Mathematisches Seminar

Numerik

Leitung: Andreas Müller

Benjamin Bouhafs-Keller, Daniel Bucher, Manuel Cattaneo
Patrick Elsener, Reto Fritsche, Nicolò Galliani, Tobias Grab
Thomas Kistler, Kistler Thomas, Fabio Marti, Joël Rechsteiner
Cédric Renda, Michael Schmid, Mike Schmid,
Michael Schneeberger, Martin Stypinski, Manuel Tischhauser
Nicolas Tobler, Raphael Unterer, Severin Weiss

Hochschule für Technik, Rapperswil, 2020

Inhaltsverzeichnis

Vorwort

Dieses Buch entstand im Rahmen des Mathematischen Seminars im Frühjahrssemester 2020 an der Hochschule für Technik Rapperswil. Die Teilnehmer, Studierende der Abteilungen für Elektrotechnik, Informatik, Bauingenieurwesen und Erneuerbare Energie und Umwelttechnik der HSR, erarbeiteten nach einer Einführung in das Themengebiet jeweils einzelne Aspekte des Gebietes in Form einer Seminararbeit, über deren Resultate sie auch in einem Vortrag informierten.

Im Frühjahr 2020 war das Thema des Seminars die Numerik, also die Berechnung von mathematischen Resultaten mit Hilfe von Computern.

In einigen Arbeiten wurde auch Code zur Demonstration der besprochenen Methoden und Resultate geschrieben, soweit möglich und sinnvoll wurde dieser Code im Github-Repository dieses Kurses[1] [3] abgelegt. Im genannten Repository findet sich auch der Source-Code dieses Skriptes, es wird hier unter einer Creative Commons Lizenz zur Verfügung gestellt.

[1] `https://github.com/AndreasFMueller/SeminarNumerik.git`

Teil I

Grundlagen

Einleitung

Kapitel 1

Berechnung

Die numerischen Datentypen eines digitalen Computers sind Approximationen für die abstrakten Zahlensysteme $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$. Man kann sie daher verwenden, die in der Analysis und der linaren Algebra definierten Konzepte wie Grenzwerte, Integrale oder inverse Matrizen zu berechnen. Da sie jedoch nur Annäherungen sind, werden die gewohnten Rechenregeln nicht immer gültig sein. In $\mathbb{R}$ kann die Addition dreier Zahlen zum Beispiel in beliebiger Reihenfolge durchgeführt werden, für `double`-Zahlen auf einem Computer ist dies nicht der Fall. Zum Beispiel kann man in Octave die folgende Rechnung durchführen[1]:

```
octave:1> a = 1;
octave:2> b = 0.0000000000000001;
octave:3> x = a+(b+b)
x =  1.0000
octave:4> y = (a+b)+b
y =  1
octave:5> x-y
ans =    2.2204e-16
octave:6> x-1
ans =    2.2204e-16
octave:7> y-1
ans = 0
```

Das Assoziativgesetz verlangt, dass $a + (b + b) = (a + b) + b$ ist und zunächst will es auch den Anschein haben, dass x und y tatsächlich gleich sind. Ungewöhnlich ist nur, dass der Wert von x mit auf den ersten Blick unnötigen Nachkommastellen angezeigt wird. Diese Nullen deuten jedoch an, dass $x \neq 1$ aber $y = 1$, man in Schritt 6 und 7 sieht, wenn man die Differenz zu 1 bestimmt.

Das Beispiel verdeutlicht, dass auf einem Computer zur Verfügung stehende numerische Datentypen die gewohnten Rechengesetze verletzen und neue Unzulänglichkeiten wie Rundungsfehler in die Berechnung injizieren. Die Numerik muss sich daher mit der Frage befassen, welcher Art diese neuen Effekte sind und wie ihnen effektiv begegnet werden kann. Nur so kann die Zuverlässigkeit numerisch gefundener Resultate garantiert werden.

[1] Das Beispiel zeigt eine Problem für Zahlen vom Typ `double`. Das Programm `assoziativ.cpp` im Verzeichnis `buch/chapters/experiments/assoziativ` findet analoge Beispiel für die anderen Floatingpoint Typen `float` und `long double`.

Dieses Kapitel befasst sich mit den Eigenschaften von Computer-Zahlensystemen. Im ersten Abschnitt werden die Zahlensysteme beschrieben und ihre Vor- und Nachteile gegeneinander abgewogen. Im zweiten Abschnitt wird gezeigt, wie Resultate bei unvorsichtiger Vorgehensweise verfälscht werden können. Rundungsfehler sind unvermeidlich, das Ziel muss daher sein, ihre Grösse unter Kontrolle zu halten. Numerische Instabilität liegt vor, wenn die Berechnung aufgrund von numerischen Effekten völlig aus dem Ruder läuft und sinnlose Resultate liefert, dies wird in Abschnitt 1.5 untersucht.

1.1 Zahlensysteme

Auf modernen Allzweck-Prozessoren steht eine ganze Reihe verschiedener numerischer Datentypen mit unterschiedlichen Eigenschaften bezüglich Geschwindigkeit und Fehlerverhalten zur Verfügung. In diesem Abschnitt sollen sie vorgestellt und miteinander verglichen werden. Es gilt, den für eine Berechnung zweckmässigsten Typen zu wählen, wobei Speicherbedarf, Laufzeit und Parallelisierbarkeit wesentliche Aspekte sind.

Microcontroller sind im Vergleich zu Allzweckprozessoren oft stark eingeschränkt. Meist sind nur Ganzahltypen mit oft sehr beschränkter Länge implementiert. Manchmal kann die Arithmetik-Einheit des Prozessors nicht einmal eine Multiplikation in Hardware ausführen, sie muss in Software nachgebildet werden. Für Floatingpoint Operationen muss oft Bibliotheken zurückgegriffen werden, die den Speicherbedarf erhöhen und langsam sind. Die Implementation von numerischen Berechnungen in eingebetteten Anwendungen ist daher mit besonderen Herausforderungen konfrontiert.

Dieselbe Schwierigkeit haben auch Allzweck-Prozessoren wenn die Genauigkeitsanforderungen die Möglichkeiten der von der Prozessor-Hardware implementierten Typen übersteigt. Dieser Fall tritt beispielsweise bei Berechnungen in der Kryptographie auf, wo oft mit Ganzzahlen mit Tausenden von Stellen gerechnet werden muss. Im Abschnitt 1.1.4 mit der GNU Multiprecision-Library ein Beispiel einer Bibliothek vorgestlelt.

1.1.1 Zahlendarstellung bezüglich verschiedener Basen

Allen Zahlensystemen gemeinsam ist die Positionsdarstellung. Eine Zahl wird als Zeichenkette $x = x_n x_{n-1} \dots x_3 x_2 x_1 x_0$ mit wobei die Zeichen x_i Ziffern mit $0 \leq x_i < b$ sind. b ist die Basis des Zahlensystems. Der Wert der Zahl x ist dann

$$x = \sum_{k=1}^{n} x_k b^k.$$

Wir hängen die Basis also Index an eine Zahlendarstellung an, um die Basis deutlich zu machen. Es ist also zum Beispiel

$$1291_{10} = 10100001011_2 = 1202211_3 = 2413_8 = 508_{16}.$$

Bruchzahlen können analog dargestellt werden. Die Zeichenkette

$$x = x_n x_{n-1} \dots x_2 x_1 x_0 \,.\, x_{-1} x_{-2} x_{-3} \dots x_{-m} \dots$$

hat den Wert

$$x = \sum_{k=-m}^{n} x_k b^k.$$

Die Zahl π hat daher die Darstellungen

$$\begin{aligned}\pi &= 11.00100100001111110110 11_2 \\ &= 10.010211012222 01_3 \\ &= 3.11037552_8 \\ &= 3.1415926_{10} \\ &= 3.243F6A_{16}\end{aligned}$$

in den Basen 2, 3, 8, 10 und 16. Eine grössere Basis erlaubt zwar eine kompaktere Darstellung, aber für die Rechnung ermöglicht die binäre Darstellung die einfachste und damit auch schnellste Implementation.

Man beachte, dass endliche Dezimalbrüche in anderen Basen durchaus nicht mehr endlich zu sein brauchen. So ist zum Beispiel

$$\begin{aligned}&\frac{1}{2}0.5_{10} = 0.1_2, \\ &\frac{1}{5} = 0.2_{10} = 0.001100110011\cdots = 0.\overline{0011}_2, \\ &\frac{1}{3} = 0.\overline{3}_{10} = 0.\overline{01}_2.\end{aligned}$$

Eine Konsequenz dieser Beobachtung ist, dass nur schon die Umwandlung einer Dezimalzahl ins Binärsystem und die Rückumwandlung in eine Dezimalzahl den Wert verändern kann. Zum Beispiel[2] bewirkt der Code

```
double x = 0.2;
std::cout << x;
```

dass der Compiler zunächst die Dezimalzahl 0.2 in eine Binärzahl verwandelt, diese Form wird im ausführbaren Code gespeichert. Zur Laufzeit des Programms muss die I/O-Bibliothek dann die gespeicherte Zahl wieder in eine Dezimalzahl verwandeln. Für $x = 0.5$ ist das unproblematisch, da diese Zahl sowohl dezimal wie auch binär ein endlicher Dezimalbruch ist. Für $x = 0.2$ tritt jedoch eine Abweichung auf, weil die im Code gespeichert Binärzahl nicht exakt in 0.2 zurückgewandelt werden kann. Stattdessen erhält man abhängig vom Datentyp die folgenden abweichenden Werte:

Typ	0.5	0.2
`float`	0.50000000000000000000	0.20000000298023223877
`double`	0.50000000000000000000	0.20000000000000001110
`long double`	0.50000000000000000000	0.20000000000000001110

Dass kein Unterschied zwischen `double` und `long double` ist nur scheinbar. Multipliziert man x mit 5 vor dem Output, wird plötzlich ein Unterschied sichtbar.

1.1.2 Festkommazahlen

Die bekanntesten Festkommazahlen sind die Ganzzahltypen, die jeder Prozessor zum Beispiel für Adressierung, Zähler und Indizierung benötigt. Damit ist auch bereits klar, dass man immer damit

[2] Dieses Beispiel wurde mit dem Programm `format.cpp` aus dem Verzeichnis `buch/chapters/experiments/limits` von [3] gerechnet.

rechnen kann, dass mindestens die Addition und die Subtraktion von ganzen Zahlen mit Wortlängen implementiert sind, die der Prozessor zum Beispiel für relative Adressierung braucht. Ebenso kann man davon ausgehen, dass jeder Kern eines Prozessors eine Einheit für ganzzahlige Operationen hat, denn er könnte sonst nicht einmal die minimal notwendigen Adressberechnungen durchführen.

Addition

Das Verfahren der "schriftlichen Addition", welches man in der Primarschule lernt, funktioniert auch für die Berechnung einer Summe in jeder beliebigen anderen Basis

Vorzeichen

Ganzzahlen mit Vorzeichen können auf verschiedene Arten binär dargestellt werden, weitgehend durchgesetzt hat sich für Festkommazahlen jedoch die Zweierkomplement-Darstellung[3]. In ihr werden 8-bit Zeichenketten wie folgt als Zahlen interpretiert:

$$
\begin{aligned}
\texttt{01111111} &= 127\\
\texttt{01111110} &= 126\\
&\vdots \qquad\quad \vdots\\
\texttt{00000010} &= 2\\
\texttt{00000001} &= 1\\
\texttt{00000000} &= 0\\
\texttt{11111111} &= -1\\
\texttt{11111110} &= -2\\
\texttt{11111101} &= -3\\
&\vdots \qquad\quad \vdots\\
\texttt{10000010} &= -126\\
\texttt{10000001} &= -127\\
\texttt{10000000} &= -128
\end{aligned}
$$

Diese Codierung ist in Hardware besonders leicht implementiertbar. Ein Zähler für eine Vorzeichenlose Ganzzahl von 8 bit Länge, initialisiert mit `10000000` wird beim Hochzählen genau die Zahle von -128 bis 127 aufzählen.

Die entgegengesetzte einer Zahl kann nach der folgenden Regel gefunden werden:

1. Man nehme das Komplement jedes einzelnen Bits einer Zahl
2. Addiere 1.

Beispiel. Die Zahl -1291 soll als 16-bit Ganzzahl in Zweier-Komplement-Darstellung geschrieben werden. Zunächst wird die Binärdarstellung benötigt: $1291_{10} = \texttt{0000010100001011}_2$.

1. Bits komplementieren: `1111101011110100`

[3] Der Exponent einer Gleitkommanzahl ist zwar auch eine Ganzzahl, er wird aber gemäss Standard IEEE 754 nach einem anderen Verfahren codiert, siehe dazu auch Abschnitt 1.1.3

2. 1 addieren: `1111101011110101` ○

Die Addition vorzeichenbehafteter Ganzzahlen funktioniert für die Zweierkomplementdarstellung nach dem bekannten Algorithmus für die Addition. Die Differenz $111-88$ kann man als Summe $111+(-88)$ schreiben. Als 8-bit Binärzahlen sind die beiden Operanden `01101111` und `10101000`. Ihre Summe ist

$$\begin{array}{r} \texttt{01101111} \\ \texttt{10101000} \\ \hline \texttt{00010111} \end{array}$$

Dabei ist zwar ein Überlauf aufgetreten, aber dieser kann ignoriert werden. Tatsächlich ist $23_{10} = 10111_2$. Der grosse Vorteil dieser Vorzeichenkonvention ist also, dass für die Addition vorzeichenbehafteter Ganzzahlen in Zweierkomplement-Darstellung die gleiche vorhandene Hardware verwendet werden kann wie für die Addition von vorzeichenloser Ganzzahlen.

Multiplikation und Division

Man kann allerdings nicht davon ausgehen, dass ein Prozessor auch die Multiplikation von ganzen Zahlen und erst recht die Division von ganzen Zahlen implementiert. In den meisten Fällen benötigt der Prozessor nur die Multiplikation mit kleinen Zweierpotenzen, die sich viel effizienter als Verschiebeoperationen durchführen lassen.

Nachkommateil

Bisher wurden ausschliesslich Ganzzahlen betrachtet. Man kann diese Ganzzahlen aber auch als rationale Zahlen mit einem Nachkommateil fester Länge betrachten. Man könnte sich zum Beispiel nach den ersten 8 bit einer 16-bit Zahl ein Komma denken und die nachfolgenden Bits als Bruchteil betrachten. Die Bitfolge `0000000110010000` muss dann als

$$00000011.00100000_2 = 3.125_{10}$$

interpretiert werden. An den Algorithmen für Addition und Subtraktion ändert sich nichts, es ist daher keine neue Hardware für die Implementation dieser Operationen notwendig, die Ganzzahloperationen reichen aus.

Etwas komplizierter ist die Sache bei der Multiplikation. Nehmen wir an, dass wir mit 8-bit Festkomma-Zahlen arbeiten mit einem Nachkommateil von 4 bits. Wir möchten das Produkt $3.125 \cdot 2.0625$ berechnen. Die Binärdarstellungen dieser Zahlen sind $3.125_{10} = 11.001_2$ und $2.0625_{10} = 10.0001_2$. Das Produkt der 8-bit Ganzzahlen `00110010` und `00100001` wird mehr Platz beanspruchen, im schlimmsten Fall 16 bit:

$$\texttt{00110010}_2 \cdot \texttt{00100001}_2 = \texttt{0000011001110010}_2.$$

In den Faktoren sind die letzten 4 Stellen jeweils als Nachkommateil zu unterpretieren, also sind im Produkt die letzten 8 Stellen als Nachkommateil zu interpretieren. Das Produkt, wieder als 8-bit Festkommazahl geschrieben ist daher

$$\texttt{0011.0010}_2 \cdot \texttt{0010.0001}_2 = \texttt{00000110.01110010}_2 \simeq \texttt{0110.0111}_2$$

Auch für die Multiplikation ist keine neue Hardware erforderlich, doch muss das Resultat entsprechend mit Schiebeoperationen wieder so formattiert werden, dass das Komma an der "richtigen" Stelle landet.

Mit den Ganzzahl-Operationen einer CPU lassen sich also auch sehr schnelle Festkomma-Operationen realisieren.

Vor- und Nachteile

⊕ Der absoluter Fehler ist konstant.

⊕ Operationen sind typischerweise deutlich schneller als mit Gleitkommazahlen vergleichbarer Grösse. Dies gilt selbst dann, wenn die Operationen zum Teil in Software realisiert werden müssen.

⊕ Die Addition und Subtraktion sind von derart elementarer Bedeutung für einen CPU-Kern, dass jeder Kern mindestens eine Einheit für Ganzzahl-Operationen hat. In einer Multicore-CPU kann man daher davon ausgehen, dass Festkomma-Operationen sich auf verschiedenen Cores nicht gegenseitig behindern.

⊖ Kleine Zahlen können nur mit wenigen signifikanten Stellen dargestellt werden.

⊖ Schon für mässig grosse Zahlen ist Überlauf möglich.

1.1.3 Gleitkommazahlen

Gleitkommazahlen erweitern den Bereich der darstellbaren Zahlen dadurch, dass sie zu einer Festkommazahl einen Exponentialfaktor hinzunehmen. Eine Gleitkommazahl x ist also von der Form

$$x = m \cdot b^k.$$

m heisst Mantisse, b ist die Basis und k ist der Exponent, üblicherweise eine kleine Ganzzahl. Die Mantisse wird typischerweise so strukturiert, dass genau eine Stelle vor dem Komma steht. Im Dezimalsystem sind also

$$1291 = 1.291 \cdot 10^3, \quad \gamma = 5.772156649 \cdot 10^{-1}, \quad N_A = 6.0221476 \cdot 10^{23}$$

korrekte Gleitkommazahlen.

Die meisten heutigen Prozessoren rechnen ausschliesslich binär. Sowohl für die Mantisse wie für den Exponenten wird daher eine Binärdarstellung verwendet, die Basis ist $b = 2$. Da die einzige Stelle vor dem Komma eine 1 sein muss, wird sie normalerweise nicht gespeichert. Das Vorzeichen der Zahl wird separat gespeichert.

Der 32 bit umfassende Gleitkommatyp `float` hat eine Mantisse von 24 bit, wovon aber nur 23 Bit gespeichert werden müssen. Von den verbleibenden 9 bit wird eines als Vorzeichen verwendet und 8 als Exponent. Mit einer 8 bit Ganzzahl lassen sich die Zahlen von 0 bis 255 darstellen. Um negative Exponenten zu ermöglichen, muss 127 subtrahiert werden. Die 8 Exponenten-bits codieren also die Exponenten −128 bis 127.

Die Zahl

$$\pi = 3.14159265_{10} = \texttt{11.001010011000101100001}0_2 = \texttt{1.100101001100010110000}10_2 \cdot 2^1$$

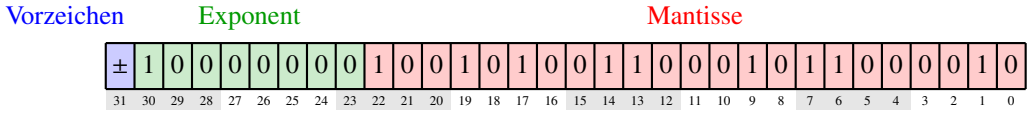
kann daher als `float`-Gleitkommazahl wie folgt gespeichert werden:

	`float`	`double`	`long double`
kleinste darstellbare Zahl	$1.17549 \cdot 10^{-38}$	$2.22507 \cdot 10^{-308}$	$3.3621 \cdot 10^{-4932}$
grösste darstellbare Zahl	$3.40282 \cdot 10^{38}$	$1.79769 \cdot 10^{308}$	$1.18973 \cdot 10^{4932}$
ε	$1.19209 \cdot 10^{-7}$	$2.22045 \cdot 10^{-16}$	$1.0842 \cdot 10^{-19}$
kleinster Exponent	-125	-1021	-16381
grösster Exponent	128	1024	16384
kleinste denormalisiert Zahl:	$1.4013 \cdot 10^{-45}$	$4.94066 \cdot 10^{-324}$	$3.6452 \cdot 10^{-4951}$

Tabelle 1.1: Eigenschaften der Gleitkommatypen `float`, `double` und `long double`. Die Zeile ε ist die Differenz zwischen 1 und der kleinsten darstellbaren Zahl, die grösser ist als 1. Einzelne Exponentenwerte haben eine spezielle Bedeutung (siehe Text), daher fallen die kleinstmöglichen Exponenten grösser aus als aufgrund ihrer Bitlänge zu erwarten ist.

Typ	Bytes	Mantisse	Exponent	IEEE-754
`half`, `binary16`	2	10	5	*
`float`, `binary32`	4	23	8	*
`extended`	5	29	10	
`double`, `binary64`	8	52	11	*
`double extended`, `long double`	10	63	15	*
`quad binary128`	16	112	15	*
`binary256`	32	236	19	*

Tabelle 1.2: Übliche Gleitkommatypen mit Länge der Mantisse und des Exponenten. Die meisten Compiler implementieren nur `float` und `double`, manchmal auch noch `long double`. Die im Standard IEEE 754-2008 definierten Typen sind in der letzten Spalte mit einen $*$ versehen.



