G. Maeß

Vorlesungen über numerische Mathematik II

Mathematische Lehrbücher und Monographien
Herausgegeben von der Akademie der Wissenschaften der DDR
Karl-Weierstraß-Institut für Mathematik

I. Abteilung
Mathematische Lehrbücher
Band 37

Vorlesungen über numerische Mathematik II

von G. Maeß

Vorlesungen über numerische Mathematik II

Analysis

von Gerhard Maeß

Mit 60 Abbildungen und 38 Tabellen

Akademie-Verlag Berlin
1988

Verfasser:

Prof. Dr. sc. nat. Gerhard Maeß

Wilhelm-Pieck-Universität Rostock
Sektion Mathematik

ISBN 3-05-500220-2
ISSN 0076-5422

Erschienen im Akademie-Verlag Berlin, DDR - 1086 Berlin, Leipziger Straße 3—4

Lizenznummer: 202 · 100/534/88
Printed in the German Democratic Republic
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", 7400 Altenburg
Lektor: Dipl.-Math. Gesine Reiher
LSV 1084
Bestellnummer: 763 540 9 (6801/2)

03900

Vorwort

Der zweite Band der Vorlesungen über numerische Mathematik ist der numerischen Analysis gewidmet. Der induktive Aufbau wurde beibehalten: Ausgehend von einführenden Beispielen aus den Technik- oder Naturwissenschaften oder aus nichtnumerischen Teilgebieten der Mathematik, die die Problemstellung rechtfertigen, führt der Weg über elementare, manchmal heuristische Lösungsansätze zu mathematisch begründeten Verfahren und zu numerischen Algorithmen. Diese sind so dargestellt, daß sie vom rechentechnisch orientierten Leser auch ohne tiefere mathematische Vorkenntnisse in eine der üblichen problemorientierten Programmiersprachen oder in einen Assembler übertragen und auf der verfügbaren Rechentechnik erprobt werden können. Konvergenzbeweise, Fehler- und Stabilitätsuntersuchungen sind für den mathematisch orientierten Leser gedacht. Die Bemerkungen am Ende jeder Vorlesung verweisen auf nicht behandelte Details und auf die weiterführende Fachliteratur. Die zahlreichen Übungsaufgaben sollen sowohl zum Testen der behandelten Verfahren als auch zu theoretischen Untersuchungen anregen.

Der Verfasser hofft, auf diese Weise nicht nur Mathematikstudenten und Lehrerstudenten der Fachkombinationen Mathematik/Physik, Chemie, Geographie sowie Studenten der Technik- und Naturwissenschaften, deren Studienplan die numerische Mathematik enthält, anzusprechen. Er möchte vielmehr auch einem durch die schnelle Verbreitung der Kleinrechentechnik stark zunehmenden Kreis von in der Praxis tätigen Mathematikern, Informatikern, Ingenieuren, Ökonomen und Naturwissenschaftlern sowie den Lehrern der Mathematik und Polytechnik den Einstieg in die numerische Mathematik erleichtern.

Das Buch besteht aus vier Kapiteln, die in relativ unabhängige Hauptabschnitte (Vorlesungen) untergliedert sind. Bei allen Numerierungen der Gestalt $i.j.k$ bezeichnen $i.j$ die Vorlesung und k die laufende Nummer. Innerhalb jeder Vorlesung werden die Unterabschnitte, Abbildungen, Tabellen, Formeln, Bemerkungen und Übungsaufgaben individuell gezählt. Definitionen, Sätze, Beispiele und Algorithmen haben dagegen eine gemeinsame Numerierung.

Das Kapitel 5 ist nichtlinearen Gleichungen und Gleichungssystemen gewidmet. Behandelt werden elementare Iterationsverfahren zur Nullstellenbestimmung einer reellen Funktion einer reellen Veränderlichen, Besonderheiten bei der Berechnung reeller und komplexer Nullstellen von Polynomen (die sowohl explizit in der Summendarstellung als auch implizit als charakteristisches Polynom einer Tridiagonalmatrix gegeben sein können) sowie einige wichtige Verfahren für nichtlineare Systeme.

Das Kapitel 6, Interpolation und Approximation, enthält neben der klassischen Polynominterpolation mit Hilfe der Formeln von LAGRANGE und NEWTON und der intervallweisen Interpolation einschließlich quadratischer und kubischer Polynomsplines

auch einen Abschnitt über stückweise polynomiale Interpolation von Flächen. Ferner werden im Rahmen der besten Quadratmittelapproximation die schnelle Fourier-Transformation und als Verfahren der besten gleichmäßigen Approximation der Remez-Algorithmus behandelt.

Das Kapitel 7, Quadratur und Kubatur, beginnt mit den Newton-Cotes-Formeln und den daraus gebildeten zusammengesetzten Quadraturverfahren. Die Richardson-Extrapolation wird zur Konvergenzbeschleunigung der numerischen Integration und der numerischen Differentiation verwendet. Behandelt werden ferner offene, halboffene und geschlossene Gauß-Formeln sowie in Gestalt von Beispielen einige Methoden der numerischen Kubatur.

Das Kapitel 8 schließlich enthält Methoden zur numerischen Integration gewöhnlicher Differentialgleichungen unter Berücksichtigung von Stabilitätsproblemen und Besonderheiten bei steifen Differentialgleichungen.

Wieder hat eine Reihe von Kollegen das Manuskript oder Teile davon gelesen und dem Autor durch Vorschläge und kritische Hinweise bei der Arbeit geholfen. Dafür sei ihnen allen, namentlich aber L. Berg (Rostock), V. Friedrich (Karl-Marx-Stadt), R. März (Berlin), W. Peters (Rostock), J. W. Schmidt (Dresden), K. Strehmel (Halle) und M. Tasche (Rostock) sehr herzlich gedankt.

Zu danken ist schließlich dem Akademie-Verlag Berlin, insbesondere der Lektorin G. Reiher für die gute Zusammenarbeit und die sorgfältige Manuskriptbearbeitung sowie den Kollegen, die den schwierigen mathematischen Formelsatz bewältigten und für die technische Herstellung verantwortlich waren.

Gerhard Maess

Inhalt

5. Nichtlineare Gleichungen und Gleichungssysteme

5.1. Elementare Iterationsverfahren

5.1.1. Problemstellung

Beispiel 5.1.1. Gesucht sei die Quadratwurzel einer positiven Zahl c, also die positive *Lösung der Gleichung*

$$x^2 = c, \quad c > 0. \tag{1}$$

Bezeichnen wir mit x^* die exakte Lösung und mit x_n eine beliebige Näherung ($x_n > 0$), so gilt stets eine der beiden Ungleichungen

$$x_n \leqq x^* \leqq c/x_n \quad \text{oder} \quad c/x_n \leqq x^* \leqq x_n, \tag{2}$$

das heißt, x_n und c/x_n bilden eine Einschließung der Lösung x^*. Denn wegen (1) ist $x^* = c/x^*$, und daraus ergibt sich mit $x_n \leqq x^*$ sofort die erste und mit $x_n \geqq x^*$ die zweite der Ungleichungen (2). Wenn aber x^* zwischen x_n und c/x_n liegt, kann man erwarten, daß das arithmetische Mittel der beiden Werte

$$x_{n+1} := (x_n + c/x_n)/2 \tag{3}$$

eine bessere Näherung und eine kleinere Einschließung für x^* liefert: $|x_{n-1} - c/x_{n-1}| < |x_n - c/x_n|$. Tatsächlich erhält man für $c := 2$ mit dem Startwert $x_0 := 1.5$ eine (quadratisch) konvergierende Folge von Einschließungen:

n	0	1	2	3
x_n	1.5	1.4166667	1.4142157	1.4142136
c/x_n	$1.\overline{3}$	1.4117647	1.4142114	1.4142136.

Beispiel 5.1.2. Ein Satellit S bewege sich auf einer elliptischen Bahn mit der (der Einfachheit halber auf Eins normierten) großen Halbachse $a := 1$, der kleinen Halbachse $b := 0.6$ und der Exzentrizität $\varepsilon := \sqrt{1 - b^2} = 0.8$. Seine Umlaufzeit betrage 90 Minuten. Zum Zeitpunkt des Periheldurchgangs sei $t_0 := 0$. Gesucht ist die Position P, in der sich der Satellit 9 Minuten später befindet (vgl. Abb. 5.1.1). Man kann sie mit Hilfe der *Keplerschen Gleichung*

$$\beta - \varepsilon \sin \beta = \alpha, \qquad 0 < \varepsilon < 1, \tag{4}$$

bestimmen. Dabei bezeichnet $\alpha := \sphericalangle P'MA$ die *mittlere Anomalie* des Satelliten. Das ist derjenige Winkel, der die Position P' eines fiktiven Satelliten S' beschreibt, der mit der gleichen Umlaufzeit wie S, aber mit konstanter Geschwindigkeit, eine Kreisbahn mit dem Radius $r := a$ durchläuft. Der aus (4) zu bestimmende Winkel $\beta := \sphericalangle QMA$ heißt *exzentrische Anomalie* des Satelliten. Ist er berechnet, so ergibt sich die gesuchte Position als Schnittpunkt der Ellipsenbahn mit dem Lot von Q auf die Abszisse. Wir setzen die Werte $\varepsilon = 0.8$ und $\alpha = \dfrac{2\pi}{90} \cdot 9 = \dfrac{\pi}{5}$ in die Gleichung (4) ein, bringen $\varepsilon \sin \beta$ auf die rechte Seite und fügen bei dem links stehenden β den Index $n + 1$ und

bei dem rechts stehenden den Index n an. Dadurch wird (4) zur Iterationsvorschrift

$$\beta_{n+1} := \pi/5 + 0.8 \sin \beta_n. \tag{5}$$

Mit dem Startwert $\beta_0 := \pi/5 \approx 0.628319$ ergibt sich hier eine konvergente Folge $\beta_1 = 1.098547$, $\beta_2 = 1.340756$, $\beta_3 = 1.407244$, ..., $\beta_8 = 1.419135$, $\beta_9 = 1.419136$ und damit die gesuchte Abszisse $x := \cos \beta = 0.1510798$. Dazu erhält man aus der Ellipsengleichung $x^2 + y^2/0.36 = 1$ die Ordinate $y = 0.6\sqrt{1 - x^2} = 0.5931129$.

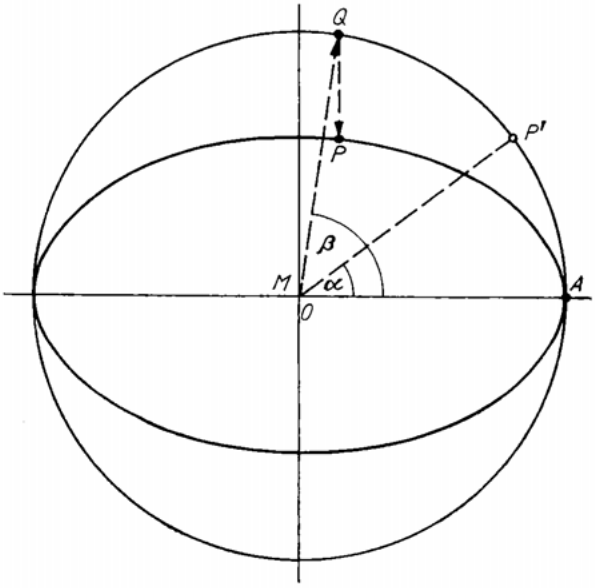

Abb. 5.1.1. Bestimmung der Satellitenposition mit Hilfe der Keplerschen Gleichung

Beispiel 5.1.3. Bei zwei aufeinanderfolgenden chemischen Reaktionen entstehe aus einem Produkt A ein Produkt X und daraus ein Produkt Y. Bezeichnen wir die Menge des Ausgangsprodukts zum Zeitpunkt $t_0 = 0$ mit $a := 1$ und die Mengen der Produkte X und Y zum Zeitpunkt t mit $x(t)$ und $y(t)$, so gilt (für Reaktionen erster Ordnung, vgl. JUST, OELSCHLÄGEL [1], S. 163)

$$\frac{dx}{dt} = h \cdot (1 - x), \qquad \frac{dy}{dt} = k \cdot (x - y), \tag{6}$$

wobei zu Beginn der Reaktionen $x(0) = y(0) = 0$ vorausgesetzt wird. Gesucht sind die Proportionalitätsfaktoren h und k, durch die die Geschwindigkeiten der beiden Reaktionen beschrieben werden. Wir beschränken uns auf eine Messung zum Zeitpunkt $t_1 := 3$, die $x(t_1) = 0.7$ und $y(t_1) = 0.3$ ergeben möge (in der Praxis nimmt man meist mehrere Messungen vor, um die Meßfehler auszugleichen). Die Lösung des linearen Differentialgleichungssystems (6) läßt sich in diesem konkreten Fall geschlossen angeben:

$$x(t) = 1 - e^{-ht}, \quad y(t) = 1 + \frac{k}{h - k} e^{-ht} - \frac{h}{h - k} e^{-kt}. \tag{7}$$

Dabei ist $h \neq k$ vorausgesetzt. Mit den vorgegebenen Randbedingungen ergibt sich für h und k das folgende *System von zwei nichtlinearen Gleichungen*

$$0.3 - e^{-3h} = 0, \qquad 0.7 + \frac{k}{h - k} e^{-3h} - \frac{h}{h - k} e^{-3k} = 0, \tag{8}$$

aus der ersten läßt sich $h = -\frac{1}{3} \ln 0.3$ bestimmen, und die zweite geht damit über in

$$0.7 + \frac{1.2}{\ln 0.3} k - e^{-3k} = 0. \tag{9}$$

Oft kann man sich einen schnellen Überblick über das Lösungsverhalten verschaffen, wenn man die linke Seite der nichtlinearen Gleichung als Differenz zweier einfacherer Funktionen auffaßt und nach Schnittpunkten der zugehörigen Kurven sucht. In Abb. 5.1.2 sind die Gerade $y = 0.7 + \frac{1.2}{\ln 0.3} k$ und die Exponentialfunktion $y = e^{-3k}$ graphisch dargestellt. In der Umgebung von $k = 0.35$ stimmen sie nahezu überein, dort sind also Lösungen der Gleichung (9) zu erwarten.

