(i)} - x_\star\right)^p}{p!}\varphi^{(p)}\left(\xi^{(i)}\right) \end{aligned}$$

mit einer Zwischenstelle $\xi^{(i)} \in \left[x_\star, x^{(i)}\right]$. Setzen wir $x^{(i+1)} = \varphi\left(x^{(i)}\right)$ und verwenden $\varphi(x_\star) = x_\star$, folgt

$$\frac{1}{p!}\varphi^{(p)}(\xi_i) = \frac{x^{(i+1)} - x_\star}{\left(x^{(i)} - x_\star\right)^p}.$$

Nun haben wir angenommen, dass $x^{(i+1)}$ und $x^{(i)}$ gegen $x_\star$ konvergiert. Da $\xi^{(i)} \in \left[x_\star, x^{(i)}\right]$ gilt, folgt daraus auch $\xi^{(i)} \to x_\star$. Grenzübergang liefert (Stetigkeit von $\varphi^{(p)}$ um $x_\star$ vorausgesetzt und Verwendung von $\varphi^{(p)}(x_\star) \neq 0$)

$$0 \neq \frac{1}{p!}\varphi^{(p)}(x_\star) = \lim_{i\to\infty} \frac{x^{(i+1)} - x_\star}{\left(x^{(i)} - x_\star\right)^p}.$$

Das bedeutet, dass die Konvergenz dann von Ordnung p ist und für die asymptotische Fehlerkonstante in (7.3a) gilt $c = \frac{1}{p!}\varphi^{(p)}(x_\star)$. Daraus erhalten wir die folgende quantitative (lokale) Konvergenzaussage für abstrakte Fixpunktiterationen.

Satz 7.13 *Sei $x_\star$ ein Fixpunkt von φ und $I_\varepsilon = \{x \in \mathbb{R} : |x - x_\star| \le \varepsilon\}$. Sei $\varphi \in C^p[I_\varepsilon]$ mit p wie in (7.21). Falls*

$$M(\varepsilon) = \max_{t\in I_\varepsilon} |\varphi'(t)| < 1$$

gilt, dann konvergiert die Fixpunktiteration gegen $x_\star$ für jeden Startwert $x_0 \in I_\varepsilon$. Die Konvergenzordnung ist p und die asymptotische Fehlerkonstante durch $\frac{1}{p!}\varphi^{(p)}(x_\star)$ gegeben.

Beweis. Die Konvergenz wird in einem allgemeineren Zusammenhang für „kontrahierende“ Abbildungen im Unterabschnitt 7.3.1 bewiesen. Die Konvergenzordnung folgt direkt aus dem bisher abgeleiteten. ■

7.3.1 Systeme nichtlinearer Gleichungen

Wir werden hier im Detail diskutieren, wie das Newton-Verfahren und die Fixpunktiteration für Systeme nichtlinearer Gleichungen verallgemeinert werden kann. Wir betrachten eine vektorwertige Funktion $f : \mathbb{R}^d \to \mathbb{R}^d$ und betrachten das Problem, die Gleichung

$$f(x) = 0 \tag{7.22}$$

zu lösen. Fixpunktiterationen sind durch eine Iterationsvorschrift φ charakterisiert. Die Konvergenz der zugehörigen Fixpunktiteration $x^{(i+1)} = \varphi\left(x^{(i)}\right)$ ist gesichert, falls die Abbildung φ kontrahierend ist.

Kontrahierende Abbildungen Wir formulieren die Aufgabe (7.22) als Fixpunktiteration:

$$x_{i+1} = \varphi(x_i)$$

mit $\varphi(x) = x - f(x)$. Die Abbildung φ ist eine Kontraktion auf einer Menge $\mathcal{D} \subset \mathbb{R}^d$, falls eine Konstante γ mit $0 < \gamma < 1$ existiert, so dass

$$\|\varphi(x) - \varphi(y)\| \leq \gamma \|x - y\|, \qquad \forall x, y \in \mathcal{D}$$

gilt in einer geeigneten Vektornorm $\|\cdot\|$.

Satz 7.14 *Sei $\mathcal{D} \subset \mathbb{R}^d$ entweder kompakt oder $\mathcal{D} = \mathbb{R}^d$. Falls die Abbildung $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ eine Kontraktion auf der Menge $\mathcal{D}$ ist und $\varphi : \mathcal{D} \to \mathcal{D}$ gilt, dann*

1. *ist die Iteration $x_{i+1} = \varphi(x_i)$ wohldefiniert und konvergiert gegen den einzigen Fixpunkt $x_\star \in \mathcal{D}$,*
2. *gilt die Fehlerabschätzung:*

$$\|x_i - x_\star\| \leq \frac{\gamma^i}{1-\gamma} \|x_1 - x_0\|$$

und

$$\|x_i - x_\star\| \leq \gamma^i \|x_0 - x_\star\|.$$

Beweis. (a) Die Iteration ist wohldefiniert, da wegen $\varphi : \mathcal{D} \to \mathcal{D}$ die Iteration mit einem Startwert aus $\mathcal{D}$ immer in $\mathcal{D}$ enthalten ist.

(b) Wir zeigen, dass x_i ein Cauchy-Folge bildet und daher konvergiert. Auf Grund der Kontraktionseigenschaft gilt für die Differenz der alten und neuen Iterierten:

$$\|x_{i+1} - x_i\| = \|\varphi(x_i) - \varphi(x_{i-1})\| \leq \gamma \|x_i - x_{i-1}\|.$$

Wiederholte Anwendung liefert:

$$\|x_{i+1} - x_i\| \leq \gamma^i \|x_1 - x_0\|.$$

Wir spalten die Differenz $x_{i+p} - x_i$ auf als Teleskopsumme

$$x_{i+p} - x_i = (x_{i+p} - x_{i+p-1}) + (x_{i+p-1} - x_{i+p-2}) + \ldots + (x_{i+1} - x_i).$$

Die Dreiecksungleichung liefert:

$$\|x_{i+p} - x_i\| \leq \sum_{j=1}^{p} \|x_{i+j} - x_{i+j-1}\| \leq \|x_1 - x_0\| \sum_{j=1}^{p} \gamma^{i+j-1} \tag{7.23}$$
$$\leq \|x_1 - x_0\| \sum_{j=1}^{\infty} \gamma^{i+j-1} = \frac{\gamma^i}{1-\gamma} \|x_1 - x_0\| .$$

Wegen $\gamma^i \to 0$ für $i \to \infty$ haben wir gezeigt, dass x_i eine Cauchy-Folge ist und daher konvergiert. Aus der Kompaktheit von $\mathcal{D}$ folgt, dass der Grenzwert $x_\star$ auch in $\mathcal{D}$ angenommen wird.

(c) Wir zeigen, dass der Grenzwert $x_\star = \lim_{i\to\infty} x_i$ auch Fixpunkt ist.

Die Kontraktionseigenschaft von φ liefert

$$\|x_i - \varphi(x_\star)\| = \|\varphi(x_{i-1}) - \varphi(x_\star)\| \leq \gamma \|x_{i-1} - x_\star\| . \tag{7.24}$$

Grenzübergang $i \to \infty$ ergibt mit der Stetigkeit der Norm

$$\|x_\star - \varphi(x_\star)\| = 0,$$

d.h. $x_\star$ ist Fixpunkt.

(d) Wir zeigen, dass der Fixpunkt eindeutig ist.

Sei $y_\star$ ein weiterer Fixpunkt in $\mathcal{D}$. Dann gilt:

$$\|x_\star - y_\star\| = \|\varphi(x_\star) - \varphi(y_\star)\| \leq \gamma \|x_\star - y_\star\| .$$

Wegen $\gamma < 1$ folgt $\|x_\star - y_\star\| = 0$.

(e) Zur Konvergenzgeschwindigkeit:

Grenzübergang $p \to \infty$ in (7.23) liefert:

$$\|x_\star - x_i\| \leq \frac{\gamma^i}{1-\gamma} \|x_1 - x_0\| ,$$

d.h. die erste Konvergenzaussage. Die zweite Aussage folgt durch wiederholte Anwendung von (7.24). ■

Dieser Satz bildet die Grundlage, um das Newtonsche Verfahren auf Systeme von nichtlinearen Gleichungen anzuwenden. Die Iterationsvorschrift wurde bereits in (7.13) angegeben:

$$x_{i+1} = x_i - \Delta_i,$$

wobei Δ_i die Lösung des d-dimensionalen Gleichungssystems:

$$\mathbf{J}(x_i)\, \Delta_i = f(x_i)$$

ist.

Bemerkung 7.15 *Zur Durchführung des Newton-Verfahrens muss keine Matrix* ***invertiert*** *werden, sondern lediglich ein lineares Gleichungssystem gelöst werden.*

Wie im eindimensionalen Fall lässt sich zeigen, dass das Newton-Verfahren unter analogen Voraussetzungen quadratisch konvergiert. Da die Berechnung der Jacobi-Matrix für viele Anwendungen eine sehr rechenintensive Aufgabe darstellt, existieren eine Reihe von Modifikationen des Newton-Verfahrens. Eine Idee hierbei ist, die Jacobi-Matrix nicht in jedem Schritt neu zu berechnen sondern nur in jedem zweiten Schritt aufzudatieren. Andere Varianten verwenden eine oder mehrere zuvor berechnete Jacobi-Matrizen, um die neue Jacobi-Matrix daraus zu extrapolieren. Derartige Modifikationen fallen unter die Klasse der Quasi-Newton-Verfahren.

Bemerkung 7.16 *Da das Newton-Verfahren in vielen Fällen nur lokal konvergiert aber andererseits quadratisch in einer Umgebung der Nullstelle, wird das Verfahren häufig mit langsameren aber robusteren Iterationen kombiniert, die eine hinreichend genaue Startnäherung für das Newtonverfahren zu erzeugen.*

8 Iterative Verfahren zur Lösung LGS

8.1 Randwertprobleme

Viele zeitunabhängige, physikalische Probleme lassen sich durch „Potentialgleichungen“ beschreiben. Ein Beispiel hierfür ist die Berechnung des elektrischen Feldes in einem unendlich dünnen Leiter. In einer Dimension ist der Leiter durch ein Intervall $\Omega = (a, b)$ beschrieben. Das elektrische Potential (Spannung) in Ω wird durch die Gleichung

$$-u''(x) = f(x), \qquad \forall x \in \Omega \tag{8.1}$$

beschrieben. Dabei stellt f ein äusseres elektrisches Feld dar. In vielen Anwendungen ist die physikalisch relevante Grösse nicht die Spannung sondern das erzeugte elektrische Feld, d.h., der Ableitung von u nach der/den Ortsvariablen..

Aus der Theorie gewöhnlicher Differentialgleichungen ist bekannt, dass man für derartige Differentialgleichungen noch geeignete *Randbedingungen* vorgeben muss. Beispielsweise lassen sich die Werte von u in den Endpunkten des Leiters vorgeben

$$u(a) = g(a), \qquad u(b) = g(b)$$

mit einer Funktion $g : \Gamma \to \mathbb{R}$ auf dem Rand $\Gamma = \{a, b\}$ von Ω. Die Lösung dieses Problems ist aus numerischer Sicht äquivalent zu einem Quadraturproblem. Die exakte Lösung besitzt die Darstellung:

$$u(x) = \frac{(b-x)\, g(a) + (x-a)\left(g(b) + \int_a^b \int_a^s f(t)\, dtds\right)}{b-a} - \int_a^x \int_a^s f(t)\, dtds.$$

Falls das Integral $\int_a^x \int_a^s f(t)\, dtds$ nicht explizit auszuwerten ist, müssen numerische Quadraturverfahren zur Approximation von Integralen eingesetzt werden.

Mit Hilfe eindimensionaler Modelle lassen sich nur sehr wenige physikalische Probleme beschreiben. Interessanter sind mehrdimensionale Probleme, beispielsweise die Berechnung eines elektrischen Feldes, welches von Ladungen erzeugt werden. Das mehrdimensionale Analogon zur Gleichung (8.1) ist die Poisson-Gleichung (Potentialgleichung). Sei dazu $\Omega \subset \mathbb{R}^2$ ein zweidimensionales (allgemein d-dimensionales) Gebiet. Wir suchen eine Potentialfunktion $u : \Omega \to \mathbb{R}$, welche die Poisson-Gleichung

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \qquad \forall \mathbf{x} \in \Omega \tag{8.2a}$$

erfüllt. Hierbei bezeichnet Δ den Laplace-Operator, der für $u \in C^2(\Omega)$ durch

$$\Delta u(\mathbf{x}) = \frac{\partial^2 u(\mathbf{x})}{\partial x_1^2} + \frac{\partial^2 u(\mathbf{x})}{\partial x_2^2}$$

gegeben ist. Hier wurde $\mathbf{x} = (x_1, x_2)$ gesetzt.

Der Vorgabe der Funktionswerte in den Intervallenden entsprechen nun Randbedingungen auf dem Rand von Ω:

$$u(\mathbf{x}) = g(\mathbf{x}), \qquad \forall \mathbf{x} \in \Gamma := \partial\Omega \tag{8.2b}$$

für eine gegebene Funktion $g : \Gamma \to \mathbb{R}$.

Problem (8.2) ist ein Beispiel eines „Randwertproblems“, welches sich aus einer partiellen Differentialgleichung (8.2a) und aus gegebenen Randbedingungen (8.2b) zusammensetzt.

Im Gegensatz zum eindimensionalen Modellproblem lässt sich die Lösung dieser Gleichung nicht mehr geschlossen angeben. Numerische Diskretisierungsmethoden müssen daher zur (näherungsweisen) Lösung eingesetzt werden.

Die numerische Diskretisierung partieller Differentialgleichungen führt typischerweise auf sehr grosse lineare Gleichungssysteme. Diese besitzen die charakteristische Eigenschaft, dass nur sehr wenig (5-10) Matrixkoeffizienten pro Zeile von Null verschieden sind. Für diese Klasse von Matrizen sind direkte Eliminationsverfahren ungeeignet, da deren Aufwand kubisch mit der Dimension wächst. *Iterative Gleichungslöser*, die in diesem Kapitel behandelt werden, sind für diese Klasse von Problemen wesentlich effizienter.

8.2 Differenzenverfahren

In diesem Abschnitt werden wir das Poisson-Problem mit Differenzenverfahren diskretisieren. Wir werden sehen, dass derartige Diskretisierungen stets auf grosse, schwachbesetzte lineare Gleichungssysteme führen.

Die einfachste Möglichkeit, eine Ableitung durch eine Approximation zu ersetzen, ist durch einen Differenzenquotienten gegeben. Wir hatten gesehen, dass die symmetrische Differenz eine gute Approximation der ersten Ableitung darstellt

$$u'(x) \approx D_h^{\mathrm{sym}}[u](x) = \frac{u(x+h/2) - u(x-h/2)}{h}.$$

Ableitungen höherer Ordnung lassen sich durch Hintereinanderausführung der Approximation erster Ordnung approximieren. Für die zweite Ableitungen erhalten wir:

$$\begin{aligned} u''(x) &\approx D_h^{\mathrm{sym}}[u'](x) = \frac{u'(x+h/2) - u'(x-h/2)}{h} \\ &\approx \frac{D_h^{\mathrm{sym}}[u](x+h/2) - D_h^{\mathrm{sym}}[u](x-h/2)}{h} \\ &= \frac{u(x+h) - u(x)}{h^2} - \frac{u(x) - u(x-h)}{h^2} = \frac{1}{h^2}u(x+h) - \frac{2}{h^2}u(x) + \frac{1}{h^2}u(x-h). \end{aligned}$$

Der Laplace-Operator enthält zweite Ableitungen in x_1- und x_2- Richtung. In beiden Richtungen lässt sich der obige Differenzenoperator anwenden. Sei $\mathbf{e}_1 = (1,0)$ und $\mathbf{e}_2 = (0,1)$. Damit erhalten wir die folgende Approximation des Laplace-Operators:

$$\begin{aligned} -\Delta u(\mathbf{x}) &\approx -\frac{1}{h^2}u(\mathbf{x}+h\mathbf{e}_1) + \frac{2}{h^2}u(\mathbf{x}) - \frac{1}{h^2}u(\mathbf{x}-h\mathbf{e}_1) \\ &\quad - \frac{1}{h^2}u(\mathbf{x}+h\mathbf{e}_2) + \frac{2}{h^2}u(\mathbf{x}) - \frac{1}{h^2}u(\mathbf{x}-h\mathbf{e}_2) \\ &=: \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} u(\mathbf{x}) =: L_{\mathbf{x}}[u]. \end{aligned} \tag{8.3}$$

Der sogenannten 5-Punkt-*Sternnotation* aus der unteren Zeile liegt die geometrische Anschauung zugrunde, dass der Laplace-Operator $\Delta u(\mathbf{x})$ durch eine Linearkombination der Werte von

u in $\mathbf{x}$ und den kartesischen Nachbarpunkten $\mathbf{x} \pm h\mathbf{e}_1$ und $\mathbf{x} \pm h\mathbf{e}_2$ approximiert wird. Die Gewichtungsfaktoren (Koeffizienten) sind durch die Einträge der Sternnotation definiert.