Die Ganzzahl $\texttt{10000000}_2 = 128_{10}$ im Exponentenfeld muss um 127 verringert werden um den Exponenten 1 zu ergeben.

Die grösste und kleinste mit einem `float` darstellbare Zahl ist somit

$$\texttt{1.11111111111111111111111}_2 \cdot 2^{128} = 3.4028232_{10} \cdot 10^{38}$$
$$\texttt{1.00000000000000000000000}_2 \cdot 2^{-127} = 5.8774717_{10} \cdot 10^{-39}$$

Den 24 binären Mantissenbits entsprechen gut 7 Dezimalstellen.

Gebräuchliche Formate

Der 1985 verabschiedete Standard IEEE 754 beschreibt die heute gebräuchlichen Implementation von Gleitkommatypen abschliessend. Damit ist gewährleistet, dass numerische Rechnungen auf verschiedenen Prozessoren reproduzierbare Resultate geben.

Die C++-Standardbibliothek bietet im `<limits>` Header die Möglichkeit, Informationen über die Datentypen zu erhalten. In Tabelle 1.1 sind die Resultate für die gebräuchlichsten Typen zusam-

mengestellt. Allerdings offenbart sich hier auch ein Problem dieser Implementation von `<limits>`. Wie wir weiter unten sehen werden, definiert der IEEE 754 Standard Werte für $\pm\infty$, die kleiner sind als die angegebenen maximalen Werte.

Aktuelle Compiler unterstützen typischerweise die Gleitkomma-Typen `float`, `double` und `long double`. Bei Mikrokontrollern, wo Berechnungen mit der hohen Präzision eines `double` nur schon wegen des Platzedarfs der Werte und des Zeitbedarfs für die Operationen kaum sinnvoll sind, ist oft nur der `float`-Typ implementiert. Der GNU-Compiler für die 8-bit AVR-Prozessorfamilie stellt behandelt zum Beispiel `double` genau gleich wie `float`. Graphikkarten unterstützen oft noch einen halben Gleitkommatyp `binary16`, dessen Genauigkeit für die Darstellung von 3D-Objekten ausreichend ist.

Rundung

Der IEEE 754 Standard schreibt auch vor, wie Resultate gerundet werden müssen. Bei vielen Operationen entstehen Resultate, die einen längeren Nachkommateil haben als in das gegeben Gleitkommaformat passt, das Resultat muss gerundet werden. Der Standard kennt fünf verschiedene Rundungsverfahren und empfiehlt, dass die Funktionen wie Wurzeln, Exponentialfunktionen, trigonometrische Funktionen und viele weitere so implementiert werden müssen, dass das Resultat korrekt gerundet ist. Damit ist gemeint, dass das Resultat entsteht, in dem auf das mathematisch exakte Resultat die gewählt Rundungs-Regel angewendet wird.

Der Standard bezweckt mit diesen, dass man sich im Rahmen der Rundungsgenauigkeit auch auf die letzte Stelle eines Gleitkommawertes verlassen kann. Dies ist keinen Selbstverständlichkeit. Gleitkomma-Implementationen von GPUs beispielsweise erfüllen diese Bedingung oft nicht und garantieren in ihren Spezifikation manchmal nur korrekte Resultate mit Ausnahme der letzten 1-2 bits.

Die Null

Die Gleitkomma-Darstellung definiert, dass die Mantisse ein implizites führends 1-bit hat. In diesem Format lässt sich die Null aber nicht darstellen, es ist also eine separate Definition nötig:

- Vorzeichen `0` oder `1` erlaubt die Unterscheidung zwischen $+0$ und -0.
- Alle Exponentenbits = `0`.
- Alle Mantissenbits = `0`.

Dies ist ein Spezialfall einer denormalisierten Zahl (siehe weiter unten)

Unendlich grosse Werte

Im Laufe einer numerischen Berechnung kann es vorkommen, dass die Resultate so gross werden, dass nicht mehr im gegebenen Gleitkommatyp gespeichert werden könne. Der Standard verlangt, dass diese Situation durch einen spezielle Wert für unendlich grosse Zahlen wiedergegeben werden kann, der wie folgt definiert ist:

- Vorzeichenbit `0` oder `1` um zwischen $+\infty$ und $-\infty$ unterscheiden zu können.
- Alle Exponenten-Bits = `1`
- Alle Mantissenbits = `0`

Dies bedeutet, dass der grösste nutzbare Exponent des `float`-Typs nur noch 127 ist.

Denormalizierung

Die Mantisse einer Gleitkommazahl beginnt immer mit einer 1, die aber nicht gespeichert wird. Wird eine Zahl kleiner als mit dem zur Verfügung stehenden Exponenten-Bereich darstellbar, kann sie nicht mehr in diesem Format gespeichert werden. Um solche Zahlen darzustellen, wurde vom Standard einem Exponenten aus lauter `0` eine besondere Bedeutung gegeben. Für den Typ `float` entspricht er nicht mehr dem Exponenten -127 sondern -126 und das implizite führend Bit der Mantisse ist jetzt 0. Mit diesen sogenannten denormalisierten Zahlen lassen sich noch kleinere Wert darstellen, die aber nicht mehr so präzis sind, weil sie weniger signifikante Stellen aufweisen.

Vor- und Nachteile

- ⊕ Konstanter relativer Fehler
- ⊕ Dank des grossen Wertebereiches sind Über- und Unterlauf unwahrscheinlich.
- ⊖ Der kleinste Gleitkommatyp ist bereits so gross wie der gebräuchlichste Ganzzahltyp, Gleitkommazahlen brauchen mehr Platz.
- ⊖ Geschwindigkeit: sofern keine Hardware-Beschleunigung zur Verfügung steht sind Gleitkomma-Operationen deutlich langsamer als Operationen mit Festkomma-Zahlen.
- ⊖ In einer Multi-Core CPU hat nicht unbedingt jeder Kern eine Gleitkomma-Einheit. Gleitkommaoperationen in verschiedenen Threads können sich also gegenseitig behindern.

1.1.4 Hochpräzisionsbibliotheken

TODO: GMP

1.2 Numerische Effekte

Die Unzulänglichkeiten der in Computern verwendeten Zahlensysteme haben zwei Effekte zur Folge, denen bei der Konzeption eines numerischen Lösungsverfahrens Rechnung getragen werden muss.

1.2.1 Auslöschnung

Auslöschung tritt auf, wenn die Differenz zweier ähnlich grosser Zahlen gebildet wird. Als Beispiel betrachten wir die beiden Zahlen $a = \pi$ und $b = \sqrt{10}$. Berechnen wir deren Differenz in Octave, erhalten wir:

```
octave:1> a=pi
a =  3.14159265358979
octave:2> b=sqrt(10)
b =  3.16227766016838
octave:3> a-b
ans = -0.0206850065785864
```

Die ersten zwei Stellen von a und b stimmen überein. Octave zeigt sowohl von a als auch von b 15 signifikante Stellen an. Ein Vergleich mit einer Berechnung mit noch mehr Stellen zeigt, dass diese 15 Stellen auch zuverlässig sind. Für die Differenz zeigt Octave ebenfalls 15 Stellen an, doch die letzte Stelle is falsch, wie zum Beispiel die Berechnung mit 20 Stellen Genauigkeit zeigt[4]

```
scale=20
a=4*a(1)
a
3.14159265358979323844
b=sqrt(10)
b
3.16227766016837933199
a-b
-.02068500657858609355
```

Schreiben wir die Subtraktion in der tabellarischen Form

```
  3.16227766016838
 -3.14159265358979
 ------------------
  0.02068500657859
 ------------------
```

wird erkennbar, dass nur 13 Stellen der Differenz tatsächlich bekannt sind.

Die Rechnung wird in Binärdarstellung etwas klarer. In der folgenden Tabelle sind die Werte in der mittleren Spalte in binärer Gleitkommadarstellung gezeigt, Vorzeichen, Exponent und Mantisse sind zur Verdeutlichung durch ein Leerzeichen getrennt. Zu Beginn der Mantisse muss man sich eine implizite 1 denken, die nicht gespeichert wird. In der dritten Spalte werden die gleichen Zahlen als binäre Festkommawerten geschrieben. Zur Berechnung der Differenz muss der Prozessor die Mantissen ja zunächst so schieben, dass sie den gleichen Exponenten bekommen, der Prozessor berechnet also die Differenz implizit in einer Festkommadarstellung.

	Gleikommawert	Festkommawert
$\sqrt{10}$	0 10000000 10010100110001011000010	11.00101001100010110000010
π	0 10000000 10010010000111111011011	11.00100100001111110110011
$\sqrt{10}-\pi$	0 01111001 01010010111001110000000	0.00000101010010111001111

Man kann gut erkennen, dass die Differenz nur 17 signifkante Stellen hat. Bei der nachfolgenden Darstellung als Gleitkommazahl werden 7 Nullen hinzugefügt, die aber nichts mit der tatsächlichen Differenz $\sqrt{10}-\pi$ zu tun haben. Aus zwei Zahlenwerten mit einer Genauigkeit von 24 Binärstellen ist ein Wert mit einer Genauigkeit von nur 17 Binärstellen geworden. Es sind 7 Binärstellen Genauigkeit ausgelöscht worden.

Beispiel. Sei X ein standardnormalverteilte Zufallsvariable, es soll die Wahrscheinlichkeit dafür berechnet werden, dass $a \leq X \leq b$ ist. In der Wahrscheinlichkeitsrechnung lernt man, dass man dazu die Verteilungsfunktion

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-s^2/2}\, ds$$

[4]Diese Berechnung wurde mit dem Linux-Kommandozeilenprogramm `bc` durchgeführt, welches mit einstellbarer Festkommapräzision von tausenden von Stellen rechnen kann. Es ist Teil jeder Linux-Distribution.

Datentyp	Rechnung mit erf(x)	Rechnung mit erfc(x)
`float`	0.000000	$6.271826 \cdot 10^{-17}$
`double`	$1.110223 \cdot 10^{-16}$	$6.271826 \cdot 10^{-17}$
`long`	$6.272110 \cdot 10^{-17}$	$6.271826 \cdot 10^{-17}$

Tabelle 1.3: Berechung der Wahrscheinlichkeit $P(4.18 \leq X \leq 5.18)$ einer standardnormalverteilten Zufallsvariable mit Hilfe der Bibliotheksfunktionen erf(x) und erfc(x). Starke Auslöschung macht die Berechnung mit erf(x) unbrauchbar.

der Standardnormalverteilung verwenden kann:

$$P(a \leq X \leq b) = \Phi(b) - \Phi(a).$$

Die Funktion $\Phi(x)$ wird in vielen Bibliotheken nicht direkt zur Verfügung gestellt, oft ist nur die sogenannte Fehlerfunktion

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

verfügbar. Die Variablentransformation $t = s/\sqrt{2}$ oder $s = \sqrt{2}t$ macht aus dem Integral für $\Phi(x)$ den Ausdruck

$$\begin{aligned}\Phi(x) &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-s^2/2}\, ds = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{2}x} e^{-t^2}\, \sqrt{2}\, dt = \frac{1}{2} + \frac{2}{\pi} \int_0^{\sqrt{2}x} e^{-t^2}\, dt \\ &= \frac{1}{2}\Big(1 + \operatorname{erf}\Big(\frac{x}{\sqrt{2}}\Big)\Big),\end{aligned}$$

die Fehlerfunktion kann also zur Berechnung der gesuchten Wahrscheinlichkeit verwendet werden.

In der C-Bibliothek stehen Funktionen zur Berechnung von $\sqrt{2}$ und erf(x) für alle zur Verfügung stehenden Datentypen zur Verfügung. Die Tabelle 1.3 zeigt die Resultate[5] für $a = 4.18$ und $b = a+1$.

Da die Werte von erf($\sqrt{2}b$) und erf($\sqrt{2}b$) fast gleich gross sind, findet starke Auslöschung statt. Beim Datentyp `float` ist überhaupt kein Unterschied mehr feststellbar.

Um dieses Problem in den Griff zu bekommen, stellt die C-Bibliothek zusätzlich die sogenannte komplementäre Fehlerfunktion

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \qquad \Rightarrow \qquad \operatorname{erf}(x) = 1 - \operatorname{erfc}(x)$$

zur Verfügung. Damit kann die Wahrscheinlichkeit natürlich auch berechnet werden:

$$P(a \leq X \leq b) = \operatorname{erf}(\sqrt{2}b) - \operatorname{erf}(\sqrt{2}a) = (1 - \operatorname{erfc}(\sqrt{2}b)) - (1 - \operatorname{erfc}(\sqrt{2}a)) = \operatorname{erfc}(\sqrt{2}a) - \operatorname{erfc}(\sqrt{2}b).$$

Für grosse Werte von x streben die Werte dieser Funktion gegen 0, es kann also nicht mehr passieren, dass man einen kleinen Wert zu finden versucht, indem man zwei vergleichsweise grosse Zahlen voneinander subtrahiert. In der dritten Spalte von Tabelle 1.3 sind die Resultate der Berechnung mit Hilfe von erfc(x) gezeigt. Der Auschlöschungseffekt ist vollständig verschwunden. Man kann sogar ablesen, dass die Verwendung des Datentyps `long double` dem Problem der Auslöschung ebenfalls nicht begegnen konnte. Der mit erf(x) berechnete Wert hat selbst bei Verwendung dieses längsten verfügbaren Typs nur drei korrekte Dezimalstellen. ○

[5] Diese Resultate wurden mit dem Programm `normal.cpp` Im Verzeichnis `buch/chapters/experimente/ausloeschung` von [3] berechnet.

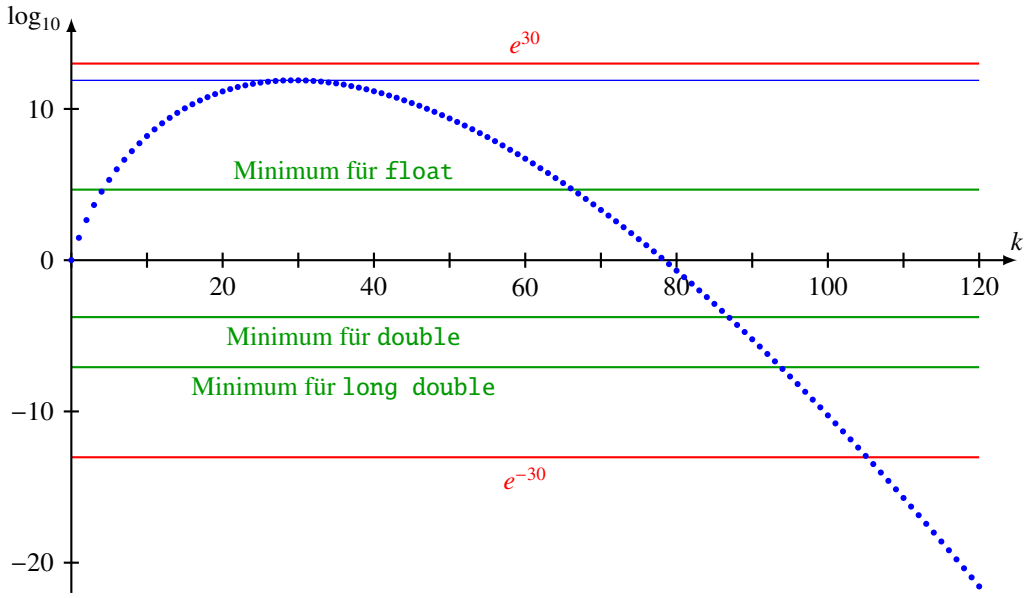


Abbildung 1.1: Verschmierung bei der Berechnung von e^{-30} und e^{30} mit der Exponentialreihe. Auf der vertikalen Achse ist der Zehnerlogarithmus der verschiedenen Grössen abgetragen. Die beiden Werte e^{30} und e^{-30} sind als rote Linien am oberen und unteren Rand eingetragen. Blau ist der absolute Betrag des Terms $s_k = x^k/k!$ in der Exponentialreihe. Die grünen Linien zeigen den kleinsten Unterschied, der zwischen zwei Termen möglich ist, die die Grösse des grössten Summanden in der Exponentialreihe haben.

1.2.2 Verschmierung

Auslöschung kann nicht nur auftreten, wenn zwei fast gleich grosse Zahlen subtrahiert werden. Sie kann in etwas weniger offensichtlicher Form stattfinden, wenn bei der Summation einer Reihe im Vergleich zum Resultat grosse Zwischenresultate entstehen. Diesen Verlust an Genauigkeit infolge grosser Zwischenresultate wird *Verschmierung* genannt.

Beispiel: Exponentialreihe e^x

Die Taylorreihe

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

der Exponentialfunktion ist sehr gut geeignet, Werte von e^x für positive x zu berechnen. Da der Nenner $k!$ exponentiell schnell anwächst, werden späte Terme in der Reihe sehr schnell vernachlässigbar klein.

Für negative Exponenten alternieren die Terme, werden zwischenzeitlich sehr gross, das Resultat ist aber ein sehr kleiner Wert. Im Laufe der Rechnung müssen sich also grosse Terme wieder wegheben. Dies ist in Abbildung 1.1 illustriert. Dort sind die Werte e^{30} und e^{-30} auf einer logarithmischen Skala vertikal als rote Linien eingezeichnet. Die einzelnen Summanden der Reihe sind in blau dargestellt.

Datentyp	e^{-30}
`float`	$-7.2959523438 \cdot 10^{-4}$
`double`	$6.1030424789 \cdot 10^{-6}$
`long double`	$-1.2489259417 \cdot 10^{-8}$
exakt	$9.3576229688 \cdot 10^{-14}$

Die grünen Linien zeigen, welche Genauigkeit mit verschiedenen Datentypen überhaupt noch möglich ist. Der `float`-Typ hat eine Mantisse von 24 bit, eine Zahl m ist daher nur unterscheidbar von $m(1+\epsilon)$, wenn $\epsilon > 2^{-24}$. Dies entspricht $24 * \log_{10}(2) = 7.22$ Dezimalstellen. Die grüne Linie für den `float`-Typ ist daher 7.22 unter dem Maximum der Terme Exponentialreihe eingetragen.

Beim `double`-Typen ist die Mantiesse 52 bit lang, beim `long double` sind es 63 bit. Doch selbst beim `long double` ist die Verschmierung vollständig, das Resultat für e^{-30} hat nichts mit der Realität zu tun. Im Gegenteil, sie zeigen eher an, wie gross der verwendete Datentyp ist. Die grünen Linien in Abbildung 1.1 befinden sich ungefähr dort, wo die gefundenen Werte eingetragen werden müssten.

Summationsreihenfolge

Das Beispiel zur Exponentialfunktion hat gezeigt, wie Verschmierung die Präzision des Resultates vollständig ausgelöscht hat. Ein weniger dramatischer Genauigkeitsverlust kann schon bei wenigen Summanden stark unterschiedlicher Grössenordnung auftreten, wie das folgende Beispiel aus [2] zeigt.

exakte Rechnung	nacheinander
10000.0	10000.0
+ 3.14159	+ 3.14159
	10003.14159
	10003.1
+ 2.71828	+ 2.71828
10005.85987	10005.81828
10005.9	10005.8

Links ist die exakte Addition mit Rundung auf sechs signifikante Stellen am Schluss durchgeführt, rechts steht die Addition mit Rundung nach jedem Term. Bereits nach drei Termen zeigt sich ein Unterschied in der letzten Stelle, der Verschmierung zuzuschreiben ist.

Die unvermeidliche Rundung nach jeder Addition beinträchtigt die Berechnung jeder unendlichen Summe

$$s_n = \sum_{k=0}^{n} a_k.$$

Die naive Berechnung in der Reihenfolge aufsteigender k wird in den meisten Fällen die Reihenfolge kleiner werdender Terme sein. Für genügend grosses k werden die zusätzlichen Terme von ähnlicher Grössenordnung sein wie die Rundungsfehler, die sich bei den ersten Termen bereits gebildet haben.

Die Reihenfolge abnehmender k bietet die Chance, dass Rundungsfehler, die bei der Addition kleiner Terme entstanden sind, bei der Addition grösserer Terme mit kleinerem k verschwinden im neuen Rundungsfehler verschwinden. Dies geschieht allerdings nur, wenn die Beträge der Terme schneller wachsen als die Rundungsfehler in den kleinen Termen.

Methode	`float`	`long double`
vorwärts	15.40368270	18.99789641385390515
rückwärts	18.80791855	18.99789641385389798
Kahan-Summation	18.99789619	18.99789641385389827

Tabelle 1.4: Berechnung der Summe h_n der harmonischen Reihe für $h = 10^8$ mit verschiedenen Methoden.

Kahan-Summationsalgorithmus

Der Kahan-Summationsalgorithmus versucht, über die sich ansammelnden Rundungsfehler in den niederwertigen Stellen in einer separaten Variable Buch zu führen. Um das Prinzip zu verstehen, sei also s_{n-1} eine Teilsumme und a_n der nächste Term der zur Summe hinzugefügt werden soll. Die Summe $s_n = s_{n-1} + t_n$ wird natürlich gerundet. Die Differenz $c_n = (s_n - s_{n-1}) - a_n$, genau in der Reihenfolge der Klammern ausgewertet, enthält die durch Rundung verlorenen Stellen.