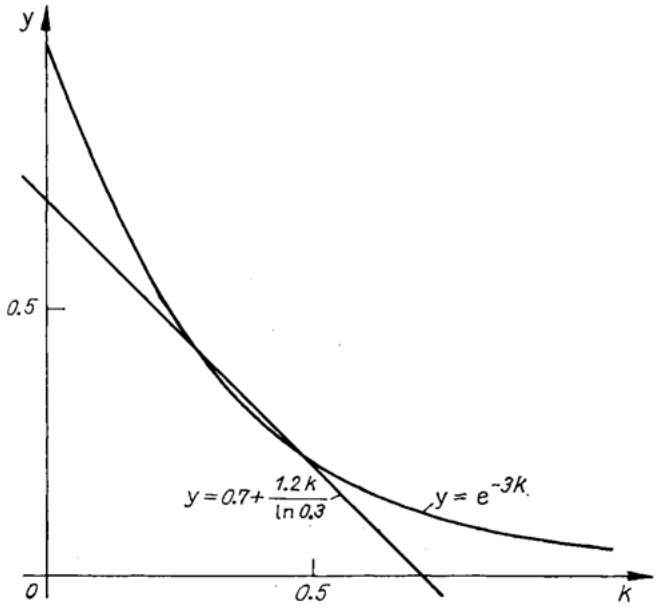

Abb. 5.1.2. Zum Nullstellenproblem aus Beispiel 5.1.3

Durch Umstellen und Anfügen von Indizes wird (9) zur Iterationsvorschrift

$$k_{n+1} = 0.702317 - 1.003311\, e^{-3k_n}. \tag{10}$$

Mit dem Startwert $k_0 := 0.35$ ergibt sich die Folge $k_1 = 0.351221$, $k_2 = 0.352505$, $k_3 = 0.353850$, $k_4 = 0.355253, \ldots$, die anscheinend nicht konvergiert. Die Differenzen aufeinanderfolgender Näherungen nehmen zu. Aus diesem Grund stellen wir (9) nach dem k in der e-Funktion um und erhalten die neue Iterationsvorschrift

$$k_{n+1} = -\frac{1}{3} \ln (0.7 - 0.996700 k_n). \tag{10'}$$

Sie liefert für $k_0 := 0.35$ die Näherungswerte

$$k_1 = 0.348843, \quad k_2 = 0.347749, \quad k_3 = 0.346720, \quad k_4 = 0.345754.$$

Jetzt werden die Differenzen zwar kleiner, die Konvergenz ist aber immer noch unbefriedigend langsam.

Alle drei Beispiele führen auf ein *Nullstellenproblem*, d. h. auf die Aufgabe, Werte zu finden, für die eine nichtlineare Funktion f verschwindet, also

$$f(x) = 0. \tag{11}$$

Dabei ist f im einfachsten Fall eine reellwertige Funktion einer reellen Veränderlichen, z. B. $f(x) := x^2 - c$ im ersten und $f(x) := x - \varepsilon \sin x - \alpha$ im zweiten Beispiel, f kann

aber auch ein Vektor von Funktionen mehrerer Veränderlicher sein. Dann stellt (11) ein *nichtlineares Gleichungssystem* der Gestalt

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_N) &= 0, \\ f_2(x_1, x_2, \ldots, x_N) &= 0, \\ &\vdots \\ f_M(x_1, x_2, \ldots, x_N) &= 0 \end{aligned} \tag{11'}$$

dar. Im dritten Beispiel ist $f_1(x_1, x_2) := 0.3 - e^{-3x_1}$, $f_2(x_1, x_2) := 0.7 + \frac{x_2}{x_1 - x_2} e^{-3x_1} - \frac{x_1}{x_1 - x_2} e^{-3x_2}$. Häufig treten nichtlineare Gleichungssysteme als Teilprobleme bei komplizierteren mathematischen Modellen auf, z. B. bei der Diskretisierung von nichtlinearen gewöhnlichen oder partiellen Differentialgleichungen, von Integralgleichungen, von Variationsaufgaben oder von Optimierungs- und Steuerungsproblemen, die in der Mechanik (bei nichtlinearen Stoffgesetzen und großen Deformationen) in der Optik, der Kernphysik, der Chemie (bei Umwandlungsprozessen mit zustandsabhängigen Eigenschaften) in der Elektronik und der Biologie (bei der Beschreibung biologischer oder ökologischer Systeme) zur Modellierung komplexer Zusammenhänge benutzt werden.

Für lineare Gleichungssysteme lassen sich allgemeine Bedingungen angeben, unter denen eine, keine oder unendlich viele Lösungen existieren (vgl. Vorlesung 2.1), für nichtlineare dagegen, bei denen auch noch der Fall endlich vieler voneinander verschiedener Lösungen auftreten kann ($x_1 = \sqrt{c}$ und $x_2 = -\sqrt{c}$ bei der Gleichung (1)), gibt es keine solchen globalen *Existenz- und Eindeutigkeitsaussagen*. Man muß zufrieden sein, wenn man für relativ enge Problemklassen Teilgebiete des $\mathbb{R}^N$ angeben kann, in denen eine eindeutige Lösung existiert. Schon im Eindimensionalen können die Lösungen von Nullstellenproblemen nur in seltenen Fällen in geschlossener Form angegeben werden. Die bereits aus der Schule bekannte quadratische Gleichung

$$p_2(x) := a_0 x^2 + a_1 x + a_2 = 0, \qquad a_0 \neq 0, \tag{12}$$

mit den beiden Lösungen $x_{1,2} = \left(-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}\right)/(2a_0)$ ist ein solcher Fall. Auch die Polynomgleichungen

$$p_N(x) := a_0 x^N + a_1 x^{N-1} + \cdots + a_{N-1} x + a_N = 0 \tag{13}$$

mit $N = 3$ und $N = 4$ zählen noch dazu. Allerdings sind die Auflösungsformeln so kompliziert (vgl. Kleine Enzyklopädie Mathematik [1], S. 108—113), daß man sie in der Praxis kaum verwendet. Für $N > 4$ ist, wie man seit den grundlegenden Untersuchungen von ABEL und GALOIS weiß, eine explizite Lösungsdarstellung mit Hilfe von Wurzelausdrücken (Radikalen) nicht mehr möglich. Man ist also auf Näherungsverfahren angewiesen und kann nicht mehr wie im Fall linearer Gleichungssysteme zwischen direkten und iterativen Lösungsmethoden wählen. Das gleiche gilt für die meisten transzendenten Gleichungen, bei denen f wie in den Beispielen 5.1.2 und 5.1.3 Ausdrücke mit transzendenten Funktionen ($\sin x$, e^x, $\ln x$, ...) enthält.

Für die iterative Behandlung ist die Formulierung als Nullstellenproblem unzweckmäßig. Die Gleichung (11) muß in eine *iterierfähige Gestalt* (*Fixpunktgestalt*)

$$x = g(x) \tag{14}$$

transformiert werden, damit aus einem Näherungswert x_n eine neue Näherung x_{n+1} berechnet werden kann. Durch (14) wird die *Iterationsvorschrift*

$$x_{n+1} = g(x_n) \tag{15}$$

nahegelegt. Bei der Transformation von (11) in (14) ist dafür zu sorgen, daß beide Gleichungen, zumindest in einem geeignet eingeschränkten Definitionsgebiet von f und g, ein und dieselbe Lösung x^* besitzen (*Konsistenz*) und daß die von (15) zu einem beliebigen Startwert x_0 (aus einer möglichst großen Teilmenge des Definitionsgebietes) gelieferte Folge von Näherungen $x_1, x_2, \ldots$ gegen diese Lösung x^* strebt (*Konvergenz*).

Eng verwandt mit dem Nullstellenproblem ist die in vielen Problemen der Praxis auftretende Aufgabe, *Extrema einer Funktion* F zu bestimmen. Ist F eine mindestens einmal stetig differenzierbare Funktion einer reellen Veränderlichen, so ist

$$f(x) := F'(x) = 0 \tag{16}$$

notwendig für das Vorliegen eines relativen Extremums im Inneren des Definitionsgebietes. Hängt F von mehreren Veränderlichen ab, so lauten die notwendigen Bedingungen

$$f_n(x_1, x_2, \ldots, x_N) := \frac{\partial F(x_1, x_2, \ldots, x_N)}{\partial x_n} = 0, \qquad n = 1(1)N. \tag{17}$$

In beiden Fällen erhält man also *Nullstellenprobleme*. Umgekehrt kann man (11) oder (11′) auch als *Extremalproblem* formulieren. Das ist insbesondere sinnvoll, wenn die Anzahl der Gleichungen die der Unbekannten übersteigt, (11′) also überbestimmt ist. Dann stellt der Vektor $\boldsymbol{x} := (x_1, x_2, \ldots, x_N)$, für den die Funktion

$$G(\boldsymbol{x}) := \sum_{m=1}^{M} f_m^2(\boldsymbol{x}), \qquad M \geqq N, \tag{18}$$

ein Minimum annimmt, eine „Lösung" von (11′) im Sinne der *Ausgleichsrechnung* (*Methode der kleinsten Quadrate*) dar. Ist (11′) lösbar, so hat G an der Minimalstelle $\boldsymbol{x}$ den Wert Null.

Im folgenden beschränken wir uns zunächst auf *Funktionen einer reellen Veränderlichen*. Sie haben den Vorteil, daß man sich durch eine Wertetabelle und eine graphische oder Bildschirmdarstellung leicht einen Überblick über das Lösungsverhalten von (11) verschaffen kann.

5.1.2. Bisektionsverfahren

Die Funktion f sei stetig, $f \in \mathbf{C}[a, b]$, und wechsle im Intervall $[a, b]$ ihr Vorzeichen. Solche Intervalle nennen wir Einschließungen.

Definition 5.1.4. Ein Intervall $[a, b]$ heißt *Einschließung* einer Nullstelle x^* von f, $f \in \mathbf{C}[a, b]$, wenn $f(a)\, f(b) < 0$ ist.

Halbiert man eine Einschließung, so ist eines der beiden Teilintervalle wegen der Stetigkeit der Funktion f wieder eine Einschließung, und man kann das Vorgehen wiederholen (Abb. 5.1.3). Diese Methode heißt *Bisektionsverfahren* (*Methode der Intervallhalbierung, Bolzano-Verfahren*). Sie liefert im Fall $a < b$ eine *Intervallschachtelung* der Gestalt

$$a = a_0 \leqq a_1 \leqq \cdots \leqq a_j \leqq \cdots \leqq x^* \leqq \cdots \leqq b_j \leqq \cdots \leqq b_1 \leqq b_0 = b. \tag{19}$$

Im Fall $a > b$ ist die Ungleichungskette umzukehren.