Der Einfachheit halber beschränken wir uns im Folgenden auf das Einheitsquadrat $\Omega = (0,1)^2$. Die Methoden und Ergebnisse übertragen sich direkt auf beliebige, rechtwinklig berandete Gebiete.

Wir definieren die Menge der kartesischen Gitterpunkte zur Schrittweite h. Sei dazu $n \in \mathbb{N}$ und $h = (n+1)^{-1}$. Für $0 \leq i, j \leq n+1$ definieren wir die Gitterpunkte $\mathbf{x}_{ij} = (ih, jh)$ ein und fassen diese in der Menge der Knotenpunkte $\Theta = \{\mathbf{x}_{ij} : 0 \leq i, j \leq n+1\}$ zusammen.

Ziel: Approximiere die Lösung u von (8.2a) und (8.2b) in den Gitterpunkten näherungsweise durch Ersetzen des Laplace-Operators durch (8.3).

Da die Lösung auf dem Rand bereits bekannt ist, muss dort nichts berechnet werden, sondern die Randdaten können direkt als Lösung übernommen werden. Die unbekannte Approximation ist also nur in den *inneren* Gitterpunkten zu bestimmen, die zur Menge $\Theta_0 := \Theta \backslash \Gamma$ zusammengefasst werden. Die Näherungslösung in den Gitterpunkten wird mit $u_{ij}^h = u^h(\mathbf{x}_{ij})$ bezeichnet. Ersetzt man nun die Gleichung (8.2a) in den *inneren* Gitterpunkten durch die Approximation (8.3) erhalten wir:

$$L_{\mathbf{x}}\left[u^h\right] = f(\mathbf{x}), \qquad \forall \mathbf{x} \in \Theta_0. \tag{8.4}$$

Aufgrund der Definition des 5-Pkte-Differenzensterns treten auf der linken Seite in (8.4) auch Gitterpunkte auf, die auf dem Rand Γ liegen. Ersetzen der Werte von u_{ij}^h in diesen Randpunkten durch $g_{ij} = g(\mathbf{x}_{i,j})$ für $\mathbf{x}_{i,j} \in \Theta \cap \Gamma$ eliminiert diese Werte, und wir erhalten modifizierte Sterne auf der zweitäussersten Gitterlinie. Dies ergibt die folgenden Gleichungen für die unbekannten Werte $\left(u_{ij}^h\right)_{i,j=1}^n$.

- Für $\mathbf{x} \in \Theta_0$ und $\operatorname{dist}(\mathbf{x}, \Gamma) \geq 2h$:

$$M_{\mathbf{x}} u := L_{\mathbf{x}} u = f(\mathbf{x}) =: r(\mathbf{x}).$$

- Für $x = (h, h)$ setzen wir

$$M_{\mathbf{x}} u := h^{-2} \begin{bmatrix} & -1 & \\ 0 & 4 & -1 \\ & 0 & \end{bmatrix} u(\mathbf{x}) = f(\mathbf{x}) + h^{-2} g(h,0) + h^{-2} g(0,h) =: r(\mathbf{x})$$

  und verfahren analog mit den anderen Punkten

$$\mathbf{x} \in \{(h,h), (1-h,h), (h,1-h), (1-h,1-h)\}.$$

- Für $x = (x_1, x_2) \in \Theta_0$ mit $2h \leq x_1 \leq 1-2h$ und $x_2 = h$ setzen wir

$$M_{\mathbf{x}} u := h^{-2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & 0 & \end{bmatrix} u(\mathbf{x}) = f(\mathbf{x}) + h^{-2} g(x_1, 0) =: r(\mathbf{x})$$

  und verfahren analog für die anderen Punkte

$$\mathbf{x} \in \{(ih,h), (h,ih), (ih,1-h), (1-h,ih) : 2 \leq i \leq n-1\}.$$

Zusammen ergibt sich das lineares Gleichungssystem

$$M_{\mathbf{x}}[u] = r(\mathbf{x}), \qquad \forall \mathbf{x} \in \Theta_0 \tag{8.5}$$

für die unbekannten Komponenten $u_{i,j}$, $1 \leq i,j \leq n$. Dies sind n^2 Gleichungen für n^2 unbekannte Funktionswerte. Man beachte, dass dieses lineare Gleichungssystem noch nicht in Matrixform ist, da keine Numerierung der Gitterpunkte eingeführt wurde. Um zu einer Numerierung und damit zu einem linearen Gleichungssystem zu gelangen, wählen wir eine lexikographische Numerierung von links unten nach rechts oben. Ein Gitterpunkt $\mathbf{x}_{ij} = (ih, jh)$ besitzt dann die Nummer

$$k = (i-1)\,n + j.$$

Die zugehörige Komponente des Näherungslösung $\mathbf{u}^h \in \mathbb{R}^{n^2}$ lautet $\mathbf{u}^h_k := u^h(\mathbf{x}_{ij})$, $1 \leq k \leq N = n^2$. Die Gleichung in einem inneren Punkt $\mathbf{x}_k \in \Theta$ mit $\operatorname{dist}(\mathbf{x}, \Gamma) \geq 2h$ besitzt die Form

$$\begin{aligned} M_{\mathbf{x}_k}[u] &= h^{-2}\left(-u(\mathbf{x}_{i-1,j}) - u(\mathbf{x}_{i,j-1}) - u(\mathbf{x}_{i,j+1}) - u(\mathbf{x}_{i+1,j}) + 4u_{ij}\right) \\ &= h^{-2}\left(-u_{k-n} - u_{k-1} - u_{k+1} - u_{k+n} + 4u_k\right) = f(\mathbf{x}_k) \end{aligned}$$

und in den übrigen Gleichungen werden Koeffizienten mit Indizes ausserhalb des Bereiches $\{1, 2, \ldots N\}$ weggelassen. Die zugehörige Matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ lässt sich schreiben als $n \times n$-Blockmatrix mit $n \times n$ Matrizen als Einträge:

$$\mathbf{M} := h^{-2}\begin{bmatrix} \mathbf{T} & -\mathbf{I} & \mathbf{0} & \ldots & \mathbf{0} \\ -\mathbf{I} & \mathbf{T} & -\mathbf{I} & \ddots & \vdots \\ \mathbf{0} & -\mathbf{I} & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & & -\mathbf{I} \\ \mathbf{0} & \ldots & \mathbf{0} & -\mathbf{I} & \mathbf{T} \end{bmatrix} \quad \text{mit} \quad \mathbf{T} := \begin{bmatrix} 4 & -1 & 0 & \ldots & 0 \\ -1 & 4 & -1 & \ddots & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & & -1 \\ 0 & \ldots & 0 & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

der Einheitsmatrix $\mathbf{I} \in \mathbb{R}^{n\times n}$ und der Nullmatrix $\mathbf{0} \in \mathbb{R}^{n\times n}$.

Der zughörige Vektor der rechten Seite $\mathbf{r} \in \mathbb{R}^N$ ist wie folgt gegeben

$$\mathbf{r}_k := r(\mathbf{x}_{\ell,m}) \qquad \text{mit } k =: (\ell-1)\,n + m, \qquad 1 \leq k \leq N.$$

Die Lösung des Gleichungssystems

$$\mathbf{M}\mathbf{u}^h = \mathbf{r} \tag{8.6}$$

ergibt einen Koeffizientenvektor $\mathbf{u}^h \in \mathbb{R}^N$ dessen Komponente $\mathbf{u}^h_k$ die Näherung im Gitterpunkt $x_{i,j}$ mit $k = (i-1)\,n + j$ darstellt. Offensichtlich ist die Matrix $\mathbf{M}$ schwach besetzt, pro Zeile sind maximal 5 Einträge von Null verschieden. Auf der anderen Seite ist die Bandbreite der Matrix nicht klein. Sie beträgt n und wächst daher mit wachsender Anzahl der Gitterpunkte. Direkte Eliminationsverfahren werden daher sehr aufwendig, selbst Varianten, die lediglich auf dem einhüllenden Matrixband operieren, werden bei grossen Problemen zu aufwendig.

Das Problem (8.6) besitzt praktisch alle charakteristischen Eigenschaften von LGS, die durch Diskretisierung von Randwertproblemen entstehen. Wir bezeichnen dieses Problem daher als „Poisson-Modellproblem“ und werden es zum Testen der Iterationsverfahren häufig verwenden.

8.3 Iterative Verfahren zur Lösung schwachbesetzter LGS

Bezeichnungen: Sei $\mathbf{A}$ eine reguläre $n \times n$-Matrix mit Koeffizienten a_{ij}, $1 \leq i, j \leq n$.

Ziel ist es, für eine gegebene rechte Seite $\mathbf{b} \in \mathbb{R}^n$ die Lösung des linearen Gleichungssystems

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{8.7}$$

zu bestimmen.

Bei der iterative Lösung eines linearen Gleichungssystems wird, ausgehend von einem Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$, eine Folge von „Iterierten“ gebildet:

$$\mathbf{x}^{(0)} \to \mathbf{x}^{(1)} \to \ldots \to \mathbf{x}^{(m)} \to \mathbf{x}^{(m+1)} \to \ldots,$$

die gegen die exakte Lösung konvergieren sollen. In den folgenden Beispielen ist $\mathbf{x}^{(m+1)}$ lediglich von der vorigern Iterierten $\mathbf{x}^{(m)}$ abhängig. Für Iterationsverfahren wollen wir generell verlangen, dass die exakte Lösung ein Fixpunkt dieser Iteration ist.

Die Konvergenzgeschwindigkeit von Iterationsverfahren hängt von der Iterationsvorschrift *und* dem gewählten Startwert ab. Da sich herausstellen wird, dass die Konvergenz von Iterationsverfahren in vielen Fällen (fast) unabhängig vom Startwert ist und die Wahl $\mathbf{x}^{(0)} = \mathbf{0}$ in vielen Fällen zu befriedigenden Ergebnissen führt, werden wir uns in diesem Abschnitt auf die Konstruktion von Iterationsverfahren beschränken.

Praktisch alle Iterationsverfahren basieren auf einer geeigneten Zerlegung der Matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F} \tag{8.8}$$

mit einer Diagonalmatrix $\mathbf{D}$, einer strikten unteren Dreiecksmatrix $\mathbf{E}$ und einer strikten oberen Dreiecksmatrix $\mathbf{F}$

$$\forall j : j \geq i : e_{i,j} = 0, \qquad f_{j,i} = 0.$$

Das Gleichungssystem (8.7) ist dann äquivalent zu

$$(\mathbf{D} - \mathbf{E})\,\mathbf{x} = \mathbf{b} + \mathbf{F}\mathbf{x}. \tag{8.9}$$

Ähnlich wie bei der Konstruktion einer geeigneten Iterationsvorschrift für nichtlineare Gleichungen ersetzt man nun $\mathbf{x}$ auf der linken Seite von (8.9) durch $\mathbf{x}^{(m+1)}$ und auf der rechten Seite durch $\mathbf{x}^{(m)}$. Die neue Iterierte erhält man dann als Lösung des linearen Gleichungssystems:

$$(\mathbf{D} - \mathbf{E})\,\mathbf{x}^{(m+1)} = \mathbf{b} + \mathbf{F}\mathbf{x}^{(m)}. \tag{8.10}$$

Diese Vorschrift definiert das *Gauss-Seidel-Verfahren*. Da die Matrix $\mathbf{D} - \mathbf{E}$ eine linke untere Dreiecksmatrix ist, lässt sich dieses LGS (8.10) durch einfaches Rückwärtseinsetzen lösen. Rekursiv erhalten wir:

$$x_i^{(m+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{i,j} x_j^{(m)} \right) / a_{i,i}. \tag{8.11}$$

Übungsaufgabe 8.1 *Das Gauss-Seidel-Verfahren hängt von der gewählten Numerierung der Gitterpunkte ab.*

Numerische Tests für die 5-Pkte-Diskretisierung des Poisson-Modellproblems zeigen, dass die Konvergenzgeschwindigkeit des Gauss-Seidel-Verfahrens sehr langsam ist und mit grösser werdender Dimension immer schlechter wird. Bevor wir dessen Konvergenz analysieren geben wir noch zwei andere, häufig verwendete Verfahren an.

Eine andere Darstellung des Gauss-Seidel-Verfahrens stellt das Verfahren als Korrekturmethode dar: Die Iterationsvorschrift (8.11) ist äquivalent zu

$$x_i^{(m+1)} = x_i^{(m)} - \left(\sum_{j=1}^{i-1} a_{i,j} x_j^{(m+1)} + \sum_{j=i}^{n} a_{i,j} x_j^{(m)} - b_i \right) / a_{i,i}.$$

Eine Verallgemeinerung dieses Verfahrens stellt die Einführung eines eindimensionalen Parameters ω dar:

$$x_i^{(m+1)} = x_i^{(m)} - \omega \left(\sum_{j=1}^{i-1} a_{i,j} x_j^{(m+1)} + \sum_{j=i}^{n} a_{i,j} x_j^{(m)} - b_i \right) / a_{i,i}.$$

Dieses Verfahren wird SOR-Verfahren („successive overrelaxation“) genannt, und es kommt darauf an, den Parameter ω „möglichst optimal“ zu wählen. Wir werden sehen, dass für das betrachtete Poisson-Modellproblem die Wahl $\omega = 2/(1 + \sin(\pi h))$ ein geeigneter Wert ist. Numerische Tests bestätigen, dass dieses Verfahren wesentlich schneller für das Poisson-Modell-Problem konvergiert als das ursprüngliche Gauss-Seidel-Verfahren. Die Analyse dieser Modellprobleme ist Gegenstand des folgenden Abschnitts.

Ein weiteres Verfahren ist das Jacobi-Verfahren, welches wiederum auf der Zerlegung (8.8) beruht. Dieses Mal wird die neue Iterierte gemäss

$$\mathbf{D}\mathbf{x}^{(m+1)} = \mathbf{b} + (\mathbf{F} + \mathbf{E})\,\mathbf{x}^{(m)}$$

bestimmt. Die Koeffizientendarstellung lautet in Korrekturdarstellung

$$x_i^{(m+1)} = x_i^{(m)} - \left(\sum_{j=1}^{n} a_{ij} x_j^{(m)} - b_i \right) / a_{ii}.$$

Die gedämpfte Variante verwendet einen Dämpfungsparameter ω:

$$x_i^{(m+1)} = x_i^{(m)} - \omega \left(\sum_{j=1}^{n} a_{ij} x_j^{(m)} - b_i \right) / a_{ii}.$$

8.3.1 Konvergenzanalyse für Jacobi, Gauss-Seidel und SOR-Verfahren

Die Konvergenzanaylse der bisher vorgestellten Iterationsverfahren zur Lösung linearer Gleichungssysteme basiert auf einer Eigenwertdarstellung der diskreten Operatoren. Die Eigenvektoren/werte des Differenzenoperators zur 5-Pkte-Diskretisierung zum Poisson-Modellproblem können explizit angeben werden.