Der nächste Summand a_{n+1} dürfte kleiner als die Summe s_n, man kann ihn daher um den Betrag c_n korrigieren und damit verlorene Genauigkeit wiederherstellen. Wir addieren daher $\bar{a}_{n+1} = a_{n+1} - c_n$ anstelle von a_{n+1} und erhalten die Summe $s_{n+1} = s_n + \bar{a}_{n+1}$. Natürlich können wir auch hier wieder den Fehler als $c_{n+1} = (s_{n+1} - s_n) - \bar{a}_{n+1}$ berechnen. Damit erhalten wir den folgenden Algorithmus.

```
double  s = 0;
double  c = 0;
for (int i = 1; i <= n; i++) {
        double  y = a(i);
        y = y - c;
        double  t = s + y;
        c = (t - s) - y;
        s = t;
}
```

Er verwendet die Funktion `a(i)`, welche den Term mit Index i der Summe berechnet. Die Kahan-Summation vervierfacht die Anzahl der Additionen hat aber das Potential die Akkumulation von Rundungsfehlern fast vollständig zu eliminieren.

Beispiel. Die harmonische Reihe

$$h_n = \sum_{k=1}^{n} \frac{1}{k} > \int_1^{n+1} \frac{1}{x}\,dx = \log(n+1)$$

divergiert, aber sehr langsam. Beim numerischen Aufsummieren gibt es ausgiebig Gelegenheit für Genauigkeitsverlust. Am akutesten ist der Verlust, wenn man die Summe mit dem grössten Term $k = 1$ beginnt, eine kleine Verbesserung ist von der umgekehrten Reihenfolge zu erwarten.

Die Resultate der Berechnung von h_{10^8} mit dem `float` Datentyp sind in Tabelle 1.4 zusammengestellt. Die naheliegende Summe beginnend beim ersten Term führt auf ein unbrauchbares Resultat. Beginnt man mit dem kleinsten Term, könnten die Rundungsfehler durch die späteren grösseren Terme übertönt werden, leider divergiert die Summe so langsam, dass damit nicht alle Fehler zum Verschwinden gebracht werden können. Es sind nur gerade die Stellen vor dem Komma korrekt. Die Kahan-Summation vermeidet dieses Problem vollständig.

Zum Vergleich ist in der Spalte rechts in Tabelle 1.4 die Summe berechnet mit dem Typ `long double` dargestellt. Dies zeigt, dass alle signifikanten Stellen der Kahan-Summation korrekt sind. ○

1.3 Iteration

Die meisten numerischen Problemlösungen mit einem Computer nutzen deren Fähigkeit aus, dieselbe Rechnung immer wieder zu wiederholen, bis zum Beispiel die gewünschte Genauigkeit erziehlt ist. Es lohnt sich daher ganz unabhängig irgendwelchen Einschränkungen der Computer-Hardware zu überlegen, was passiert, wenn man eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ immer wieder auf ihren eigenen Output anwendet, wie das zum Beispiel der Code in einer Schleife bei jedem Durchlauf macht.

In diesem Abschnitt ist also

$$f\colon \mathbb{R} \to \mathbb{R} : x \mapsto f(x)$$

eine differenzierbar Funktion. Mit einem gegebenen Startwert $x_0 \in \mathbb{R}$ lässt sich durch wiederholte Anwendung von f die Folge

$$x_0,\ x_1 = f(x_0),\ x_2 = f(x_1),\ x_3 = f(x_2),\ x_4 = f(x_3), \dots$$

konstruieren.

Ist der Punkt x^* ein Fixpunkt der Funktio $f(x)$, ist also $f(x^*) = x^*$, dann ist die mit $x_0 = x^*$ gebildete Iterationsfolge konstant. Es stellt sich damit automatisch die Frage, was mit einem von x^* abweichenden Startwert passiert. Entfernen sich die Werte x_k von x^* oder konvergiert die Folge am Ende gegen x^*?

1.3.1 Beispiele

Wir illustrieren die verschiedenen Situation, die beim Iterieren der Funktion f auftreten können an einigen Beispielen.

Beispiel. Wir betrachten die Funktion

$$f\colon \mathbb{R} \to \mathbb{R} : x \mapsto \sqrt{x+2}$$

Die Iterationsfolge ausgehend vom Startwert $x_0 = 0$ ist in Tabelle 1.5 dargestellt.

Die Werte konvergieren offenbar gegen den Wert 2, In der dritten Spalte steht die Abweichung δ_k des k-ten Folgengliedes vom Grenzwert 2. Mit jeder Iteration wird der Fehler um den Faktor 4 kleiner, wie die vierte Spalte zeigt, in der Quotient aufeinanderfolgender Fehler berechnet ist.

Dieses Verhalten des Fehlers kann man auch analytisch verstehen. Nehmen wir an, dass $x_n = 2 + \delta_n$ und versuchen wir x_{n+} zu berechnen. Indem wir die Funktion $f(x)$ im Punkt $x = 2$ mit Hilfe der Ableitung linear approximieren, erhalten wir wegen

$$f'(x) = \frac{1}{2\sqrt{x+2}}$$

$$x_{n+1} = f(x_n) = f(2+\delta_n) \simeq f(2) + f'(2)\cdot\delta_n 2 + \underbrace{\frac{1}{4}\cdot\delta_n}_{\simeq \delta_{n+1}},$$

der Fehler von x_{n+1} ist also $\delta_{n+1} \simeq \frac{1}{4}\delta_n$. In jedem Iterationsschritt gewinnen wir daher etwa 2 bit Genauigkeit. Für die 52 bit Mantisse des `double` Typs brauchen wir also etwa 26 Iterationen. ◯

k	x_k	$\delta_k = 2 - x_k$	δ_k/δ_{k-1}
0	0.0000000000000000	2.0000000000000000	
2	1.4142135623730951	0.5857864376269049	3.4142
3	1.8477590650225735	0.1522409349774265	3.8478
4	1.9615705608064609	0.0384294391935391	3.9616
5	1.9903694533443939	0.0096305466556061	3.9904
6	1.9975909124103448	0.0024090875896552	3.9976
7	1.9993976373924085	0.0006023626075915	3.9994
8	1.9998494036782890	0.0001505963217110	3.9998
9	1.9999623505652022	0.0000376494347978	4.0000
10	1.9999905876191524	0.0000094123808476	4.0000
11	1.9999976469034038	0.0000023530965962	4.0000
12	1.9999994117257645	0.0000005882742355	4.0000
13	1.9999998529314358	0.0000001470685642	4.0000
14	1.9999999632328584	0.0000000367671416	4.0000
15	1.9999999908082147	0.0000000091917853	4.0000
16	1.9999999977020537	0.0000000022979463	4.0000
17	1.9999999994255133	0.0000000005744867	4.0000
18	1.9999999998563782	0.0000000001436218	4.0000
19	1.9999999999640945	0.0000000000359055	4.0000
20	1.9999999999910236	0.0000000000089764	4.0000
21	1.9999999999977558	0.0000000000022442	3.9998
22	1.9999999999994389	0.0000000000005611	3.9996
23	1.9999999999998597	0.0000000000001403	3.9984
24	1.9999999999999649	0.0000000000000351	4.0000
25	1.9999999999999911	0.0000000000000089	3.9500
26	1.9999999999999978	0.0000000000000022	4.0000
27	1.9999999999999993	0.0000000000000007	3.3333
28	1.9999999999999998	0.0000000000000002	3.0000
29	2.0000000000000000	0.0000000000000000	
30	2.0000000000000000	0.0000000000000000	

Tabelle 1.5: Iterationsfolge für die Funktion $f(x) = \sqrt{x+2}$ ausgehend vom Startwert $x_0 = 0$.

Beispiel. Wir betrachten die Funktion

$$f(x) = 3x(1-x).$$

Sie hat zwei Fixpunkte, die man durch lösen der quadratischen Gleichung

$$f(x^*) = x^* \quad \Rightarrow \quad x^* = -3x^{*2} + 3x^* \quad \Rightarrow \quad 3x^{*2} - 2x^* = 3x^*(x^* - \tfrac{2}{3}) = 0 \quad \Rightarrow \quad x^* = \begin{cases} 0 \\ \frac{2}{3} \end{cases}$$

findet.

Für einen Startwert x_0 nahe des Fixpunktes $x^* = 0$ gilt

$$f(\delta) = 3\delta(1-\delta) = 3\delta - 3\delta^2.$$

Für kleine Werte von δ kann man den quadratischen Wert vernachlässigen und sieht, dass der Fehler durch die Iteration verdreifacht wird. Konvergenz zu diesem Fixpunkt ist also nicht möglich.

Für den Fixpunkt $x =^*= \frac{2}{3}$ finden wir

$$f(\tfrac{2}{3}+\delta) = 3(\tfrac{2}{3}+\delta)(\tfrac{1}{3}-\delta) = \frac{(2+3\delta)(1-3\delta)}{3} = \frac{2-3\delta+9\delta^2}{3} = \frac{2}{3} - \delta + 9\delta^2 \tag{1.1}$$

Der Fehler δ wird zu $-\delta(1-9\delta)$, er ändert also sein Vorzeichen. Ist $\delta > 0$ wird der Fehlerbetrag wird nur um den Faktor $1 - 9\delta < 1$ reduziert. Ist aber $\delta < 0$, dann ist $1 - 9\delta > 0$, der Fehlerbetrag wird wieder vergrössert.

Sei jetzt $\delta > 0$, wir wollen den Fehler nach zwei Iterationsschritten berechnen. Nach dem ersten ist der Fehler $\delta(1-9\delta)$, nach dem zweiten

$$\delta(1-9\delta)(1-\delta(1-9\delta)) = \delta(1 - 18\delta + 162\delta^2 - 729\delta^3).$$

Für kleines δ können die Terme höhrer als erster Ordnung vernachlässigt werden und man kann schliessen, dass der Fehler nach zwei Iterationen tatsächlich um den Faktor $(1-18\delta)$ kleiner geworden ist?

Wie viele Iterationsschritte sind mindestens nötig, um den Fehler von δ_0 auf δ_{2n} zu verbessern? Wie wir soeben berechnet haben, nimmt der Betrag des Fehlers zwischen den Schritten k und $k+2$ gemäss

$$\delta_{k+2} = \delta_k(1 - 18\delta_k)$$

um den Faktor $(1-18\delta_k)$ ab. Es ist übersichtlicher, mit dem Logarithmus des Fehlers zu rechnen. Dieser verändert sich gemäss

$$\log \delta_{k+2} = \log \delta_k + \log(1 - 18\delta_k) \simeq \log \delta_k - 18\delta_k,$$

wobei wir für den zweiten Term die lineare Approximation aus der Taylorreihe

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

von $\log x$ an der Stelle $x = 1$ verwendet haben. Zwischen δ_0 und δ_{2n} bedeutet dies

$$\log \delta_{2n} = \log \delta_0 - 18 \sum_{k=0}^{n-1} \delta_{2k} \qquad \Rightarrow \qquad \frac{1}{18} \log \frac{\delta_0}{\delta_{2n}} = \sum_{k=0}^{n-1} \delta_{2k}.$$

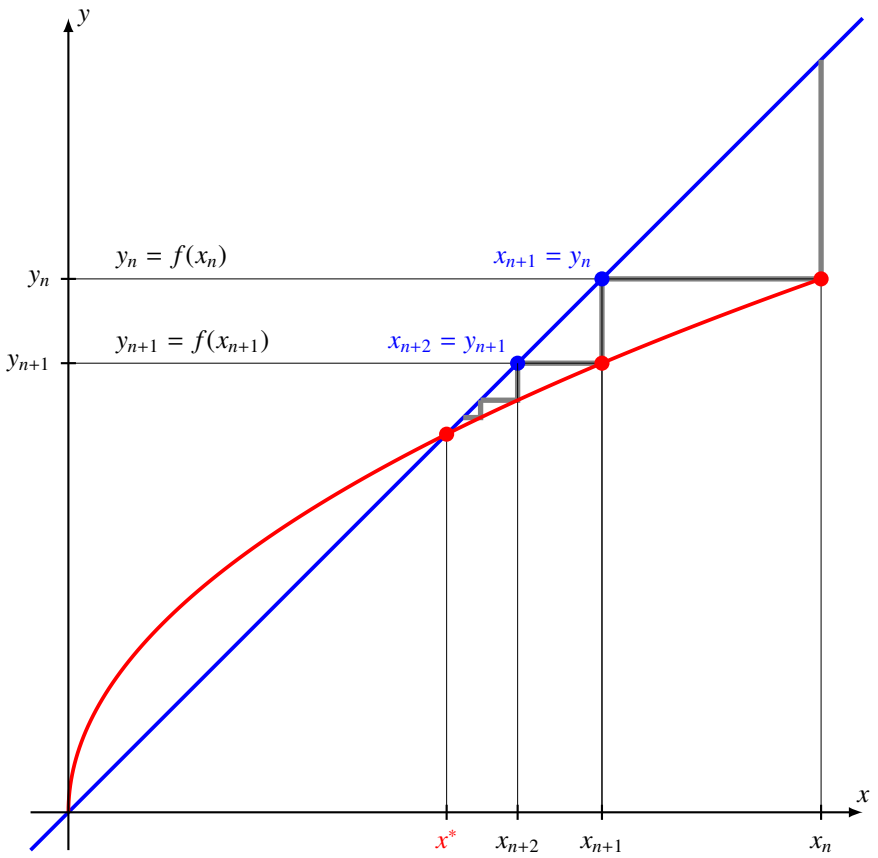


Abbildung 1.2: Ein Fixpunkt x^* der Funktion $f(x)$ manifestiert sich als Schnittpunkt des Graphen $y = f(x)$ mit der 45°-Geraden. Die Iterationsfolge ausgehend von einem Startwert x_0 wird als Treppenkurve zwischen dem Graphen von f und der 45°-Geraden sichtbar.

Da der Fehler immer kleiner wir, kann man für eine erste grobe Abschätzung der Summe auf der rechten Seite die grössten und kleinsten Terme Damit lässt sich die Summe nach oben Abschätzen

$$n\delta_{2n-2} \leq \sum_{k=0}^{n-1} \delta_{2k} \leq n\delta_0.$$

Einsetzen ergibt

$$n\delta_{2n-2} \leq \frac{1}{18} \log \frac{q_0}{q_{2n}} \leq n\delta_0 \quad \Rightarrow \quad \frac{1}{18q_0} \log \frac{q_0}{q_{2n}} \leq n \leq \frac{1}{18q_{2n-2}} \log \frac{q_0}{q_{2n}} < \frac{1}{18q_{2n}} \log \frac{q_0}{q_{2n}}$$

Eine Verbesserung um zwei Stellen von $q_0 = 0.01$ auf $q_{2n} = 0.0001$ braucht also zwischen 25 und 2558 Iterationen. ○

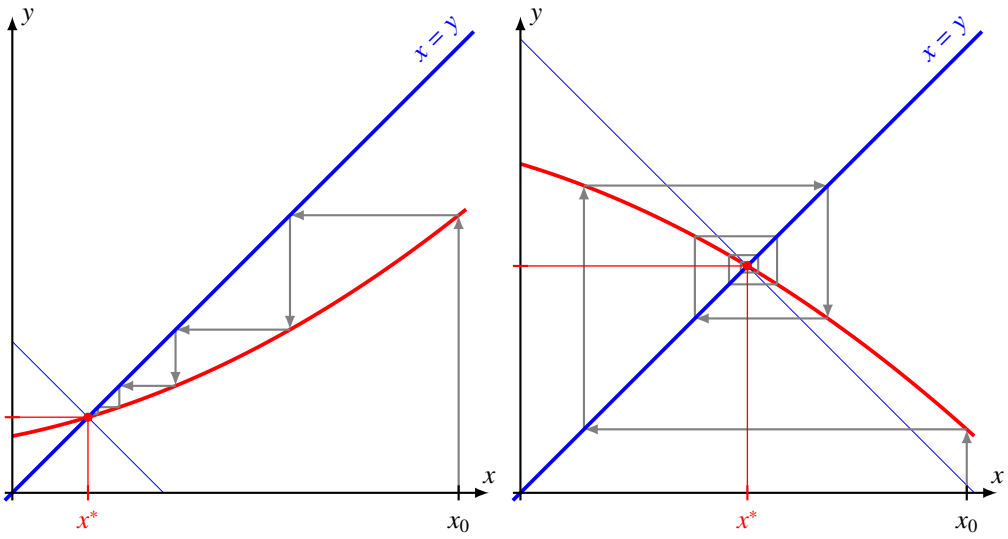


Abbildung 1.3: Die Fixpunktiteration $x_{n+1} = f(x_n)$ konvergiert gegen den Fixpunkt x^* falls $|f'(x^*)| < 1$ mit mindestens linearer Konvergenzgeschwindigkeit.

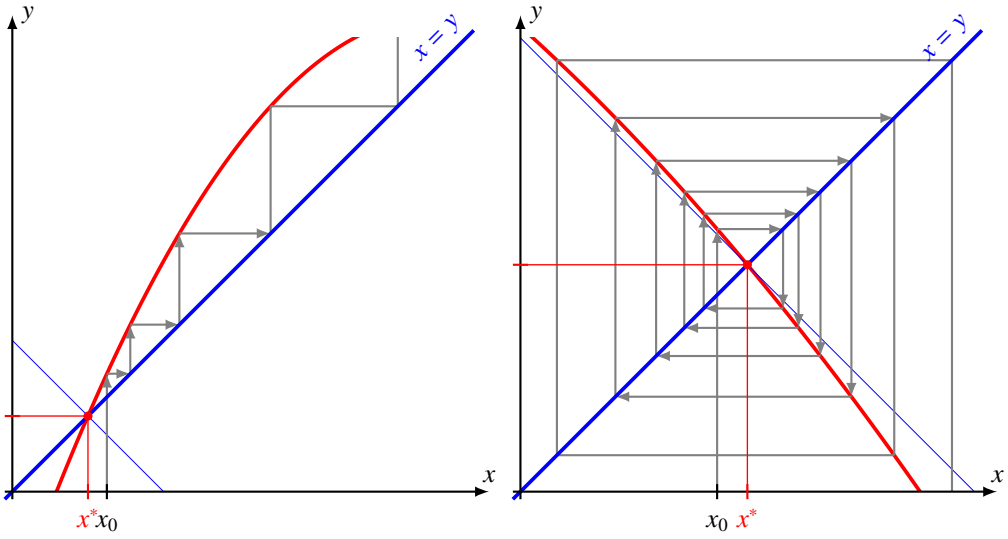


Abbildung 1.4: Die Fixpunktiteration $x_{n+1} = f(x_n)$ mit Fixpunkt x^* divergiert für $|f'(x^*)| > 1$.

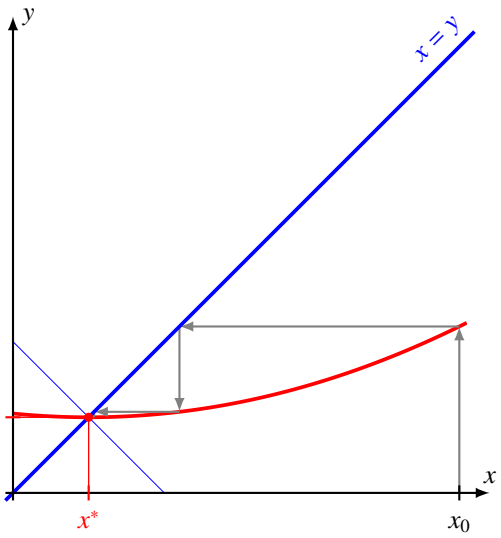


Abbildung 1.5: Die Fixpunktiteration $x_{n+1} = f(x_n)$ konvergiert quadratisch gegen den Fixpunkt x^* für $f'(x^*) = 0$.

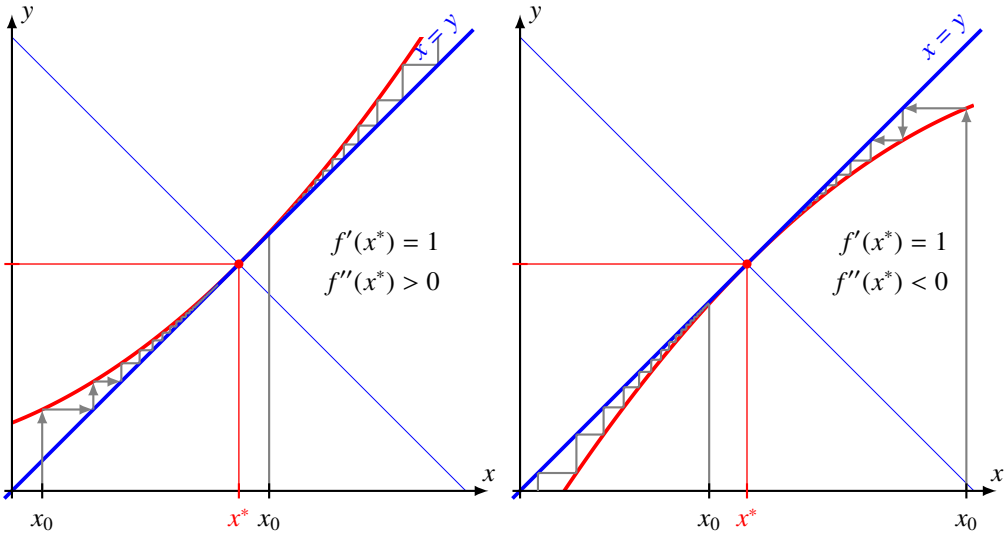


Abbildung 1.6: Konvergenzverhalten der Fixpunktiteration $x_{n+1} = f(x_n)$ in der Umgebung des Fixpunktes x^* für $f'(x^*) = 1$. Falls $f''(x^*) > 0$ konvergiert die Iteration sehr langsam für einen Startwert $x_0 > x^*$ falls $f''(x^*) > 0$, divergiert aber für einen Startwert $x_0 > x^*$. Bei negativer zweiter Ableitung ist es umgekehrt.