Algorithmus 5.1.5. Bisektionsverfahren BISE (f, a, b, δ)

f		Funktion, muß als Unterprogramm oder als Tabelle vorliegen
a, b		Intervallgrenzen mit $f(a) > 0$, $f(b) < 0$
δ		Genauigkeitsschranke
1	$x := b + (a - b) \cdot 0.5$	Intervallhalbierung
2	falls $f(x) \leqq 0$, dann $b := x$ sonst $a := x$	Auswahl des neuen Intervalls
3	falls $\lvert b - a \rvert > \delta$, dann Schritt 1	Rücksprung, falls Intervall noch zu groß
	drucke a, b	

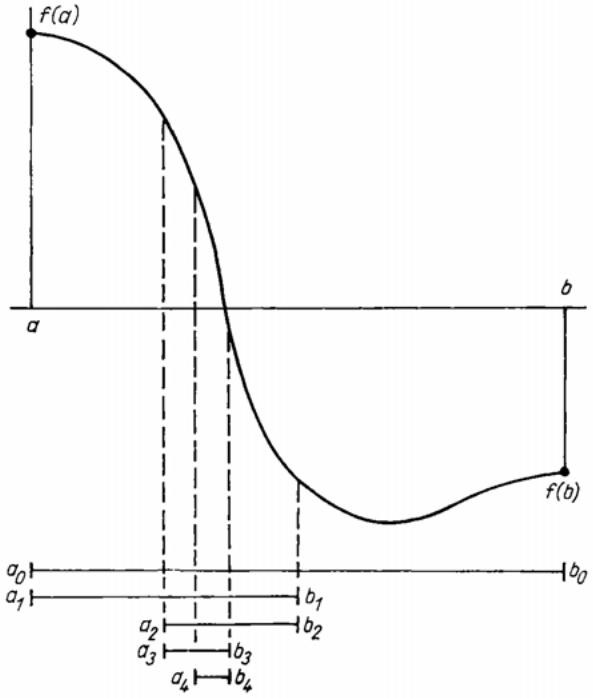

Abb. 5.1.3. Bisektionsverfahren

Das Ergebnis ist eine Einschließung der Lösung, $a \leqq x^* \leqq b$, mit $|b - a| \leqq \delta$. Der *Rechenaufwand*, d. h. die Anzahl der benötigten Funktionsberechnungen (vgl. B 5.1.3), kann von vornherein angegeben werden. Er hängt nur von der Länge d des Startintervalls ($d := |b - a|$) und nicht von der Funktion f ab. Aus der Forderung $2^{-n}d \leqq \delta$ ergibt sich die benötigte Schrittanzahl

$$n := 1 + [-\log_2 (\delta/d)] = 1 + \left[-\frac{\ln \delta/d}{\ln 2}\right].$$

Dabei bezeichnet $[x]$ die größte ganze Zahl, die kleiner oder gleich x ist. Wegen $2^{10} = 1024 \approx 10^3$ gewinnt man in 10 Schritten etwa 3 Dezimalen. Das Bisektionsverfahren konvergiert also nicht sehr schnell. Dafür stellt es aber nur sehr geringe Anforderungen an die Funktion f.

Beispiel 5.1.6. Gegeben seien eine Folge von monoton zunehmend angeordneten Werten x_n, $n = 1(1)N$,

n	1	2	3	4	5	6	7	8	9	10
x_n	0.5	1.1	5.0	6.0	9.0	9.5	9.8	10.0	11.0	11.2

und ein Wert $c := 5.3$, der in diese Folge einsortiert werden soll. Wir definieren eine *diskrete* (nur für ganze Zahlen erklärte) *Funktion*

$$f(n) := \begin{cases} c - x_n, & \text{falls } n \leqq N, \\ c - x_N, & \text{falls } n \geqq N, \end{cases}$$

und wählen b so, daß $b - 1$ die kleinste Zweierpotenz oberhalb von $N - 1$ ist, $2^{k-1} + 1 < N \leqq b := 2^k + 1$, das bedeutet in unserem Beispiel $b := 17$. Rufen wir nun mit $a := 1$, $b := 17$ und $\delta := 1$ den Bisektionsalgorithmus auf, so ergibt sich die Folge der Einschließungen [1, 17], [1, 9], [1, 5], [3, 5], [3, 4]. Aus der letzten Einschließung ist abzulesen, daß $c = 5.3$ zwischen x_3 und x_4 eingeordnet werden muß.

Das Verfahren eignet sich auch als *Such- und Sortieralgorithmus* in der *nichtnumerischen Datenverarbeitung*, zum Beispiel für die alphabetische Anordnung von Stichwörtern, und wird in diesem Zusammenhang *binäres* (*logarithmisches*) *Suchen* genannt (KNUTH [1], S. 407).

5.1.3. Einfache Iteration

Das Bisektionsverfahren konvergiert für alle Funktionen gleich langsam. Das liegt daran, daß man als einzige Information das Vorzeichen der Funktion benutzt. Besser angepaßte Iterationsverfahren sind zu erwarten, wenn man die Funktionswerte selbst mit in die Rechnung einbezieht. In den Beispielen 5.1.1, 5.1.2 und 5.1.3 haben wir die Ausgangsgleichungen (1), (4) und (9) durch verschiedene Umformungen auf die Gestalt einer Fixpunktgleichung $x = g(x)$ gebracht und diese als Iterationsvorschrift $x_{n+1} = g(x_n)$ geschrieben. Offen blieb bisher die Frage, unter welchen Voraussetzungen eine solche *einfache Iteration* (sukzessive Approximation, gewöhnliches Iterationsverfahren) eine konvergente Folge von Näherungen liefert.

Beispiel 5.1.7. In den ersten beiden Beispielen ergab sich

$$g(x) := \frac{x}{2} + \frac{1}{x} \text{ beziehungsweise } g(x) := \frac{\pi}{5} + 0.8 \sin x, \tag{20}$$

und die Näherungswerte konvergierten gut. In Abb. 5.1.4 a), b) sind die Kurven $y = g(x)$ graphisch dargestellt. Ihr Schnittpunkt mit der Geraden $y = x$ ist der gesuchte Fixpunkt. Die Pfeile deuten den Verlauf der Iteration an. Im dritten Beispiel, in dem wir gleich zwei Möglichkeiten ausprobierten (vgl. (10) und (10′)),

$$\begin{aligned} g(x) &:= 0.702317 - 1.003311\, e^{-3x}, \\ g(x) &:= -\frac{1}{3} \ln (0.7 - 0.996700\, x), \end{aligned} \tag{21}$$

erhielten wir eine divergente und eine extrem langsam konvergierende Folge (Abb. 5.1.4, c), d)).

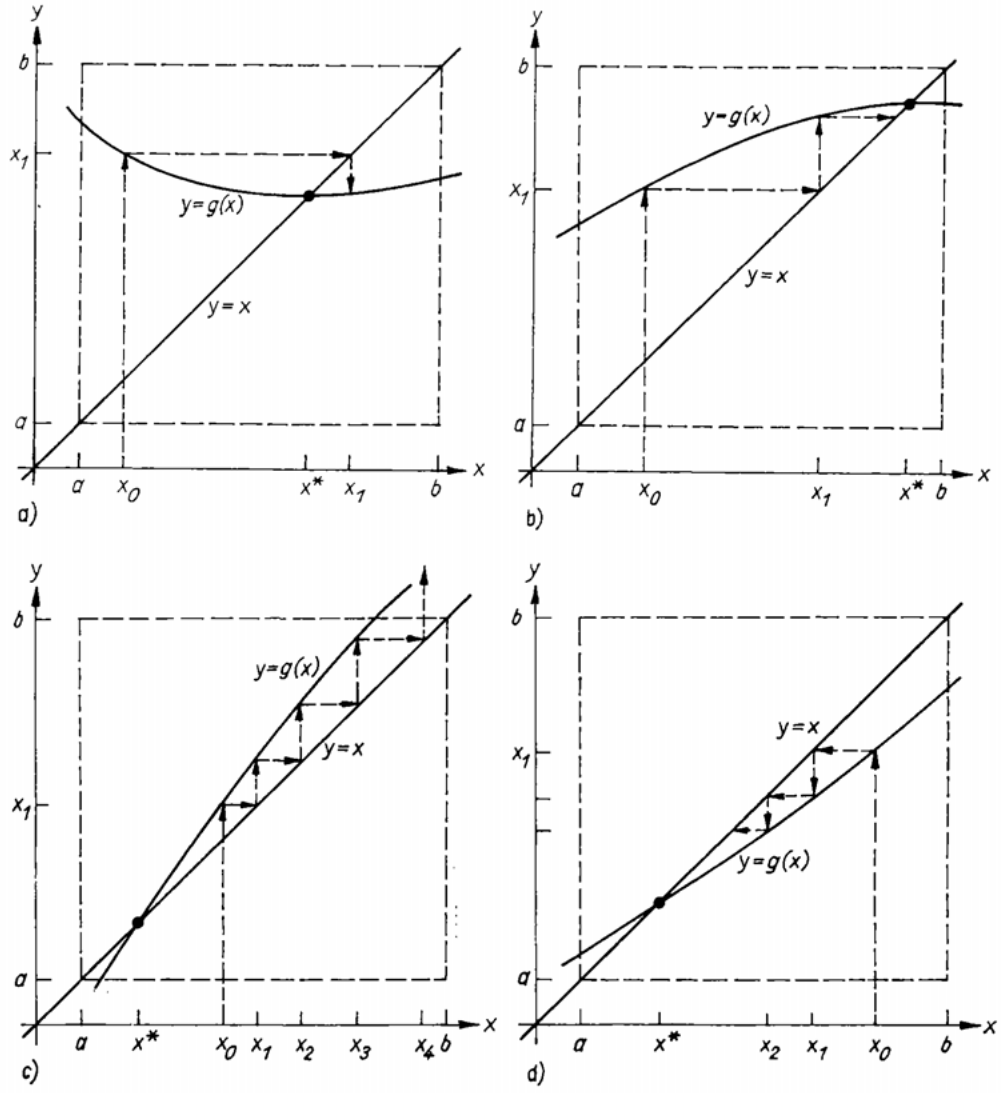

Abb. 5.1.4. Einfache Iteration: a), b), d) konvergent, c) divergent

Eine konvergente Folge von Näherungen ergibt sich offenbar dann, wenn man ein Intervall $[a, b]$ mit der Eigenschaft findet, daß für jedes x aus $[a, b]$ auch $g(x)$ in $[a, b]$ liegt und die Funktion g dort nicht schneller wächst als die Gerade $y = x$. Das heißt, genauer gesagt, alle Sekantenanstiege

$$\left|\frac{g(x) - g(t)}{x - t}\right| \leqq L < 1, \qquad x, t \in [a, b], \tag{22}$$

oder, falls g differenzierbar ist, alle Tangentenanstiege (Ü 5.1.2)

$$|g'(x)| \leqq L < 1, \qquad x \in [a, b], \tag{23}$$

müssen betragsmäßig echt kleiner als Eins bleiben. (Im Fall der Abb. 5.1.4c) sind die Voraussetzungen verletzt.) Daß diese Konvergenzbedingungen auch für allgemeinere Klassen von Iterationsverfahren gelten, zeigt der aus der Analysis bekannte *Kontraktionssatz* (*Banachsche Fixpunktsatz*). Um ihn allgemein formulieren zu können, erinnern wir an die Definition des Banach-Raumes:

Definition 5.1.8. Eine Menge $\mathbb{X}$ von Elementen $x, y, \ldots$ heißt *Banach-Raum* über dem Körper der reellen (komplexen) Zahlen, wenn

a) $\mathbb{X}$ ein linearer Raum ist, das heißt mit x auch $c \cdot x$, c reell (komplex), und mit x, y auch $x + y$ in $\mathbb{X}$ liegt, wobei die üblichen Rechenregeln gelten,

b) $\mathbb{X}$ normiert ist, das heißt für jedes x eine Norm $\|x\|$ mit den üblichen Eigenschaften erklärt ist (vgl. Definition 2.1.10 und Definition 6.4.2), und

c) $\mathbb{X}$ vollständig ist, das heißt zu jeder in sich konvergenten Folge (Cauchy-Folge) (x_n) von Elementen aus $\mathbb{X}$ auch der Grenzwert x^* existiert und in $\mathbb{X}$ liegt.

Einfache Beispiele für Banach-Räume sind der uns hier interessierende Fall der Zahlengeraden $\mathbb{X} := \mathbb{R}^1$, in dem der Betrag als Norm dient, $\|x\| := |x|$, oder der Fall des N-dimensionalen Vektorraums $\mathbb{X} := \mathbb{R}^N$, in dem irgendeine der früher besprochenen Vektornormen (vgl. Abschnitt 2.1.3.) verwendet werden kann. Aber auch die Menge $\mathbb{P}^{N+1}[-1, 1]$ der Polynome von höchstens N-tem Grad über $[-1, 1]$ (vgl. Beispiel 6.4.5) oder die Menge $\mathbf{C}[a, b]$ der über $[a, b]$ stetigen Funktionen bilden Banach-Räume, wenn man geeignete Normen einführt (vgl. (6.4.5.) und (6.4.6)).