Lemma 8.2 *Zum Differenzenoperator M aus (8.5) existieren n^2 Eigenpaare e_{ij}, $1 \leq i, j \leq n$. Die Eigenfunktionen sind Abbildungen $e_{ij} : \Theta_0 \to \mathbb{R}$ die*

$$e_{ij}(\mathbf{y}) := \frac{h}{2} \sin(i\pi y_1) \sin(j\pi y_2), \qquad \mathbf{y} \in \Theta_0.$$

Die zugehörigen Eigenwerte lauten $\lambda_{ij} := 4h^{-2}\left(\sin^2\frac{\pi ih}{2} + \sin^2\frac{\pi jh}{2}\right)$, *d.h.*

$$Me_{ij} = \lambda_{ij} e_{ij}.$$

Die Eigenfunktionen e_{ij} *bilden eine Orthonormalbasis des* $\mathbb{R}^{\Theta_0}$ *(versehen mit dem Skalarprodukt):*

$$\langle u, v\rangle := \sum_{\mathbf{x}\in\Theta_0} u(x)\, v(x). \tag{8.12}$$

Beweis. Übungsaufgabe. ■

Korollar 8.3 *Die Eigenwerte von* M *sind nicht alle verschieden. Die Vielfachheit der Eigenwerte ist durch die Anzahl der verschiedenen Indexpaare* (i,j), (i',j') *gegeben, für die* $\lambda_{ij} = \lambda_{i'j'}$ *gilt. Der minimale Eigenwert wird für* $(i,j) = (1,1)$ *angenommen und der maximale für* $(i,j) = (n,n)$*:*

$$\lambda_{\min} = 8h^{-2}\sin^2\frac{\pi h}{2},$$
$$\lambda_{\max} = 8h^{-2}\cos^2\frac{\pi h}{2}.$$

Bemerkung 8.4 *Die Operatornorm von* M *ist durch* $\lambda_{\max}$ *beschränkt und die Norm der Inversen durch* $\lambda_{\min}^{-1}$*.*

Beweis. Für die Operatornorm gilt:

$$\|M\| = \sup_{v\in\mathbb{R}^{\Theta_0}\setminus\{0\}} \frac{\|Mv\|}{\|v\|},$$

wobei hier $\|\cdot\|$ die zum Skalarprodukt (8.12) gehörende Norm bezeichnet ($\|\cdot\| := \langle\cdot,\cdot\rangle^{1/2}$). Entwicklung von v nach Eigenfunktionen:

$$v = \sum_{i,j=1}^{n} \alpha_{ij} e_{ij}$$

liefert:

$$\|Mv\|^2 = v\ M\ Mv = \sum_{i,j,\nu,\mu=1}^{n} \alpha_{ij} e_{ij} \lambda_{i,j} \lambda_{\nu\mu} e_{\nu\mu} \alpha_{\nu\mu} = \sum_{i,j=1}^{n} \lambda_{ij}^2 \alpha_{ij}^2 \le \lambda_{\max}^2 \sum_{i,j=1}^{n} \alpha_{ij}^2,$$
$$\|v\|^2 = \sum_{i,j=1}^{n} \alpha_{ij}^2.$$

Daraus folgt:

$$\|M\| \le \lambda_{\max}.$$

Setzen wir $v = e_{\max}$, wobei $e_{\max}$ die Gitterfunktion zum Eigenwert $\lambda_{\max}$ bezeichnet erhalten wir

$$\|M\| \ge \lambda_{\max}$$

und daraus die Abschätzung für die Operatornorm von M. Um die Norm der Inversen abzuschätzen, verwendet man den Ansatz:

$$\left\|M^{-1}\right\| = \sup_{v\in\mathbb{R}^{\Theta_0}\setminus\{0\}} \frac{\|M^{-1}v\|}{\|v\|} \overset{v=Mw}{=} \sup_{v\leq\mathbb{R}^{\Theta_0}\setminus\{0\}} \frac{\|w\|}{\|Mw\|} = \frac{1}{\inf \frac{\|Mw\|}{\|w\|}}$$

und eine analoge Entwicklung von w in Eigenfunktionen. ■

Korollar 8.5 *Da der Operator M symmetrisch ist und alle Eigenwerte positiv sind, ist M ein positiv definiter Operator:*

$$\langle v, Mv\rangle > 0, \qquad \forall v \in \mathbb{R}^{\Theta_0}\setminus\{0\}\,.$$

Konvergenzanalyse für lineare Iterationsverfahren Das einfachste (sinnvolle) Iterationsverfahren ist das Richardson-Verfahren. In der gedämpften Version lautet die Iterationsvorschrift:

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} - \omega\left(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\right).$$

Eine andere Darstellung, die für die Konvergenzanalyse geeigneter ist, ist durch

$$\mathbf{x}^{(m+1)} = (\mathbf{I} - \omega\mathbf{A})\,\mathbf{x}^{(m)} + \omega\mathbf{b}$$

gegeben. Die Matrix $\mathbf{K}_\omega^{Rich} := (\mathbf{I} - \omega\mathbf{A})$ wird Iterationsmatrix der *Richardson-Iteration* genannt. Für die Konvergenz von Iterationsverfahren spielt der *Spektralradius* der Iterationsmatrix die entscheidende Rolle.

Definition 8.6 *Der Spektralradius einer quadratischen Matrix $\mathbf{A}$ ist der betragsmässig grösste Eigenwert von $\mathbf{A}$:*

$$\rho(\mathbf{A}) := \max\{|\lambda| : \lambda \text{ Eigenwert von } \mathbf{A}\}.$$

Der Spektralradius eines Differenzenoperators ist analog definiert.

Lemma 8.7 *Seien $\lambda_{\min}$ bzw. $\lambda_{\max}$ der kleinste, bzw. grösste Eigenwert einer positiv definiten Matrix $\mathbf{A}$. Für das Richardson-Verfahren gilt:*

$$\rho\left(\mathbf{K}_\omega^{Rich}\right) = \max\{|1-\omega\lambda_{\min}|, |1-\omega\lambda_{\max}|\}.$$

Beweis. Die Eigenwerte von $\mathbf{K}_\omega^{Rich}$ haben die Bauart: $1-\omega\lambda_i$ mit den Eigenwerten λ_i von $\mathbf{A}$. Da die Funktion $p(x) = |1-\omega x|$ kein lokales Maximum besitzt, wird dieses in einem der beiden Intervallenden $\lambda_i \in [\lambda_{\min}, \lambda_{\max}]$ angenommen. ■

In Satz 8.16 werden wir zeigen, unter welchen Voraussetzungen das Richardson-Verfahren konvergiert. Wesentlich wichtiger für die *Bewertung* des Verfahrens ist jedoch die Frage, wie schnell das Verfahren konvergiert. Dies wollen wir zunächst abstrakt definieren und dann auf das Richardson-Verfahren anwenden.

Für $\mathbf{A} \in \mathbb{R}^{n\times n}$ ($\mathbf{A}$ regulär) und $\mathbf{b} \in \mathbb{R}^n$ betrachten wir das Problem

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{8.13}$$

Dazu verwenden wir Iterationsverfahren der Form:

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{N}\left(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\right) \tag{8.14}$$

mit einer Matrix $\mathbf{N}$, die das Iterationsverfahren charakterisiert. Die Iterationsmatrix dieses Verfahren ist dann durch $\mathbf{K} := \mathbf{I} - \mathbf{N}\mathbf{A}$ gegeben.

Definition 8.8 *Ein Iterationsverfahren der Form (8.14) heisst regulär, falls* $\mathbf{N}$ *regulär ist.*

Definition 8.9 *Die Konvergenzrate eines Iterationsverfahrens der Form (8.14)*

$$\mathbf{x}^{(m+1)} = \mathbf{K}\mathbf{x}^{(m)} + \mathbf{N}\mathbf{b}$$

ist der Spektralradius $\rho(\mathbf{K})$ *der Iterationsmatrix* $\mathbf{K}$*.*

Um die Konvergenzgeschwindigkeit definieren zu können, müssen wir geeignete Normen auf $\mathbb{R}^n$ wählen.

Lemma 8.10 *Sei* $\|\cdot\|$ *eine Norm auf* $\mathbb{R}^n$*. Dann gilt für den Fehler* $\mathbf{e}^{(m)} := \mathbf{x}^{(m)} - \mathbf{x}$ *(wobei* $\mathbf{x}$ *die exakte Lösung von* $\mathbf{A}\mathbf{x} = \mathbf{b}$*) bezeichnet:*

$$\left\|\mathbf{e}^{(m)}\right\| \leq \|\mathbf{K}\|^m \left\|\mathbf{e}^{(0)}\right\|.$$

Beweis. Übungsaufgabe ■

Die Konvergenzgeschwindigkeit hängt von der Wahl einer geeigneten Norm ab. Wir werden zeigen, dass (i) die Konvergenz des Verfahrens mit Hilfe des Spektralradius der Iterationsmatrix beschrieben werden kann und (ii) dass der Spektralradius der Iterationsmatrix eine *untere* Schranke für $\|\mathbf{K}\|$ darstellt. Daher kann der Spektralradius als Mass für die Güte eines Iterationsverfahrens verwendet werden.

Lemma 8.11 *Sei* $\|\cdot\|$ *eine Norm auf* $\mathbb{R}^n$*. Dann gilt für alle Eigenwerte einer quadratischen Matrix* $\mathbf{A}$*:*

$$\begin{aligned} |\lambda| &\leq \|\mathbf{A}\|, && \text{für alle Eigenwerte von } \mathbf{A} \\ \rho(\mathbf{A}) &\leq \|\mathbf{A}\|, && \text{für alle Matrizen } \mathbf{A}. \end{aligned}$$

Beweis. Sei $\mathbf{e}$ der (auf Eins normierte) Eigenvektor zum Eigenwert λ. Dann gilt:

$$|\lambda| = \|\lambda\mathbf{e}\| = \|\mathbf{A}\mathbf{e}\| \leq \|\mathbf{A}\| \|\mathbf{e}\| = \|\mathbf{A}\|.$$

Die zweite Aussage folgt direkt aus der ersten. ■

Das folgende Lemma zeigt, dass man für jedes $\varepsilon > 0$ eine Norm auf $\mathbb{R}^n$ definieren kann, so dass die Norm von $\mathbf{K}$ nur um maximal ε vom Spektralradius abweicht.

Lemma 8.12 *Für jede Matrix* $\mathbf{B}$ *und jedem* ε *existiert eine Vektornorm* $\|\cdot\|$ *auf* $\mathbb{R}^n$ *mit*

$$\rho(\mathbf{B}) \leq \|\mathbf{B}\| \leq \rho(\mathbf{B}) + \varepsilon \tag{8.15}$$

gilt.

Der Beweis dieses Lemmas findet sich in [J. Stoer, R. Bulirsch: Numerische Mathematik II, Satz (6.9.2)] und wird hier weggelassen.

Im folgenden, abstrakten Konvergenztheorem beschränken wir uns auf Iterationsverfahren der Form (8.14) zur Lösung des linearen Gleichungssystems (8.13). Wir beginnen mit der präzisen Definition von Konvergenz.

Definition 8.13 *Ein Iterationsverfahren der Form (8.14) heisst konvergent, falls für alle rechten Seiten* $\mathbf{b} \in \mathbb{R}^n$ *in (8.13) ein Grenzwert* $\mathbf{x}^\star$ *existiert, der unabhängig vom Startwert* $\mathbf{x}^{(0)}$ *ist.*

Satz 8.14 *Ein Iterationsverfahren der Form (8.14) ist konvergent genau dann, wenn für den Spektralradius der Iterationsmatrix gilt:*

$$\rho(\mathbf{K}) < 1.$$

Es konvergiert gegen die exakte Lösung von (8.13), falls es regulär ist.

Beweis. (i) Iterationsverfahren konvergent$\Rightarrow \rho(\mathbf{K}) < 1$.
Wir wählen $\mathbf{b} = \mathbf{0}$ in (8.13). Dann lautet das Iterationsverfahren:

$$\mathbf{x}^{(m)} = \mathbf{K}^m \mathbf{x}^{(0)}.$$

Der Startwert $\mathbf{x}^{(0)} = \mathbf{0}$ liefert den Grenzwert $\mathbf{x}^\star = \mathbf{0}$, der nach Definition von konvergenten Iterationsverfahren für alle Startwerte gelten muss. Sei $\mathbf{e}$ der Eigenvektor zum (betragsmässig) maximalen Eigenwert $\lambda_{\max}$. Wählen wir diesen als Startwert erhalten wir:

$$\mathbf{0} = \lim_{m\to\infty} \mathbf{K}^m \mathbf{e} = \lim_{m\to\infty} \lambda_{\max}^m \mathbf{e} = \left(\lim_{m\to\infty} \lambda_{\max}^m\right) \mathbf{e}.$$

Wegen $\mathbf{e} \neq \mathbf{0}$ muss $\lim_{m\to\infty} \lambda_{\max}^m = 0$ gelten, was $|\lambda_{\max}| < 1$ impliziert.
(ii) $\rho(\mathbf{K}) < 1 \Rightarrow$Iterationsverfahren konvergent.
(ii a) In Hilfssatz 8.15 wird bewiesen, dass für jede zugehörige Matrixnorm $|||\cdot|||$ gilt:

$$\rho(\mathbf{A}) = \lim_{m\to\infty} |||\mathbf{A}|||^{1/m}. \tag{8.16}$$

(ii b) Wir zeigen

$$\lim_{m\to\infty} |||\mathbf{K}^m||| = 0. \tag{8.17}$$

Sei $\rho(\mathbf{K}) =: \overline{\rho} < 1$. Sei $\bar{\rho} < \rho' < 1$. Wegen (8.16) gilt für hinreichend grosses m_0:

$$|||\mathbf{K}^m|||^{1/m} \leq \rho', \qquad \forall m \geq m_0$$

bzw.

$$|||\mathbf{K}^m||| \leq (\rho')^m, \qquad \forall m \geq m_0.$$

Für $m \to \infty$ geht aber wegen $\rho' < 1$ die rechte Seite gegen Null und daher gilt (8.17).
(ii c) Wir verwenden die Darstellung

$$\mathbf{x}^{(m+1)} = \mathbf{K}\mathbf{x}^{(m)} + \mathbf{N}\mathbf{b}$$

der Iterationsvorschrift. Rekursiv ergibt sich

$$\begin{aligned}
\mathbf{x}^{(m)} &= \mathbf{K}\mathbf{x}^{(m-1)} + \mathbf{N}\mathbf{b} = \mathbf{K}\left(\mathbf{K}\mathbf{x}^{(m-2)} + \mathbf{N}\mathbf{b}\right) + \mathbf{N}\mathbf{b} = \mathbf{K}^2\mathbf{x}^{(m-2)} + \mathbf{K}\mathbf{N}\mathbf{b} + \mathbf{N}\mathbf{b} \\
&= \mathbf{K}^3\mathbf{x}^{(m-3)} + \mathbf{K}^2\mathbf{N}\mathbf{b} + \mathbf{K}\mathbf{N}\mathbf{b} + \mathbf{N}\mathbf{b} = \dots \\
&= \mathbf{K}^m\mathbf{x}^{(0)} + \left(\sum_{i=0}^{m-1} \mathbf{K}^i\right) \mathbf{N}\mathbf{b}.
\end{aligned} \tag{8.18}$$

Aus (ii a,b) folgt:

$$\left\|\mathbf{K}^m\mathbf{x}^{(0)}\right\| \overset{m\to\infty}{\to} 0$$

und daher $\mathbf{K}^m\mathbf{x}^{(0)} \overset{m\to\infty}{\to} \mathbf{0}$ für **alle** Startwerte $\mathbf{x}^{(0)}$. Für die Summe rechnet man leicht nach, dass

$$\sum_{i=0}^{m-1}\mathbf{K}^i\left(\mathbf{I}-\mathbf{K}\right) = \mathbf{I}-\mathbf{K}^m$$

gilt. Wegen $\rho\left(\mathbf{K}\right) < 1$ ist 1 kein Eigenwert von $\mathbf{K}$ und daher $\mathbf{I}-\mathbf{K}$ regulär. Daraus folgt:

$$\sum_{i=0}^{m-1}\mathbf{K}^i = \left(\mathbf{I}-\mathbf{K}^m\right)\left(\mathbf{I}-\mathbf{K}\right)^{-1}.$$

Wegen $\mathbf{K}^m \overset{m\to\infty}{\to} 0$ konvergiert die geklammerte Summe in (8.18) gegen $\left(\mathbf{I}-\mathbf{K}\right)^{-1}$. Für **jede** rechte Seite $\mathbf{b}$ erhalten wir daher den Grenzwert der Iteration

$$\mathbf{x}^\star = \left(\mathbf{I}-\mathbf{K}\right)^{-1}\mathbf{N}\mathbf{b} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x},$$

wobei

$$\mathbf{K} = \mathbf{I}-\mathbf{N}\mathbf{A} \Leftrightarrow \mathbf{N}\mathbf{A} = \left(\mathbf{I}-\mathbf{K}\right) \Leftrightarrow \left(\mathbf{I}-\mathbf{K}\right)^{-1} = \mathbf{A}^{-1}\mathbf{N}^{-1}$$

verwendet wurde. ■

Es bleibt der Hilfssatz nachzutragen.