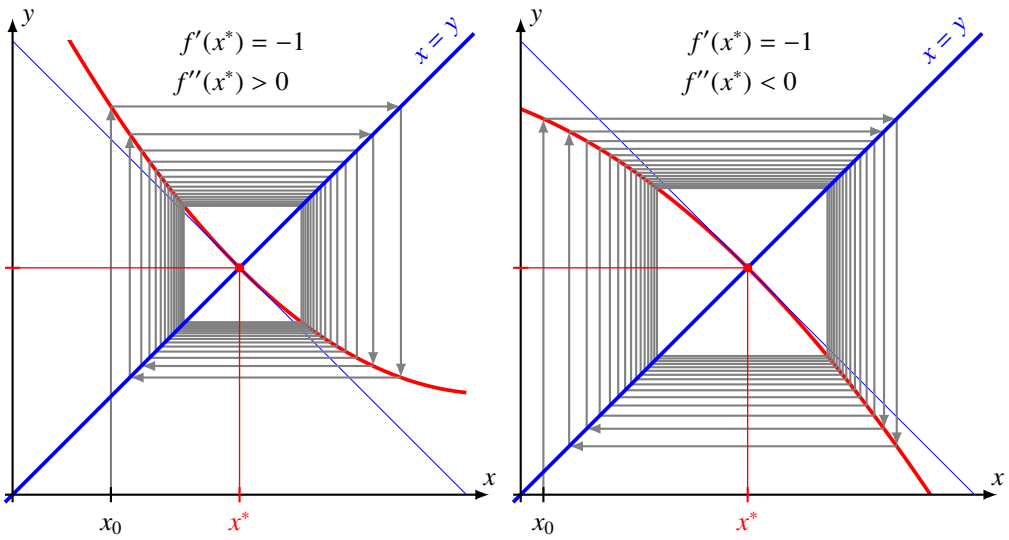


Abbildung 1.7: Die Fixpunktiteration $x_{n+1} = f(x_n)$ konvergiert sehr langsam in der Umgebung des Fixpunktes x^* für $f'(x^*) = -1$ für $f''(x^*) \neq 0$.

1.3.2 Graphische Analyse

Ein Fixpunkt der Funktion $f(x)$ manifestiert sich in einem Schnittpunkte des Graphen von f mit der 45°-Geraden $y = x$ wie in Abbildung 1.2. Die Iterationsfolge x_n kann graphisch wie folgt korrigiert werden. Ausgehend vom Wert x_n auf der x-Achse folgt man der vertikalen, bis man auf den Graphen der Funktion f trifft, dies liefert den Wert $y_n = f(x_n)$. Dieser soll jetzt als neuer x-Wert verwendet werden. Dazu folgt man der Horizontalen bis zur 45°-Geraden, die x-Koordinate des Schnittpunktes ist der x_{n+1}. So entsteht Treppenlinie in Abbildung 1.2.

1.3.3 Konvergenzbedingung

Aus der graphischen Analyse von Abschnitt 1.3.2 kann man jetzt leicht Kriterien ableiten, wann die Iterationsfolge konvergent ist. Die Situation in Abbildung 1.2 tritt zum Beispiel immer ein, wenn es in einer Umgebung von x^*

$$\begin{aligned} x^* < f(x) < x \qquad & \text{für } x > x^* \text{ und} \\ x^* > f(x) < x \qquad & \text{für } x < x^* \end{aligned}$$

gilt.

Ist dagegen der Graph von f in einer Umgebung von x^* steiler als die 45°-Gerade, gilt also

$$f(x) \begin{cases} > x & x > x^* \\ < x & x < x^* \end{cases} \tag{1.2}$$

dann folgt

$$x_{n+1} = f(x_n) \begin{cases} > x_n & x_n > x^* \\ < x_n & x_n < x^*. \end{cases}$$

In beiden Fällen ist x_{n+1} weiter vom Fixpunkt x^* entfernt als x_n, die Folge kann also nicht konvergieren.

Andererseits ist die Bedignung 1.2 zu speziell. Der Funktionswert $f(x_n)$ darf durchaus auch kleiner als x^* werden, wenn nur der Betrag der Differenz zu x^* abnimmt, wenn also gilt

$$|x^* - f(x_n)| < |x^* - x_n| \tag{1.3}$$

Diese Bedingung ist genügend nahe bei x^* erfüllt, wenn $|f'(x^*)| < 1$ ist. Andererseits liegt Divergenz vor, wenn

$$|x^* - f(x_n)| > |x^* - x_n|,$$

was nahe bei x^* erfüllt ist, wenn $|f'(x^*)| > 1$ gilt. Wir fassen diese Resultate zusammen im folgenden Satz

Satz 1.1. *Die Iterationsfolge $x_{n+1} = f(x_n)$ der Funktion $f(x)$ mit Fixpunkt $x^* = f(x^*)$ konvergiert für einen Startwert x_0 nahe genug bei x^*, wenn $|f'(x^*)| < 1$, sie divergiert, wenn $|f'(x^*)| > 1$.*

1.3.4 Die logistische Gleichung

Die logistische Gleichung ist

$$f_\lambda(x) = \lambda x(1 - x)$$

mit dem positiven Parameter λ. Im Beispiel auf Seite 23 haben wir diese Funktion bereits einmal angetroffen mit dem Parameterwert $\lambda = 3$. Der Graph von $f_\lambda(x)$ ist eine Parabel, die die x-Achse in den Punkten $(0, 0)$ und $(1, 0)$ schneidet. Das Maximum wird bei $x = \frac{1}{2}$ angenommen und ist $f_\lambda(\frac{1}{2}) = \lambda/4$. Die Funktion bildet also das Interval $[0, 1]$ wieder in das selbe Interval ab, wenn $0 \le \lambda \le 4$.

Wir berechnen die Fixpunkte von $f_\lambda(x)$ im Interval $[0, 1]$, also die Lösungen der quadratischen Gleichung

$$x^* = \lambda x^*(1 - x^*) \quad \Rightarrow \quad \lambda x^{*2} + (1 - \lambda)x^* = x^*(\lambda x^* + 1 - \lambda) = 0 \quad \Rightarrow \quad x^* = \begin{cases} 0 \\ \dfrac{\lambda - 1}{\lambda}. \end{cases}$$

Für $\lambda < 1$ gibt es keinen weitere Fixpunkt im Interval $[0, 1]$, für $\lambda > 1$ ist $(\lambda - 1)/\lambda < 1$ immer im Interval.

Zur Beurteilung der Konvergenz der Iterationsfolge müssen wir die Ableitung von f_λ im Fixpunkt berechnen. Die Ableitung ist $f'_\lambda(x) = \lambda(1 - 2x)$, also gilt in den Fixpunkten

$$\begin{aligned} f'_\lambda(0) &= \lambda \\ f_\lambda\Big(\frac{\lambda - 1}{\lambda}\Big) &= \lambda\Big(1 - 2\Big(\frac{\lambda - 1}{\lambda}\Big)\Big) = \lambda - 2(\lambda - 1) = 2 - \lambda. \end{aligned}$$

Daraus liest man mit dem Konvergenzkriterium von Satz 1.1 ab, dass Konvergenz zum Punkte $(\lambda - 1)/\lambda$ vorliegt für $\lambda \in (1, 3)$. Für $\lambda > 3$ divergiert die Iterationsfolge.

Im Beispiel auf Seite Seite 23 haben wir wir den Grenzfall $\lambda = 3$ untersucht und festgestellt, dass die Iterationsfolge gerade noch konvergiert, allerdings sehr langsam. Man beachte, dass das Kriterium (1.3) in diesem Fall nicht erfüllt ist.

Für $\lambda < 1$ konvergiert die Iterationsfolge zum Nullpunkt. Wie verhält sich die Folge im Grenzfall $\lambda = 1$. Auch in diesem Grenzfall ist das Kriterium nicht erfüllt, wir müssen das Problem als wieder gesondert analysieren. Es gilt

$$x_{n+1} = x_n(1 - x_n) = x_n - x_n^2.$$

Für $x_n > 0$ folgt also, dass $0 < x_{n+1} < x_n$ ist, die Folge wird also konvergieren. Allerdings ist die Konvergenz wieder ähnlich langsam wie im Falle $\lambda = 3$. Für einen Wert $x_n < 0$ ist allerdings $x_{n+1} = x_n - x_n^2 < x_n$, d. h. die Folge divergiert.

Dieses Verhalten lässt sich auch allgemein formulieren.

Satz 1.2. *Falls $f'(x^*) = 1$, dann konvergiert die Iterationsfolge für einen Startwert x_0 genügend nahe und grösser als x^* wenn $f''(x^*) < 0$ und sie divergiert für $f''(x^*) > 0$. Für einen Startwert genügend nahe und kleiner als x^* dagegen konvergiert die Iterationsfolge für $f''(x^*) < 0$ und divergiert für $f''(x^*) > 0$.*

Beweis. Wir entwickeln die Funktion f im Punkt x^* die Potenzreihe

$$\begin{aligned} f(x^* + \delta) &= f(x^*) + f'(x^*) \cdot \delta + \frac{1}{2} f''(x^*) \cdot \delta^* + O(\delta^3) \\ &= x^* + \delta + \frac{1}{2} f''(x^*)\delta^2 \end{aligned}$$

Die Entfernung zu Fixpunkt ist

$$|f(x^* + \delta) - x^*| \simeq |\delta + \frac{1}{2} f''(x^*)\delta^2| = |\delta| \cdot |1 + \frac{1}{2} f''(x^*)\delta|.$$

Ob die Entfernung wird genau dann grösser, wenn $f''(x^*)\delta$ positiv ist. sie wird kleiner, wenn $f''(x^*)\delta$ negativ sind. Der erste Fall tritt ein, wenn δ und $f''(x^*)$ gleiches Vorzeichen haben. Für Werte grösser als x^* heisst das, dass $f''(x^*) > 0$ sein muss. Analog folgen alle anderen Fälle. □

Für den Fall $\lambda = 3$ der logistischen Gleichung können wir den folgenden Satz formulieren:

Satz 1.3. *Falls $f'(x^*) = -1$, dann konvergiert die Iterationsfolge für einen Startwert genügend nahe bei x^*, wenn $f''(x^*) < 0$, sie divergiert, wenn $f''(x^*) > 0$.*

Beweis. Wir verwenden wieder die Entwicklung

$$\begin{aligned} f(x_n) = f(x^* + \delta_n) &= f(x^*) + f'(x^*) \cdot \delta_n + \frac{1}{2} f''(x^*) \cdot \delta_n^2 + \cdots = x^* - \delta_n + \frac{1}{2} f''(x^*)\delta_n^2 + \dots \\ \delta_{n+1} &= -\delta + \frac{1}{2} f''(x^*)\delta^2 \dots \end{aligned}$$

Daraus kann man zunächst erkennen, dass der Fehler bei jedem Iterationsschritt das Vorzeichen wechselt. Wir berechnen den Fehler nach zwei solchen Schritten.

$$\begin{aligned} \delta_{n+2} &= -\delta_{n+1} + \frac{1}{2} f''(x^*)\delta_{n+1}^2 = \delta_n - \frac{1}{2} f''(x^*)\, \delta_n^2 + \frac{1}{2} f''(x^*)(-\delta_n + f''(x^*)\delta_n^2)^2 \\ &= \delta_n - \frac{1}{2} f''(x^*)\delta_n^2 + \frac{1}{2} f''(x^*)(\delta_n^2 + 2f''(x^*)\delta_n^3 + f''(x^*)^2\delta_n^4) \\ &= \delta_n + f''(x^*)^2\delta_n^3 + \frac{1}{2} f''(x^*)^3\delta_n^4 + \dots \\ &= \delta_n(1 + f''(x^*)\delta_n^2) \end{aligned}$$

Die Terme höherer Ordnung als 3 können für kleines δ_n vernachlässigt werden. Nach zwei Iterationsschritten hat also der Fehler wieder das gleiche Vorzeichen, aber der Betrag hat sich im den Faktor

$$1 + f''(x^*)\delta_n^2$$

verändert. Dieser Faktor ist genau dann < 1, wenn $f''(x^*) < 0$ ist. □

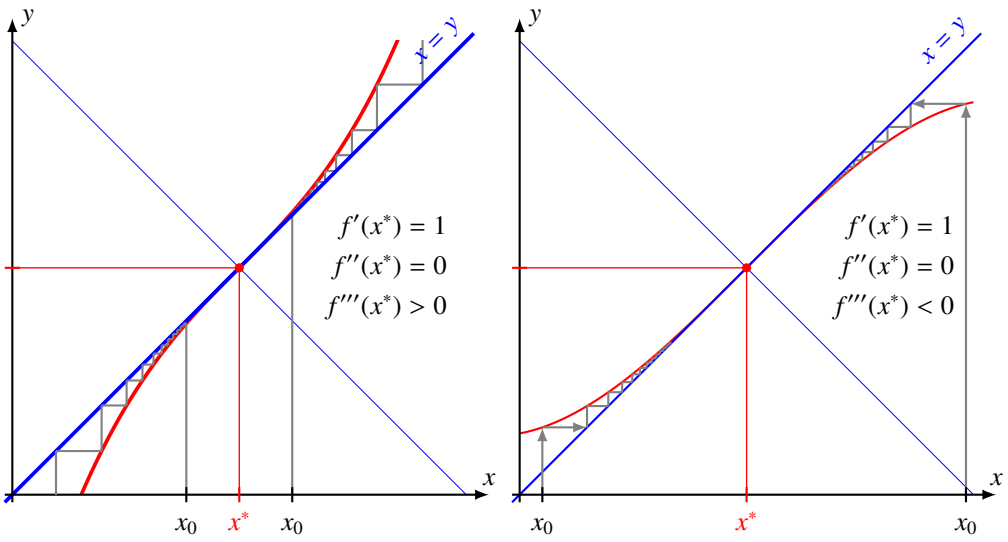


Abbildung 1.8: Die Fixpunktiteration $x_{n+1} = f(x_n)$ mit Fixpunkt x^* mit $f'(x^*) = 1$ und $f''(x^*) = 0$ divergiert für $f'''(x^*) > 0$ und konvergiert sehr langsam für $f'''(x^*) < 0$.

Ähnlich wie im Beispiel auf Seite 23 kann man auch in diesem Fall zeigen, dass die Konvergenz der Folge sehr langsam ist.

1.3.5 Dritte Ableitung im Fixpunkt

Die bisher zusammengetragenen Sätze decken die Fälle $|f'(x^*)| = 1$ mit nicht verschwindender zweiter Ableitung $f''(x^*) \neq 0$ ab. In einigen Fällen ist Konvergenz nicht gesichert, doch wenn Konvergenz vorliegt, dann ist sie sehr langsam. Speziell an dieser Situation ist, dass $f(x) - (x^* + f'(x^*)(x - x^*))$ das Vorzeichen in einer Umgebung von x^* nicht wechselt.

Noch nicht untersucht wurde der Fall $f''(x^*) = 0$. Es ist klar, dass die Konvergenzgeschwindigkeit nicht schneller werden kann. Für $f'''(x^*) \neq$ folgt zudem, dass in diesem Fall $f(x) - (x^* + f'(x^*)(x - x^*))$ das Vorzeichen in einer Umgebung von x^* wechselt. Da aber das Iterationsverfahren sehr empfindlich darauf regiert, ob der Graph von f oberhalb oder unterhalb der Geraden $y = \pm x$ liegt, erwarten wir ein anderes Verhalten als im Fall $f''(x^*) \neq 0$. Alle möglichen Situationen sind in den Abbildungen 1.8 und 1.9 dargestellt.

Wir gehen also davon aus, dass sich f in einer Umgebung von x^* in der Form

$$f(x^* + \delta) \simeq \tilde{f}(x^* + \delta) = x^* + f'(x^*)\delta + \frac{1}{6}f'''(x^*)\delta^3 = x^* + s\delta + a\delta^3$$

entwickeln lässt, wobei $s = \pm 1$ und $a \neq 0$ ist.

Der Iterationsfehler δ_{n+1} verhält sich wie

$$x^* + \delta_{n+1} = f(x_n) = f(x^* + \delta_n) = x^* + s\delta_n + a\delta_n^3 + O(\delta_n^4).$$

$$\delta_{n+1} = s\delta_n + a\delta_n^3 + O(\delta_n^4) = \delta_n(s + a\delta_n^2) + O(\delta_n^4) = s\delta_n\left(1 + \frac{a}{s}\delta_n^2\right) + O(\delta_n^4)$$

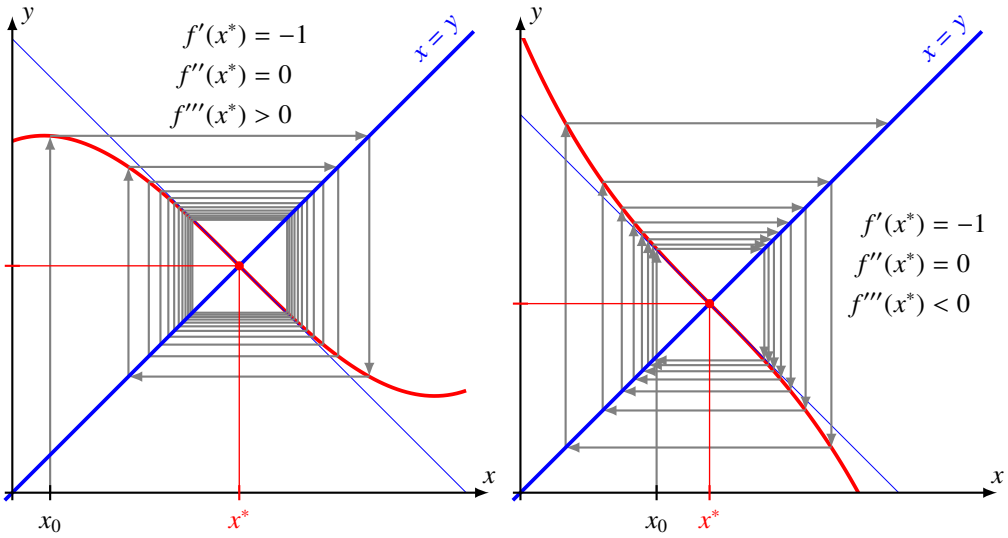


Abbildung 1.9: Die Fixpunktiteration $x_{n+1} = f(x_n)$ mit Fixpunkt x^* mit $f'(x^*) = -1$ und $f''(x^*) = 0$ konvergiert sehr langsam für $f'''(x^*) > 0$ und divergiert für $f'''(x^*) < 0$.

Da $\delta_n^2 > 0$ ist, ist der Klammerausdruck genügend nahe bei x^* immer positiv. Der Betrag des Fehlers verhält sich dann wie

$$|\delta_{n+1}| = |\delta_n| \cdot \left(1 + \frac{a}{s}\delta_n^2\right) + O(\delta_n^4).$$

Ob der Fehler bei der Iteration grösser oder kleiner wird, hängt also einzig vom Vorzeichen von a/s ab. Ist $a/s > 0$, wird der Fehler grösser, die Iterationsfolge ist divergent, ist $a/s < 0$, wird der Fehler kleiner, die Iterationsfolge konvergiert. Wir fassen diese Resultate zusammen im folgenden Satz.

Satz 1.4. *Falls $f : \mathbb{R} \to \mathbb{R}$ den Fixpunkt $x^* \in \mathbb{R}$ hat, und $f'(x^*) = \pm 1$, $f''(x^*) = 0$ und $f'''(x^*) \neq 0$ ist, dann ist die Iterationsfolge $x_{n+1} = f(x_n)$ konvergent, falls $f'(x^*) \cdot f'''(x^*) < 0$ ist, gleichbedeutend damit, wenn $f'(x^*)$ und $f'''(x^*)$ verschiedene Vorzeichen haben. Wenn $f'(x^*)$ und $f'''(x^*)$ das gleiche Vorzeichen haben, dann ist die Fixpunktiterationsfolge divergent.*

Abbildung 1.8 zeigt die Situation $f'(x^*) = 1$ während Abbildung 1.9 die Situation $f'(x^*) = -1$ abdeckt.

1.4 Konvergenzgeschwindigkeit

Sind die Fehler einer iterativen numerischen Berechnungsmethoden bekannt, kann die Gesetzmässigkeit der Fehler ausgenutzt werden, sie weitgehend zu eliminieren, und so die Konvergenz des Verfahrens zu beschleunigen. In diesem Abschnitt wollen wir das Prinzip veranschaulichen, es soll später bei der Berechnung von Integralen systematisch angewendet werden.