Satz 5.1.9 (Kontraktionssatz). *Bezeichnet $\mathbb{X}$ einen Banach-Raum oder eine abgeschlossene Teilmenge eines solchen, und ist $g(x)$ eine Abbildung, die $\mathbb{X}$ in sich abbildet,*

$$g(x) \in \mathbb{X} \text{ für alle } x \text{ aus } \mathbb{X}, \tag{24}$$

und einer Kontraktionsbedingung

$$\|g(x) - g(t)\| \leqq L \|x - t\| \qquad \text{für alle } x, t \text{ aus } \mathbb{X} \tag{25}$$

mit einer Zahl L aus dem Intervall $[0, 1)$ (d. h. einer Lipschitz-Bedingung mit einer Lipschitz-Konstanten, die echt kleiner als Eins ist) genügt, so besitzt g in $\mathbb{X}$ genau einen Fixpunkt x^*, und die Folge der durch (15) *gelieferten Näherungen x_n konvergiert für jeden Startwert x_0 aus $\mathbb{X}$ gegen x^*.*

Zum Beweis der Konvergenz benutzen wir das *Cauchysche Konvergenzkriterium*, d. h., wir zeigen, daß der *Abstand* (also die Norm der Differenz) zweier Elemente x_n und x_{n+m} der Folge für beliebiges m und hinreichend groß gewähltes n kleiner als jedes positive ε wird. Dazu untersuchen wir zunächst den Abstand zweier aufeinanderfolgender Elemente x_n und x_{n+1}. Wegen (15) und (25) ist

$$\begin{aligned} \|x_{n+1} - x_n\| = \|g(x_n) - g(x_{n-1})\| &\leqq L \|x_n - x_{n-1}\| = L \|g(x_{n-1}) - g(x_{n-2})\| \\ &\leqq \cdots \leqq L^n \|x_1 - x_0\|. \end{aligned} \tag{26}$$

Für die Differenz zweier beliebiger Folgenelemente gilt, wenn wir die dazwischenliegenden Elemente subtrahieren und wieder addieren und die Dreiecksungleichung verwenden

$$\begin{aligned} \|x_{n+m} - x_n\| &= \|x_{n+m} - x_{n+m-1} + x_{n+m-1} - + \cdots - x_{n+1} + x_{n+1} - x_n\| \\ &\leqq \|x_{n+m} - x_{n+m-1}\| + \cdots + \|x_{n+2} - x_{n+1}\| + \|x_{n+1} - x_n\|. \end{aligned}$$

Schätzt man nun jeden der auf der rechten Seite der Ungleichung stehenden Terme nach (26) ab, so folgt

$$\|x_{n+m} - x_n\| \leqq (L^{n+m-1} + \cdots + L^{n+1} + L^n)\,\|x_1 - x_0\| = L^n \frac{1 - L^m}{1 - L}\,\|x_1 - x_0\|$$

$$\leqq \frac{L^n}{1 - L}\,\|x_1 - x_0\|, \tag{27}$$

wobei wir die L-Potenzen mit Hilfe der bekannten Formel für Teilsummen der geometrischen Reihe aufsummiert haben. Damit ist die Konvergenz der Folge (x_n) nachgewiesen, denn L^n strebt wegen $L < 1$ für $n \to \infty$ gegen Null, so daß der rechts stehende Ausdruck durch Wahl eines hinreichend großen n für jedes m und jeden festen Startwert x_0 unter jedes positive ε gedrückt werden kann. Man braucht dazu bloß

$$n \geqq N(\varepsilon) := \ln\left(\frac{\varepsilon(1 - L)}{\|g(x_0) - x_0\|}\right) \Big/ \ln L$$

zu wählen (vgl. (2.7.27)). Bezeichnen wir den Grenzwert der Folge (x_n) mit x^*, so erhalten wir für $m \to \infty$ aus (27) die für lineare Iterationsverfahren (vgl. (2.7.25)) bereits bekannte A-priori-Abschätzung

$$\|\delta_n\| := \|x_n - x^*\| \leqq \frac{L^n}{1 - L}\,\|x_1 - x_0\|. \tag{28}$$

Mit ihrer Hilfe läßt sich zeigen, daß der Grenzwert x^* Fixpunkt von $g(x)$ ist. Wir ergänzen die Differenz $x^* - g(x^*)$ durch $-x_{n+1} + g(x_n) = 0$ und erhalten unter Verwendung der Dreiecksungleichung

$$\|x^* - g(x^*)\| = \|x^* - x_{n+1} + g(x_n) - g(x^*)\| \leqq \|x^* - x_{n+1}\| + \|g(x_n) - g(x^*)\|$$

$$\leqq 2\,\frac{L^{n+1}}{1 - L}\,\|x_1 - x_0\|.$$

Die rechte Seite der letzten Ungleichung kann kleiner als jedes positive ε gemacht werden. Also ist $\|x^* - g(x^*)\| = 0$ und wegen des ersten Normaxioms $x^* = g(x^*)$. Die Eindeutigkeit von x^* beweisen wir indirekt: Gäbe es noch einen zweiten Fixpunkt x', so gälte $\|x^* - x'\| = \|g(x^*) - g(x')\| \leqq L\,\|x^* - x'\|$ und folglich $(1 - L)\,\|x^* - x'\| \leqq 0$. Der Faktor $1 - L$ ist positiv, also muß $\|x^* - x'\|$ gleich Null und damit $x^* = x'$ sein. ∎

Ist $\mathbb{X} := \mathbb{R}^N$, $\boldsymbol{x}$ ein Vektor des $\mathbb{R}^N$ und g eine lineare Abbildung des $\mathbb{R}^N$ auf sich, $g(\boldsymbol{x}) := \boldsymbol{T}\boldsymbol{x} + \boldsymbol{v}$, so erhält man aus (15) das Iterationsverfahren (2.7.16) zur Lösung linearer Gleichungssysteme. Wegen $\|g(\boldsymbol{x}) - g(\boldsymbol{t})\| = \|\boldsymbol{T}\boldsymbol{x} + \boldsymbol{v} - \boldsymbol{T}\boldsymbol{t} - \boldsymbol{v}\| = \|\boldsymbol{T}(\boldsymbol{x} - \boldsymbol{t})\|$ $\leqq \|\boldsymbol{T}\|_M\,\|\boldsymbol{x} - \boldsymbol{t}\|$, wo $L := \|\boldsymbol{T}\|_M$ eine mit der Vektornorm $\|\boldsymbol{x}\|$ verträgliche Matrixnorm bezeichnet, ist Satz 2.7.10a) als Spezialfall im Kontraktionssatz enthalten. Satz 2.7.11 bleibt auch im allgemeineren Fall der Iteration (15) gültig. Insbesondere gilt also die A-posteriori-Abschätzung (2.7.24)

$$\frac{1}{1 + L}\,\|x_{n+1} - x_n\| \leqq \|\delta_n\| \leqq \frac{1}{1 - L}\,\|x_{n+1} - x_n\| \leqq \frac{L}{1 - L}\,\|x_n - x_{n-1}\|. \tag{29}$$

Für nichtlineare Funktionen g ist es u. U. schwierig, eine für den Kontraktionssatz geeignete Teilmenge des Definitionsgebiets $\mathbb{D}$ der Funktion zu finden. Gelegentlich hilft der folgende

Satz 5.1.10 (Konvergenzkugel). *Enthält das Definitionsgebiet $\mathbb{D}$ der Funktion g eine Kugel*

$$\mathbb{K}(t, r) := \{x; \|x - t\| \leqq r, \qquad x, t \in \mathbb{D}\} \tag{30}$$

mit dem Mittelpunkt t aus $\mathbb{D}$ und dem Radius r, genügt g in $\mathbb{K}$ einer Kontraktionsbedingung der Gestalt (25) *und gilt*

$$\|g(t) - t\| \leqq (1 - L)\, r, \tag{31}$$

so erfüllt g mit $\mathbb{X} := \mathbb{K}(t, r)$ die Voraussetzungen des Kontraktionssatzes.

Beweis. Die Kugel ist nach Definition (30) abgeschlossen, die Kontraktionsbedingung ist nach Voraussetzung erfüllt. Es bleibt also lediglich zu zeigen, daß g die Kugel in sich abbildet. Ist x aus $\mathbb{K}$, so gilt für g wegen (25), (30) und (31)

$$\begin{aligned}\|g(x) - t\| &= \|g(x) - g(t) + g(t) - t\| \leqq \|g(x) - g(t)\| + \|g(t) - t\| \\ &\leqq L\,\|x - t\| + (1 - L)\, r \leqq Lr + (1 - L)\, r = r.\end{aligned}$$

Das Bildelement liegt also ebenfalls in der Kugel. ∎

Aus dem Satz 5.1.10 ergibt sich die

Folgerung 5.1.11 (Beispiel einer Konvergenzkugel). *Genügt g in $\mathbb{D}$ einer Kontraktionsbedingung* (25) *und liegt die Kugel $\mathbb{K}(t, r)$ mit dem Mittelpunkt $t := x_1$ und dem Radius $r := \dfrac{L}{1 - L}\,\|x_1 - x_0\|$ in $\mathbb{D}$, so ist $\mathbb{K}(t, r)$ eine Konvergenzkugel.*

Beweis. Es ist lediglich zu zeigen, daß (31) erfüllt ist. Aus der Kontraktionsbedingung folgt aber unmittelbar

$$\|g(x_1) - x_1\| = \|g(x_1) - g(x_0)\| \leqq L \cdot \frac{1 - L}{1 - L}\,\|x_1 - x_0\| = (1 - L)\, r. \;\blacksquare$$

Wir kehren nun zum Spezialfall *reeller Funktionen* einer reellen Veränderlichen zurück, $\mathbb{X}$ ist dann ein Intervall der Zahlengeraden, und die Norm wird zum Betrag reeller Zahlen. Die Kontraktionsbedingung geht in die Ungleichung (22) über. Die Sekantenanstiege abzuschätzen ist i. allg. sehr aufwendig. Für differenzierbare Funktionen verwendet man deshalb besser die Kontraktionsbedingung (23), bestimmt also das *Betragsmaximum der Ableitung $g'(x)$*.

Beispiel 5.1.12. Für die erste Funktion aus (20), also die Iterationsvorschrift aus Beispiel 5.1.1, ergibt sich $g'(x) = \dfrac{1}{2} - \dfrac{1}{x^2}$ und $g''(x) = 2/x^3$. Über der positiven Halbachse ist $g'(x)$ demnach monoton wachsend. Auf Grund der numerischen Tests in Beispiel 5.1.1 wählen wir $\mathbb{D} := [1.4, 1.5]$. Für dieses Intervall erhält man aus $|g'(x)| \leqq \max(|g'(1.4)|, |g'(1.5)|) < 0.056$ eine Lipschitz-Konstante $L := 0.056$, die wesentlich kleiner als Eins ist. Das bedeutet schnelle Konvergenz. Nehmen wir $x_0 := 1.5$ als Startwert und die Näherung $x_1 = 1.41\bar{6}$ als Mittelpunkt, so liegt die „Kugel" $\mathbb{K}(x_1, r)$ mit dem aus $0.056 \cdot |1.41\bar{6} - 1.5|/0.944 < 0.005 =: r$ bestimmten Radius in $\mathbb{D}$, ist also nach Folgerung 5.1.11 eine Konvergenzkugel. Für die Näherung x_2 liefert (28) die *A-priori-Abschätzung* $|\delta_2| \leqq 0.056 \cdot r < 2.77 \cdot 10^{-4}$ und (29) die *A-posteriori-Abschätzung* $2.01 \cdot 10^{-6} \leqq |\delta_2| \leqq 2.25 \cdot 10^{-6}$. Der wahre Fehler ist $\delta_2 := x_2 - \sqrt{2} \approx 2.12 \cdot 10^{-6}$.

Für die zweite Funktion aus (21), also die Iterationsvorschrift (10′) aus Beispiel 5.1.3, ergibt sich $g'(x) = 1/(2.106953 - 3x)$. Bis zu ihrer Polstelle $x_\infty \approx 0.7$ ist die Ableitung positiv und

monoton wachsend. Über jedem Intervall $[a, b]$ mit $b < 0.7$ wird das Maximum also im rechten Randpunkt angenommen, für $[0, 0.35]$ ergibt sich aus $g'(x) < g'(0.35) < 0.95 =: L$ eine Lipschitz-Konstante, die dicht bei Eins liegt. Das bedeutet schlechte Konvergenz. Nehmen wir $t := 0.34$ als Mittelpunkt, so liefert $|g(t) - t|/(1 - L) < 0.01 =: r$ das Konvergenzintervall $\mathbb{K}(0.34, 0.01) := [0.33, 0.35]$. Die A-priori-Abschätzung hat für den Startwert $x_0 := 0.35$ die Gestalt $|\delta_n| \leqq (0.95)^n \cdot 0.024$. Danach benötigt man etwa 45 Schritte, um die Genauigkeit um eine Dezimale zu verbessern. Die Iterationsvorschrift ist also praktisch unbrauchbar.

Wenn die Iterationsfunktion $g(x)$ in $[a, b]$ monoton ist, kann man mit Hilfe von (15) Intervallschachtelungen der Gestalt (19) konstruieren.