Hilfssatz 8.15 *Sei* $\mathbf{B}$ *eine reguläre Matrix. Weiter sei* $\|\cdot\|$ *eine Vektornorm auf* $\mathbb{R}^n$ *und* $|||\cdot|||$ *eine zugordnete Matrixnorm. Dann gilt*

$$\rho\left(\mathbf{B}\right) = \lim_{m\to\infty}|||\mathbf{B}^m|||^{1/m}.$$

Beweis. Die Regularität von $\mathbf{B}$ impliziert $\rho := \rho\left(\mathbf{B}\right) > 0$. Wir definieren die Hilfsmatrix $\mathbf{C} := \frac{1}{\rho}\mathbf{B}$. Die Behauptung ist dann äquivalent zu

$$\lim_{m\to\infty}|||\mathbf{C}^m|||^{1/m} = 1.$$

Wir wählen nun eine Norm $|||\cdot|||_{\mathbf{C},\varepsilon}$ die von $\varepsilon > 0$ und $\mathbf{C}$ abhängt und die Eigenschaft (8.15) ($\mathbf{B}$ ersetzt durch $\mathbf{C}$) besitzt. Dann gilt[14]:

$$1 = \rho\left(\mathbf{C}\right) = \rho\left(\mathbf{C}^m\right)^{1/m} \le |||\mathbf{C}^m|||_{\mathbf{C},\varepsilon}^{1/m} \le |||\mathbf{C}|||_{\mathbf{C},\varepsilon} \le \rho\left(\mathbf{C}\right)+\varepsilon = 1+\varepsilon.$$

Diese Ungleichung gilt für alle m und daher auch für den Limes superior:

$$\overline{\lim_{m\to\infty}}|||\mathbf{C}^m|||_{\mathbf{C},\varepsilon}^{1/m} \le 1+\varepsilon.$$

[14]Die zweite Gleichheit folgt aus dem Satz, dass $\rho\left(P\left(\mathbf{A}\right)\right) = P\left(\rho\left(\mathbf{A}\right)\right)$ gilt für jede quadratische Matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ und jedes Polynom der Form $P(x) = x^m$. Der Beweis dieser Aussage basiert aus der *Schur-Zerlegung* einer $n\times n$ Matrix $\mathbf{A}$, d.h., $\mathbf{A} = \mathbf{Q}\mathbf{R}\mathbf{Q}^{\mathrm{H}}$ mit einer unitären Matrix $\mathbf{Q}$ und einer rechten oberen Dreiecksmatrix $\mathbf{R}$. Man rechnet dann leicht nach, dass $P\left(\mathbf{A}\right) = \mathbf{Q}P\left(\mathbf{R}\right)\mathbf{Q}^{\mathrm{H}}$ gilt, und daher stimmen die charakteristischen Polynome (und Spektren) von $P\left(\mathbf{A}\right)$ und von $P\left(\mathbf{R}\right)$ überein. Die Matrix $P\left(\mathbf{R}\right)$ ist wiederum eine Dreiecksmatrix mit den Diagonalelementen $P\left(R_{ii}\right)$. Da die Eigenwerte einer Dreicksmatrix die Elemente auf der Hauptdiagonalen sind, folgt die Behauptung.

Grenzübergang $\varepsilon \to 0$ liefert die Behauptung, falls wir zeigen, dass der Limes superior auf der linken Seite von ε unabhängig ist. Sei ε_0 eines der auftretenden ε-Werte. Wir betrachten zwei assoziierte Matrixnormen:

$$\begin{aligned} |||\mathbf{C}|||_1 &:= |||\mathbf{C}|||_{\mathbf{C},\varepsilon_0} \\ |||\mathbf{C}|||_2 &:= |||\mathbf{C}|||_{\mathbf{C},\varepsilon} . \end{aligned}$$

Aus der Äquivalenz von Normen im Endlichdimensionalen folgt, dass eine Konstante existiert $0 < c < 1$, so dass

$$c\,|||\mathbf{C}|||_1 \le |||\mathbf{C}|||_2 \le \frac{1}{c}\,|||\mathbf{C}|||_1$$

gilt. Daraus folgt:

$$\begin{aligned} &\overline{\lim_{m\to\infty}} |||\mathbf{C}^m|||_1^{1/m} \overset{c^{1/m}\to 1}{=} \overline{\lim_{m\to\infty}} \left(c\,|||\mathbf{C}^m|||_1\right)^{1/m} \\ &\le \overline{\lim_{m\to\infty}} \left(|||\mathbf{C}^m|||_2\right)^{1/m} \le \overline{\lim_{m\to\infty}} \left(c^{-1}\,|||\mathbf{C}^m|||_1\right)^{1/m} \\ &= \overline{\lim_{m\to\infty}} |||\mathbf{C}^m|||_1^{1/m} , \end{aligned}$$

und daher gilt überall das Gleichheitszeichen. Daher ist der Limes superior nicht abhängig von der gewählten Norm. ■

Satz 8.16 *Annahme: Die Matrix* $\mathbf{A}$ *besitze nur positive Eigenwerte (mit maximalem (bzw. minimalem) Eigenwert* $\lambda_{\max}$ *(bzw.* $\lambda_{\min}$*)). Der Dämpfungsparameter* ω *sei reell. Dann konvergiert das Richardson-Verfahren genau dann, wenn*

$$0 < \omega < 2/\lambda_{\max} \tag{8.19}$$

gilt. Die Konvergenzrate ist durch $\rho\left(\mathbf{K}_\omega^{Rich}\right) = \max\left\{|1-\omega\lambda_{\min}|, |1-\omega\lambda_{\max}|\right\}$ *gegeben.*

Beweis. (i) (8.19) $\Rightarrow$ Konvergenz.
Für $0 < \omega < 2/\lambda_{\max}$ gilt $-1 < 1-\omega\lambda_{\max} \le 1-\omega\lambda_{\min} < 1$. Aus Lemma 8.7 folgt

$$\rho\left(\mathbf{K}_\omega^{Rich}\right) < 1.$$

Satz 8.14 impliziert Konvergenz.
(ii) Konvergenz$\Rightarrow$(8.19).
Konvergenz impliziert $\rho\left(\mathbf{K}_\omega^{Rich}\right) < 1$. Aus

$$1 > \rho\left(\mathbf{K}_\omega^{Rich}\right) \ge |1-\omega\lambda_{\max}| \ge 1-\omega\lambda_{\max}$$

schliesst man $\omega > 0$. Umgekehrt folgt aus

$$-1 < -\rho\left(\mathbf{K}_\omega^{Rich}\right) \le -|1-\omega\lambda_{\max}| \le 1-\omega\lambda_{\max},$$

dass $\omega\lambda_{\max} < 2$ gilt, und dies ist die Abschätzung nach oben. ■

Da wir den Spektralradius des Richardson-Verfahrens explizit angeben können, lässt sich die Wahl des Dämpfungsparameters ω so bestimmen, dass der Spektralradius möglichst klein wird.

Satz 8.17 *Die Matrix* $\mathbf{A}$ *habe nur positive Eigenwerte.* $\lambda_{\max}$ *bezeichnet den grössten und* $\lambda_{\min}$ *den kleinsten. Die optimale Konvergenzrate des Richardson-Verfahrens ergibt sich für*

$$\omega_{opt} = \frac{2}{\lambda_{\max} + \lambda_{\min}}$$

zu

$$\rho\left(\mathbf{K}_{\omega_{opt}}^{Rich}\right) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

Beweis Der Beweis ergibt sich direkt aus der Darstellung des Spektralradius von $\mathbf{K}_{\omega}^{Rich}$. ■

Korollar 8.18 *Für die 5-Punkt-Diksretisierung des Poisson-Modellproblems ergibt sich*

$$\rho\left(\mathbf{K}_{\omega_{opt}}^{Rich}\right) = 2\cos^2\frac{\pi h}{2} - 1 = 1 - \frac{\pi^2 h^2}{2} + O\left(h^4\right).$$

Das bedeutet, dass sich die Konvergenzrate mit feiner werdender Diskretisierung verschlechtert.

Die Frage, welche Eigenwerte eine Matrix $\mathbf{A}$ (und damit die zugehörige Iterationsmatrix $\mathbf{K}$) besitzt, ist typischerweise wesentlich komplizierter, als die Frage nach der Lösung des linearen Gleichungssystems. Abschätzungen, die man für die grössten und kleinsten Eigenwerte herleitet sind häufig ziemlich pessimistisch. Für symmetrische Matrizen stimmt der grösste Eigenwert einer Matrix mit der euklidischen Norm der Matrix überein. Eine leicht berechenbare Abschätzung nach oben liefert die Maximumsnorm.

Korollar 8.19 *Sei* $\mathbf{A}$ *positiv definit und* ω *reell. Das Richardson-Verfahren konvergiert genau dann, wenn*

$$0 < \omega < \frac{2}{\lambda_{\max}}.$$

Hinreichend für Konvergenz ist die Bedingung

$$0 < \omega < \frac{2}{\|\mathbf{A}\|_\infty}$$

mit

$$\|\mathbf{A}\|_\infty := \sup_{\mathbf{v}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\mathbf{A}\mathbf{v}\|_\infty}{\|\mathbf{v}\|_\infty} = \max_{1\le i\le n} \sum_{j=1}^{n} |A_{i,j}|.$$

Beweis. Die letzte Gleichheit ist in einer Übungsaufgabe zu zeigen. Es genügt also $\|\mathbf{A}\|_\infty \geq \lambda_{\max}$ zu zeigen, und das folgt aus Lemma 8.11. ■

Im Gegensatz zum grössten Eigenwert von $\mathbf{A}$, lässt sich die Maximumnorm einer schwachbesetzten Matrix leicht exakt berechnen.

Jacobi-Iteration

Wir beginnen mit dem zentralen Konvergenzsatz. Für eine positiv definite Matrix $\mathbf{B}$ definieren wir die Norm $\|\cdot\|_{\mathbf{B}}$ auf dem $\mathbb{R}^n$ durch

$$\|\cdot\|_{\mathbf{B}} = \langle\cdot, \mathbf{B}\cdot\rangle^{1/2}.$$

Die zugeordnete Matrixnorm wird ebenfalls mit $\|\cdot\|_{\mathbf{B}}$ bezeichnet, wobei man am Argument abliest, ob es sich um eine Vektor- oder um eine zugeordnete Matrixnorm handelt. Wir werden die Notation verwenden, dass für zwei Matrizen $\mathbf{A}, \mathbf{B}$ die Schreibweise $\mathbf{A} > \mathbf{B}$ bedeutet: $\mathbf{A}-\mathbf{B}$ ist positiv definit. Für positiv definite Matrizen lassen sich bruchzahlige Potenzen bilden. Aus der linearen Algebra ist bekannt, dass für eine positiv definite Matrix $\mathbf{B}$ eine unitäre Transformation $\mathbf{Q}$ existiert mit

$$\mathbf{B} = \mathbf{Q}^{\mathrm{H}}\mathbf{D}\mathbf{Q}$$

und einer Diagonalmatrix $\mathbf{D}$, welche die (positiven) Eigenwerte von $\mathbf{B}$ als Diagonaleinträge besitzt. Dann definiert man für $s \in \mathbb{R}$

$$\mathbf{B}^s := \mathbf{Q}^{\mathrm{H}}\mathbf{D}^s\mathbf{Q},$$

wobei

$$(\mathbf{D}^s)_{i,j} = \begin{cases} (d_{ii})^s & i = j, \\ 0 & \text{sonst} \end{cases}$$

gesetzt wird. Damit lassen sich Quadratwurzeln für positiv definiten Matrizen definieren. Da $\mathbf{Q}$ unitär ist, gilt

$$\mathbf{B}^{1/2}\mathbf{B}^{1/2} = \mathbf{Q}^{\mathrm{H}}\mathbf{D}^{1/2}\mathbf{Q}\mathbf{Q}^{\mathrm{H}}\mathbf{D}^{1/2}\mathbf{Q} = \mathbf{Q}^{\mathrm{H}}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{Q} = \mathbf{Q}^{\mathrm{H}}\mathbf{D}\mathbf{Q} = \mathbf{B}.$$

Man überlegt sich leicht, dass $\mathbf{B}^s$ wiederum positiv definit ist.

Jede Norm $\|\cdot\|_{\mathbf{B}}$ lässt sich mittels der euklidischen Norm ausdrücken. Es gilt (da $\mathbf{B}$ und $\mathbf{B}^{1/2}$ positiv definit und damit auch hermitesch sind)

$$\begin{aligned} \|\mathbf{v}\|_{\mathbf{B}} &= \langle \mathbf{v}, \mathbf{B}\mathbf{v}\rangle^{1/2} = \left\langle \mathbf{v}, \mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{v}\right\rangle^{1/2} = \left\langle \left(\mathbf{B}^{1/2}\right)^{\mathrm{H}}\mathbf{v}, \mathbf{B}^{1/2}\mathbf{v}\right\rangle^{1/2} \\ &= \left\langle \mathbf{B}^{1/2}\mathbf{v}, \mathbf{B}^{1/2}\mathbf{v}\right\rangle^{1/2} = \left\|\mathbf{B}^{1/2}\mathbf{v}\right\| \end{aligned}$$

und

$$\begin{aligned} \|\mathbf{A}\|_{\mathbf{B}} &= \sup_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{v}\|_{\mathbf{B}}}{\|\mathbf{v}\|_{\mathbf{B}}} = \sup_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\left\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{v}\right\|}{\|\mathbf{B}^{1/2}\mathbf{v}\|} \\ &\overset{\mathbf{v}=\mathbf{B}^{-1/2}\mathbf{w}}{=} \sup_{\mathbf{w}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\left\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{w}\right\|}{\|\mathbf{w}\|} = \left\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{-1/2}\right\|, \end{aligned}$$

wobei $\langle\cdot,\cdot\rangle$ (bzw. $\|\cdot\|$) das euklidische Skalarprodukt (euklidische Norm) bezeichnet.