1.4.1 Lineare und quadratische Konvergenz

Wir betrachten zwei Beispiele von Iterationsfolgen genauer, die deutlich unterschiedlich schnell konvergieren. Ziel ist, diesen Unterschied zu quantifizieren und ein Kriterium zu finden, welches

schnelle Konvergenz von Iterationsfolgen garantiert.

Lineare Konvergenz

Am Beispiel der Iterationsfolge der Funktion $f(x) = \sqrt{x+2}$ von Seite 21 wurde berechnet, wie sich der Fehler von einer Iteration zur nächsten entwickelt. Dabei wurde gefunden, dass der Fehler δ_n in der n-ten Iteration zu $\delta_{n+1} \simeq f'(x^*) \cdot \delta_n$ in der $(n+1)$-ten Iteration wird. Sind in der n-ten Iteration k Nachkommabits korrekt, ist der Fehler also von der Grössenordnung 2^{-k}, wird der Fehler in der $(n+1)$-ten Iteration von der Grössenordnung $f'(x^*) \cdot 2^{-k} = 2^{-k+\log_2 f'(x^*)}$ sein. In jeder Iteration werden also $\log_2 f'(x^*)$ Binärstellen Genauigkeit gewonnen.

Man sagt, eine Folge ist *linear* konvergent, wenn die Anzahl korrekter Stellen in jeder Iteration um die gleiche Anzahl zunimmt. Die Anzahl korrekter Stellen wächst in diesem Fall linear mit der Anzahl Iterationen.

Für eine konvergente Iterationsfolge einer Funktion f liegt also normalerweise lineare Konvergenz vor, es werden $\log_2 f'(x^*)$ Binärstellen oder $\log_1 0 f'(x^3)$ Dezimalstellen Genauigkeit in jeder Iteration gewonnen. Falls die Ableitung $f'(x^*)$ verschwindet liegt offenbar ein Spezialfall vor, er wird im übernächsten Abschnitt genauer untersucht.

Quadratische Konvergenz

Ist $x = \sqrt{a}$, dann muss gelten $x = a/x$, also auch

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right) = f_a(x).$$

Mit der Funktion $f_a(x)$ kann man jetzt ein Iterationsfolge konstruieren, die gegen den Fixpunkt

$$x^* = f_a(x^*) \quad \Rightarrow \quad x^* = \frac{1}{2}\left(x^* + \frac{a^*}{x}\right) \quad \Rightarrow \quad x^{*2} - a = 0 \quad \Rightarrow \quad x^* = \sqrt{a}$$

von $f_a(x)$ konvergiert. Wir untersuchen die Konvergenzgeschwindigkeit, indem wir $f_a(x)$ mit Hilfe der Taylorreihe approximieren:

$$f_a(x^* + \delta) \simeq f_a(x^*) + f_a'(x^*) \cdot \delta + \frac{1}{2} f_a''(x^*) \cdot \delta^2$$

Die Ableitungen von f_a sind

$$f_a'(x) = \frac{1}{2} - \frac{a}{2x^2}$$
$$f_a''(x^*) = \frac{a}{x^3}.$$

Mit $x = x^* = \sqrt{a}$ folgt für den Fehler

$$f_a(x^* + \delta) \simeq x^* + \underbrace{\left(\frac{1}{2} - \frac{a}{2x^{*2}}\right)}_{=0}\delta + \frac{a}{2x^{*3}}\delta^2 = x^* + \frac{1}{2\sqrt{a}}\delta^2.$$

Der Fehler wird also im Wesentlichen quadriert. Sind k Nachkommabits korrekt, liegt ein Fehler von der Grössenordnung 2^{-k} vor. Daraus wird in der nächsten Iteration ein Fehler von der Grössenordnung 2^{-2k}, die Anzahl korrekter Nachkommabits ist $2k$, hat sich also verdoppelt.

Man sagt, eine Folge konvergiert *quadratisch* gegen x^*, wenn der Fehler $x_n - x^*$ für $n \to n+1$ quadriert wird, die Anzahl korrekter Stellen also verdoppelt wird. Konvergiert eine Folge quadratisch, werden N korrekte Stellen innert $\log_2 N$ Iterationen erreicht. Quadratische Konvergenz ist also im Vergleich zu linearer Konvergenz exponentiell schneller und wird daher in Anwendungen angestrebt.

Konvergenzgeschwindigkeit von Iterationsfolgen

Früher wurde gezeigt, dass die Iterationsfolge $x_{n+1} = f(x_n)$ einer Funktion f für genügend Nahe bei einem Fixpunkt x^* liegende Startwerte x_0 gegen x^* konvergiert, wenn $|f'(x^*)| < 1$. Wir sind jetzt auch in der Lage, die Konvergenzgeschwindigkeit zu quantifizieren. Dazu entwickeln wird $f(x)$ um x^* bis zur zweiten Ordnung und erhalten für den Fehler

$$f(x^* + \delta) \simeq f(x^*) + f'(x^*) \cdot \delta + \frac{1}{2} f''(x^*) \cdot \delta^2 \tag{1.4}$$

Sofern $f'(x^*) \neq 0$ ist, ist der zweite Term dominant, der Fehler wird in jeder Iteration mit $f'(x^*)$ multipliziert, es werden in jeder Iteration $\log_2 f'(x^*)$ Genauigkeit gewonnen. Es liegt also *lineare* Konvergenz vor.

Verschwindet die Ableitung $f'(x^*) = 0$, dann fällt der zweite Term auf der rechten Seite von (1.4) weg, der Fehler ist im wesentlich quadratisch kleiner, es liegt *quadratische* Konvergenz vor. Wenden wir das Kriterium auf das Beispiel

$$f(x) = \frac{1}{2}\left(x + \frac{a}{x}\right) \qquad \text{mit der Ableitung} \qquad f'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right),$$

finden wir für $f'(x^*) = f'(\sqrt{a}) = 0$. Die Iterationsfolge muss also quadratisch konvergieren, wie wir im vorangegangenen Abschnitt auch direkt nachgerechnet haben.

1.4.2 Konvergenzbeschleunigung

Die genaue Kenntnis des Fehlers kann auch ermöglichen, einen Teil des Fehlers zu eliminieren. Wir illustrieren dieses Prinzip wieder am Beipsiel der Iterationsfolge

$$x_0,\ x_1 = f(x_0),\ x_2 = f(x_1),\ x_3 = f(x_2), \ldots$$

mit der Funktion $f(x) = \sqrt{x+a}$. Die Folge konvergiert gegen den Fixpunkt $x^* = f(x^*)$, der sich durch quadrieren und lösen der quadratische Gleichung

$$\begin{aligned} x^{*2} &= x^* + a \\ x^{*2} - x^* - a &= 0 \\ \Rightarrow \qquad x^* &= \frac{1}{2} + \sqrt{\frac{1}{4} + a} \end{aligned}$$

finden lässt. Die negative Wurzel kommt nicht in Frage, weil dies auf eine negative Zahl führen würde, die nicht Fixpunkt der Wurzelfunktion sein kann, die nur positive Werte hat.

Wie im früheren Beispiel mit $a = 2$ können wir das Verhalten des Fehlers $\delta_n = x_n - x^*$ mit Hilfe der linearen Approximation

$$x_n = x^* + \delta_n \qquad \Rightarrow \qquad x_{n+1} = f(x_n) = f(x^* + \delta_n) = f(x^*) + f'(x^*) \cdot \delta_n = x^* + \underbrace{f'(x^*) \cdot \delta_n}_{\simeq \delta_{n+1}}.$$

k	Fixpunkt	Fehler	q
1	2.039426062845019	0.039426062845019	0.306
2	2.000008018463113	0.000008018463113	0.249
3	2.000000000000335	0.000000000000335	0.249
4	2.000000000000000	0.000000000000000	0.249

Tabelle 1.6: Resultate des beschleunigten Verfahrens zur Bestimmung eines Fixpunktes der Iterationsfolge der Funktino $f(x) = \sqrt{x+a}$ mit $a = 2$.

Der Fehler wird also in jeder Iteration um den Faktor

$$f'(x^*) = \frac{1}{2\sqrt{x^*+a}} = \frac{1}{2x^*}$$

kleiner. Wir wissen somit, dass der Fehler in jeder Iteration um den gleichen Faktor kleiner wird, aber da wir x^* noch nicht kennen, kennen wir den Wert des Faktors noch nicht.

Drei aufeinanderfolgende Folgenglieder sind

$$\begin{aligned} x_{n-1} &= x^* + \quad \delta \\ x_n \quad &= x^* + q\ \delta \qquad (1.5) \\ x_{n+1} &= x^* + q^2\delta \qquad (1.6) \end{aligned}$$

Darin sind alle drei Grössen auf der rechten Seite unbekannt. Die Differenzen aufeinanderfolgender Folgenglieder sind

$$\begin{aligned} x_n \quad - x_{n-1} &= \delta(q-1) \\ x_{n+1} - x_n \quad &= \delta(q^2-q) = \delta q(q-1). \end{aligned}$$

Der Differenzenquotient

$$\frac{x_{n+1}-x_n}{x_n \quad - x_{n-1}} = q$$

kann verwendet werden, eine bessere Approximation zu bestimmen. Dazu multipliziert man (1.5) mit q und subtrahiert das Resultate von (1.6). Man erhalt

$$x_{n+1} - qx_n = (1-q)x^* \qquad \Rightarrow \qquad x^* = \frac{x_{n+1}-qx_n}{1-q}.$$

Damit können wir eine neue Iteration definieren:

1. Berechne aus x_0 die Werte $x_1 = f(x_0)$ und $x_2 = f(x_1)$.
2. Berechne
$$q = \frac{x_2-x_1}{x_1-x_0} \qquad \text{und} \qquad x^* = \frac{x_2-qx_1}{1-q}.$$
3. Setze $x_0 = x^*$ und beginne bei 1.

Das neue Verfahren zur Berechnung des Fixpunktes konvergiert jetzt viel schneller, wie die Resulate in Tabelle 1.6 zeigen. Die Konvergenz ist sehr rasch, es scheint als ob sich in jeder Iteration die Anzahl der korrekten Stellen verdoppelt, dass also quadratische Konvergenz vorliegt. Dies lässt sich auch mit Hilfe einer analytischen Rechnung besetätigen. Dazu approximiert man $f(x^* + \delta)$ bis zu Termen zweiter Ordnung, also als

$$f(x^* + \delta) = f(x^*) + f'(x^*) \cdot \delta + \frac{1}{2} f''(x^*) \cdot \delta^2 + \ldots, = x^* + q\delta + s\delta^2$$

verwendet dies für die Analyse des neuen Iterationsverfahrens. Wir schreiben also

$$\begin{aligned} x_0 &= x^* + \delta \\ x_1 &= x^* + q\delta + s\delta^2 \\ x_2 &= x^* + q(q\delta + s\delta^2) + s(q\delta + s\delta^2)^2 = x^* + q^2\delta + qs\delta^2 + sq^2\delta^2 + 2qs^2\delta^3 + s^3\delta^4 \\ &= x^* + q^2\delta + qs(1+q)\delta^2 \end{aligned} \tag{1.7}$$

und wenden den neuen Algorithmus darauf an. Der Quotient ist

$$\begin{aligned} \frac{x_2 - x_1}{x_1 - x_0} &= \frac{x^* + q^2\delta + qs(1+q)\delta^2 - (x^* + q\delta + s\delta^2)}{x^* + q\delta + s\delta^2 - (x^* + \delta)} = \frac{q(q-1)\delta + s(q^2+q-1)\delta^2}{(q-1)\delta + s\delta^2} \\ &\simeq \frac{q(q-1)\delta}{(q-1)\delta} = q \end{aligned}$$

In der zweiten Zeile gehen wir davon aus, dass $\delta^2 \ll \delta$ und dass daher die Terme mit δ^2 im Bruch vernachlässigt werden dürfen. Nache dem zweiten Schritt des Algorithmus Algorithmus ist der neue Wert von x^*

$$\frac{x_2 - qx_1}{1-q} = \frac{x^* + q^2\delta + qs(1+q)\delta^2 - q(x^* + q\delta + s\delta^2)}{1-q} = \frac{(1-q)x^* + qs(1+q)\delta^2 - qs\delta^2}{1-q} \tag{1.8}$$

$$= x^* + \frac{q^2 s}{1-q}\delta^2 \tag{1.9}$$

Im Vergleich zum ursprünglichen Fehler in durch die Iteration in im wesentlichen quadriert worden. Damit ist die quadratische Konvergenz bestätigt.

Ein fairer Vergleich der Konvergenzgeschwindigkeit der usprünglichen Iterationsfolge mit dem neuen Algorithmus sollte die Anzahl der Auswertungen der Funktion f mit zählen. In der ursprünglichen Iterationsfolge wird in jeder Iteration die Funktions einmal ausgewertet, der neue Algorithmus braucht zwei Funktionsauswertungen pro Iteration. In der ursprünglichen Iterationsfolge wurden 52 bit Genauigkeit nach 26 Schritten und damit nach 26 Funktionsauswertungen erreicht. Die neue Folge erreicht Konvergenz in 4 Iterationen und damit 8 Funktionsauswertungen. Auch unter Berücksichtigung der Anzahl Funktionsauswertungen ist das neue Verfahren bedeutend weniger aufwendig.

Trotz dieses spektakulären Erfolgs der Konvergenzbeschleunigung weist das Verfahren für praktische Rechnungen einige entscheidende Mängel auf. Da die Folge gegen x^* konvergiert, werden die Werte x_0, x_1 und x_2 fast gleich gross sein, so dass es zu Auslöschung und damit zu Werten sehr geringer Genauigkeit für q kommen kann. Ein solcher Fehler in q schlägt sich wegen (1.8) sofort auch im nächsten Approximationswert für x^* nieder.

Im Abschnitt 4.2 wird am Beispiel des Romberg-Integrationsverfahrens gezeigt, wie die Konvergenz der Integralberechnung mit der Trapezregel beschleunigt werden kann.

1.5 Numerische Instabilität

Die Untersuchungen zum Verhalten von Iterationsfolgen sind davon ausgegangen, dass alle Rechnung ohne Fehler ausgeführt werden können. Dies ist aber nicht realistisch, in den meisten numerischen Berechnungen müssen Zwischenresultate mit der beschränkten verfügbaren Genauigkeit der Gleitkommaformate gespeichert werden, die die Maschine zur Verfügung stellt. Es ist daher zu untersuchen, wie sich die Konvergenz eines Lösungsverfahrens ändert, wenn es durch Rundungsfehler im Laufe der Rechnung gestört wird.

1.5.1 Eine instabile Quadratwurzel

Die Quadratwurzel von a ist ein Fixpunkt der Funktion $f(x) = x^3/a$ denn $f(\sqrt{a}) = a^{\frac{3}{2}}/a = \sqrt{a}$. Das Konvergenzkriterium Satz 1.1 besagt, dass Konvergenz garantiert ist, wenn der Betrag der Ableitung von $f'(x)$ am Fixpunkt kleiner als 1 ist. Aber

$$f'(x) = \frac{3x^2}{a} \quad \Rightarrow \quad f'(\sqrt{a}) = \frac{3\sqrt{a}^2}{a} = 3,$$

die Iterationsfolge wird also niemals konvergieren. Im Gegenteil, der Fehler wird in jeder Iteration um den Faktor 3 anwachsen.

Der `float`-Gleitkommatyp verewendet eine Mantisse von 23 bit, das niederwertigste Bit hat also die Wertigkeit 2^{-23}. Eine Approximation von $\sqrt{a}$ wird daher, sofern sie nicht exakt ist, beim Speichern als `float`-Zahl einen Fehler von der Grössenordnung 2^{-24} aufnehmen. Nach k Iterationen wird dieser Fehler auf $3^k 2^{-24} = 2^{k \log_2 3 - 24}$ angewachsen sein. Nach $k \geq 24/\log_2 3 = 15.142$ Iterationen ist also der Fehler von der gleichen Grössenordnung wie das gesuchte Resultat.

Wir illustrieren dies mit einer Rechnung[6], deren Resultate in Tabelle 1.7 zusammengestellt sind. Zunächst berechnen wir die Quadratwurzel $\sqrt{2}$ als `float`-Zahl. Dann bilden wir die nächsten Nachbarn, die sich nur in den letzten zwei Bits unterscheiden, die letzten fünf Bits der Mantisse dieser Zahlen sind in der obersten Kopfzeile der Tabelle gezeigt. In der zweiten Kopfzeile sieht man, dass diese Unterschiede so klein sind, dass sie nur Gerade die Rundung in der sechsten dezimalen Nachkommastelle beinflussen können. Der Unterschied dieser Startwerte zum Maschinen-Wert für $\sqrt{2}$ ist in der dritten Kopfzeile dargestellt.

Im unteren Teil der Tabelle wird dann ausgehend von jedem dieser Startwerte die Iterationsfolge mit der Funktion $f(x) = x^3/2$ gebildet. Spätestens nach der Iteration beginnen sich die Werte vom Fixpunkt zu entfernen, nach 16 Iterationen sind die Abweichungen von der Grössenordnung 1 wie in der Überschlagsrechnung weiter oben vorhergesagt.

1.5.2 Numerische Instabilität

Von *numerischer Instabilität* spricht man, wenn ein Berechnungsverfahren allein wegen der unvermeidlichen Rundungsfehler keine sinnvollen Resultate liefern kann.

Beispiel. Es soll das Integral

$$I_n = \int_0^1 \frac{x^n}{x+a}\,dx$$

für festes $a > 1$ und für ganze Zahlen n mit $1 \leq n \leq 15$ berechnet werden. Da im Interval $[0, 1]$ gilt $x^{n+1} > x^n$ ist I_n eine monoton abnehmende Folge. Es ist auch klar, dass $\lim_{n\to\infty} = 0$.

[6] Das C++-Programm hierzu ist `sqrt.cpp` im Verzechnis `buch/chapters/experimente/sqrt` von [3].

	10001	10010	10011	10100	10101
	1.414213	1.414213	1.414214	1.414214	1.414214
	$-2.384 \cdot 10^{-7}$	$-1.192 \cdot 10^{-7}$	0	$1.192 \cdot 10^{-7}$	$2.384 \cdot 10^{-7}$
1	1.414213	1.414213	1.414213	1.414214	1.414214
2	1.414211	1.414212	1.414213	1.414214	1.414216
3	1.414207	1.414210	1.414212	1.414216	1.414220
4	1.414193	1.414204	1.414210	1.414220	1.414232
5	1.414152	1.414184	1.414204	1.414232	1.414268
6	1.414029	1.414126	1.414184	1.414268	1.414377
7	1.413660	1.413951	1.414126	1.414377	1.414705
8	1.412553	1.413427	1.413951	1.414705	1.415687
9	1.409239	1.411856	1.413427	1.415687	1.418640
10	1.399341	1.407152	1.411856	1.418640	1.427533
11	1.370064	1.393135	1.407152	1.427533	1.454550
12	1.285858	1.351915	1.393135	1.454550	1.538706
13	1.063039	1.235429	1.351915	1.538706	1.821534
14	0.600645	0.942808	1.235429	1.821534	3.021912
15	0.108348	0.419024	0.942808	3.021912	13.797982
16	$6.359 \cdot 10^{-4}$	$3.678 \cdot 10^{-2}$	0.419024	13.797982	$1.313 \cdot 10^{3}$
17	$1.286 \cdot 10^{-10}$	$2.489 \cdot 10^{-5}$	$3.678 \cdot 10^{-2}$	$1.313 \cdot 10^{3}$	$1.132 \cdot 10^{9}$
18	$1.063 \cdot 10^{-30}$	$7.710 \cdot 10^{-15}$	$2.489 \cdot 10^{-5}$	$1.132 \cdot 10^{9}$	$7.271 \cdot 10^{26}$
19	0.000000	$2.298 \cdot 10^{-43}$	$7.710 \cdot 10^{-15}$	$7.271 \cdot 10^{26}$	∞
20	0.000000	0.000000	$2.298 \cdot 10^{-43}$	∞	∞

Tabelle 1.7: Iterationsfolge der Funktion $f(x) = x^3/2$ für verschiedene Näherungen des Fixpunktes $x^* = \sqrt{2}$. Die Startwerte unterscheiden sich nur in den letzten zwei Bits ihrer Darstellung als `float`-Gleitkommazahl, die letzten fünf Bits der Mantisse sind in der ersten Kopfzeile dargestellt. In der dritten Zeile stehen die Differenzen zur besten Approximation von $\sqrt{2}$ durch eine `float`-Zahl.