Satz 5.1.13 (Monotone Einschließung). *Die Funktion g sei aus $\mathbf{C}[a, b]$ und genüge über $[a, b]$ den Voraussetzungen des Fixpunktsatzes. Ist g monoton fallend und der Startwert x_0 nicht größer als x^*, so genügen die durch* (15) *gelieferten Näherungen der Ungleichung*

$$a \leqq x_0 \leqq x_2 \leqq \cdots \leqq x_{2n} \leqq \cdots \leqq x^* \leqq \cdots \leqq x_{2n+1} \leqq \cdots \leqq x_3 \leqq x_1 \leqq b. \quad (32)$$

Ist g monoton steigend und startet man zwei Iterationsfolgen (x_n) und (x'_n) mit $a \leqq x_0 \leqq x^* \leqq x'_0 \leqq b$, so gilt

$$a \leqq x_0 \leqq x_1 \leqq \cdots \leqq x_n \leqq \cdots \leqq x^* \leqq \cdots \leqq x'_n \leqq \cdots \leqq x'_1 \leqq x'_0 \leqq b. \quad (33)$$

Beweis. Aus $x_0 \leqq x^*$ folgt für monoton fallende Funktionen $x_1 = g(x_0) \geqq g(x^*) = x^*$ und für monoton steigende Funktionen $x_1 = g(x_0) \leqq g(x^*) = x^*$. Wenn man noch die Kontraktionsbedingung berücksichtigt, ergeben sich daraus durch vollständige Induktion die Ungleichungen (32) und (33). ∎

Zum Vergleich von Iterationsverfahren benutzen wir den bereits früher eingeführten Begriff der *Konvergenzordnung* (vgl. Definition 2.7.9). Sie kann, wenn g hinreichend oft differenzierbar ist, mit Hilfe der Taylor-Entwicklung von g an der Stelle $x = x^*$ bestimmt werden:

Satz 5.1.14 (Konvergenzordnung). *Konvergiert die durch* (15) *und einen Startwert x_0 erklärte Folge (x_n) gegen x^*, ist die Funktion g aus $\mathbf{C}^p[a, b]$ und gilt $g^{(k)}(x^*) = 0$ für $k = 1(1)p - 1$ und $g^{(p)}(x^*) \neq 0$, so hat die Folge (x_n) die Konvergenzordnung p und den Konvergenzfaktor (asymptotischen Fehlerkoeffizienten)*

$$q := \frac{1}{p!} \sup_{x \in [a,b]} |g^{(p)}(x)|. \quad (34)$$

Beweis. Wegen $g \in \mathbf{C}^p[a, b]$ gilt für g im Punkt $x = x^*$ die Taylor-Formel

$$g(x) = \sum_{n=0}^{p-1} \frac{1}{n!} (x - x^*)^n g^{(n)}(x^*) + \frac{1}{p!} (x - x^*)^p g^{(p)}\big(x^* + \vartheta(x - x^*)\big),$$

$$0 < \vartheta < 1. \quad (35)$$

Setzt man $x := x_n$ und $\delta_n := x_n - x^*$ und berücksichtigt $x^* = g(x^*)$ und $x_{n+1} = g(x_n)$, so ergibt sich

$$|\delta_{n+1}| = |\delta_n|^p \frac{1}{p!} |g^{(p)}(x^* + \vartheta\delta_n)| \leqq q \cdot |\delta_n|^p. \quad (36)$$ ∎

Beispiel 5.1.15. Für die erste Funktion aus (20) ist $g'(x^*) = 0$, und $g''(x^*) = 2/x^{*3} \neq 0$. Also besitzt die Iteration (3) die Konvergenzordnung 2 und, wenn man das Intervall [1.4, 1.5] zugrunde legt, den Konvergenzfaktor $\frac{1}{2!} \sup_{x \in [1.4,1.5]} (2/x^3) < 0.37 =: q$. Diese *überlineare Konvergenz* erklärt die

bereits in Beispiel 5.1.1 beobachtete schnelle Zunahme der Genauigkeit: Hat δ_n die Größenordnung 10^{-k}, so ist δ_{n+1} bereits von der Größenordnung 10^{-2k}. Abgesehen von einer gewissen Anlaufrechnung verdoppelt sich also bei quadratischer Konvergenz, grob gesprochen, bei jedem Schritt die Anzahl der richtigen Ziffern.

5.1.4. Newton-Verfahren und Regula falsi

Das heuristische Vorgehen des Umstellens der Nullstellengleichung (11) zu einer Fixpunktgleichung (14) kann, wie wir gesehen haben, zu schlecht konvergierenden oder gar divergierenden Näherungsfolgen führen. Systematischer gelangt man von (11) durch *lokale Linearisierung* zu einer iterierfähigen Gestalt (14). Dazu setzen wir zunächst voraus, daß f mindestens einmal stetig differenzierbar ist, bereits eine Einschließung $[a, b]$ für eine Nullstelle x^* gefunden wurde und diese Nullstelle einfach ist, das heißt, daß $f'(x^*) \neq 0$ ist und somit

$$f(x) = (x - x^*) \cdot f'\big(x^* + \vartheta(x - x^*)\big) \text{ mit einem } \vartheta \text{ aus } (0, 1) \tag{37}$$

gilt, wobei die Ableitung $f'(x)$ in der Umgebung von $x = x^*$ nicht verschwindet (vgl. B 5.1.5). Wir bezeichnen nun mit x_n eine beliebige Näherung und mit $f(x_n)$ den zugehörigen Funktionswert und ersetzen die Funktion $y = f(x)$ durch eine Gerade

$$y = \tilde{f}(x) := f(x_n) + \mu(x - x_n) \tag{38}$$

durch den Punkt $\big(x_n, f(x_n)\big)$. Den Anstieg μ legen wir dabei so fest, daß die Gerade in der Umgebung dieses Punktes als Näherung der Funktion angesehen werden kann, z. B. wählen wir den *Anstieg der Sekante* durch die Punkte $\big(a, f(a)\big)$ und $\big(b, f(b)\big)$ (der auch *Steigung erster Ordnung* genannt und durch $f[a, b]$ abgekürzt wird, vgl. (6.1.22))

$$\mu := f[a, b] := \big(f(b) - f(a)\big)/(b - a), \tag{39}$$

oder den *Anstieg der Tangente* im Punkt $\big(x_n, f(x_n)\big)$

$$\mu := f'(x_n), \tag{40}$$

oder den Anstieg der Sekante durch $\big(x_n, f(x_n)\big)$ und einen festen Punkt $\big(x_0, f(x_0)\big)$

$$\mu := f[x_0, x_n], \tag{41}$$

oder den Anstieg der Sekante durch die letzten beiden Näherungen $\big(x_n, f(x_n)\big)$ und $\big(x_{n-1}, f(x_{n-1})\big)$

$$\mu := f[x_n, x_{n-1}]. \tag{42}$$

Wenn die Gerade (38) eine hinreichend gute Näherung für die Funktion $y = f(x)$ darstellt, kann man erwarten, daß die Nullstelle der Geraden (wir bezeichnen sie mit x_{n+1}) eine bessere Näherung für x^* liefert als x_n. Um sie zu erhalten, setzen wir in (38) $y := 0$ und $x := x_{n+1}$ und stellen nach x_{n+1} um:

$$x_{n+1} = x_n - \frac{1}{\mu} f(x_n). \tag{43}$$

Dieses Verfahren heißt *vereinfachtes Newton-Verfahren*, wenn man μ über mehrere Schritte konstant läßt, (*klassisches*) *Newton-Verfahren* (*Tangenten-Näherungsverfahren*), wenn man jeweils den aktuellen Tangentenanstieg (40) wählt, und *Regula falsi* (*1. Form*) beziehungsweise *Regula falsi* (*2. Form*) (*Sekanten-Näherungsverfahren*), wenn man je-

weils den Sekantenanstieg (41) bzw. (42) verwendet. In Abb. 5.1.5 ist das Vorgehen für alle vier Fälle graphisch veranschaulicht. Es zeigt sich, daß man für hinreichend gute Startwerte x_0 (bzw. x_0 und x_1) jeweils konvergente Näherungsfolgen erhält. Bei ungeschickter Wahl der Anfangsnäherung (sie ist in den Abbildungen mit t_0 oder s_0 bezeichnet) können dagegen in allen vier Varianten divergente Folgen auftreten. Um solche Fälle auszuschließen, konstruieren wir wie beim Bisektionsverfahren eine Folge von ineinandergeschachtelten Einschließungen. Dazu benutzen wir abwechselnd die Regula

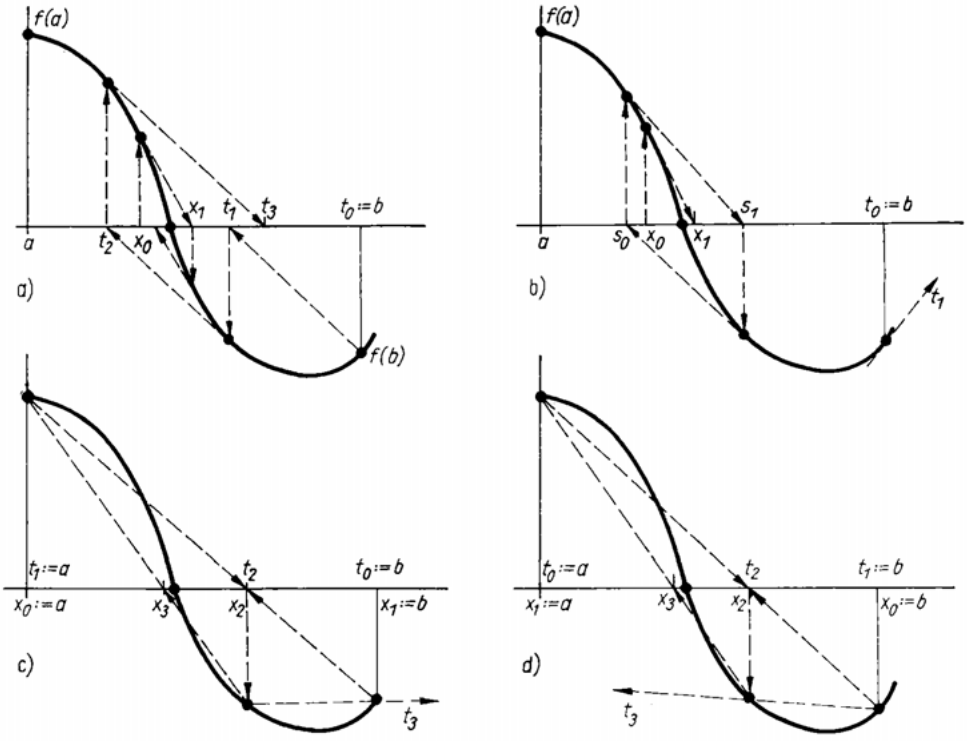

Abb. 5.1.5. Lokale Linearisierung: a) vereinfachtes Newton-Verfahren, b) Newton-Verfahren, c) Regula falsi (1. Form), d) Regula falsi (2. Form)

falsi und das Newton-Verfahren (Abb. 5.1.6): Ausgehend von einer Einschließung $[a, b]$ wird mit Hilfe der Regula falsi ein Zwischenpunkt $x := a - f(a)/f[a, b]$ berechnet und derjenige Randpunkt überspeichert, dessen Funktionswert das gleiche Vorzeichen wie $f(x)$ hat. Anschließend wird mit dem anderen Randpunkt als Startwert (wir bezeichnen ihn mit x) ein Newton-Schritt ausgeführt, $x := x - f(x)/f'(x)$. Liegt die Newton-Näherung außerhalb der Einschließung oder verkleinert sie die Einschließung nur unwesentlich, so verwenden wir an ihrer Stelle die Bisektionsnäherung $x := (a + b)/2$. Als Kriterium wird dabei

$$(x - a)\,(b - x) > \alpha(b - a)^2, \qquad 0 \leqq \alpha < 1/4,$$

verwendet, wo α einen Steuerparameter bezeichnet. Der Newton-Schritt verkleinert die Intervallänge dann mindestens auf $\left(1 + \sqrt{1 - 4\alpha}\right)|b - a|/2$. Im Fall $\alpha = 0$, in dem jede

Newton-Näherung zugelassen ist, die nicht außerhalb von $[a, b]$ liegt, kann die Abnahme beliebig klein werden, im Grenzfall $\alpha = 1/4$ reduziert sich das Verfahren auf eine Kombination der Regula falsi mit dem Bisektionsverfahren (die Berechnung von f_1 und die Schritte 2 und 3 sind dann überflüssig, vgl. Ü 5.1.6).

Algorithmus 5.1.16. Regula-falsi-Newton-Verfahren zur Einschließung von Nullstellen RENE (f, f', a, b, δ, MAX, α)

f	differenzierbare Funktion, muß als Unterprogramm vorliegen	
f'	erste Ableitung von f, muß als Unterprogramm vorliegen	
a, b	Intervallgrenzen mit $f(a) > 0$, $f(b) \leqq 0$	
δ, MAX	Genauigkeitsschranke, maximale Schrittanzahl	
α	Steuerparameter, $0 \leqq \alpha < 0.25$	
	$f_a := f(a), \quad f_b := f(b)$	Funktionswerte
	$n = 1(1)\mathrm{MAX}$	Schrittzähler
1	$x := a + \dfrac{f_a}{f_a - f_b}(b - a)$	Regula falsi
	$f_x := f(x)$	Funktionswert
	falls $f_x \leqq 0$, dann $b := x, f_b := f_x, x := a, f_x := f_a$, sonst $a := x, f_a := f_x, x := b, f_x := f_b$	
	$f_1 := f'(x)$	Wert der 1. Ableitung
	falls $\lvert f_1 \rvert < \delta$, dann Schritt 4	
2	$x := x - f_x/f_1$	Newton-Schritt
	$q_1 := (x - a)/(b - a), q_2 := (b - x)/(b - a)$	
3	falls $q_1 \cdot q_2 > \alpha$, dann Schritt 5	
4	$x := a + (b - a) \cdot 0.5$	Intervallhalbierung, falls Newton-Näherung ungeeignet
5	$f_x := f(x)$	Funktionswert zur Newton- oder Bisektionsnäherung
	falls $f_x \leqq 0$, dann $b := x, f_b := f_x$, sonst $a := x, f_a := f_x$	
	drucke n, a, b	
6	falls $\lvert b - a \rvert < \delta$, dann stop	reguläres Ende
	drucke: Geforderte Genauigkeit nicht erreicht.	