Satz 8.20 *Für die Matrix* $\mathbf{A}$ *mit Diagonalen* $\mathbf{D}$ *gelte:*

$$\mathbf{A} \text{ und } 2\mathbf{D} - \mathbf{A} \text{ sind positiv definit}$$

Dann gilt:

$$\rho\left(\mathbf{K}^{Jac}\right) = \left\|\mathbf{K}^{Jac}\right\|_{\mathbf{A}} = \left\|\mathbf{K}^{Jac}\right\|_{\mathbf{D}} < 1.$$

Beweis. Der Beweis ergibt sich aus dem folgenden allgemeineren Satz, indem $\mathbf{N} = \mathbf{D}^{-1}$ gesetzt wird. ■

Satz 8.21 *Wir betrachten die Iteration in der Form:*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{N}\left(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\right) \tag{8.20}$$

mit positiv definiten $\mathbf{N}$*.*

1. Unter der Voraussetzung
$$2\mathbf{N}^{-1} > \mathbf{A} > \mathbf{0} \tag{8.21}$$
konvergiert die Iteration (8.20). Der Spektralradius stimmt mit der Energienorm $\|\cdot\|_{\mathbf{A}}$ und der Norm $\|\cdot\|_{\mathbf{N}^{-1}}$ überein:
$$\rho(\mathbf{K}) = \|\mathbf{K}\|_{\mathbf{A}} = \|\mathbf{K}\|_{\mathbf{N}^{-1}} < 1 \tag{8.22}$$
mit $\mathbf{K} := \mathbf{I} - \mathbf{N}\mathbf{A}$.

2. Seien λ, Λ reell mit $0 < \lambda \leq \Lambda$. Dann sind die Aussagen (8.23) und (8.24) äquivalent:
$$\mathbf{0} < \lambda\mathbf{N}^{-1} \leq \mathbf{A} \leq \Lambda\mathbf{N}^{-1}. \tag{8.23}$$
Das Spektrum von $\mathbf{K}$ ist reell und enthalten in
$$\sigma(\mathbf{K}) \subset [1-\Lambda, 1-\lambda]. \tag{8.24}$$
Die Konvergenzrate beträgt
$$\rho(\mathbf{K}) = \|\mathbf{K}\|_{\mathbf{A}} = \|\mathbf{K}\|_{\mathbf{N}^{-1}} \leq \max\{1-\lambda, \Lambda-1\}. \tag{8.25}$$

Beweis. (i) Um Aussagen über die Eigenwerte einer Matrix zu beweisen, verwendet man häufig *ähnliche* Matrizen. Dabei heissen zwei Matrizen $\mathbf{A}, \mathbf{B}$ ähnlich, falls eine reguläre Matrix $\mathbf{T}$ existiert mit $\mathbf{A} = \mathbf{T}^{-1}\mathbf{B}\mathbf{T}$. Ähnliche Matrizen besitzen das gleiche Spektrum:

$$\mathbf{B}\mathbf{e} = \lambda\mathbf{e} \Leftrightarrow \mathbf{T}^{-1}\mathbf{B}\mathbf{e} = \lambda\mathbf{T}^{-1}\mathbf{e} \Leftrightarrow \underbrace{\mathbf{T}^{-1}\mathbf{B}\mathbf{T}}_{\mathbf{A}}\mathbf{e}' = \lambda\mathbf{e}'.$$

Daher ist $(\mathbf{e}, \lambda)$ Eigenpaar von $\mathbf{A}$ genau dann, wenn $(\mathbf{e}', \lambda)$ Eigenpaar zu $\mathbf{B} := \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ mit $\mathbf{T}\mathbf{e}' = \mathbf{e}$.

Multipliziert man die Iterationsmatrix $\mathbf{K}$ von links mit $\mathbf{A}^{1/2}$ (bzw $\mathbf{N}^{-1/2}$) und von rechts mit $\mathbf{A}^{-1/2}$ (bzw. $\mathbf{N}^{1/2}$) erhält man (Beachte: $\mathbf{K} = \mathbf{I} - \mathbf{N}\mathbf{A}$)

$$\begin{aligned} \mathbf{K}' &= \mathbf{A}^{1/2}\mathbf{K}\mathbf{A}^{-1/2} = \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}, \\ \mathbf{K}'' &= \mathbf{N}^{-1/2}\mathbf{K}\mathbf{N}^{1/2} = \mathbf{I} - \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2} \end{aligned} \tag{8.26}$$

und die Spektralradien stimmen mit $\rho(\mathbf{K})$ überein.
Da $\mathbf{K}'$ und $\mathbf{K}''$ hermitesch sind, gilt:

$$\begin{aligned} \rho(\mathbf{K}') &= \|\mathbf{K}'\| = \|\mathbf{K}\|_{\mathbf{A}}, \\ \rho(\mathbf{K}'') &= \|\mathbf{K}''\| = \|\mathbf{K}\|_{\mathbf{N}^{-1}}. \end{aligned} \tag{8.27}$$

(ii) Multiplikation von (8.21) und (8.23) mit $\mathbf{N}^{1/2}$ von beiden Seiten liefert:

$$2\mathbf{I} > \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2} > \mathbf{0}, \qquad \lambda\mathbf{I} \leq \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2} \leq \Lambda\mathbf{I}. \tag{8.28}$$

Hilfsaussage: Sei $\mathbf{B}$ eine hermitesche Matrix mit $\alpha_1\mathbf{I} < \mathbf{B} < \alpha_2\mathbf{I}$. Dann gilt $\sigma(\mathbf{B}) \subset (\alpha_1, \alpha_2)$.
Beweis der Hilfsaussage: $\mathbf{B}$ hermitesch$\Rightarrow$ $\mathbf{B}$ wird durch eine unitäre Matrix diagonalisiert:

$$\mathbf{B} = \mathbf{Q}^{\mathrm{H}}\mathbf{D}\mathbf{Q}.$$

Daher ist die Hilfsaussage äquivalent zu

$$\alpha_1\mathbf{I} < \mathbf{D} < \alpha_2\mathbf{I}.$$

Da $\mathbf{D}$ auf der Diagonalen die Eigenwerte von $\mathbf{B}$ enthält, bedeutet „$\mathbf{D} - \alpha_1\mathbf{I}$ positiv definit“, dass alle Eigenwerte von $\mathbf{D}$ grösser als α_1 sein müssen und die Abschätzung nach oben ergibt sich analog. Damit ist die Hilfsaussage bewiesen.

Aus (8.28) folgt daher für das Spektrum der Matrix $\mathbf{A}' := \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}$ (vgl. (8.26) und (8.28)):

$$\sigma(\mathbf{A}') \subset (0,2), \qquad \text{und } \sigma(\mathbf{A}') \subset [\lambda, \Lambda].$$

Aus $\mathbf{K}'' = \mathbf{I} - \mathbf{A}'$ folgert man für das **Spektrum** $\sigma(\mathbf{K}'') \subset (-1,1)$ und $\sigma(\mathbf{K}'') \subset [1-\Lambda, 1-\lambda]$. Aus der ersten Inklusion für $\sigma(\mathbf{K}'')$ folgt

$$\rho(\mathbf{K}) = \rho(\mathbf{K}'') < 1$$

und mit (8.27) die Behauptung (8.22). Die zweite Inklusion für $\sigma(\mathbf{K}'')$ zeigt (8.24).

(iii) Die Abschätzung des Spektralradius (8.25) aus der Abschätzung des Spektrums ergibt sich aus

$$\begin{aligned}\rho(\mathbf{K}) = \rho(\mathbf{K}'') &= \max\{|\lambda| : \lambda \text{ Ew von } \mathbf{K}''\} \le \max\{|\xi| : \xi \in [1-\Lambda, 1-\lambda]\} \\ &= \max\{|1-\Lambda|, |1-\lambda|\}.\end{aligned}$$

Aus $0 < \lambda \le \Lambda$ folgt $\max\{|1-\Lambda|, |1-\lambda|\} = \max\{\Lambda - 1, 1-\lambda\}$ und daher gilt (8.25). (Hier sind die einzelnen Fälle $0 < \lambda \le \Lambda \le 1$, $0 < \lambda \le 1 \le \Lambda$ und $1 \le \lambda \le \Lambda$ getrennt zu verifizieren.)

(iv) Da das Spektrum von $\mathbf{K}$ mit dem von $\mathbf{K}''$ übereinstimmt und letzteres unter der Voraussetzung $\mathbf{N} > 0$, $\mathbf{A} = \mathbf{A}^{\mathrm{H}}$ reell ist, gibt es einen maximalen Eigenwert $\rho_{\max}$ und einen minimalen $\rho_{\min}$:

$$\sigma(\mathbf{K}) = \sigma(\mathbf{K}'') \subset [\rho_{\min}, \rho_{\max}].$$

Daher ist $\lambda := 1 - \rho_{\max}$ der minimale und $\Lambda := 1 - \rho_{\min}$ der maximale Eigenwert von $\mathbf{A}' = \mathbf{I} - \mathbf{K}''$. Aus $\sigma(\mathbf{A}') \subset [\lambda, \Lambda]$ und „$\mathbf{A}'$ positiv definit“ folgert man, dass

$$\lambda\mathbf{I} \le \mathbf{A}' \le \Lambda\mathbf{I}$$

gilt und damit auch (8.23) mit den oben genannten Eigenwerten. ■

Der obige Satz und zugehöriges Kriterium sichert die Konvergenz des (ungedämpften) Jacobi-Verfahrens unter geeigneten Voraussetzungen. Die Voraussetzung $2\mathbf{D} - \mathbf{A} > 0$ kann bei geeigneter Dämpfung entfallen.

Satz 8.22 *Sei $\mathbf{A}$ positiv definit. Die mit ω gedämpfte Jacobi-Iteration konvergiert für*

$$0 < \omega < 2/\Lambda \qquad \textit{mit } \Lambda := \left\|\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right\| = \rho\left(\mathbf{D}^{-1}\mathbf{A}\right).$$

Eine äquivalente Formulierung dieser Bedingung lautet:

$$0 < \omega\mathbf{A} < 2\mathbf{D}. \tag{8.29}$$

Beweis. Das gedämpfte Jacobi-Verfahren besitzt die Darstellung:

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \omega \mathbf{D}^{-1} \left(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b} \right),$$

d.h. die Iterationsmatrix $\mathbf{N} = \omega \mathbf{D}^{-1}$. Das Kriterium (8.21) lautet mit diesem $\mathbf{N}$:

$$2\omega^{-1}\mathbf{D} > \mathbf{A} > \mathbf{0}$$

und das ist äquivalent zu $\mathbf{0} < \omega\mathbf{A} < 2\mathbf{D}$. Diese Bedingung ist äquivalent zur ersten Bedingung im Satz, da (8.29) äquivalent zu

$$\begin{aligned} &\mathbf{0} < \omega \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} < 2\mathbf{I} \Leftrightarrow 0 < \omega \rho \left(\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} \right) < 2 \\ &\Leftrightarrow 0 < \omega \left\| \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} \right\| < 2. \end{aligned}$$

■

Gauss-Seidel-Verfahren

Satz 8.23 *Das Gauss-Seidel-Verfahren konvergiert für positiv definite Matrizen* $\mathbf{A}$*. Die Energienorm* ($\|\cdot\|_{\mathbf{A}}$) *der Iterationsmatrix ist kleiner als* 1*.*

Beweis. Da $\mathbf{A}$ positiv definit ist, besitzt die Diagonalmatrix $\mathbf{D}$ nur positive Diagonaleinträge. Die Iterationsmatrix ist durch $\mathbf{K} = \mathbf{I} - \mathbf{W}^{-1}\mathbf{A}$ gegeben mit $\mathbf{W} := \mathbf{D} - \mathbf{E}$. Sie erfüllt

$$\mathbf{W} + \mathbf{W}^{\mathrm{H}} = (\mathbf{D} - \mathbf{E}) + (\mathbf{D} - \mathbf{E})^{\mathrm{H}} = 2\mathbf{D} - \mathbf{E} - \mathbf{F} = \mathbf{A} + \mathbf{D} > \mathbf{A}.$$

Die Konvergenz folgt dann aus dem folgenden Satz. ■

Satz 8.24 *Die Iterationsmatrix eines linearen Iterationsverfahren besitze die Form* $\mathbf{K} = \mathbf{I} - \mathbf{W}^{-1}\mathbf{A}$ *mit*

$$\mathbf{W} + \mathbf{W}^{\mathrm{H}} > \mathbf{A} > \mathbf{0}. \tag{8.30}$$

Dann konvergiert die Iteration und die Energienorm ist kleiner als 1*.*

Beweis. Es genügt $\|\mathbf{K}\|_{\mathbf{A}} < 1$ zu zeigen, da der Spektralradius kleiner als alle zugehörigen Normen ist. Wir haben bereits die Umrechnungsformel

$$\|\mathbf{K}\|_{\mathbf{A}} = \left\| \mathbf{A}^{1/2} \mathbf{K} \mathbf{A}^{-1/2} \right\|$$

bewiesen. Für die Hilfsmatrix $\hat{\mathbf{K}} := \mathbf{A}^{1/2}\mathbf{K}\mathbf{A}^{-1/2}$ gilt: $\hat{\mathbf{K}} = \mathbf{I} - \mathbf{A}^{1/2}\mathbf{W}^{-1}\mathbf{A}^{1/2}$ und

$$\begin{aligned} \hat{\mathbf{K}}^{\mathrm{H}}\hat{\mathbf{K}} &= \mathbf{I} - \mathbf{A}^{1/2} \left(\mathbf{W}^{-H} + \mathbf{W}^{-1} \right) \mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{W}^{-H}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^{1/2} \\ &= \mathbf{I} - \mathbf{A}^{1/2}\mathbf{W}^{-H} \left(\mathbf{W} + \mathbf{W}^{\mathrm{H}} \right) \mathbf{W}^{-1}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{W}^{-H}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^{1/2} \\ &\overset{(8.30)}{<} \mathbf{I} - \mathbf{A}^{1/2}\mathbf{W}^{-H}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{W}^{-H}\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^{1/2} \\ &= \mathbf{I}. \end{aligned}$$

Man prüft leicht nach, dass für jede Matrix $n \times n$-Matrix $\mathbf{B}$ gilt

$$\|\mathbf{B}\| = \rho\left(\mathbf{B}^{\mathrm{H}}\mathbf{B}\right)^{1/2}$$

sowie für positiv definite Matrizen $\mathbf{B}, \mathbf{C}$ mit $\mathbf{C} > \mathbf{B}$:

$$\rho(\mathbf{B}) < \rho(\mathbf{C}).$$

Daraus folgt

$$\|\mathbf{K}\|_{\mathbf{A}} = \left\|\hat{\mathbf{K}}\right\| = \left(\rho\left(\hat{\mathbf{K}}^{\mathrm{H}}\hat{\mathbf{K}}\right)\right)^{1/2} < 1.$$

■

SOR-Verfahren

Das SOR-Verfahren konnte als Verallgemeinerung des Gauss-Seidel-Verfahrens aufgefasst werden. Die Iterationsvorschrift lautete:

$$x_i^{(m+1)} = x_i^{(m)} - \omega\left(\sum_{j=1}^{i-1} a_{i,j}x_j^{(m+1)} + \sum_{j=i}^{n} a_{i,j}x_j^{(m)} - b_i\right)/a_{i,i}.$$

In Matrixform lautet diese Iteration

$$\mathbf{D}\mathbf{x}^{(m+1)} = \mathbf{D}\mathbf{x}^{(m)} + \omega\mathbf{E}\mathbf{x}^{(m+1)} - \omega(\mathbf{D}-\mathbf{F})\mathbf{x}^{(m)} + \omega\mathbf{b}$$

bzw.

$$(\mathbf{D}-\omega\mathbf{E})\mathbf{x}^{(m+1)} = ((1-\omega)\mathbf{D} + \omega\mathbf{F})\mathbf{x}^{(m)} + \omega\mathbf{b}$$

und schliesslich:

$$\begin{aligned}
\mathbf{x}^{(m+1)} &= (\mathbf{D}-\omega\mathbf{E})^{-1}((1-\omega)\mathbf{D} + \omega\mathbf{F})\mathbf{x}^{(m)} + \omega\mathbf{N}\mathbf{b} \\
&= \left(\mathbf{I} - \omega\underbrace{\mathbf{D}^{-1}\mathbf{E}}_{=:\mathbf{L}}\right)^{-1}\left((1-\omega)\mathbf{I} + \omega\underbrace{\mathbf{D}^{-1}\mathbf{F}}_{=:\mathbf{U}}\mathbf{x}^{(m)}\right) + \omega\mathbf{N}\mathbf{b} \\
&=: (\mathbf{I}-\omega\mathbf{L})^{-1}((1-\omega)\mathbf{I} + \omega\mathbf{U})\mathbf{x}^{(m)} + \omega\mathbf{N}\mathbf{b}, \qquad (8.31)
\end{aligned}$$

mit

$$\mathbf{N} := (\mathbf{D}-\omega\mathbf{E})^{-1}, \qquad \mathbf{L} := \mathbf{D}^{-1}\mathbf{E}, \qquad \mathbf{U} := \mathbf{D}^{-1}\mathbf{F}.$$

Diese Iteration besitzt für $\omega \neq 1$ nicht die Standardform $\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{N}\left(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\right)$. Die Iterationsmatrix lässt sich jedoch angeben und lautet:

$$\begin{aligned}
\mathbf{K}_\omega^{\mathrm{SOR}} &= (\mathbf{D}-\omega\mathbf{E})^{-1}((1-\omega)\mathbf{D} + \omega\mathbf{F}) = (\mathbf{D}-\omega\mathbf{E})^{-1}(\mathbf{D} - \omega\mathbf{E} - \omega(\mathbf{D}-\mathbf{E}-\mathbf{F})) \\
&= \mathbf{I} - \omega(\mathbf{D}-\omega\mathbf{E})^{-1}\mathbf{A} = \mathbf{I} - \left(\frac{1}{\omega}\mathbf{D} - \mathbf{E}\right)^{-1}\mathbf{A} = \mathbf{I} - \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{-1}\mathbf{A}
\end{aligned}$$

mit $\mathbf{W}_\omega^{\mathrm{SOR}} := \frac{1}{\omega}\mathbf{D} - \mathbf{E}$.