Das Integral ist im Prinzip nicht schwierig zu berechnen, wenn man im Integranden die Polynom-Division ausführt:

$$\frac{x^n}{x+a} = x^{n-1} - ax^{n-2} + a^2x^{n-3} - \cdots + (-1)^n xa^{n-2} + (-1)^{n-1}a^{n-1} + (-1)^n \frac{a^n}{a+x}$$

$$\int_0^1 \frac{x^n}{x+a}\,dx = \frac{1}{n} - \frac{a}{n-1} + \frac{a^2}{n-2} - \cdots + (-1)^n \frac{a^{n-2}}{2} + (-1)^{n-1}a^{n-1} + \log(1+a) - \log a. \qquad \bigcirc$$

Wenn n gross ist, ist dies eine ziemlich aufwendig Rechnung. Da man das Integral für alle n haben will, bietet sich eine Rekursionsformel an. Es ist nämlich

$$\begin{aligned} I_n &= \int_0^1 \frac{x^n}{x+a}\,dx = \int_0^1 \frac{x^{n-1}(x+a-a)}{x+a}\,dx = \int_0^1 x^n - \frac{ax^{n-1}}{x+a}\,dx \\ &= \int_0^1 x^n\,dx - aI_{n-1} = \frac{1}{n+1} - aI_{n-1}. \end{aligned}$$

n	I_n	Rückwärtsiteration
0	0.0953101798043249	0.09531017980432486
1	0.0468982019567507	0.04689820195675140
2	0.0310179804324935	0.03101798043248600
3	0.0231535290083985	0.02315352900847329
4	0.0184647099160155	0.01846470991526711
5	0.0153529008398455	0.01535290084732894
6	0.0131376582682119	0.01313765819337729
7	0.0114805601750236	0.01148056092337000
8	0.0101943982497636	0.01019439076629997
9	0.0091671286134746	0.00916720344811137
10	0.0083287138652537	0.00832796551888631
11	0.0076219522565537	0.00762943572022779
12	0.0071138107677959	0.00703897613105546
13	0.0057849692451174	0.00653331561252233
14	0.0135788789773972	0.00609541530334817
15	−0.0691221231073049	0.00571251363318499
16	0.753721231073049	0.00537486366815009
17	−7.47838878131872	0.00507489273026384
18	74.8394433687428	0.00480662825291712
19	−748.341802108480	0.00456529641819719
20	7483.46802108480	0.00434703581802811

Tabelle 1.8: Instabile Iteration zur Berechnung der Integrale I_n mit $a = 10$. In jedem Schritt wird der Fehler mit a multipliziert. In der dritten Spalte die Resultate der stabilen Rückwärtsiteration.

Wenn man also erst I_0 berechnet hat, kann man mit dieser Rekursionsformel alle anderen I_n berechnen. I_0 ist nicht schwierig zu berechnen, es ist

$$I_0 = \int_0^1 \frac{1}{x+a}\,dx = \log(1+a) - \log a = \log\frac{1+a}{a}.$$

Nehmen wir an, dass ausschliesslich bei der Berechnung von I_0 ein unvermeidlicher Rundungsfehler der Grösse ϵ passiert und dass alle anderen Operationen exakt ausgeführt werden können. Da der Logarithmus eine transzendente Funktion ist, werden fast alle Werte des Logrithmus bei Speichern als Gleitkommazahl gerundet werden müssen. Wir müssen jetzt also die Rekursionsfolge I_n^*

ausgehend von $I_0^* = I_0 + \varepsilon$ bilden:

$$\begin{aligned}
I_0^* &= I_0 + \varepsilon \\
I_1^* &= \frac{1}{2} - a(I_0^*) = \frac{1}{2} - aI_0 - a\varepsilon && = I_1 - a\varepsilon \\
I_2^* &= \frac{1}{3} - a(I_1^*) = \frac{1}{3} - aI_1 + a^2\varepsilon && = I_2 + a^2\varepsilon \\
&\vdots \quad \vdots \\
I_n^* &= \frac{1}{n+1} - aI_{n-1}^* = \frac{1}{n+1} - aI_{n-1} + (-a)^n\varepsilon && = I_{n-1} + (-a)^n\varepsilon.
\end{aligned}$$

In jedem Iterationsschritt wird der Fehler mit a multipliziert. In jedem Schritt gehen als $\log_2 a$ Binärstellen Genaugikeit verloren. Für $a = 10$ bedeutet dies, dass in jedem Schritt eine Dezimalstelle verloren geht.

Die Instabilität rührt daher, dass der Fehler in jedem Schritt mit a multipliziert wird. Indem wir nach I_{n-1} auflösen, erhalten wir die Rekursionsformel

$$I_{n-1} = \frac{1}{a}\Big(\frac{1}{n+1} - I_n\Big).$$

In dieser Rekursion wird in jedem Schritt durch a dividiert, man darf daher davon ausgehen, dass in jeder Iteration $\log_2 a$ Binärstellen Genauigkeit gewonnen werden.

Allerdings muss man für diese Iteration einen der Werte I_n bereits kennen. Man weiss aber, dass $\lim_{n\to\infty} I_n = 0$ ist, d. h. wenn man I_n durch 0 ersetzt macht man einen Fehler exakt von der Grössenordnung I_n. Die Rekursion reduziert diesen Fehler in jedem Schritt um $\log_2 a$ Binärstellen. Da die Werte I_n alle kleiner als 1 sind, macht man also niemals einen Fehler grösser als 1 und nach k Rückwärtsiterationsschritten ist der Fehler kleiner als a^k. Man kann also selbst ohne die Kenntnis eines Startwertes ausreichend genaue Werte von I_n bestimmen. Dies ist in der dritten Spalte von Tabelle 1.8 durchgeführt.

1.6 Kondition

Die Beispiele zur numerischen Instabilität haben deutlich gemacht, dass Instabilität dadurch entstehen kann, dass der Fehler im Laufe der Rechnung grösser wird und dass diese Rechnung vielfach wiederholt wird.

Man sagt, ein Problem sei schlecht konditioniert, wenn eine kleine Änderung der Eingangsdaten eine grosse Änderung der Resultate zur Folge hat. Solche Problem verlangen, dass von Anfang an mit hoher Genauigkeit gerechnet wird, und sie lassen sich schlecht iterieren, da sich die Rundungsfehler mit der Zeit derart aufschaukeln werden, dass man kein Vertrauen mehr in die gefunden Resultate haben kann.

Gut konditioniert Problem sind dagegen solche, bei denen kleine Störungen der Eingangsdaten nur geringe Fehler in den Resultaten zur Folge haben. In einem gut konditionierten Problem werden sich während der Rechnung auftretende Rundungsfehler kaum gravierend auswirken.

Übungsaufgaben

1.1. Betrachten Sie die Funktion

$$f(x) = \frac{1 - \cos x}{x}.$$

a) Berechnen Sie $f(10^{-10})$.

b) Berechnen Sie $\lim_{x \to 0} f(x)$.

c) Was für ein Problem tritt bei der numerischen Berechnung von $f(x)$ für kleine Werte von x auf?

d) Schätzen Sie ab, wie klein x maximal werden darf, damit die naive Berechnung von $f(x)$ gemäss obiger Formel mit `float`- oder `double`-Zahlen einen vom in a) berechneten Grenzwert verschiedenen Wert liefert.

e) Finden Sie eine Berechnungsformel für $f(x)$, die auch für kleine Werte von x funktioniert.

Lösung. a) Mit jedem Datentyp findet man $f(10^{-10}) = 0$.

b) Der Grenzwert kann mit Hilfe der Regel von de l'Hospital berechnet werden:

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{1 - \cos x}{x} = \lim_{x \to 0} \frac{\frac{d}{dx}(1 - \cos x)}{\frac{d}{dx} x} = \lim_{x \to 0} \frac{\sin x}{1} = 0.$$

c) Der Wert von $\cos x$ ist sehr nahe bei 1, daher tritt Auslöschung auf.

d) Die Taylor-Reihe für $\cos x$ ist

$$\cos x = 1 - \frac{1}{2!} x^2 + \frac{1}{4!} x^4 - \frac{1}{6!} x^6 + \ldots$$

Der Funktionswert von $f(x)$ lässt sich nicht mehr von 0 unterscheiden, wenn der zweite Term der Reihe sich nicht mehr von 1 unterscheiden lässt. Dies ist genau der Wert für ε für den verwendeten Datentyp, den man der Tabelle 1.1 entnehmen kann. Daraus leitet man ab:

$$\frac{1}{2} x^2 = \varepsilon \qquad \Rightarrow \qquad x = \sqrt{2\varepsilon}.$$

Für `float` findet man $x \simeq 0.000488281$, für `double` dagegen $x \simeq 2.10734 \cdot 10^{-8}$.

e) Die Halbwinkelformel für die sin-Funktion liefert

$$\sin^2 \frac{x}{2} = \frac{1 - \cos x}{2} \qquad \Rightarrow \qquad f(x) = \frac{1 - \cos x}{x} = \frac{2 \sin^2(x/2)}{x} = \frac{2}{x} \sin^2 \frac{x}{2} =: g(x).$$

Der Ausdruck $g(x)$ leidet nicht unter Auslöschung. ○

Abbildung 1.10 vergleicht die beiden Ausdrücke $f(x)$ und $g(x)$ zur Berechnung der gegebenen Funktion.

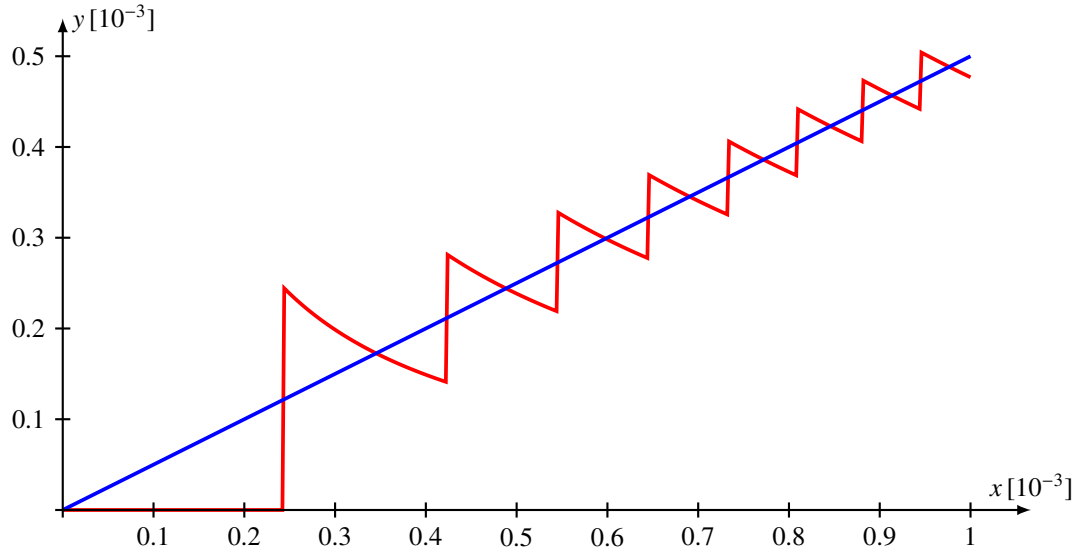


Abbildung 1.10: Numerische Berechnung der Funktion $f(x)$ von Aufgabe 1.1 für kleine Werte von x. Die rote Kurve zeigt die Unzuverlässigkeit der Resultate infolge Auslöschung. Die blaue Kurve zeigt die Berechnung mit der verbesserten Funktion $g(x)$, in der Auslöschung vermieden wird.

1.2. Man kann zeigen, dass die durch

$$x_{n+1} = 2^{n+1}(\sqrt{1 + 2^{-n}x_n} - 1) \tag{1.10}$$

definierte Folge für Startwerte $x_0 > -1$ gegen $\log(x_0 + 1)$ konvergiert.

a) Warum tritt in der Rekursionsformel (1.10) Auslöschung auf?

b) Formulieren Sie (1.10) derart, dass keine Auslöschung auftritt.

Hinweis. Verwenden Sie die Halbwinkelformel

$$\tan\frac{\alpha}{2} = \frac{\sqrt{1 + \tan^2\alpha} - 1}{\tan\alpha}. \tag{1.11}$$

Lösung. a) Da die Folge x_n konvergiert, geht $2^{-n}x_n$ gegen 0 und damit geht $\sqrt{1 + 2^{-n}x_n}$ gegen 1. Die Differenz mit 1 führt dann zu Auslöschung.

b) Im Zähler der rechten Seite der Halbwinkelformel (1.11) für den Tangens kommt genau der Ausdruck vor, der für die Auslöschung verantwortlich ist. Um die Formel verwenden zu können muss

$$\tan\ \alpha = \sqrt{\frac{x_n}{2^n}} \qquad \Rightarrow \qquad \alpha_n = \arctan\sqrt{\frac{x_n}{2^n}}.$$

Die Rekursionsformel besagt dann

$$\frac{x_{n+1}}{2^{n+1}} = \tan\alpha \cdot \tan\frac{\alpha}{2} = \sqrt{\frac{x_n}{2^n}}\tan\Big(\frac{1}{2}\arctan\sqrt{\frac{x_n}{2^n}}\Big)$$

n	x_n nach (1.10)	x_n nach (1.12)
0	0.5000000000	0.5000000000
1	0.4494898319	0.4494897127
2	0.4267277718	0.4267276525
3	0.4159164429	0.4159160256
4	0.4106464386	0.4106463492
5	0.4080467224	0.4080447853
6	0.4067535400	0.4067522287
7	0.4061126709	0.4061080217
8	0.4057922363	0.4057863951
9	0.4056396484	0.4056257606
10	0.4055175781	0.4055454433
11	0.4055175781	0.4055052698
12	0.4052734375	0.4054851830
13	0.4052734375	0.4054751098
14	0.4042968750	0.4054700732
15	0.4023437500	0.4054675400
16	0.3984375000	0.4054662883
17	0.3906250000	0.4054656327
18	0.3750000000	0.4054653049
19	0.3750000000	0.4054651558
20	0.3750000000	0.4054650664
21	0.2500000000	0.4054650068
∞	0.4054650962	0.4054650962

Tabelle 1.9: Auslöschung in der Folge x_n berechnet mit den zwei verschiedenen Rekursionsformeln (1.10) und (1.12) mit Hilfe von `float`-Zahlen.

oder

$$x_{n+1} = 2^{n+1} \sqrt{\frac{x_n}{2^n}} \tan\Big(\frac{1}{2} \arctan \sqrt{\frac{x_n}{2^n}}\Big). \tag{1.12}$$

In dieser Form ist die Iteration ohne Auslöschung durchführbar, wie man in Tabelle 1.9 sehen kann. Allerdings benötigt die Iteration jetzt zusätzlich die Auswertung eines Arkustangens und eines Tangens, was den Rechenaufwand beträchtlich erhöht, selbst auf modernen Floatingpoint-Hardware, die diese Operationen sehr schnell ausführen kann. ○

Kapitel 2

Gleichungen lösen

Im Januar 1535 stellten sich Niccolò Tartaglia und Antonio Maria Fior öffentlich je 30 kubische Gleichungen mit der Form $x^3 + px = q$ oder $x^3 = px + q$ und forderten sich gegenseitig heraus, diese Gleichungen innert 50 Tagen zu lösen. Für moderne Leser scheint es zwischen diesen Gleichungen keinen Unterschied zu geben, aber negative Zahlen waren damals noch nicht in Gebrauch. Fior war ein Schüler von Scipione dal Ferro, der ein Lösungsmethode für einige Typen dieser kubischen Gleichungen aufgestellt hatte. Tartaglia strengte sich darauf hin besonders an und fand 12. Februar 1535 eine Lösungsformel für beide Typen und am darauffolgenden Tag auch eine für die Gleichung $x^3 + q = px$.

Der Wettbewerb ging sehr ungleich aus: mit seiner Lösungsformel konnte Tartaglia alle gestellten Aufgaben lösen, während Fior keine einzige lösen konnte. Solche öffentlichen Wettbewerbe unter Gelehrten waren in der Renaissance durchaus üblich, sie waren Teil des Marketings mit dem Gelehrte bekannt werden und neue Kunden gewinnen konnten. Tartaglia zum Beispiel verdiente seinen Lebensunterhalt vorwiegend als kaufmännischer Rechner und Privatlehrer. Tartaglia ist auch der Autor eines Buches über Ballistik, seine mathematischen Forschungen waren durchaus auch von konkreten Anwendungen motiviert.

Die Lösung der kubischen Gleichung durch Tartaglia und die spätere Verallgemeinerung durch Gerolamo Cardano (1501–1576) sowie die Lösung der Gleichung vierten Grades durch Lodovico Ferrari (1522–1565) waren Lösungsformeln wie sie heute jeder Schüler für die quadratische Gleichung kennelernt. Wie sieht die Lösungsformel für allgemeine Gleichungen fünften Grades aus? Die überraschende Antwort gab 1824 Niels Henrik Abel, er zeigte, dass es eine allgemeine Lösungsformel nicht geben kann. Dies war eines der ersten Resultate in einer langen Reihe von Unmöglichkeitsaussagen. So wissen wir zum Beispiel heute, dass die Funktion e^{-x^2} keine analytische Darstellung mit Hilfe von Potenzfunktionen, Brüchen, Exponential- und Logarithmus-Funktionen hat. Es gibt sogar einen Algorithmus[1] von Risch, mit dem man entscheiden kan, ob eine solche Darstellung für einen vorgegebenen Integranden möglich ist.

Diese Beispiel zeigen uns, dass die Lösung einer Gleichung mit einer Lösungsformel eher die Ausnahme als die Regel darstellt. Gefragt sind daher numerische Methoden, die Gleichungen effizient und zuverlässig lösen können. Dieses Kapitel befasst sich mit den besonderen Schwierigkeiten dieser Aufgabenstellung.

[1] Eigentlich handelt es sich um einen Pseudo-Algorithmus, denn einzelne Schritte des Algorithmus verlangen, dass entschieden werden muss, ob zwei Terme identisch sind. Auch dies ist ein Problem, welches von einem Computer nicht in voller Allgemeinheit gelöst werden kann, allerdings aus ganz anderem Grund.

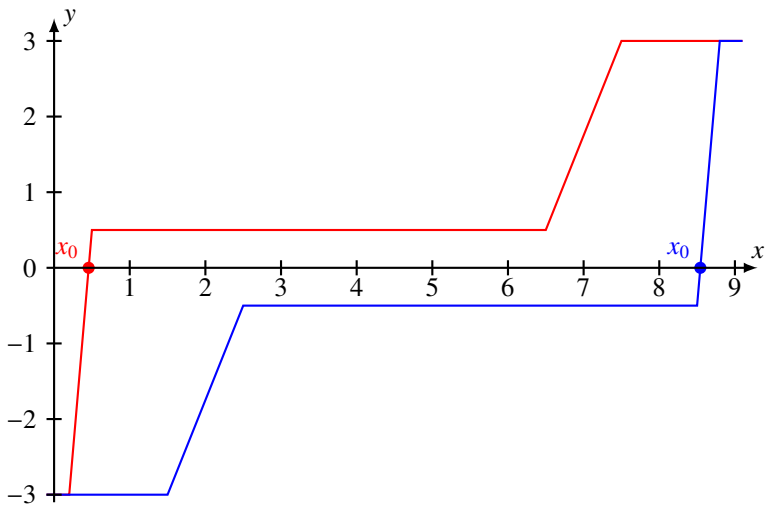


Abbildung 2.1: Die Werte einer stetigen Funktion an den Intervallenden verraten nichts über die Lage einer Nullstelle im Interval. Die beiden Graphen gehören zu Funktionen, die den gleichen Wert an den Intervallenden haben, aber die Nullstelle x_0 liegt an völlig unterschiedlichen Stellen

2.1 Nullstellen von Funktionen

Die Aufgabe, eine Gleichung der Form $f(x) = g(x)$ zu lösen, also ein $x \in \mathbb{R}$ zu finden derart, dass die Gleichung erfüllt wird, ist gleichbedeutend damit, eine Nullstelle der Funktion $f(x) - g(x)$ zu finden. Es ist also gar nicht nötig, allgemeine Gleichungslösungsverfahren zu entwickeln, es reicht völlig aus, Nullstellen finden zu können.

Es muss davon ausgegangen werden, dass die Funktion f nicht einfach algebraisch invertiert werden kann. Sie wird also als Black-Box behandelt, man kann nur Funktionswerte zu vorgegebnen x ermitteln. Bei einem Versuch mit einem Wert x_0 gibt der Funktionswert $f(x_0)$ nur die Information, ob der Versuch erfolgreich war oder nicht. Grundsätzlich können wir nicht einmal schliessen, dass ein grosser Funktionswert bedeutet, dass x_0 weit von einer Nullestelle entfernt liegt. Dazu sind weitere Annahmen über die Funktion notwendig.

In diesem Abschnitt untersuchen wir, wie verschiedene ergänzende Annahmen über die Funktion f die Möglichkeiten erweitern, Nullstellen effizient zu finden. In keinem Fall werden wir allerdings Differenzierbarkeit von f voraussetzen.

2.1.1 Intervallhalbierung

Ist die Funktion stetig, sagt die grösse eines Funktionswertes immer noch nichts darüber aus, wie weit weg von der Nullstelle das Argument entfernt ist. Zwei Werte mit unterschiedlichem Vorzeichen zeigen dagegen klar an, dass sich eine Nullstellen zwischen Argumenten befinden muss. Dies ist der Inhalt des folgenden Spezialfalls des Zwischenwertsatzes.