Wie das Bisektionsverfahren findet der Algorithmus eine Einschließung für eine Nullstelle aus dem Startintervall. Das gilt auch noch, wenn (37) nicht erfüllt ist oder f in $[a, b]$ mehrere Nullstellen besitzt. Der *Rechenaufwand* besteht aus drei Funktionsauswertungen je Doppelschritt (zweimal $f(x)$ und einmal $f'(x)$).

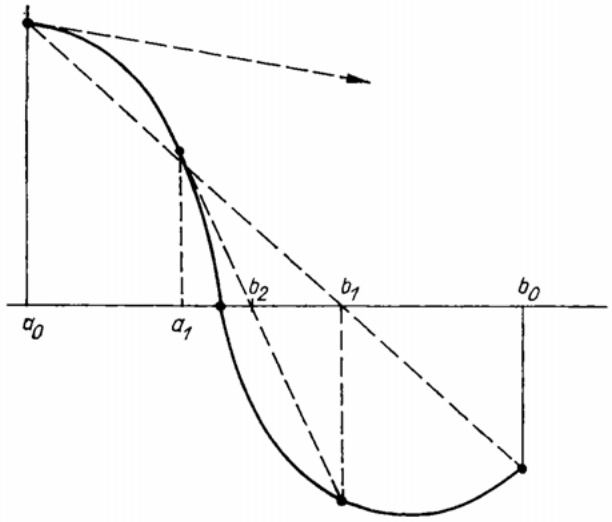

Abb. 5.1.6. Kombination von Regula falsi und Newton-Verfahren

Beispiel 5.1.17. Wir verwenden den Algorithmus zur Lösung des Nullstellenproblems (9) aus Beispiel 5.1.3. Mit $k := x$ ergibt sich

$$f(x) := 0.7 + \frac{1.2}{\ln 0.3}\, x - \mathrm{e}^{-3x},\ f'(x) := \frac{1.2}{\ln 0.3} + 3\mathrm{e}^{-3x}. \tag{44}$$

Startet man den Algorithmus mit der Einschließung $a := 0.4$, $b := 0.3$ und den Parametern $\alpha := 0$ und $\delta := 10^{-6}$, so erhält man bei zehnstelliger Gleitpunktrechnung im vierten Doppelschritt die Einschließung

$$b_4 := 0.3344065 \leqq x^* \leqq 0.3344073 =: a_4.$$

Ein mit $x_0 := 0.4$ gestartetes Newton-Verfahren würde aus dem Intervall [0.3, 0.4] herausführen, die Regula falsi (1. Form) konvergiert (wenn auch sehr langsam, Ü 5.1.4), liefert aber keine Einschließungen, die sich auf Null zusammenziehen.

Für eine genauere Analyse des Konvergenzverhaltens der einzelnen Verfahren benutzen wir zunächst den Kontraktionssatz. Im Fall des *vereinfachten Newton-Verfahrens* haben die Iterationsfunktion g und ihre erste Ableitung die Gestalt

$$g(x) := x - \frac{1}{\mu} f(x), \qquad g'(x) = 1 - \frac{1}{\mu} f'(x). \tag{45}$$

Die Ableitung $g'(x)$ bleibt demnach nur dann betragsmäßig unterhalb von Eins, wenn man $\operatorname{sgn} \mu = \operatorname{sgn} f'(x)$ wählt. Ein für das gesamte Intervall verwendbares μ findet sich also nur, wenn f' in diesem Intervall sein Vorzeichen nicht wechselt, das heißt, wenn f dort monoton ist. (In Abb. 5.1.5a) erweist sich das Intervall $[a, b]$ deshalb als ungeeignet, $[a, t_1]$ dagegen ist ein Monotonieintervall.) Im folgenden setzen wir zusätzlich

$$0 < m_1 \leqq |f'(x)| \leqq M_1, \qquad x \in [a, b], \tag{46}$$

voraus und erhalten damit für $g'(x)$ die Abschätzung

$$-1 < 1 - \frac{M_1}{|\mu|} \leqq g'(x) \leqq 1 - \frac{m_1}{|\mu|} < 1.$$

Die rechte Ungleichung ist stets erfüllt, die linke gilt, wenn $|\mu| > M_1/2$ gewählt wird. Die Lipschitz-Konstante wird durch

$$L := \max\left(\left|1 - \frac{M_1}{|\mu|}\right|, \left|1 - \frac{m_1}{|\mu|}\right|\right) \tag{47}$$

bestimmt. Die Ausdrücke $1 - M_1/|\mu|$ und $1 - m_1/|\mu|$ nehmen mit wachsendem $|\mu|$ monoton zu und sind für $0 < |\mu| < m_1$ beide negativ und für $M_1 < |\mu| < \infty$ beide positiv. Im Intervall (m_1, M_1) haben sie entgegengesetztes Vorzeichen. Die kleinste Lipschitz-Konstante ergibt sich also, wenn man μ so wählt, daß $-(1 - M_1/|\mu|) = 1 - m_1/|\mu|$, also $|\mu| := (M_1 + m_1)/2$ ist. Die *Konvergenzordnung* des vereinfachten Newton-Verfahrens ist nach Satz 5.1.14 mindestens gleich Eins, da $g'(x^*) = 1 - f'(x^*)/\mu$ im allgemeinen von Null verschieden sein wird. Für $\mu = f'(x^*)$ ergibt sich sogar quadratische Konvergenz. (Wenn also für den Anstieg $f'(x^*)$ ein Näherungswert bekannt ist, empfiehlt es sich, für μ diesen Wert zu benutzen.) Wir fassen zusammen:

Satz 5.1.18 (Konvergenz des vereinfachten Newton-Verfahrens). *Ist die Funktion f aus $C^1[a, b]$, wobei $[a, b]$ eine Einschließung einer einfachen Nullstelle von f bezeichnet, und genügt f' einer Abschätzung der Gestalt* (46), *so konvergiert die durch* (43) *erklärte Folge von Näherungswerten x_n für jedes μ mit $|\mu| > M_1/2$, $\operatorname{sgn} \mu = \operatorname{sgn} f'(x)$ und jeden Startwert x_0 aus $[a, b]$ mit der Konvergenzordnung $p = 1$ und dem durch* (47) *bestimmten Konvergenzfaktor $q := L$ gegen x^*. Für $|\mu| = (M_1 + m_1)/2$ nimmt die Lipschitz-Konstante ihren kleinsten Wert $L_{\min} := (M_1 - m_1)/(M_1 + m_1)$ an.*

Für das *klassische Newton-Verfahren* (43), (40) ist

$$g(x) := x - \frac{f(x)}{f'(x)}, \qquad g'(x) = \frac{f(x)\, f''(x)}{[f'(x)]^2}. \tag{48}$$

Damit ergibt sich im Fall einfacher Nullstellen $g'(x^*) = 0$, so daß man unter der Voraussetzung $f \in C^2[a, b]$ nach Satz 5.1.14 die Konvergenzordnung $p = 2$ und den Konvergenzfaktor $q = \frac{1}{2} \sup_{x \in [a,b]} |g''(x)|$ erhält (vgl. Ü 5.1.7). Der folgende Satz kommt mit einer schwächeren Glattheitsvoraussetzung aus und enthält einen handlicheren Ausdruck für den Konvergenzfaktor und dazu eine A-priori-Abschätzung, die in einer hinreichend kleinen Umgebung von x^* wesentlich bessere Fehlerschranken liefert als (28).

Satz 5.1.19 (Konvergenz des Newton-Verfahrens). *Ist die Funktion f aus $C^2[a, b]$, wobei $[a, b]$ eine Einschließung einer einfachen Nullstelle x^* von f bezeichnet, und genügt die Intervallänge $|b - a|$ der Abschätzung*

$$L := |b - a|\, M_2/(2m_1) < 1 \text{ mit } 0 < m_1 \leqq |f'(x)| \text{ und } |f''(x)| \leqq M_2,\ x \in [a, b], \tag{49}$$

so konvergiert die durch (43), (40) *erklärte Folge von Näherungswerten x_n für jeden Startwert x_0 aus $[a, b]$ mit der Konvergenzordnung $p = 2$ und dem Konvergenzfaktor $q := M_2/(2m_1)$ gegen x^*. Für den Fehler der n-ten Näherung gilt die A-priori-Abschätzung*

$$|\delta_n| \leqq \frac{2m_1}{M_2}\left(\frac{|b - a|\, M_2}{2m_1}\right)^{2^n}. \tag{50}$$

Zum Beweis entwickeln wir $f(x)$ an der Stelle $x = x_n$ mit Hilfe der Taylor-Formel

$$f(x^*) = f\big(x_n - (x_n - x^*)\big) = f(x_n) - (x_n - x^*)\, f'(x_n) + \frac{1}{2!}(x_n - x^*)^2 f''\big(x_n + \vartheta(x_n - x^*)\big), \tag{51}$$

dividieren durch $f'(x_n)$, benutzen $f(x^*) = 0$ und $x_{n+1} = x_n - f(x_n)/f'(x_n)$, und stellen nach $\delta_{n+1} := x_{n+1} - x^*$ um. Dann ergibt sich die Ungleichung

$$|\delta_{n+1}| \leqq \delta_n^2 \frac{M_2}{2f'(x_n)} \leqq \delta_n^2 \cdot \frac{M_2}{2m_1} := \delta_n^2 \cdot q. \tag{52}$$

Konvergenzordnung und Konvergenzfaktor sind damit bestimmt. Ist nun x_0 aus $[a, b]$, so folgt mit (49)

$$|\delta_0| \leqq |b - a| = \frac{L \cdot 2m_1}{M_2} = L/q$$

und mit (52) weiter

$$|\delta_1| \leqq L^2/q, \qquad |\delta_2| \leqq L^4/q, \ldots, \qquad |\delta_n| \leqq L^{2^n}/q.$$

Die letzte Ungleichung ist identisch mit (50). Da $L < 1$ vorausgesetzt wurde, beweist sie die Konvergenz. *A-posteriori-Fehlerabschätzungen* findet man in Ü 5.1.11 und Ü 5.1.12. ∎

Beispiel 5.1.20. Wir benutzen das Newton-Verfahren zur Bestimmung der positiven Nullstelle der Funktion $f(x) := x^2 - c$, $c > 0$. Es ist $f'(x) = 2x$, so daß sich mit $g(x) := x - \frac{f(x)}{f'(x)} = \frac{x}{2} + \frac{c}{2x}$ die in Beispiel 5.1.1 heuristisch gefundene Iteration (3) ergibt. Für $c := 2$ und $[a, b] = [1.4, 1.6]$ ist $m_1 := \inf |f'(x)| = 2.8$ und $M_2 := \sup |f''(x)| = 2$, also $L := 0.2/2.8 \approx 0.07143$, und mit (50) weiter $|\delta_n| < 2.8 \cdot (0.07143)^{(2^n)}$. Wählt man also einen beliebigen Startwert x_0 aus dem Intervall [1.4, 1.6], so gelten für die Fehler der Näherungen x_1, x_2, x_3 (bei rundungsfehlerfreier Rechnung!) die Abschätzungen $|\delta_1| < 0.0143$, $|\delta_2| < 7.29 \cdot 10^{-5}$, $|\delta_3| < 1.90 \cdot 10^{-9}$. Daran ist die quadratische Konvergenz des Newton-Verfahrens gut zu erkennen. Für die zum Startwert $x_0 := 1.5$ in Beispiel 5.1.1 berechneten Näherungen x_1, x_2 und x_3 liefert die A-posteriori-Abschätzung aus Ü 5.1.11 die Fehlereinschließungen $2.17 \cdot 10^{-3} \leqq |\delta_1| \leqq 2.48 \cdot 10^{-3}$, $1.89 \cdot 10^{-6} \leqq |\delta_2| \leqq 2.16 \cdot 10^{-6}$, $3.31 \cdot 10^{-8} \leqq |\delta_3| \leqq 3.79 \cdot 10^{-8}$.

Die *Regula falsi (1. Form)* konvergiert in jeder hinreichend kleinen Umgebung einer einfachen Nullstelle von $f, f \in \mathbf{C}^1[a, b]$, wenn f' einer Abschätzung (46) genügt. Ihre Konvergenzordnung ist $p = 1$ (vgl. Ü 5.1.10).

Die *Regula falsi (2. Form)* läßt sich nicht in der Form (15) schreiben, da die neue Näherung x_{n+1} nicht nur von x_n, sondern auch noch von x_{n-1} abhängt. Mit (42) wird (43) zu

$$x_{n+1} = g(x_n, x_{n-1}) := x_n - f(x_n)/f[x_n, x_{n-1}]. \tag{53}$$

Das ist ein Beispiel für ein *zweistufiges Iterationsverfahren* (B 5.1.2). Der Kontraktionssatz kann darauf nicht unmittelbar angewendet werden, so daß eine gesonderte Konvergenzuntersuchung notwendig wird.