Die Wahl $\omega = 1$ liefert genau das Gauss-Seidel-Verfahrens.

Das SOR-Verfahren divergiert, falls $\omega \notin (0,2)$ erfüllt. Die Details finden sich in folgendem Lemma.

Lemma 8.25 *Es gilt*

$$\rho\left(\mathbf{K}_\omega^{\mathrm{SOR}}\right) \geq |\omega - 1|, \qquad \forall \omega \in \mathbb{C}.$$

Beweis. Sei $n := \dim \mathbf{A}$. Da $\mathbf{I} - \omega\mathbf{L}$ bzw. $(1-\omega)\,\mathbf{I} + \omega\mathbf{U}$ aus (8.31) untere bzw. obere Dreiecksmatrizen sind mit konstanten Diagonaleinträgen (1, bzw. $(1-\omega)$) gilt:

$$\det\left(\mathbf{I} - \omega\mathbf{L}\right) = 1, \qquad \det\left((1-\omega)\,\mathbf{I} + \omega\mathbf{U}\right) = (1-\omega)^n,$$

d.h.

$$\det\left(\mathbf{K}_\omega^{\mathrm{SOR}}\right) = \frac{1}{\det\left(\mathbf{I} - \omega\mathbf{L}\right)} \det\left((1-\omega)\,\mathbf{I} + \omega\mathbf{U}\right) = (1-\omega)^n. \tag{8.32}$$

Für den Spektralradius von $\mathbf{K}_\omega^{\mathrm{SOR}}$ sind die Eigenwerte von $\mathbf{K}_\omega^{\mathrm{SOR}}$ massgeblich. Wir verwenden die Identität:

$$\det\left(\lambda\mathbf{I} - \mathbf{K}_\omega^{\mathrm{SOR}}\right) = \prod_{\nu=1}^{n} (\lambda - \lambda_\nu)$$

mit den Eigenwerten λ_ν von $\mathbf{K}_\omega^{\mathrm{SOR}}$. Indem in dieser Gleichung $\lambda = 0$ gesetzt wird, erhalten wir:

$$(-1)^n \det\left(\mathbf{K}_\omega^{\mathrm{SOR}}\right) = \det\left(-\mathbf{K}_\omega^{\mathrm{SOR}}\right) = (-1)^n \prod_{\nu=1}^{n} \lambda_\nu. \tag{8.33}$$

Kombination von (8.32) und (8.33) liefert:

$$(1-\omega)^n = \prod_{\nu=1}^{n} \lambda_\nu$$

und weiter:

$$|1-\omega|^n = \prod_{\nu=1}^{n} |\lambda_\nu|.$$

Daraus folgt, dass mindestens ein Eigenwert mit $|\lambda_\nu| \geq |1-\omega|$ existieren muss. ■

Auf der anderen Seite ist die Bedingung $\omega \in (0,2)$ auch hinreichend für die Konvergenz des Verfahrens.

Satz 8.26 *Es gelte* $\mathbf{A}$ *ist positiv definit und* $0 < \omega < 2$. *Dann konvergiert das SOR-Verfahren und für die Konvergenzrate gilt:*

$$\rho\left(\mathbf{K}_\omega^{\mathrm{SOR}}\right) \leq \left\|\mathbf{K}_\omega^{\mathrm{SOR}}\right\|_{\mathbf{A}} < 1.$$

Beweis. Die Iterationsmatrix des SOR-Verfahrens besitzt die Darstellung

$$\mathbf{K}_\omega^{\mathrm{SOR}} = \mathbf{I} - \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{-1} \mathbf{A} \qquad \text{mit} \qquad \mathbf{W}_\omega^{\mathrm{SOR}} := \frac{1}{\omega}\mathbf{D} - \mathbf{E}.$$

Das Kriterium (8.30) ist daher wegen $\left(\omega \in (0,2) \implies \frac{2}{\omega} - 1 > 0\right)$ erfüllt:

$$\mathbf{W}_\omega^{\mathrm{SOR}} + \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}} = \frac{2}{\omega}\mathbf{D} - \mathbf{E} - \mathbf{F} = \mathbf{A} + \left(\frac{2}{\omega} - 1\right)\mathbf{D} > \mathbf{A} > \mathbf{0}.$$

Damit das Verfahren definiert ist, müssen wir noch die Regularität von $\mathbf{W}_\omega^{\mathrm{SOR}}$ beweisen. Sei dazu $\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{x} = \mathbf{0}$. Daraus folgt:

$$0 = \left\langle \mathbf{x}, \mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{x} \right\rangle + \left\langle \mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{x}, \mathbf{x} \right\rangle = \left\langle \mathbf{x}, \mathbf{W}_\omega^{\mathrm{SOR}} + \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}} \mathbf{x} \right\rangle$$

und wegen $\mathbf{W}_\omega^{\mathrm{SOR}} + \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}} > 0$ auch $\mathbf{x} = \mathbf{0}$. ■

In der Praxis kann das SOR-Verfahren bei geeigneter Wahl des Parameters ω *wesentlich* schneller als das eng verwandte Gauss-Seidel-Verfahren konvergieren.

Lemma 8.27 *Sei* $\mathbf{A}$ *positiv definit und* $\omega \in (0, 2)$. *Dann gilt*

$$\left\|\mathbf{K}_\omega^{\mathrm{SOR}}\right\|_{\mathbf{A}} = \sqrt{1 - \frac{2/\omega - 1}{\|\mathbf{A}^{-1/2}\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{D}^{-1/2}\|_2^2}}$$

mit $\mathbf{W}_\omega^{\mathrm{SOR}} = (\omega\mathbf{N})^{-1} = (1/\omega\mathbf{D} - \mathbf{E})$. *Falls*

$$\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{D}^{-1}\left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}} \leq c\mathbf{A} \tag{8.34}$$

für ein $c \in \mathbb{R}_{>0}$ *gilt, folgen*

$$\left\|\mathbf{A}^{-1/2}\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{D}^{-1/2}\right\|^2 \leq c, \qquad \left\|\mathbf{K}_\omega^{\mathrm{SOR}}\right\|_{\mathbf{A}} \leq \sqrt{1 - \frac{2/\omega - 1}{c}}.$$

Beweis. Wir verwenden

$$\mathbf{K}_{\mathbf{A}} := \mathbf{A}^{1/2}\mathbf{K}_\omega^{\mathrm{SOR}}\mathbf{A}^{-1/2} = \mathbf{I} - \omega\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}.$$

Aus der Gleichung

$$\mathbf{A} - \mathbf{W}_\omega^{\mathrm{SOR}} - \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}} = \mathbf{D} - \mathbf{E} - \mathbf{E}^{\mathrm{H}} - \frac{1}{\omega}\mathbf{D} + \mathbf{E} - \frac{1}{\omega}\mathbf{D} + \mathbf{E}^{\mathrm{H}} = \left(1 - \frac{2}{\omega}\right)\mathbf{D}$$

folgt dann

$$\begin{aligned}
\mathbf{K}_{\mathbf{A}}^{\mathrm{H}}\mathbf{K}_{\mathbf{A}} &= \left(\mathbf{I} - \omega\mathbf{A}^{1/2}\mathbf{N}^{\mathrm{H}}\mathbf{A}^{1/2}\right)\left(\mathbf{I} - \omega\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}\right) \\
&= \mathbf{I} - \mathbf{A}^{1/2}\left(\omega\mathbf{N}^{\mathrm{H}} + \omega\mathbf{N}\right)\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\omega\mathbf{N}^{\mathrm{H}}\mathbf{A}\omega\mathbf{N}\mathbf{A}^{1/2} \\
&= \mathbf{I} - \mathbf{A}^{1/2}\left(\omega\mathbf{N}^{\mathrm{H}}\mathbf{W}_\omega^{\mathrm{SOR}}\omega\mathbf{N} + \omega\mathbf{N}^{\mathrm{H}}\left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}}\omega\mathbf{N}\right)\mathbf{A}^{1/2} \\
&\quad + \mathbf{A}^{1/2}\omega\mathbf{N}^{\mathrm{H}}\mathbf{A}\omega\mathbf{N}\mathbf{A}^{1/2} \\
&= \mathbf{I} + \mathbf{A}^{1/2}\omega\mathbf{N}^{\mathrm{H}}\left(\mathbf{A} - \mathbf{W}_\omega^{\mathrm{SOR}} - \left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}}\right)\omega\mathbf{N}\mathbf{A}^{1/2} \\
&= \mathbf{I} + \left(1 - \frac{2}{\omega}\right)\mathbf{A}^{1/2}\omega\mathbf{N}^{\mathrm{H}}\mathbf{D}\omega\mathbf{N}\mathbf{A}^{1/2} = \mathbf{I} - \left(\frac{2}{\omega} - 1\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{H}}\right)^{-1}
\end{aligned}$$

mit der Matrix

$$\mathbf{X} := \mathbf{A}^{-1/2}\left(\omega\mathbf{N}\right)^{-1}\mathbf{D}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{D}^{-1/2}.$$

Um die gewünschte Abschätzung zu erhalten, müssen wir den kleinsten Eigenwert von $\left(\mathbf{X}\mathbf{X}^{\mathrm{H}}\right)^{-1}$ nach unten abschätzen. Dieser Eigenwert ist gerade

$$\frac{1}{\rho\left(\mathbf{X}\mathbf{X}^{\mathrm{H}}\right)} = \frac{1}{\|\mathbf{X}\|^2},$$

also erhalten wir

$$\left\|\mathbf{K}_\omega^{\mathrm{SOR}}\right\|_{\mathbf{A}}^2 = \|\mathbf{K}_{\mathbf{A}}\|^2 = \rho\left(\mathbf{K}_{\mathbf{A}}^{\mathrm{H}}\mathbf{K}_{\mathbf{A}}\right) = 1 - \left(\frac{2}{\omega} - 1\right)\frac{1}{\|\mathbf{X}\|_2^2}.$$

Das ist die gesuchte obere Schranke für die Konvergenzrate.

Jetzt bleibt noch die Abschätzung der Norm zu zeigen. Wir haben

$$\mathbf{X}\mathbf{X}^{\mathrm{H}} = \mathbf{A}^{-1/2}\mathbf{W}_\omega^{\mathrm{SOR}}\mathbf{D}^{-1}\left(\mathbf{W}_\omega^{\mathrm{SOR}}\right)^{\mathrm{H}}\mathbf{A}^{-1/2} \leq c\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = c\mathbf{I},$$

also folgt $\sigma\left(\mathbf{X}\mathbf{X}^{\mathrm{H}}\right) \subseteq [0, c]$ und damit $\|\mathbf{X}\|^2 = \rho\left(\mathbf{X}\mathbf{X}^{\mathrm{H}}\right) \leq c$. ■

9 Anfangswertprobleme für gewöhnliche Differentialgleichungen

9.1 Einleitung

Viele Probleme aus verschiedenen Anwendungsgebieten der Mathematik führen auf gewöhnliche Differentialgleichungen. Im einfachsten Fall ist dabei eine differenzierbare Funktion $y = y(x), x \in \mathbb{R}$ gesucht, deren Ableitung $y'(x)$ einer Gleichung der Form

$$y'(x) = f(x, y(x)) \quad \text{oder kürzer} \quad y' = f(x, y) \tag{9.1}$$

genügt. Die Gleichung (9.1) wird als gewöhnliche Differentialgleichung bezeichnet. Diese besitzen im Allgemeinen unendliche viele Lösungen. Durch zusätzliche Forderungen an y, kann man jedoch einzelne Lösungen auszeichnen. Bei einem Anfangswertproblem sucht man beispielsweise eine Lösung y, die für gegebene x_0 und y_0 eine Anfangsbedingung der Form

$$y(x_0) = y_0 \tag{9.2}$$

erfüllt.

Allgemeiner betrachtet man auch Systeme von n gewöhnlichen Differentialgleichungen, die man vektoriell in der Form

$$y' = f(x, y) \quad \text{mit} \quad y' := \begin{bmatrix} y_1' \\ \vdots \\ y_n' \end{bmatrix}, \quad f(x, y) := \begin{bmatrix} f_1(x, y_1, y_2, \ldots, y_n) \\ \vdots \\ f_n(x, y_1, y_2, \ldots, y_n) \end{bmatrix} \tag{9.3}$$

darstellen kann. Der Anfangsbedingung (9.2) entspricht dann eine Bedingung der Form

$$y(x_0) = y_0 = \begin{bmatrix} y_{10} \\ \vdots \\ y_{n0} \end{bmatrix}.$$

Die Gleichungen (9.3) werden gewöhnliche Differentialgleichungen erster Ordnung genannt, da nur erste Ableitungen der unbekannten Funktionen $y(x)$ auftreten. Allgemeiner kann man auch gewöhnliche Differentialgleichungen m-ter Ordnung der Form

$$y^{(m)}(x) = f(x, y(x), y^{(1)}(x), \ldots, y^{(m-1)}(x))$$

betrachten. Diese lassen sich jedoch stets in ein äquivalentes System von Differentialgleichungen erster Ordnung transformieren, weshalb wir uns im Folgenden nur mit Anfangswertproblemen für Differentialgleichungen erster Ordnung befassen werden.
Verfahren für gewöhnliche Differentialgleichungen berechnen typischerweise Näherungswerte der exakten Lösung an gewissen Punkten x_i. Oft sind diese Punkte äquidistant gewählt, d.h. $x_i = x_0 + ih$, wobei h die Schrittweite des Verfahrens bezeichnet. Ein wichtiges Problem wird sein, zu prüfen, ob und wie schnell die Näherungswerte für kleiner werdende Schrittweite gegen die exakte Lösung des Problems konvergieren.
Im Folgenden betrachten wir das Anfangswertproblem

$$y' = f(x, y), \quad a \le x \le b, \quad y(a) = y_0. \tag{9.4}$$

Es gilt folgendes Existenz- und Eindeutigkeitstheorem:

Satz 9.1 *Sei $f(x,y)$ stetig in der ersten Variablen für $x \in [a,b]$. Bezüglich der zweiten Variablen sei die Lipschitzbedingung*

$$\|f(x,y) - f(x,y^*)\| \leq L\|y - y^*\|, x \in [a,b], \quad y, y^* \in \mathbb{R}^n$$

erfüllt, wobei $\|\cdot\|$ eine Vektornorm bezeichnet. Dann besitzt das Anfangswertproblem (9.4) eine eindeutige Lösung $y(x), x \in [a,b]$, für einen beliebigen Anfangswert y_0. Ausserdem hängt $y(x)$ stetig von x_0 und y_0 ab.