Satz 2.1 (Zwischenwertsatz für Nullstellen)**.** *Ist die Funktion* $f\colon [a,b] \to \mathbb{R}$ *stetig mit* $f(a) < 0$ *und* $f(b) > 0$*, dann hat* f *eine Nullstelle im Inneren des Intervals.*

Man beachte, dass der Betrag der Funktionswerte an den Intervallenden keine Information darüber liefert, wo im Interval die Nullstelle zu finden ist. Abbildung 2.1 zeigt zwei Funktionen mit identischen Funktionswerten an den Intervallenden aber völlig verschiedener Lage der Nullstellen. Erst zusätzliche Annahmen über die Steigung oder Krümmung der Kurve können die Lage der Nullstelle besser eingrenzen.

Der Zwischenwertsatz 2.1 liefert trotzdem genug Information, um die Nullstelle zu finden. Für die folgende Diskussion nehmen wir der Einfachheit an, dass $f(a) < 0$ und $f(b) > 0$ ist. Wir wissen bereits, dass die Nullstelle im Interval $[a, b]$ liegen muss. Sei $m = \frac{1}{2}(a + b)$ der Mittelpunkt des Intervals. Wenn $f(m) > 0$ ist, können wir schliessen, dass eine Nullstelle im Teilinterval $[a, m]$ liegen muss. Wenn $f(m) < 0$ ist, liegt eine Nullstelle in $[m, b]$. Damit ist ein neues Interval halber Länge gefunden, welches eine Nullstelle von f enthält. Durch Wiederholen dieses Prozesses können wir ein beliebige kleines Interval erhalten, welches eine Nullstelle enthält. Nach dem Intervalschachtelungsprinzip der Analysis definiert dies die Nullstelle.

Satz 2.2 (Intervallhalbierung)**.** *Sei $f\colon [a, b] \to \mathbb{R}$ eine stetige Funktion mit $f(a) < 0$ und $f(b) > 0$. Sei $I_k = [a_k, b_k]$ die Folge von Intervallen rekursiv definiert wie folgt.*

1. *Das Startinterval hat Intervallenden $a_0 = a$ und $b_0 = b$.*
2. *Sei $m_k = \frac{1}{2}(a_k + b_k)$. Das Intervall I_{k+1} ist*

$$I_{k+1} = [a_{k+1}, b_{k+1}] = \begin{cases} [a_k, m_k] & f(m_k) > 0 \\ [m_k, b_k] & f(m_k) < 0. \end{cases}$$

Die Intervalle I_k haben die folgenden Eigenschaften

1. *Die Länge der Intervalle halbiert sich in jeder Iteration: $|I_k| = 2^{-k}(b - a)$.*
2. *Für jedes Intervall gilt $f(a_k) < 0$ und $f(b_k) > 0$.*
3. *Die Schnittmenge $\bigcap_{k=0}^{\infty} I_k$ enthält nur den Wert*

$$x_0 = \lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = \lim_{k\to\infty} m_k,$$

der eine Nullstelle der Funktion f ist.

Beweis. Es ist nur noch die Ausssage über die Konvergenz der Folgen a_k und b_k zu beweisen. Da aber die Intervall-Länge $|I_k| = 2^{-k}(b-a)$ ist, kann man die Entfernung der Folgenglieder voneinander abschätzen. Ist $k = \min\{m, n\}$, dann folgt

$$|a_m - a_n| < 2^{-k}(b - a), \quad |b_m - b_n| < 2^{-k}(b - a), \quad |m_m - m_n| < 2^{-k}(b - a).$$

Wählt man N so gross, dass $2^{-N}(b - a) < \varepsilon$ ist, dann folgt für jedes beliebige $\varepsilon > 0$, dass

$$|a_m - a_n| < \varepsilon, \quad |b_m - b_n| < \varepsilon, \quad |m_m - m_n| < \varepsilon,$$

die Folgen sind also Cauchy-Folgen. □

Die Konvergenzgeschwindigkeit des Intervall-Halbierungs-Verfahrens ist sehr beschränkt. Der Fehler halbiert sich in jeder Iteration, man gewinnt also genau 1 Bit Genauigkeit. Für die volle Genauigkeit eines Gleitkomma-Typs der Maschine braucht man also etwa soviele Iterationen, wie die Mantisse Binärstellen hat, aber auch nur, wenn das erste Bit und der Exponent bereits richtig sind.

k	a_k	b_k	$b_k - a_k$
0	0.00000000	2.00000000	2.00000000
1	1.00000000	2.00000000	1.00000000
2	1.00000000	1.50000000	0.50000000
3	1.00000000	1.25000000	0.25000000
4	1.00000000	1.12500000	0.12500000
5	1.00000000	1.06250000	0.06250000
6	1.00000000	1.03125000	0.03125000
7	1.01562500	1.03125000	0.01562500
8	1.01562500	1.02343750	0.00781250
9	1.01953125	1.02343750	0.00390625
10	1.02148438	1.02343750	0.00195312
11	1.02246094	1.02343750	0.00097656
12	1.02294922	1.02343750	0.00048828
13	1.02319336	1.02343750	0.00024414
14	1.02319336	1.02331543	0.00012207
15	1.02325439	1.02331543	0.00006104
16	1.02328491	1.02331543	0.00003052
17	1.02328491	1.02330017	0.00001526
18	1.02329254	1.02330017	0.00000763
19	1.02329254	1.02329636	0.00000381
20	1.02329254	1.02329445	0.00000191
21	1.02329254	1.02329350	0.00000095
22	1.02329254	1.02329302	0.00000048
23	1.02329278	1.02329302	0.00000024
24	1.02329290	1.02329302	0.00000012

Tabelle 2.1: Bestimmung von $\sqrt[100]{10}$ mit Hilfe des Intervallhalbierungsverfahrens. Die jeweils neue Intervallgrenze ist rot markiert.

Beispiel. Als Beispiel berechnen wir $\sqrt[100]{10}$. Die Potenzfunktion $p(x) = x^{100}$, die wir dazu invertieren müssen, hat extrem grosse Steigung in der Näher der Lösung. Es muss eine Nullstelle der Funktion $f(x) = x^{100} - 10$ gefunden werden. Wegen $f(0) = -10$ und $f(2) \simeq 1.2677 \cdot 10^{30}$ muss das Intervall $[0, 2]$ eine Nullstelle enthalten. Weil die Funktion monoton wachsend ist, ist es auch die einzige. Der Wert $f(2)$ ist klein genug für den `float` Typ, daher kann man die Berechnung in `float` durchführen.

Tabelle 2.1 zeigt den Gang der Berechnung. Die Implementation in C++ verwendet als Abbruchkriterium die Länge des Intervals. Die Iteration endet, wenn die Intervallänge $b_k - a_k$ den ε-Wert erreicht, der für den Typ `float` im Header `limits` definiert ist. Wie erwartet sind soviele Iterationen nötig, wie die Mantisse des `float`-Typs Bits hat.

Das Programm `nullstellen.cpp`[2] implementiert den Algorithmus für jeden Maschinentypen. Damit kann man sehen, dass die berechnete Laufzeit auch für die grösseren Gleitkommatypen durch die Bitlänge der Mantisse gegeben ist. ○

[2]Das Programm kann im Verzeichnis `buch/chapters/experiments/nullstellen` von [3] gefunden werden.

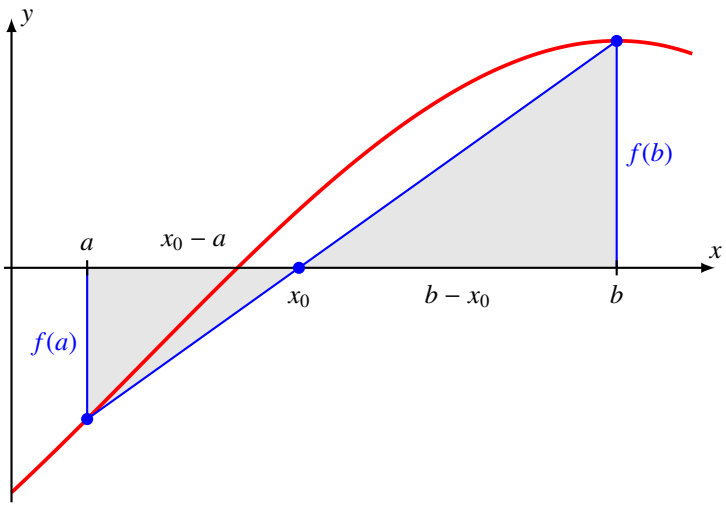


Abbildung 2.2: Bestimmung einer neuen Schätzung x_0 für die Nullstelle mit Hilfe der Sekante. Nach dem Strahlensatz ist $(a - x_0) : f(a) = (b - x_0) : f(b)$, woraus sich x_0 bestimmen lässt (siehe auch (2.1)).

2.1.2 Sekanten-Verfahren

Das Intervallhalbierungsverfahren verwendet nicht mehr als die Stetigkeit der Funktion. Dies führt zu sehr langsamer Konvergenz, weil die relative Grösse der Funktionswerte in den Intervallenden keine Information über die Lage der Nullstelle im Intervall gegen kann. Dazu wird Information darüber benötigt, wie schnell sich Funktionswerte ändern können. Insbesondere muss davon ausgegangen werden können, dass die Steigung der Funktion über das betrachtete Interval nicht zu stark schwankt. Die Ableitung einer differenzierbaren Funktion kann diese Information liefern, es ist aber ausreichend, eine Lipschitz-Bedingung zu verlangen, die wie folgt definiert ist.

Definition 2.3. *Die Funktion $f\colon [a,b] \to \mathbb{R}$ erfüllt eine Lipschitz-Bedingung zum Exponenten α, wenn es eine Konstante L gibt derart, dass*

$$|f(x) - f(y)| < L|x - y|^\alpha$$

ist.

Eine Lipschitz-Bedingung limitiert, wie schnell sich der Wert einer Funktion ändern kann. Eine stetig differenzierbare Funktion erfüllt automatisch eine Lipschitz-Bedingung für $\alpha = 1$, die Umkehrung gilt allerdings nicht.

Sei jetzt $f\colon [a,b] \to \mathbb{R}$ wieder eine Funktion mit $f(a) < 0$ und $f(b) > 0$. Wenn f eine Lipschitz-Bedingung erfüllt, dann geben die Werte $f(a)$ und $f(b)$ zusätzlich Information über die Lage der Nullstelle, die im Interval liegen muss. Da sich Funktionswerte nicht beliebig schnell ändern können, dürfte die Nullstelle näher beim kleineren Funktionswert sein.

Nehmen wir an, die Funktion f sei linear zwischen den beiden Funktionswerten, dann ist die Nullstelle der Schnittpunkt der Geraden durch $(a, f(a))$ und $(b, f(b))$ mit der x-Achse. Der Strahlen-

k	m_k
−1	0.00000000
0	1.50000000
2	0.55041450
3	0.52616811
4	0.52383864
5	0.52362108
6	0.52360088
7	0.52359897
8	0.52359879

Tabelle 2.2: Sekanten-Verfahren zur Bestimmung von $\arcsin \frac{1}{2}$. Die Konvergenz ist unbefähr linear, aber deutlich schneller als beim Intervallhalbierungsverfahren.

satz zeigt

$$\begin{aligned} & (a - x_0) : f(a) = (b - x_0) : f(b) \\ \Leftrightarrow \quad & a\,f(b) - b\,f(a) = x_0(f(b) - f(a)) \\ \Rightarrow \quad & x_0 = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}. \end{aligned}$$

Man kann diese Formel für x_0 auch bekommen als mit den Funktionswerten gewichtetes Mittel der Endpunkte wie folgt. Das Gewicht m_a von a ist $f(b)$, das Gewicht m_b von b ist $-f(a)$, hier verwenden wir $f(a) < 0$. Das gewichtete Mittel ist dann

$$\frac{a\,m_a + b\,m_b}{m_a + m_b} = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)} = x_0. \tag{2.1}$$

Indem die Intervallhalbierung durch eine Teilung des Intervalls im Punkt x_0 ersetzt wird, kann ein neuer Algorithmus gewonnen werden, der möglicherweise schneller konvergiert, weil Intervalle schneller kleiner werden können.

Leider ist dem nicht so. Im Intervallhalbierungsverfahren ist sichergestellt, dass das neue Intervall hochstens halb so gross ist. Eine solche Garantie gibt es bei Verwendung der Formel (2.1) nicht. Ist zum Beispiel im Interval $[a, b]$ die erste Ableitung $f'(x) > 0$ und die zweite Ableitung $f''(x) < 0$, dann ist der neue Teilpunkt immer rechts von der Nullstelle. Das neue Intervall hat daher immer den gleichen linken Endpunkt a, die Folge a_k konvergiert nicht.

Beispiel. Die Funktion $f(x) = \sin x - \frac{1}{2}$ hat im Interval $[0, \frac{\pi}{2}]$ positive Steigung und negative zweite Ableitung. Wie erwartet bleibt a_k konstant, aber b_k konvergiert monton gegen die Nullstelle. In Tabelle 2.2 sind die Resultate der Rechnung gezeigt. Die Werte der Teilpunkte m_k konvergiert linear gegen den gesuchten Wert $\arcsin \frac{1}{2}$. ○

Satz 2.4 (Sekanten-Verfahren)**.** *Sei $f\colon [a, b] \to \mathbb{R}$ eine stetig Funktion mit $f(a) < 0$ und $f(b) > 0$, die eine Lipschitz-Bedingung mit $\alpha = 1$ erfüllt. Setzt man $x_{-1} = a$ und $x_0 = b$ und konstruiert die Folge*

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{2.2}$$

Falls a und b nahe genug bei einer Nullstelle der Funktion f ist, dann konvergiert die Folge x_{n+1} gegen die Nullstelle.

Beweis. TODO □

Die Iterationsformel (2.2) hat den gravierenden Nachteil, dass in der Nähe der Lösung die beiden Grössen im Zähler und im Nenner fast gleich gross sind und damit starke Auslöschung auftreten wird. Die Umformung

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1}f(x_n) - x_{n-1}f(x_{n-1}) + x_{n-1}f(x_{n-1}) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \tag{2.3}$$

$$= x_{n-1} - f(x_{n-1})\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \tag{2.4}$$

ist algebraisch identisch, jedoch wird in der letzten Form die Approximation x_{n-1} um einen Getrag korrigiert, der umso kleiner wird, je besser x_{n-1} die Nullstelle bereits approximiert und damit je kleiner $f(x_{n-1})$ ist. Der Bruch in (2.4) kann jedoch immer noch erratische Werte annehmen und damit die Konvergenz des Verfahrens gefährden.

Man beachte, dass die Bedingung an die Vorzeichen von $f(a)$ und $f(b)$ nur dazu dient, die Existenz einer Nullstelle im Intervall zu garantieren. Ein anderes Problem dieses Verfahrens ist, dass mit genauer werdender Approximation x_n der Nullstelle die Werte $f(x_n)$ und $f(x_{n-1})$ sehr nahe beieinander liegen und damit die Differenz stark von Auslöschung betroffen sein wird. Im Intervallhalbierungsalgorithmus ist dies kein Problem, weil keine Differenzen von Funktionswerten gebildet werden und ausschliesslich das Vorzeichen des Funktionswertes verwendet wird.

2.2 Newton-Verfahren

Die bisher vorgestellten Verfahren zur Bestimmung einer Nullstelle x^* der Funktion f, $f(x^*) = 0$, sind eher langsam. Dafür war nicht mehr als Stetigkeit nötig, um Konvergenz des Verfahrens sicherzustellen.

2.2.1 Analytischer Ansatz für ein quadratisch konvergentes Verfahren

In Abschnitt 1.4.1 wurde dargestellt, dass nach Möglichkeit quadratische Konvergenz angestrebt werden sollte. Quadratische Konvergenz könnte in einer Fixpunktiteration $x_{n+1} = g(x_n)$ erreicht werden, wenn die Ableitung $g'(x^*) = 0$ ist.

Je grösser $f(x_n)$ ist, desto weiter dürfte x_n von der Nullstelle entfernt sein. Wir versuchen daher, die Approximation x_{n+} proportional zum Wert von $f(x_n)$ zu korrigieren mit Hilfe der Funktion

$$x_{n+1} = g(x_n) = x_n - a(x_n) \cdot f(x_n).$$

Die Funktion $a(x_n)$ muss noch bestimmt werden. Sie soll so gewählt werden, dass die Konvergenz quadratisch wird, was mit $g'(x^*) = 0$ erreicht wird.

Die Ableitung von g ist

$$g'(x) = 1 - a'(x)f(x) - a(x)f'(x).$$

An der Stelle x^* gilt

$$g'(x^*) = 1 - a'(x^*)\underbrace{f(x^*)}_{=0} - a(x^*)f'(x^*) = 0$$

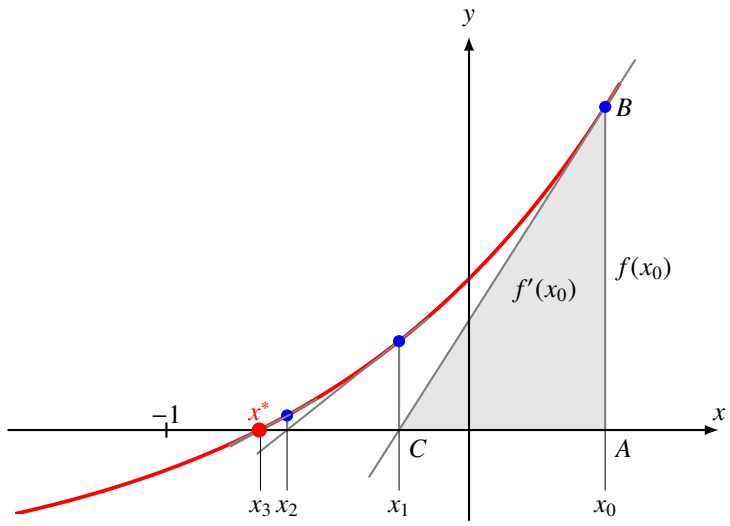


Abbildung 2.3: Graphische Interpretation des Newton-Verfahrens. In jedem Iterationsschritt wird die bisherige Approximation A mit Hilfe einer Tangente vom Funktionswert B zum Punkt C korrigiert, $\overline{AC} = f(x_0)/f'(x_0)$.

$$\Leftrightarrow \qquad 1 = a(x^*)f'(x^*) \qquad \Rightarrow \qquad a(x) = \frac{1}{f'(x)}.$$

Damit finden wir das im folgenden Satz beschriebene Verfahren mit quadratischer Konvergenz.

Satz 2.5 (Newton-Verfahren)**.** *Hat die differenzierbare Funktion f eine Nullstelle x^* und gilt $f'(x^*) \neq 0$, dann konvergiert die Iterationsfolge*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.5}$$

für Startwerte x_0 genügend nahe bei x^* quadratisch gegen die Nullstelle.

Das Sekantenverfahren hat darunter gelitten, dass die Berechnung der der nächsten Approximation mit Hilfe der Formel (2.2) von umso stärker Auslöschung geplagt ist, je näher man bereits an der Lösung ist. Die Iterationsformel (2.5) für die Newton-Iteration hat dieses Problem nicht. In (2.5) wird der aktuelle Wert x_n um einen Korrekturbetrag korrigiert, der proportional zu $f(x_n)$ verkleinert wird. Je genauer der Wert x_n schon ist, desto kleiner wird auch die Korrektur.

2.2.2 Geometrische Interpretation des Newton-Verfahrens

Die Iterationsformel (2.5) lässt sich sehr schön graphisch interpretieren. In Abbildung 2.3 wird die Nullstelle der Funktion $f(x) = e^x - \frac{1}{2}$ mit dem Newton-Verfahren bestimmt. Im Iterationsschritt wird die Approximation x_n korrigiert nach der Formel

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{e^{x_n} - \frac{1}{2}}{e^{x_n}} = x_n - \left(1 - \frac{e^{-x_n}}{2}\right)$$

n	x_n
0	10.00000000000000
1	8.00020000000000
2	6.40064823242493
3	5.12171019598693
4	4.10027465454472
5	3.28729556684556
6	2.64696320430731
7	2.15831219143923
8	1.81881622015378
9	1.63781027109793
10	1.58820394873794
11	1.58490696686523
12	1.58489319270054
13	1.58489319246111
∞	$\sqrt[5]{10} = 1.58489319246111$

Tabelle 2.3: Berechnung von $\sqrt[5]{10}$ mit dem Newton-Verfahren. Der Startwert $x_0 = 10$ ist sehr weit von der Lösung entfernt, so dass es einige Iterationen braucht, bis die Konvergenz quadratisch wird.

Die Korrektur für $n = 0$ ist in Abbildung 2.3 als Grundseite des rechtwinkligen Dreiecks ABC erkennbar. Die Hypothenuse hat die Steigung $f'(x_0)$, daher ist

$$\overline{AC} \cdot f'(x_0) = f(x_0) \qquad \Rightarrow \qquad \overline{AC} = \frac{f(x_0)}{f'(x_0)}.$$

Die vom Newton-Verfahren berechnete Korrektur ist also die optimale Korrektur, die sich berechnen lässt aus Funktionswert und erster Ableitung an der Stelle x_n.