Satz 5.1.21 (Konvergenz der Regula falsi, 2. Form). *Unter den Voraussetzungen von Satz* 5.1.19 *konvergiert die durch* (53) *erklärte Folge von Näherungswerten für jedes Paar von (nicht zusammenfallenden) Startwerten x_0, x_1 aus $[a, b]$ entweder in endlich vielen Schritten oder mit der Konvergenzordnung $p := \big(1 + \sqrt{5}\big)/2 \approx 1.618$ und dem Konvergenz-*

faktor $q := \left(M_2/(2m_1)\right)^{1/p}$ *gegen* x^*. *Für den Fehler der n-ten Näherung gilt die A-priori-Abschätzung*

$$|\delta_n| \leqq \frac{2m_1}{M_2}\left(\frac{|b-a|\,M_2}{2m_1}\right)^{k_{n+1}}, \tag{54}$$

wobei k_{n+1} *die* $(n+1)$*-te der durch*

$$k_0 := 0, \qquad k_1 := 1, \qquad k_{n+1} := k_n + k_{n-1}, \qquad n = 1, 2, \ldots, \tag{55}$$

rekursiv definierten und durch

$$k_n := \frac{1}{\sqrt{5}}\,(t_1^n - t_2^n), \qquad t_{1,2} := \frac{1}{2}\left(1 \pm \sqrt{5}\right) \approx \begin{cases} 1.618 \\ -0.618 \end{cases} \tag{56}$$

explizit darstellbaren Fibonacci-Zahlen bezeichnet.

Beweis. Wenn man in (53) rechts und links x^* und auf dem Bruchstrich zusätzlich $f(x^*) = 0$ subtrahiert, so ergibt sich mit $\delta_n := x_n - x^*$

$$\begin{aligned}
\delta_{n+1} &= \delta_n\left(1 - \frac{1}{f[x_{n-1}, x_n]} \cdot \frac{f(x_n) - f(x^*)}{x_n - x^*}\right) \\
&= \delta_n\delta_{n-1}\left(\frac{f[x_{n-1}, x_n] - f[x^*, x_n]}{x_{n-1} - x^*}\right) \cdot \frac{1}{f[x_{n-1}, x_n]} \\
&= \delta_n\delta_{n-1} \cdot f[x^*, x_{n-1}, x_n]/f[x_{n-1}, x_n].
\end{aligned} \tag{57}$$

Dabei haben wir den in der vorletzten Zeile auftretenden Differenzenquotienten zweier Differenzenquotienten (Steigungen) als *Steigung zweiter Ordnung* mit $f[x^*, x_{n-1}, x_n]$ abgekürzt (vgl. (6.1.23)). Für $n = 1$ ist die Umformung offenbar zulässig, wenn die Startwerte x_0 und x_1 voneinander und von x^* verschieden sind, denn wegen (46) sind dann auch die zugehörigen Funktionswerte paarweise voneinander verschieden. Das bleibt auch für $n > 1$ richtig, wenn wir den trivialen Fall ausschließen, daß (53) nach n Schritten den exakten Wert $x_{n+1} = x^*$ liefert. Nach dem Mittelwertsatz der Differentialrechnung gibt es zu jeder Steigung erster Ordnung einer Funktion aus $\mathbf{C}^1[a, b]$ einen Zwischenwert ξ_1 mit $f[x_{n-1}, x_n] = \frac{1}{1!}\,f'(\xi_1)$. Entsprechendes gilt für hinreichend glatte Funktionen auch für die Steigungen höherer Ordnung (vgl. (6.1.51)). Wir verwenden hier $f[x^*, x_{n-1}, x_n] = \frac{1}{2!}\,f''(\xi_2)$ und erhalten damit für die Steigungen in (57) die Abschätzungen

$$m_1 \leqq |f[x_{n-1}, x_n]|, \qquad |f[x^*, x_{n-1}, x_n]| \leqq M_2/2 \tag{58}$$

und weiter

$$|\delta_{n+1}| \leqq |\delta_n|\,|\delta_{n-1}| \cdot \frac{M_2}{2m_1} =: |\delta_n|\,|\delta_{n-1}| \cdot c. \tag{59}$$

Für beliebige Startwerte x_0, x_1 aus $[a, b]$ gilt nun wegen (49)

$$|\delta_0| := |x_0 - x^*| \leqq |b - a| = L/c = L^{k_1}/c, \qquad |\delta_1| \leqq L/c = L^{k_2}/c$$

und mit (59) weiter

$$|\delta_2| \leqq \frac{1}{c} L^{k_1+k_2} = \frac{1}{c} L^{k_2}, \ldots, |\delta_n| \leqq \frac{1}{c} L^{k_{n-1}+k_n} = \frac{1}{c} L^{k_{n+1}}.$$

Damit ist die Ungleichung (54) und wegen $L < 1$ auch die Konvergenz der Folge (x_n) bewiesen. Die Lösung der *homogenen linearen Differenzengleichung* $k_{n+1} = k_n + k_{n-1}$ (vgl. (55)) bestimmt man mit Hilfe eines *Potenzansatzes* $k_n := t^n$. Er führt auf $t^{n+1} = t^n + t^{n-1}$ und damit auf die *charakteristische Gleichung* $t^2 - t - 1 = 0$. Sie hat die beiden in (56) angegebenen Lösungen t_1 und t_2, so daß die allgemeine Lösung von (55) die Gestalt $k_n = c_1 t_1^n + c_2 t_2^n$ erhält. Für die noch unbekannten Koeffizienten c_1 und c_2 ergibt sich aus den Anfangsbedingungen $k_0 = 0$ und $k_1 = 1$ schließlich $c_1 = -c_2 = 1/\sqrt{5}$, womit auch (56) bewiesen ist. Zur Abschätzung der *Konvergenzordnung* schreiben wir die Konstante L aus (49) in der Form β^{t_1}, wo $t_1 := \left(1 + \sqrt{5}\right)/2$ die größere der beiden Wurzeln von $t^2 - t - 1 = 0$ bezeichnet. Da $L < 1$ vorausgesetzt wurde, gilt $L = \beta^{t_1} < \beta = L^{1/t_1} < 1$. Aus den bereits früher benutzten Ungleichungen $|\delta_0| \leqq L/c$ und $|\delta_1| \leqq L/c$ ergibt sich damit $|\delta_0| \leqq \frac{1}{c} \beta^{t_1^0} =: \Delta_0$, $|\delta_1| \leqq \frac{1}{c} \beta^{t_1^1} =: \Delta_1$. Nehmen wir im Sinne einer vollständigen Induktion an, daß

$$|\delta_n| \leqq \frac{1}{c} \beta^{t_1^n} =: \Delta_n \tag{60}$$

gilt, so folgt aus (59) wegen $t_1^2 = t_1 + 1$ sofort $|\delta_{n+1}| \leqq \frac{1}{c} \beta^{t_1^n + t_1^{n-1}} = \frac{1}{c} \beta^{t_1^{n+1}} =: \Delta_{n+1}$, womit (60) für jedes natürliche n bewiesen ist. Für die Folge der Betragsschranken Δ_n der Fehler δ_n ergeben sich wegen $\Delta_{n+1} = c^{t_1 - 1} \Delta_n^{t_1} = c^{1/t_1} \Delta_n^{t_1}$ schließlich wie behauptet die Konvergenzordnung $p := t_1$ und der Konvergenzfaktor $q := c^{1/p}$ (vgl. B 5.1.6). ∎

Die Regula falsi (53) konvergiert also ebenfalls *überlinear*. Obwohl ihre Konvergenzordnung etwas niedriger ist als die des Newton-Verfahrens, ist sie bezüglich des *Rechenaufwands* vorteilhafter. Bezeichnet nämlich n die Anzahl der Newton-Schritte und r die Anzahl der Regula-falsi-Schritte, die zum Erreichen ein und derselben Genauigkeitsschranke erforderlich sind, so gilt wegen (50) und (54) $2^n \approx k_{r+1} \approx \frac{1}{\sqrt{5}} t_1^{r+1}$, also $r + 1 \approx 1.44n$. Im Vergleich zum Newton-Verfahren benötigt die Regula falsi etwa die $1^1/_2$-fache Schrittanzahl. Da aber beim Newton-Verfahren je Schritt zwei Werte ($f(x_n)$ und $f'(x_n)$), bei der Regula falsi dagegen jeweils nur ein neuer Funktionswert bestimmt werden muß ($f(x_{n-1})$ ist bereits vom vorhergehenden Schritt bekannt), beträgt der Gesamtaufwand der Regula falsi (2. Form) im Vergleich zum Newton-Verfahren nur etwa 75% (vgl. B 5.1.3).

5.1.5. Bemerkungen

B 5.1.1. Ausführlicher wird die Lösung nichtlinearer Gleichungen und Gleichungssysteme in einer Reihe von Numerik-Lehrbüchern (vgl. Bachvalov [1], Beresin, Shidkow [1], Band 2, Collatz [1], Henrici [1], Isaacson, Keller [1], Ralston [1], Schmeisser, Schirmeier [1], Stoer [1], Stummel [1], Werner [1], Zurmühl [1]) und in der Spezialliteratur behandelt. Wir verweisen auf Schwetlick [1] sowie auf Dennis, Schnabel [1], Householder [1], Ortega, Rheinboldt [1], Ostrowski [1] und Traub [1].

B 5.1.2. Allgemein kann man ein *nichtlineares Iterationsverfahren* durch eine Folge von Abbildungen g_n beschreiben (vgl. Ortega, Rheinboldt [1], S. 236, Schwetlick [1], S. 73, Young [1],

S. 64)

$$\begin{aligned} x_1 &:= g_1(x_0), & g_1 &: \mathbb{D}_1 \to \mathbb{R}^N,\ N \geqq 1, \\ x_2 &:= g_2(x_0, x_1), & g_2 &: \mathbb{D}_2 \to \mathbb{R}^N, \\ &\vdots & &\vdots \\ x_n &:= g_n(x_0, x_1, \ldots, x_{n-1}), & g_n &: \mathbb{D}_n \to \mathbb{R}^N. \\ &\vdots & &\vdots \end{aligned} \tag{61}$$

Dabei müssen die Definitionsbereiche $\mathbb{D}_n$ solche Teilmengen des $(\mathbb{R}^N)^n$ sein, daß durch (61) zu jedem Startwert x_0 aus $\mathbb{D}_1$ in eindeutiger Weise eine Folge (x_n) definiert ist. Hängen die g_n nicht von allen vorangegangenen, sondern jeweils nur von einer (in der Regel der aktuellsten) Näherung ab, $x_n := g_n(x_{n-1})$, so bezeichnet man (61) als *einstufiges Iterationsverfahren* (*Einschrittverfahren*), anderenfalls als *mehrstufiges* (*Mehrschritt-*) *Verfahren*. Werden, abgesehen von einer Anlaufrechnung, jeweils die m aktuellsten Näherungen benutzt,

$$x_n := g_n(x_{n-m}, x_{n-m+1}, \ldots, x_{n-1}), \qquad n = m, m+1, \ldots, \tag{62}$$

so spricht man von einem *sequentiellen m-Stufen-* (*m-Schritt-*) *Verfahren*. Sind die g_n unabhängig von der Schrittnummer, so nennt man (62) *stationär*. Das vereinfachte und das klassische Newton-Verfahren sind beispielsweise stationär und einstufig, die Regula falsi (2. Form) ist stationär, sequentiell und zweistufig (für lineare Iterationsverfahren vgl. Abschnitt 2.7.2.).

B 5.1.3. Als Maß für den *Rechenaufwand* nichtlinearer Iterationsverfahren wird meist die Anzahl der *Funktionsberechnungen* (Unterprogrammaufrufe) benutzt, da diese in der Regel wesentlich mehr Rechenzeit benötigen als die (zumindest im eindimensionalen Fall) wenigen arithmetischen Operationen des Verfahrens. Als *Maßeinheit* verwenden einige Autoren die von OSTROWSKI vorgeschlagene Einheit *Horner* (1 Horner = 1 Funktionsberechnung). Um die Effektivität verschiedener Verfahren vergleichen zu können, führt man den Begriff des *Wirkungsgrads* (*Informationswirkungsgrad*) eines Verfahrens ein, das ist der Quotient aus der *Konvergenzordnung* p des Verfahrens und der Anzahl der Funktionsberechnungen je Schritt. Das Newton-Verfahren hat also den Wirkungsgrad 1, die Regula falsi (2. Form) dagegen 1.618. Für *überlinear konvergente Iterationsverfahren* definiert man die *asymptotische Effektivität* durch

$$\mathrm{EFF} := \frac{1}{\mathrm{RAWS}} \log p \tag{63}$$

und den *Effektivitätsindex* η durch

$$\eta := \exp(\mathrm{EFF}) = p^{1/\mathrm{RAWS}} \tag{64}$$

(SCHWETLICK [1], S. 89, vgl. auch TRAUB [1], S. 260ff.). Dabei bezeichnet RAWS den Rechenaufwand je Schritt. Mit Hilfe der asymptotischen Effektivität kann man den zum Unterschreiten einer Genauigkeitsschranke ε erforderlichen Gesamtaufwand RAW (ε) zweier Iterationsverfahren vergleichen: Für hinreichend kleines ε gilt

$$\mathrm{RAW}_1(\varepsilon)/\mathrm{RAW}_2(\varepsilon) \approx \mathrm{EFF}_2/\mathrm{EFF}_1. \tag{65}$$

Für diese Formel ist es belanglos, welche Maßeinheit für die Kosten je Schritt und welcher Logarithmus in (63) verwendet wird. Bezeichnen zum Beispiel $\mathrm{EFF}_1 := \ln\left(\frac{1+\sqrt{5}}{2}\right)$ und $\mathrm{EFF}_2 = \frac{1}{2}\ln 2$ die asymptotischen Effektivitäten der Regula falsi (2. Form) und des Newton-Verfahrens, so ergibt sich $\mathrm{RAW}_1(\varepsilon)/\mathrm{RAW}_2(\varepsilon) \approx 0.72$. Die Regula falsi benötigt also (für hinreichend kleines ε) nur 72% des Aufwandes des Newton-Verfahrens.