Bemerkung 9.2 *Für lineare Systeme von Differentialgleichungen, wobei*

$$f_i(x,y) = \sum_{j=1}^{n} a_{ij}(x)y_i + b_i(x), \quad i = 1, \ldots, n$$

und $a_{ij}(x), b_i(x)$ stetige Funktionen auf $[a,b]$ sind, ist die Lipschitzbedingung in Satz 9.1 erfüllt.

9.2 Einschrittverfahren

Im Folgenden beschränken wir uns auf den Fall nur einer gewöhnlichen Differentialgleichung erster Ordnung für nur eine unbekannte Funktion (d.h. $n = 1$). In der Regel gelten die Ergebnisse auch für Systeme ($n > 1$), sofern man Grössen wie $y, f(x,y)$ als Vektoren und $|\cdot|$ als Norm $\|\cdot\|$ interpretiert.

Sei $x \in [a,b]$ ein beliebiger Punkt und $y \in \mathbb{R}$. Ein Schritt eines Einschrittverfahrens ist durch

$$y_{\text{next}} = y + h\Phi(x,y,h) \tag{9.5}$$

definiert, wobei die Funktion $\Phi : [a,b] \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ die Methode definiert. Zusammen mit (9.5) betrachten wir die Lösung $u(t)$ der Differentialgleichung (9.4), die durch den Punkt (x,y) verläuft, also dem lokalen Anfangswertproblem

$$u'(t) = f(t,u), \quad x \leq t \leq x+h, \quad u(x) = y \tag{9.6}$$

entspricht. $u(t)$ wird Referenzlösung genannt. y_{next} in (9.5) soll $u(x+h)$ approximieren. Wie gut diese Approximation ist, wird durch den *lokalen Diskretisierungsfehler* charakterisiert.

Definition 9.3 *Der lokale Diskretisierungsfehler der Methode Φ am Punkt (x,y) ist durch*

$$T(x,y,h) = \frac{1}{h}\left[y_{next} - u(x+h)\right] = \Phi(x,y,h) - \frac{1}{h}\left(u(x+h) - u(x)\right) \tag{9.7}$$

definiert.

Die lokale Genauigkeit einer Einschrittmethode lässt sich mit dem lokalen Diskretisierungsfehler beschreiben.

Definition 9.4 *Die Methode Φ heisst konsistent, falls*

$$T(x,y,h) \to 0 \quad \textit{für} \quad h \to 0$$

uniform auf $[a,b] \times \mathbb{R}$.

Bemerkung 9.5 *Da*

$$T(x,y,h) = \Phi(x,y,h) - \frac{1}{h}\left[u(x+h) - u(x)\right]$$

ist Φ *konsistent genau dann wenn* $\Phi(x,y,0) = u'(x) = f(x,y)$ *für* $(x,y) \in [a,b] \times \mathbb{R}$.

Definition 9.6 *Die Methode* Φ *ist von der Ordnung* p*, falls*

$$|T(x,y,h)| \leq Ch^p$$

uniform auf $[a,b] \times \mathbb{R}$*, wobei* C *nicht von* x, y *oder* h *abhängt.*

9.3 Beispiele für Einschrittverfahren

9.3.1 Explizites Euler-Verfahren

Im expliziten Euler-Verfahren wählt man $\Phi(x,y,h) = f(x,y)$ und somit

$$y_{\text{next}} = y + hf(x,y).$$

Φ hängt nicht von h ab und die Methode ist offensichtlich konsistent. Für den lokalen Diskretisierungsfehler gilt mit Taylors Theorem and (9.6)

$$\begin{aligned} T(x,y,h) &= f(x,y) - -\frac{1}{h}\left[u(x+h) - u(x)\right] \\ &= u'(x) - \frac{1}{h}\left[u(x+h) - u(x)\right] \\ &= u'(x) - \frac{1}{h}\left[u(x) + hu'(x) + \frac{1}{2}h^2u''(\xi) - u(x)\right] \\ &= -\frac{1}{2}hu''(\xi), \quad \text{für } x < \xi < x+h, \end{aligned}$$

wobei wir angenommen haben, dass $u \in C^2([x, x+h])$. Totales Ableiten von (9.6) bzgl. t und Einsetzen von $t = \xi$ zeigt, dass

$$T(x,y,h) = -\frac{1}{2}h[f_x + f_y f](\xi, u(\xi)), \tag{9.8}$$

wobei f_x und f_y die partiellen Ableitungen bzgl. der x- und y-Variablen bezeichnet. Falls wir annehmen, dass f und dessen partielle Ableitungen gleichmässig beschränkt in $[a,b] \times \mathbb{R}$ sind, zeigt dies, dass eine Konstante C existiert, die unabhängig von x, y, h ist mit

$$T(x,y,h) \leq C \cdot h.$$

Das explizite Euler-Verfahren ist somit von der Ordnung 1.

9.3.2 Implizites Euler-Verfahren

Beim expliziten Euler-Verfahren wurde die Steigung der Lösung im Punkt x verwendet um das Inkrement $\Phi(x, y, h)$ und somit y_{next} zu definieren. Beim impliziten Euler-Verfahren wird hingegen die Steigung im Punkt $x + h$ verwendet. Man wählt $\Phi(x, y, h) = f(x + h, y(x + h))$ und somit

$$y_{\text{next}} = y + hf(x + h, y_{\text{next}}). \tag{9.9}$$

Die Approximation der Lösung y im Punkt $x+h$, y_{next}, kommt auf beiden Seiten der Gleichung vor und ist somit nur implizit durch (9.9) gegeben. Um y_{next} zu bestimmen, muss also eine algebraische Gleichung, beispielsweise mit dem Newton-Verfahren, gelöst werden. Die Berechnung von y_{next} ist typischerweise also deutlich aufwendiger als beim expliziten Euler-Verfahren. Wie wir sehen werden kann sich der zusätzliche Aufwand jedoch lohnen, da implizite Verfahren oft eine höhere Stabilität aufweisen. Das implizite Euler-Verfahren ist ebenfalls von der Ordnung 1.

9.3.3 Crank-Nicolson Verfahren

Beim Crank-Nicolson Verfahren wird gefordert, dass das Inkrement der Mittelwert der Steigungen im Punkt x und $x + h$ ist, also

$$y_{\text{next}} = y + h\frac{f(x, y) + f(x + h, y_{\text{next}})}{2}. \tag{9.10}$$

Auch das Crank-Nicolson Verfahren ist implizit. Eine explizite Alternative erhält man, indem man y_{next} auf der rechten Seite von (9.10) durch einen expliziten Schritt mit dem Euler-Verfahren approximiert:

$$\begin{aligned} \eta &= y + hf(x, y) \\ y_{\text{next}} &= y + h\frac{f(x, y) + f(x + h, \eta)}{2}. \end{aligned}$$

Dies wird als Verfahren von Heun bezeichnet.

9.3.4 Verbessertes explizites Euler-Verfahren

Das Euler-Verfahren nutzt die Steigung der Lösung im Punkt x um daraus eine Approximation der Lösung um Punkt $x + h$ zu berechnen. Eine genauere Schätzung des Inkrements kann durch folgenden zweistufigen Prozess erreicht werden. Man berechnet zunächst eine Näherung von $y\left(x + \frac{1}{2}h\right)$ mit dem expliziten Euler-Verfahren und verwendet die hiermit berechnete Näherung für die Steigung der Lösung im Punkt $x + \frac{1}{2}h$ für die Steigung von y im ganzen Intervall $[x, x + h]$. Formal ausgedrückt führt dies auf

$$\begin{aligned} k_1(x, y) &= f(x, y) \\ k_2(x, y) &= f\left(x + \frac{1}{2}h, y + \frac{1}{2}hk_1\right) \\ y_{\text{next}} &= y + hk_2. \end{aligned}$$

Es kann gezeigt werden, dass dieses Verfahren von der Ordnung 2 ist.

9.3.5 Runge-Kutta Methoden

Das verbesserte explizite Euler-Verfahren ist eine zweistufige Methode. Eine Verallgemeinerung dieser Idee auf r Stufen führt auf die sogenannten expliziten Runge-Kutta Methoden:

$$k_1(x,y) = f(x,y)$$
$$k_s(x,y) = f\left(x+\mu_s h, y+\sum_{j=1}^{s-1}\lambda_{sj}k_j\right), \quad s=2,\dots,r$$
$$y_{\text{next}} = y + h\Phi(x,y,h) \quad \text{mit} \quad \Phi(x,y,h) = \sum_{s=1}^{r}\alpha_s k_s.$$

Dies wird als explizites r-stufiges Runge-Kutta Verfahren bezeichnet. Es benötigt r Auswertungen der rechten Seite f um das Inkrement $\Phi(x,y,h)$ zu berechnen. Die Paramenter λ_{sj} und α_s sind so zu wählen, dass die Ordnung des Verfahrens möglichst hoch ist für alle hinreichend glatten Funktionen f. Eine natürliche Bedingung für die α_s lautet

$$\sum_{s=1}^{r}\alpha_s = 1.$$

Dies sichert die Konsistenz des Verfahrens. Die Analyse von Runge-Kutta Verfahren ist nicht trivial und geht über den Rahmen der Vorlesung hinaus.

9.4 Globale Beschreibung von Einschrittverfahren

Um Anfangswertprobleme der Form (9.4) numerisch im Intervall $[a,b]$ zu lösen, werden Gitter und Gitterfunktionen betrachtet. Ein Gitter im Intervall $[a,b]$ ist eine Menge von Punkten $\{x_n\}_{n=0}^{N}$, so dass

$$a = x_0 < x_1 < \cdots < x_N - 1 < x_N = b, \tag{9.11}$$

mit Gitterlängen

$$h_n := x_{n+1} - x_n, \quad n = 0,1,\dots,N-1.$$

Falls $h_0 = h_1 = \cdots = h_{N-1} = (b-a)/N$ wird (9.11) als uniformes Gitter bezeichnet. Die Feinheit eines Gitters wird mit Hilfe der Grösse

$$|h| = \max_{0\le n\le N-1} h_n$$

gemessen. Im Folgenden bezeichnet h die Menge der Gitterlängen $\{h_n\}$ oder, im Falle eines uniformen Gitters, die gemeinsame Gitterlänge $(b-a)/N$.
Eine Funktion $v = \{v_n\}, v_n \in \mathbb{R}$, die auf dem Gitter (9.11) definiert ist, wird als Gitterfunktion bezeichnet. Hierbei ist v_n der Wert von v im Gitterpunkt x_n. Jede Funktion $v(x)$, die im Intervall $[a,b]$ definiert ist, induziert durch Einschränkung auf das Gitter eine Gitterfunktion. Wir bezeichnen die Menge der Gitterfunktionen im Intervall $[a,b]$ als $\Gamma_h[a,b]$. Für $v \in \Gamma_h[a,b]$ definieren wir die Norm

$$\|v\|_\infty = \max_{0\le n\le N} |v_n|.$$

Eine Einschrittmethode generiert eine Gitterfunktion $u = \{u_n\}$, so dass $u \approx y$, wobei $y = \{y_n\}$ die Gitterfunktion ist, die von der exakten Lösung des Anfangswertswertproblems (9.4)

induziert ist. Das Gitter (9.11) kann hierbei vorgegeben sein (z.B. uniform) oder dynamisch, als Teil der Methode, generiert werden. Für eine gegebene Einschrittmethode Φ kann das Verfahren zur Lösung des Anfangswertproblems wie folgt angegeben werden:

$$\begin{aligned} x_{n+1} &= x_n + h_n \\ u_{n+1} &= u_n + h_n \Phi(x_n, u_n, h_n), \quad n = 0, 1, \ldots, N-1, \end{aligned} \tag{9.12}$$

wobei $x_0 = a$ und $u_0 = y_0$ gesetzt werden.

9.5 Stabilität und Konvergenz

Die Stabilität eines numerischen Verfahrens (9.12) zu Lösung eines Anfangswertproblems charakterisiert die Robustheit des Verfahrens bezüglich kleiner Störungen. Wir werden später sehen, dass Stabilität und Konsistenz eines Verfahrens ausreichen, so dass die numerische Lösung gegen die exakte Lösung konvergiert.
Um den Begriff der Stabilität einzuführen, definieren wir zunächst den diskreten Residuenoperator R_h durch

$$(R_h v)_n := \frac{1}{h_n}(v_{n+1} - v_n) - \Phi(x_n, v_n, h_n), \quad n = 0, 1, \ldots, N-1$$

für $v = \{v_n\} \in \Gamma_h[a, b]$. Mit dieser Definition kann das Problem (9.12) auch geschrieben werden als

$$R_h u = 0 \quad \text{auf } [a, b], \quad u_0 = y_0.$$

Weiterhin bemerken wir, dass der diskrete Residuenoperator eng mit dem lokalen Diskretisierungsfehler verwandt ist. Wenn man R_h auf den Punkt $(x_n, y(x_n))$, wobei y die exakte Lösung des Problems bezeichnet, anwendet, stimmt die Referenzlösung $u(t)$ mit $y(t)$ überein und

$$(R_h y)_n = \frac{1}{h_n}(y(x_n + 1) - y(x_n)) - \Phi(x_n, y(x_n), h_n) = -T(x_n, y(x_n), h_n). \tag{9.13}$$

Definition 9.7 *Die Methode heisst stabil auf $[a, b]$, falls eine Konstante $K > 0$ existiert, die unabhänging von h ist, so dass für ein beliebiges Gitter h auf $[a, b]$ und für zwei beliebige Gitterfunktionen $v, w \in \Gamma_h[a, b]$ gilt, dass*

$$\|v - w\|_\infty \leq K\left(|v_0 - w_0| + \|R_h v - R_h w\|_\infty\right) \tag{9.14}$$

für alle h mit $|h|$ hinreichend klein.

(9.14) wird als Stabilitätsungleichung bezeichnet. Der nächste Satz zeigt, dass eine Lipschitzbedingung für Φ ausreicht, um die Stabilität der Methode sicherzustellen.

Satz 9.8 *Falls $\Phi(x, y, h)$ die Lipschitzbedingung*

$$|\Phi(x, y, h) - \Phi(x, \tilde{y}, h)| \leq M|y - \tilde{y}| \quad \text{auf } [a, b] \times \mathbb{R} \times [0, h_0]$$

erfüllt, dann ist die Methode (9.12) stabil.