B 5.1.4. Bei zahlreichen *numerischen Tests* erwies sich der auf VAN WIJNGAARDEN, DEKKER und BRENT zurückgehende Algorithmus ZEROIN, eine Kombination von Regula falsi, Bisektion und inverser quadratischer Interpolation, als besonders wirtschaftlich (vgl. FORSYTHE, MALCOLM, MOLER [1], S. 177). *Testbeispiele für Nullstellenalgorithmen* findet man u. a. bei RICE [2] und bei BUS, DEKKER [1]. Eine Auswahl davon haben wir in den Tabellen 5.1.1 und 5.1.2 zusammengestellt (vgl. auch Ü 5.1.1 und für Polynomnullstellen B 5.2.9).

B 5.1.5. Ist x^* eine *mehrfache Nullstelle* und bezeichnet s ihre Vielfachheit ($s > 1$), verschwinden also an der Stelle $x = x^*$ nicht nur der Funktionswert, sondern auch alle Ableitungen bis zur Ordnung $s - 1$,

$$f(x^*) = f'(x^*) = \cdots = f^{(s-1)}(x^*) = 0, \qquad f^{(s)}(x^*) \neq 0, \tag{66}$$

Tab. 5.1.1. Testbeispiele für Nullstellenverfahren (RICE [2])

Nr.	Intervall	Funktion	Nullstelle
1	[0, 1]	$(x^2+1)\sin x-(x-1)(x^2-5)\,e^{\sqrt{x}}$	0.874511
2	[−2.5, 0.5]	$(x+1)/(x^2+2)$	−1
3	[−0.1, 0.3]	$\sin x - x/2$	0
4	[−2, 0.75]	$x\cdot e^x$	0
5	[0, 1]	$x - e^{-x}$	0.567143
6	[0, 3]	$\tan x - \cos x - 1/2$	2.74270
7	[−2, 0]	$\cos x - x\cdot e^x$	−1.86400
8	[0.01, 0.7]	$\tan x - 1.01x$	0.172175
9	[0.5, 1.5]	$\tan x - 2x$	1.16556
10	[3, 4]	$x(x-3) - 4(\sin x)^2$	3.01961
11	[0, 1]	$x\cdot e^{-1/x^2}$	0

Tab. 5.1.2. Testbeispiele für Nullstellenverfahren (BUS, DEKKER [1])

Nr.	Intervall	Funktion	Parameter	
12	[0, 1]	$x^n + (x-1)\,e^{-nx}$	$n = 1(1)10$	
13	[0, 1]	$1 + 2x\,e^{-n} - 2e^{-nx}$	$n = 1(1)10$	
14	[0, 1]	$(1+(1-n)^k)\,x - (1-nx)^k$	$n = 1(2)11,$	$k = 1(1)5$
15	[0, 1]	$x^n - (1-x)^k$	$n = 1(2)19,$	$k = 0(1)10$
16	[−0.75, 0.5]	$x^n + 10^{-k}$	$n = 3(2)19,$	$k = 0(1)10$
17	[−0.75, 0.5]	$x^n + x + 10^{-k}$	$n = 3(2)19,$	$k = 0(1)10$

so gilt für f statt (37) die Taylor-Formel

$$f(x) = h^s \cdot f^{(s)}(x^* + \vartheta_0 \cdot h)/s!, \quad h := x - x^*, \quad 0 < \vartheta_0 < 1, \tag{67}$$

und für die Ableitungen f' und f'' analog

$$\begin{aligned} f'(x) &= h^{s-1} \cdot f^{(s)}(x^* + \vartheta_1 \cdot h)/(s-1)!, \qquad 0 < \vartheta_1 < 1, \\ f''(x) &= h^{s-2} \cdot f^{(s)}(x^* + \vartheta_2 \cdot h)/(s-2)!, \qquad 0 < \vartheta_2 < 1. \end{aligned} \tag{68}$$

Daraus ergibt sich für die Newtonsche Iterationsfunktion $g(x) := x - f(x)/f'(x)$ die Ableitung

$$g'(x) = \frac{f(x)\,f''(x)}{f'(x)^2} = \frac{s-1}{s} \cdot \frac{f^{(s)}(x^* + \vartheta_0 \cdot h)\, f^{(s)}(x^* + \vartheta_2 \cdot h)}{f^{(s)}(x^* + \vartheta_1 \cdot h)^2}. \tag{69}$$

Also ist $g'(x^*) = (s-1)/s \neq 0$, und nach Satz 5.1.14 besitzt das *Newton-Verfahren im Fall mehrfacher Nullstellen* nur noch die Konvergenzordnung 1. Multipliziert man aber die Newton-Korrektur mit einem geeigneten Relaxationsfaktor ω, so kann man (theoretisch) die Konvergenzordnung 2 wiederherstellen. Die Iterationsfunktion und ihre erste Ableitung haben dann nämlich die Gestalt

$$g(x) := x - \omega \frac{f(x)}{f'(x)}, \quad g'(x) = 1 - \omega + \omega \frac{s-1}{s} \cdot \frac{f^{(s)}(x^* + \vartheta_0 \cdot h)\, f^{(s)}(x^* + \vartheta_2 \cdot h)}{f^{(s)}(x^* + \vartheta_1 \cdot h)^2}, \tag{70}$$

so daß sich für $\omega := s$ die Ableitung $g'(x^*) = 0$ ergibt: Das *Newton-Verfahren mit Relaxation* konvergiert für hinreichend gute Startwerte auch im Fall mehrfacher Nullstellen quadratisch, wenn man den Relaxationsfaktor ω gleich der Vielfachheit s der Nullstelle x^* wählt. Praktisch nutzt diese Erkenntnis wenig, da im allgemeinen vor Beginn der Rechnung kaum Informationen über die Vielfachheit der Nullstelle vorliegen. Man kann sie sich aber während der Rechnung (zumindest näherungsweise) verschaffen. Dazu führen wir die Funktion $F(x) := f(x)/f'(x)$ ein. Sie besitzt auch im Fall mehrfacher Nullstellen von f bei $x = x^*$ nur eine einfache Nullstelle, denn wegen (69) ergibt sich $F'(x^*) = \lim\limits_{x \to x^*} \left(1 - \dfrac{f(x)\,f''(x)}{f'(x)^2}\right) = 1/s \neq 0.$

Das auf $F(x)$ angewandte *Newton-Verfahren konvergiert* demzufolge *quadratisch*. Die überlineare Konvergenz wird allerdings mit dem relativ hohen *Rechenaufwand von 3 Funktionsauswertungen* je Schritt erkauft, denn die Iterationsfunktion g hat die Gestalt

$$g(x) := x - \frac{F(x)}{F'(x)} = x - \frac{f(x)}{f'(x)} \Big/ \left(1 - \frac{f(x)\, f''(x)}{f'(x)^2}\right).$$

Bezogen auf die Ausgangsfunktion f handelt es sich also um ein *Newton-Verfahren mit schrittabhängigem Relaxationsfaktor*:

$$x_{n+1} = x_n - \omega_n f(x_n)/f'(x_n), \qquad \omega_n := 1/\left(1 - f(x_n)\, f''(x_n)\, f'(x_n)^{-2}\right). \tag{71}$$

Für $x_n \to x^*$ strebt ω_n gegen die Vielfachheit s der Nullstelle. Der Effektivitätsindex des Verfahrens ist $\eta := 2^{1/3} \approx 1.26$ (vgl. Ü 5.1.17).

B 5.1.6. Im Beweis von Satz 5.1.21 wurde lediglich nachgewiesen, daß die Majorantenfolge (Δ_n) der Fehlerfolge (δ_n) die *Konvergenzordnung* (*Q-Ordnung*) p besitzt. Die Folge (δ_n) selbst hat die *R-Ordnung* p (von engl. *root* — Wurzel) (vgl. SCHWETLICK [1], S. 85, und die Arbeiten von SCHMIDT [2] und BURMEISTER, SCHMIDT [1]).

B 5.1.7. *Verfahren von höherer Konvergenzordnung* erhält man, wenn man die Funktion f in der Umgebung der aktuellen Näherung $(x_n, f(x_n))$ nicht durch eine Gerade (vgl. (38)), sondern durch ein Polynom höheren Grades, z. B. durch eine Parabel

$$y = \tilde{f}(x) := b + c \cdot (x - x_n) + \frac{d}{2} \cdot (x - x_n)^2, \quad b := f(x_n), \tag{72}$$

ersetzt und eine Nullstelle dieser Funktion als neue Näherung verwendet (vgl. auch BERG [3] und SZABO [1]). Man fordert also $\tilde{f}(x_{n+1}) = 0$ und erhält damit aus (72) für die Korrektur $\delta := x_{n+1} - x_n$ die beiden Werte $\delta_{1,2} := \left(-c \pm \sqrt{c^2 - 2bd}\right)/d$. Meist wählt man die betragsmäßig kleinere Wurzel (vgl. Alg. 1.2.4)

$$\delta_1 := -2b/\left(c + (\operatorname{sgn} c)\sqrt{c^2 - 2bd}\right) \tag{73}$$

und rechnet den nächsten Schritt mit $x_{n+1} := x_n + \delta_1$. Man kann aber auch beide δ_i bestimmen und dasjenige auswählen, für das $|f(x_n + \delta_i)|$ den kleineren Wert annimmt. Faßt man (72) als *Taylor-Entwicklung* von f an der Stelle $x = x_n$ auf, die nach dem quadratischen Term abgebrochen wurde, so ergibt sich mit $x := x_{n+1}$ das *Verfahren von* MULLER [1]

$$\begin{aligned} &b := f(x_n), \quad c := f'(x_n), \quad d := f''(x_n), \\ &x_{n+1} := x_n - 2 \cdot b/\left(c + (\operatorname{sgn} c)\sqrt{c^2 - 2 \cdot b \cdot d}\right). \end{aligned} \tag{74}$$

Will man den Wurzelausdruck vermeiden, so kann man die nichtlineare Gleichung $\tilde{f}(x_n + \delta) = 0$ auch iterativ lösen: Man startet mit der *Newton-Korrektur* $\delta_0 := -b/c$ und verbessert sie anschließend mit Hilfe der aus (72) hergeleiteten Vorschrift

$$\delta_{k+1} := -\left(b + \frac{d}{2}\,\delta_k^2\right)\Big/ c. \tag{75}$$

Begnügt man sich mit einem Iterationsschritt, so erhält man das *Verfahren von* EHRMANN [1]

$$x_{n+1} := x_n - \frac{b}{c}\left(1 + \frac{b \cdot d}{2 \cdot c^2}\right), \tag{76}$$

das wie (71) als *Newton-Verfahren mit einem schrittabhängigen Relaxationsfaktor* aufgefaßt werden kann. Im Fall einfacher Nullstellen und hinreichend guter Startwerte *konvergieren* die Verfahren von MULLER und EHRMANN *kubisch* (Ü 5.1.18). Sie kosten einen *Rechenaufwand* von 3 Funktionsberechnungen je Schritt, ihr *Effektivitätsindex* ist demnach $\eta := 3^{1/3} \approx 1.44$. Er liegt also unter dem der *Regula falsi* $\left(\eta := \left(1 + \sqrt{5}\right)/2 \approx 1.62\right)$ und nur unwesentlich über dem des *Newton-Verfahrens* $\left(\eta := \sqrt{2}\right)$. Dies und die relativ hohe *Startwertempfindlichkeit* sind Gründe dafür, daß *Verfahren höherer Ordnung* für die Praxis kaum Bedeutung erlangt haben. In der Regel werden *hybride Methoden* verwendet, die zur Verkleinerung einer Einschließung eine Kombination von Regula falsi, (inverser) quadratischer Interpolation und Bisektion benutzen und jeweils für eine feste Anzahl von Schritten eine (überlineare) Abnahme der Intervallänge garantieren (vgl. SHRAGER [1] und den auf VAN WIJNGAARDEN, DEKKER und BRENT zurückgehenden, in FORSYTHE, MALCOLM, MOLER [1], S. 177 ff. beschriebenen Algorithmus ZEROIN).