Der Beweis von Satz 9.8 benötigt folgendes Lemma:

Lemma 9.9 *Sei $\{e_n\}$ eine Folge von reellen Zahlen $e_n \in \mathbb{R}$ mit*

$$e_{n+1} \leq a_n e_n + b_n, \quad n = 0, 1, \cdots, N-1, \tag{9.15}$$

wobei $a_n > 0$ und $b_n \in \mathbb{R}$. Dann gilt

$$e_n \leq E_n, \quad E_n = \left(\prod_{k=0}^{n-1} a_k \right) e_0 + \sum_{k=0}^{n-1} \left(\prod_{\ell=k+1}^{n-1} a_\ell \right) b_k, \quad n = 0, 1, \ldots, N.$$

Beweis. Es ist leicht zu zeigen, dass

$$E_{n+1} = a_n E_n + b_n, \quad n + 0, 1, \ldots, N-1; \quad E_0 = e_0.$$

Zieht man diese Ungleichung von (9.15) ab, erhält man

$$e_{n+1} - E_{n+1} \leq a_n \left(e_n - E_n \right), \quad n = 0, 1, \ldots, N-1.$$

Da $e_0 - E_0 = 0$ gilt demnach $e_1 - E_1 \leq 0$. Da $a_1 > 0$ gilt weiterhin $e_2 - E_2 \leq a_1(e_1 - E_1) \leq 0$. Mit Induktion erhält man schliesslich $e_n - E_n \leq 0$. ■

Beweis von Satz 9.8. Sei $h = \{h_n\}$ ein beliebiges Gitter auf $[a, b]$, and $v, w \in \Gamma_h[a, b]$ zwei beliebige Gitterfunktionen. Die Definition von R_h führt zu

$$v_{n+1} = v_n + h_n \Phi(x_n, v_n, h_n) + h_n \left(R_h v \right)_n, \quad n = 0, 1, \ldots, N-1$$

und

$$w_{n+1} = w_n + h_n \Phi(x_n, w_n, h_n) + h_n \left(R_h w \right)_n, \quad n = 0, 1, \ldots, N-1.$$

Subtraktion ergibt

$$v_{n+1} - w_{n+1} = v_n - w_n + h_n \left[\Phi(x_n, v_n, h_n) - \Phi(x_n, w_n, h_n) \right] + h_n \left[\left(R_h v \right)_n - \left(R_h w \right)_n \right]$$

für $n = 0, 1, \ldots, N-1$. Wir definieren

$$e_n = |v_n - w_n|, \quad d_n = | \left(R_h v \right)_n - \left(R_h w \right)_n |, \quad \delta = \max_{0 \leq n \leq N-1} d_n$$

Dann gilt mit der Dreiecksungleichung und der Lipschitzbedingung für Φ:

$$e_{n+1} \leq e_n + h_n M e_n + h_n \delta = (1 + h_n M) e_n + h_n \delta$$

für $n = 0, 1, \ldots, N-1$. Dies entspricht Ungleichung (9.15) mit $a_n = 1 + h_n M$ und $b_n = h_n \delta$. Da für $k = 0, 1, \ldots, n-1$ und $n \leq N$

$$\prod_{\ell=k+1}^{n-1} a_\ell \leq \prod_{\ell=0}^{N-1} a_\ell = \prod_{\ell=k+1}^{N-1} (1 + h_\ell M) \leq \prod_{\ell=0}^{N-1} \mathrm{e}^{h_\ell M} = \mathrm{e}^{M \sum_{\ell=0}^{N-1} h_\ell} = \mathrm{e}^{M(b-a)}$$

gilt, folgt aus Lemma 9.9, dass

$$e_n \leq \mathrm{e}^{M(b-a)} e_0 + \mathrm{e}^{M(b-a)} \sum_{k=0}^{n-1} h_k \delta \leq \mathrm{e}^{M(b-a)} \left(e_0 + (b-a)\delta \right), \quad n = 0, 1, \ldots, N-1.$$

Falls das Maximum über alle n genommen wird erhalten wir

$$\|v - w\|_\infty \leq \mathrm{e}^{M(b-a)} \left(|v_0 - w_0| + (b-a)\|R_h v - R_h w\|_\infty\right)$$

und somit

$$\|v - w\|_\infty \leq K \left(|v_0 - w_0| + \|R_h v - R_h w\|_\infty\right)$$

mit $K = \mathrm{e}^{M(b-a)} \max\{1, b-a\}$. ■

Aus der Stabilität eines Verfahrens lässt sich fast unmittelbar die Konvergenz folgern. Dafür muss zunächst die Konvergenz eines Einschrittverfahren definiert werden.

Definition 9.10 *Sei $a = x_0 < x_1 < \cdots < x_N = b$ ein Gitter auf $[a, b]$ mit maximaler Länge $|h| = \max_{1 \leq n \leq N}(x_n - x_{n-1})$. Sei $u = \{u_n\}$ eine Gitterfunktion, die durch das Verfahren (9.12) generiert wurde. Weiterhin sei $y = \{y_n\}$ die Gitterfunktion, die durch die exakte Lösung des Anfangswertproblems induziert wird. Das Verfahren (9.12) heisst konvergent in $[a, b]$, falls*

$$\|u - y\|_\infty \to 0 \quad \textit{für } |h| \to 0.$$

Satz 9.11 *Falls das Verfahren (9.12) konsistent und stabil auf $[a, b]$ ist, dann ist es konvergent. Falls Φ von der Ordnung p ist, dann*

$$\|u - y\|_\infty = O(|h|^p) \quad \textit{für } |h| \to 0.$$

Beweis. Wir wenden die Stabilitätsungleichung (9.14) für $v = u$ und $w = y$ an. Es folgt

$$\|u - y\|_\infty \leq K \left(|u_0 - y(x_0)| + \|R_h u - R_h y\|_\infty\right) = K\|R_h y\|_\infty$$

da $u_0 = y(x_0)$ und $R_h u = 0$ wegen (9.12). Wegen (9.13) gilt aber

$$\|R_h y\|_\infty = \|T(\cdot, y, h)\|_\infty,$$

wobei T wieder den lokalen Diskretisierungsfehler bezeichnet. Da angenommen wurde, dass das Verfahren konsistent ist, gilt

$$\|T(\cdot, y, h)\|_\infty \to 0 \quad \text{für } |h| \to 0$$

woraus die Konvergenz folgt. Falls die Methode die Ordnung p besitzt gilt

$$\|T(\cdot, y, h)\|_\infty = O(|h|^p) \quad \text{für } |h| \to 0$$

woraus, wie oben, die zweite Behauptung folgt. ■

Die in Abschnitt 9.3 vorgestellten Verfahren sind alle stabil und von Ordnung $p \geq 1$. Somit sind sie auch konvergent.

9.6 Schrittweitensteuerung

Die Methode (9.12) wurde für beliebige Gitter (9.11) definiert. Falls das Anfangswertproblem nur einmalig gelöst werden muss oder auf eine hohe Effizienz des Verfahren verzichtet werden kann, wird in der Praxis häufig ein hinreichend feines Gitter (typischerweise äquidistant) vorgegeben. Falls hingegen eine Zielgenauigkeit der numerischen Approximation vorgegeben ist und diese mit möglichst wenig Funktionsauswertungen von f berechnet werden soll, muss das

Gitter problemangepasst (adaptiv) gewählt werden. Da das optimale Gitter für eine gegebene Genauigkeit nicht a priori bekannt ist, wird es schrittweise (als Teil der Methode) konstruiert. Hierzu ist des notwendig den lokalen Diskretisierungsfehler in jedem Schritt zu kontrollieren. Da typischerweise jedoch nicht eine lokale sondern einer globale Fehlertoleranz vorgegeben ist (und somit der globale Fehler kontrolliert werden muss), müssen wir eine Verbindung des lokalen Fehlers zum globalen Fehler herstellen. Dazu wird der folgende Satz benötigt, der die Stetigkeit der Lösung eines Anfangswertproblems bezüglich den Anfangsdaten quantifiziert.

Satz 9.12 *Sei $f(x,y)$ stetig in $x, a \le x \le b$ und erfülle die Lipschitzbedingung*

$$|f(x,y) - f(x,\hat{y}| \le L|y-\hat{y}|, \quad x \in [a,b], \quad y,\hat{y} \in \mathbb{R}.$$

Dann hat das Anfangswertproblem

$$\begin{aligned} y'(x) &= f(x,y), \quad a \le x \le b \\ y(c) &= y_c \end{aligned} \tag{9.16}$$

eine eindeutige Lösung auf $[a,b]$ für jedes c mit $a \le c \le b$ und für jedes $y_c \in \mathbb{R}$. Seien $y(x;s)$ und $y(x;\tilde{s})$ die Lösungen von (9.16) für $y_c = s$ und $y_c = \tilde{s}$. Dann gilt

$$|y(x;s) - y(x;\tilde{s})| \le \mathrm{e}^{L|x-c|}\,|s-\tilde{s}|. \tag{9.17}$$

Dieser Satz kann dazu benutzt werden um den globalen Fehler einer Einschrittmethode mit Hilfe der lokalen Diskretisierungsfehler abzuschätzen. Das Lösen des Anfangswertsproblems (9.4) mit einer Einschrittmethode bedeutet, dass man einer Folge von "Lösungslinien" folgt, wobei man bei jedem Gitterpunkt x_n von einer Linie zur nächsten springt. Die Höhe des Sprungs wird dabei durch den lokalen Diskretisierungsfehler am Punkt x_n bestimmt. Dies ist leicht aus der Definition des lokalen Diskretisierungsfehlers (9.7) zu entnehmen, da die Referenzlösung gerade eine solche "Lösungslinie" darstellt. Insbesondere, ist die n-te Lösungslinie (Referenzlösung) durch die Lösung des Anfangswertproblems

$$\begin{aligned} v_n'(x) &= f(x,v_n), \quad x_n \le x \le b, \\ v_n(x_n) &= u_n, \end{aligned} \tag{9.18}$$

wobei u_n wieder der n-te Schritt des Einschrittverfahrens (9.12) ist, gegeben. Mit der Definition des lokalen Diskretisierungsfehlers an der Stelle x_n,

$$T(x_n,u_n,h_n) = \frac{1}{h_n}\left[u_{n+1} - v_n(x_n+h_n)\right]$$

ergibt sich demnach

$$u_{n+1} = v_n(x_{n+1}) + h_n T(x_n,u_n,h_n), \quad n = 0,1,\ldots,N-1.$$

Da wegen (9.18) auch $u_{n+1} = v_{n+1}(x_{n+1})$ gilt, kann Satz 9.12 auf v_n und v_{n+1} (mit $c = x_{n+1}, s = u_{n+1}, \tilde{s} = u_{n+1} - h_n T(x_n,u_n,h_n)$) angewendet werden. Es gilt

$$|v_{n+1}(x) - v_n(x)| \le h_n\,\mathrm{e}^{L|x-x_{n+1}|}\,|T(x_n,u_n,h_n)|, \quad n = 0,1,\ldots,N-1.$$

Da

$$\sum_{n=0}^{N-1}[v_{n+1}(x) - v_n(x)] = v_N(x) - v_0(x) = v_N(x) - y(x)$$

und wegen $v_N(x_N) = u_N$ folgt mit $x = x_N$, dass

$$|u_N - y(x_N)| \leq \sum_{n=0}^{N-1} |v_{n+1}(x_N) - v_n(x_N)| \leq \sum_{n=0}^{N-1} h_n \, \mathrm{e}^{L|x_N - x_{n+1}|} \, |T(x_n, u_n, h_n)|.$$

Falls die Schrittweite h_n nun so gewählt wird, dass

$$|T(x_n, u_n, h_n)| \leq \varepsilon_T, \quad n = 0, 1, \ldots, N-1,$$

dann gilt

$$|u_N - y(x_N)| \leq \varepsilon_T \sum_{n=0}^{N-1} (x_{n+1} - x_n) \, \mathrm{e}^{L|x_N - x_{n+1}|} = \varepsilon_T \sum_{n=0}^{N-1} (x_{n+1} - x_n) \, \mathrm{e}^{L(b - x_{n+1})}$$

Die Summe auf der rechten Seite kann als Riemann-Summe eines bestimmten Integrals aufgefasst werden, so dass

$$|u_N - y(x_N)| \leq \varepsilon_T \int_a^b \mathrm{e}^{L(b-x)} \, dx = \frac{\varepsilon_T}{L} \left(\mathrm{e}^{L(b-a)} - 1 \right).$$

Falls die Lipschitz Konstante L bekannt ist, kann man also

$$\varepsilon_T \leq \frac{L}{\mathrm{e}^{L(b-a)} - 1} \varepsilon$$

setzen um den globalen Fehler $|u_N - y(x_N)| \leq \varepsilon$ zu garantieren. Diese Betrachtungen motivieren den folgenden Mechanismus zur Schrittweitensteuerung: Zur Bestimmung der Schrittweite h_n in jedem Schritt (von x_n nach $x_{n+1} = x_n + h_n$) werden folgende Punkte durchlaufen:

(1) Schätzung von h_n.

(2) Berechnung (bzw. Abschätzung) von $|T(x_n, u_n, h_n)|$. Falls $|T(x_n, u_n, h_n)| \leq \varepsilon_T$, gehe zu (3). Ansonsten wiederhole (2) für kleineres h_n (z.B. halb so gross) solange bis $|T(x_n, u_n, h_n)| \leq \varepsilon_T$.

(3) Berechnung von $u_{n+1} = u_n + h_n \Phi(x_n, u_n, h_n)$.

Die Schwierigkeit, die bei dieser Vorgehensweise auftritt, ist die effiziente Abschätzung des lokalen Diskretisierungsfehlers $|T(x_n, u_n, h_n)|$. Für das explizite Euler-Verfahren konnte der führende Term von $|T(x_n, u_n, h_n)|$ angegeben werden (siehe (9.8)), für komplexere Verfahren, wie das Runge-Kutta Verfahren, ist dies aber aufwendig. Auch wenn eine explizite Darstellung des führenden Terms von $|T(x_n, u_n, h_n)|$ berechnet werden kann, ist dies für die Numerik oft ungeeignet, da diese Darstellungen typischerweise partielle Ableitungen von f beinhalten, deren Auswertung teuer ist.

Abschätzung des lokalen Diskretisierungsfehlers

Sei Φ eine Methode der Ordnung p. Wie gesehen, kann der zugehörige lokale Diskretisierungsfehler durch

$$T(x, y, h) = \Phi(x, y, h) - \frac{1}{h} \left[u(x + h) - u(x) \right]$$

dargestellt werden. Da Φ von der Ordnung p ist, gilt ausserdem $T(x, y, h) = O(h^p)$. Wir wollen nun die führende Fehlerfunktion $\tau(x, y)$ abschätzen, wobei

$$T(x, y, h) = \tau(x, y)h^p + O(h^{p+1}).$$

Um einen Ausdruck für einen Fehlerschätzer $r(x, y, h)$ mit

$$r(x, y, h) = \tau(x, y) + O(h)$$

zu erhalten, betrachten wir neben Φ eine weitere Einschrittmethode $\Phi^\star$, diesmal jedoch von der Ordnung $p + 1$. Da

$$\begin{aligned}\Phi(x, y, h) - \frac{1}{h}\left(u(x + h) - u(x)\right) &= \tau(x, y)h^p + O(h^{p+1})\\ \Phi^\star(x, y, h) - \frac{1}{h}\left(u(x + h) - u(x)\right) &= O(h^{p+1})\end{aligned}$$

folgt nach Subtraktion für den Fehlerschätzer

$$r(x, y, h) = \frac{1}{h^p}[\Phi(x, y, h) - \Phi^\star(x, y, h)].$$

Der Term $h^p r(x_n, u_n, h_n)$ kann somit verwendet werden, um $|T(x_n, u_n, h_n)|$ in der Schrittweitensteuerung abzuschätzen. Die Frage ist nun, wie sich der Fehlerschätzer $r(x, y, h)$ möglichst effizient berechnen lässt. Eine Möglichkeit ist einen Runge-Kutta Prozess (der Ordnung p) in einen anderen der Ordnung $p + 1$ einzubetten, so dass möglichst viele Auswertungen von f wiederverwendet werden können. Diese Idee geht auf E. Fehlberg zurück, welcher solche Paare von eingebetteten $(p, p+1)$ Runge-Kutta Formeln in den späten 1960er Jahren entwickelt hat. Solche Verfahren sind heute als Runge-Kutta-Fehlberg-Verfahren bekannt.