2.2.3 Wurzeln

Als Beispiel berechnen wir die k-te Wurzel einer positiven reellen Zahl a, wir lösen also die Gleichung $x^k = a$. Dies ist gleichbedeutend damit, eine Nullstelle der Funktion $f(x) = x^k - a$ zu bestimmen. Die Ableitung von f ist $f'(x) = kx^{k-1}$, woraus wir die Iterationsformel des Newton-Verfahrens ablesen können:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^k - a}{kx_n^{k-1}} = \frac{1}{k}\left((k-1)x_n + \frac{a}{x_n^{k-1}}\right).$$

Im Falle $n = 2$ finden wir das bereits in Abschnitt 1.4.1 untersuchte, quadratisch Konvergente Verfahren zur Bestimmung der Quadratwurzel wieder. In Tabelle 2.3 ist das Verfahren für $a = 10$, $k = 5$ und $x_0 = a$ gezeigt. Quadratische Konvergenz stellt sich allerdings erst bei x_{10} ein, der Startwert x_0 ist zu weit von der Lösung entfernt.

2.2.4 Newton-Verfahren in $\mathbb{R}^n$

In der bisher beschriebenen Form erlaubt das Newton-Verfahren, Nullstellen von rellwertigen Funktion zu finden. Es eignet sich nicht, Vektorgleichungen zu lösen.

Sei daher im folgenden $f\colon \mathbb{R}^n \to \mathbb{R}^n$ eine Vektorfunktion mit einer Nullstelle x^*, $f(x^*) = 0$, die numerisch gefunden werden soll. Wie in Abschnitt 2.2.1 soll eine Approximation x_n für die Nullstelle x^* proportional zur Grösse von $f(x_n)$ korrigiert werden. Eine skalare Funktion $a(x)$ wird aber im allgemeinen zu wenig allgemein für eine performante Lösung des Problem sein, daher wird für $a(x)$ eine Funktion mit Werten in $\mathrm{GL}_2(\mathbb{R})$ gewählt. Wir setzen also an

$$x_{n+1} = g(x_n) = x_n - a(x_n)f(x_n)$$

und versuchen wie früher $a(x)$ so zu wählen, dass die Ableitung von g an der Stelle x^* verschwindet.

Die Ableitung von $g(x) = x - a(x)f(x)$ ist

$$Dg(x)\cdot h = h - (Da(x)\cdot h)f(x) - a(x)Df(x)\cdot h.$$

An der Stelle $x = x^*$ verschwindet der mittlere Term wegen $f(x^*) = 0$, so dass als Gleichung für $a(x)$

$$0 = h - a(x)Df(x)\cdot h \qquad \Rightarrow \qquad h = a(x)Df(x)\cdot h \qquad \forall h \in \mathbb{R}^n$$

übrigbleibt. Dies ist nur möglich, wenn $a(x)$ die inverse Matrix von $Df(x)$ ist, was den folgenden Satz motiviert.

Satz 2.6 (Newton-Verfahren für Vektorgleichungen)**.** *Hat die Funktion $f\colon \mathbb{R}^n \to \mathbb{R}^n$ eine Nullstelle $x^* \in \mathbb{R}^n$, dann ist die Folge*

$$x_{n+1} = x_n - Df(x_n)^{-1}\cdot f(x_n)$$

für Startwerte x_0 nahe genug an x^* quadratisch konvergent mit Grenzwert x^*.

Beweis. Es ist klar, das x^* ein Fixpunkt der Abbildung

$$g(x) = x - Df(x)^{-1}\cdot f(x)$$

ist. Wir müssen nur noch zeigen, dass der Fehler der Iteration quadratisch abnimmt. Dazu entwickeln wir f um den Punkt ein eine Taylor-Reihe

$$\begin{aligned} f(x^* + \delta) &= f(x^*) + Df(x^*)\cdot\delta + O(|\delta|^2) \\ &= Df(x^*)\cdot\delta + O(|\delta|^2) \end{aligned}$$

wegen $f(x^*) = 0$. Die Iteration ist

$$\begin{aligned} x_{n+1} = x^* + \delta_{n+1} = g(x^* + \delta_n) &= x^* + \delta_n - Df(x_n)^{-1}\cdot f(x_n) \\ &= x^* + \delta_n - Df(x_n)^{-1}\cdot f(x^* + \delta_n) \\ &= x^* + \delta_n - Df(x_n)^{-1}\cdot(Df(x^*)\cdot\delta_n + O(|\delta_n|^2). \end{aligned} \tag{2.6}$$

Um das zu berechnen, muss man auch $Df(x_n)$ noch entwickeln, es ist

$$Df(x_n) = Df(x^* + \delta_n) = Df(x^*) + D^2f(x^*)\cdot\delta_n.$$

Setzt man dies in (2.6) ein, erhält man

$$x_{n+1} = x^* + \delta_{n+1} = x^* + \delta_n - (Df(x^*) + D^2f(x^*)\cdot\delta_n)^{-1}(Df(x^*)\cdot\delta_n + O(|\delta_n|^2)$$

n	r_n	φ_n
0	3.1415926535897931	0.3678794411714423
1	0.8067763763745672	0.5559550343664228
2	0.9030213557712843	1.0884392177149671
3	0.9960918007116625	0.9906298199545209
4	0.9999561001841604	1.0000366265580796
5	0.9999999993292477	0.9999999983920385
6	1.0000000000000000	1.0000000000000000
∞	1.0000000000000000	1.0000000000000000

Tabelle 2.4: Quadratische Konvergenz des Iterationsverfahrens (2.7) zur Bestimmung der Polarkoordinaten

$$\begin{aligned}
&= x^* + \delta_n - (Df(x^*)^{-1} + O(|\delta_n|)) \cdot (Df(x^*) \cdot \delta_n + O(|\delta_n|^2) \\
&= x^* + \delta_n - \underbrace{Df(x^*)^{-1} Df(x^*)}_{=E} \cdot \delta_n + O(|\delta_n|^2) \\
&= x^* + O(|\delta_n|^2).
\end{aligned}$$

Der Fehler $\delta_{n+1} = O(|\delta_n|^2)$ nimmt somit quadratisch ab und damit ist gezeigt, dass die Iterationsfolge quadratrisch konvergiert. □

Beispiel. Es sollen die Polarkoordinaten des Punktes (x, y) als Lösung der Gleichung

$$\begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \qquad \Rightarrow \qquad f(r,\varphi) = \begin{pmatrix} r\cos\varphi - x \\ r\sin\varphi - y \end{pmatrix} = 0$$

bestimmt werden kann.

Die Ableitungsmatrix von f ist

$$Df(r,\varphi) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} \qquad \Rightarrow \qquad Df(r,\varphi)^{-1} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\frac{1}{r}\sin\varphi & \frac{1}{r}\cos\varphi \end{pmatrix}.$$

Die Iterationsformel wird jetzt

$$\begin{aligned}
\begin{pmatrix} r_{n+1} \\ \varphi_{n+1} \end{pmatrix} &= \begin{pmatrix} r_n \\ \varphi_n \end{pmatrix} - \begin{pmatrix} \cos\varphi_n & \sin\varphi_n \\ -\frac{1}{r_n}\sin\varphi_n & \frac{1}{r_n}\cos\varphi_n \end{pmatrix} \begin{pmatrix} r_n\cos\varphi_n - x \\ r_n\sin\varphi_n - y \end{pmatrix} \\
&= \begin{pmatrix} x\cos\varphi_n + y\sin\varphi_n \\ \varphi_n - \frac{x}{r_n}\sin\varphi_n + \frac{y}{r_n}\cos\varphi_n \end{pmatrix} \qquad (2.7)
\end{aligned}$$

Die Resultate der Iteration (2.7) für den Punkt $(x, y) = (\cos 1, \sin 1)$ ist in Tabelle 2.4 gegeben. Die quadratische Konvergenz ist wieder deutlich erkennbar. ○

2.2.5 Der Fall $f'(x^*) = 0$

Im Fall $f'(x^*) = 0$ versagt die Iterationsformel des Newtonverfahrens. Es ist damit zu rechnen, dass das Verfahren sehr langsam oder gar nicht konvergiert. Wir untersuchen dies mit Hilfe einer Entwicklung der Funktion f um den Punkt x^*:

$$f(x^* + \delta) = \frac{1}{2} f''(x^*)\delta^2 + \frac{1}{6} f'''(x^*)\delta^3 + O(\delta^4).$$

Für den Fehler δ_n der Approximation $x_n = x^* + \delta_n$ folgt die Iteration

$$\begin{aligned} x^* + \delta_{n+1} = x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} &= x^* + \delta_n - \frac{\frac{1}{2}f''(x^*)\delta_n^2 + O(\delta_n^4)}{f''(x^*)\delta_n + \frac{1}{2}f'''(x^*)\delta_n^2) + O(\delta^3)} \\ &= x^* + \delta_n - \frac{1}{2}\delta_n \frac{1 + O(\delta_n)}{1 + O(\delta_n)} = x^* + \delta_n - \frac{1}{2}\delta_n(1 + O(\delta_n)) \\ &= x^* + \frac{1}{2}\delta_n + O(\delta_n^2) \\ \Rightarrow \qquad \delta_{n+1} &= \frac{1}{2}\delta_n + O(\delta_n^2) \end{aligned}$$

Der Fehler halbiert sich in jeder Interation. Die Folge $(x_n)_{n\in\mathbb{N}}$ konvergiert also immer noch, aber die Konvergenz ist nur noch linear.

2.2.6 Vergleich mit dem Sekanten-Verfahren

Die Ähnlichkeit des Newton-Verfahrens mit dem Sekantenverfahren ist unübersehbar. Um dies deutlich zu machen, berechnen wir den Grenzfall $x_{n-1} \to x_n$ mit Hilfe der Form (2.4). Beim Grenzübergang $x_{n-1} \to x_n$ geht der Quotient auf der rechten Seite in den Kehrwert der Ableitung $f'(x_n)$ über. Der Grenzfall des Sekantenverfahrens ist daher

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

also das Newtonverfahren.

Der Vorteil des Newton-Verfahrens gegenüber dem Sekanten-Verfahren ist jedoch, dass die Ableitung nicht nur mit Hilfe eines Differenzenquotienten approximiert wird, sondern exakt zur Verfügung steht. Damit ist das Newton-Verfahren nicht anfällig auf die Auslöschung, die die Zuverlässigkeit des Sekantenverfahrens beeinträchtigt.

2.2.7 Nullstellen von Polynomen

Das Newton-Verfahren verlang, dass die Ableitung $f'(x_n)$ genau berechnet werden kann. In einigen Fällen kann dies ein Hindernis für die Anwendung des Verfahrens sein. Polynome sind jedoch einfach genug, dass die Ableitung immer berechnet werden kann. Somit ist das Newton-Verfahren besonders gut geeignet, Nullstellen von Polynomen zu finden. In diesem Abschnitt sei daher

$$f(X) = a_n X^n + a_{n-1}X^{n-1} + \cdots + a_2 X^2 + a_1 X + a_0 \tag{2.8}$$

ein Polynom mit reellen Koeffizienten, $a_k \in \mathbb{R}$. Wir gehen davon aus, dass f eine reelle Nullstelle x^* hat und dass x_0 eine ausreichend genaue Schätzung für die Nullstelle ist.

Berechnung von Funktionswerten

Die übliche Darstellung (2.8) ist nicht die effizientest Form zur Berechnung des Polynomwertes. Die Berechnung der Potenzen x^k für $1 \le k \le n$ benötigt bereits $n-1$ Multiplikationen, dazu kommen $n-1$ Multiplikationen mit Koeffizienten und n Additionen. Zudem besteht die Gefahr von Verschmierung.

Durch Ausklammern möglichst vieler Faktoren x findet man die Formel

$$f(x) = ((\ldots((a_n x + a_{n-1})x + a_{n-2})x + \ldots)x + a_1)x + a_0, \tag{2.9}$$

welche den Funktionswert in genau n Multiplikationen und n Additionen zu berechnen gestattet.

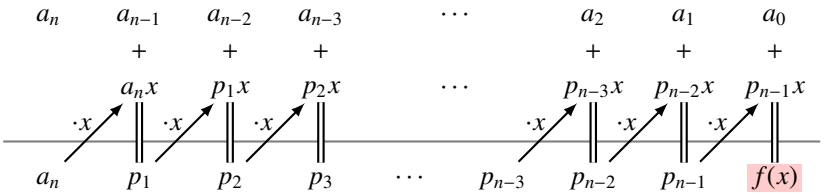
Wir bezeichnen die Teilprodukte in (2.9) mit

$$(\dots((a_n x + a_{n-1})x + a_{n-2})x + \dots)x + a_k = p_{n-k},$$

d. h.

$$\begin{aligned} p_0 &= a_n \\ p_1 &= a_n x + a_{n-1} = p_0 x + a_{n-1} \\ p_2 &= (a_n x + a_{n-1})x + a_{n-2} = p_1 x + a_{n-2} \\ &\vdots \qquad \vdots \\ f(x) = p_n &= p_{n-1} x + a_0. \end{aligned} \tag{2.10}$$

Diese Berechnung lässt sich in der folgenden, *Horner-Schema* genannten Tabelle zusammenfassen.



Beispiel. Wir berechnen den Wert des Polyoms

$$f(X) = X^6 - X^5 + X^4 - X^3 + X^2 - X + 1$$

an der Stelle $X = 2$ mit Hilfe des Horner-Schemas

1	−1	1	−1	1	−1	1
	2	2	6	10	22	42
1	1	3	5	11	21	43

Der Wert in der rechten unteren Ecke stimmt überein mit

$$\begin{aligned} f(2) &= 2^6 - 2^5 + 2^4 - 2^3 + 2^2 - 2 + 1 \\ &= 64 - 32 + 16 - 8 + 4 - 2 + 1 \\ &= 32 + 8 + 2 + 1 = 43. \end{aligned}$$

○

Deflation

Die Bedeutung der Werte $p_0, \dots, p_{n-1}$ lässt sich verstehen, wenn man den Polynom-Divisions-Algorithmus für $f(X)/(X - x)$ ausschreibt. Die Rekursionsformeln (2.10) zeigen, dass die Teilreste

der Division die Koeffizienten p_k haben:

$$\begin{array}{l}
(a_nX^n + a_{n-1}X^{n-1} + a_{n-2}X^{n-2} + \ a_{n-3}X^{n-3} \ + \ldots) : (X - x) = a_nX^{n-1} + p_1X^{n-2} + p_2X^{n-3} + \ldots \\
\underline{a_nX^n \ - \ a_nxX^{n-1}} \\
\qquad\quad p_1X^{n-1} + a_{n-2}X^{n-2} \\
\qquad\quad \underline{p_1X^{n-1} \ - \ p_1xX^{n-2}} \\
\qquad\qquad\quad p_2X^{n-2} + \ a_{n-3}X^{n-3} \\
\qquad\qquad\quad \underline{p_2X^{n-2} \ - a_{n-4}xX^{n-3}} \\
\qquad\qquad\qquad\quad p_3X^{n-3} \qquad \ldots \\
\qquad\qquad\qquad\quad \ldots \qquad\quad \ldots
\end{array} \tag{2.11}$$

Die Koeffizienten p_k sind daher auch die Koeffizienten des Quotienten

$$q(X) = p_0X^{n-1} + p_1X^{n-2} + p_2X^{n-3} + \ldots p_{n-2}X + p_{n-1}.$$

Es gilt daher

$$f(X) = (X - x) \cdot (p_0X^{n-1} + p_1X^{n-2} + p_2X^{n-3} + \ldots p_{n-2}X + p_{n-1}) + f(x).$$

Ist x ein Nullstelle, dann ist das Polynom ist $f(X)$ durch $X - x$ teilbar und $q(X)$ ist der andere Faktor, also $f(X) = (X - x) \cdot q(X)$.

Beispiel. Das Polynom

$$f(x) = x^4 - 25x^2 + 144$$

hat eine Nullstelle $x = 3$ und $x = 4$ als Nulllstellen. Man finde zwei weitere Nullstellen.

Die Polynomdivision mit dem Horner-Schema für $x = 3$

1	0	−25	0	144
	3	9	−48	−144
1	3	−16	−48	0

ergibt

$$q_1(x) = f(x)/(x - 3) = x^3 + 3x^2 - 16x - 48.$$

Weiter bekommt man

$$a_2(x) = f(x)/((x - 3)(x - 4)) = (x^2 + 7x + 12) = (x + 3)(x + 4)$$

aus

1	3	−16	−48
	4	28	48
1	7	12	0

Insbesondere schliesst man, dass $x = -3$ und $x = -4$ die verbleibenden Nullstellen von $f(x) = (x - 3)(x - 4)(x + 4)(x + 3)$ sind. ○

Berechnung der Ableitung

Für das Newton-Vefahren wird ausser dem Funktionswert auch die Ableitung benötigt. Der Funktionswert $r = f(x)$ wird mit dem Horner-Schema sofort gefunden, ebenso der Quotient $q(x)$, Es gilt also

$$f(X) = q(X)(X - x) + r,$$

was wir dazu verwenden können, die Ableitung von f mit Hilfe der Produktregel zu berechnen:

$$f'(X) = q'(X)(X - x) + q(X).$$

An der Stelle $X = x$ ist daher $f'(x) = q(x)$. Da die Koeffizienten von $q(X)$ bereits mit dem Horner-Schema berechnet worden sind, kann $f'(x)$ durch Iteration des Horner-Schemas berechnet werden.

Beispiel. Man berechne den Funktionswert und die Ableitung des Polynoms

$$f(x) = 2x^3 + x + 9$$

an der Stelle $x = 4$.

2	0	1	9
	8	32	132
2	8	33	141
	8	64	
2	16	97	

Man liest $f(4) = 141$ und $f'(4) = 97$ ab. ○

Ist x eine doppelte Nullstelle des Polynoms $f(x)$, dann ist $f'(x) = 0$. Das Horner-Schema kann daher auch dazu verwendet werden, doppelte Nullstellen zu erkennen und damit die Faktorisierung zu vereinfachen, wie das folgende Beispiel zeigt.

Beispiel. Wir betrachten das Polynom

$$f(x) = x^4 - 13x^3 + 41x^2 - 47x + 18.$$

Es hat die Nullstelle $x = 1$, das Hornerschema liefert für den Quotienten

1	−13	41	−47	18
	1	−12	29	−18
1	−12	29	−18	0
	1	−11	18	
1	−11	18	0	
	1	−10		
1	−10	8		

Daraus kann man ablesen, dass $x = 1$ eine doppelte aber nicht eine dreifache Nullstelle ist und dass sich das Polynom schreiben lässt als

$$f(x) = (x - 1)^2 \cdot (x - 11x + 18) = (x - 1)^2(x - 2)(x - 9).$$

Damit ist das Polynom $f(x)$ vollständig faktorisiert. ○

n	x_n	$f(x_n)$	$f'(x_n)$	$q_n(X)$
0	−10.000000	−182.00000	129.00000	$X^2 - X + 19$
1	−8.589148	−38.99232	75.71571	$X^2 + 0.4108524323X + 5.47112751$
2	−8.074164	−4.31028	59.24143	$X^2 + 0.9258356094X + 1.524651051$
3	−8.001407	−0.08021	57.04221	$X^2 + 0.9985933304X + 1.009848595$
4	−8.000000	−0.00005	57.00003	$X^2 + 0.9999990463X + 1.000006676$
5	−8.000000	0.00000	57.00000	$X^2 + X + 1$

Tabelle 2.5: Newton-Verfahren für das Polynom $f(X) = X^3 + 9X^2 + 9X + 8$ mit Hilfe des Horner-Schemas. Als Nebeneffekt bestimmt das Hornerschema in jeder Iteration auch den Quotienten $q_n(x) = f(x)/(x - x_n)$.

Newton-Verfahren für Nullstellen von Polynomen

Da mit dem Hornerschema sowohl Funktionswerte wie auch Ableitungen effizient berechnet werden können, kann es dazu verwendet werden, das Newton-Verfahren für Polynomnullstellen zu implementieren.

Das Horner-Schemas liefert zu jeder Nullstelle x auch immer gleich den Quotienten $q(X) = f(X)/(X - x)$, welches für die Suche nach weiteren Nullstellen verwendet werden kann. Man nennt diesen Prozess *Deflation*.

Beispiel. Die reellen Nullstellen von $f(X) = X^3 + 9X^2 + 9X + 8$ sollen mit Hilfe des Newton-Verfahrens gefunden werden. Die Tabelle 2.5 zeigt die vom Horner-Schema berechneten Funktions- und Ableitungswerte sowie das Quotientenpolynom. Die Konvergenz ist quadratisch und liefert die Nullstelle $x = -8$ sowie den Quotienten $q(X) = X^2 + X + 1$, tatsächlich ist

$$(X + 8)q(X) = (X + 8)(X^2 + X + 1) = X^3 + 9X^2 + 9X + 8 = f(X).$$

Die Diskriminante von $q(X)$ ist $b^2 - 4ac = 1^2 - 4 \cdot 1 \cdot 1 = -3 < 0$, $q(X)$ hat also keine weiteren reellen Nullstellen. ○

2.2.8 Inverse der Normalverteilungsfunktion

Das Integral der Standardnormalverteilungsdichte

$$\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

kann nicht in geschlossener Form berechnet werden und erst recht nicht invertiert werden. Für die Anwendung wird jedoch die Umkehrfunktion benötigt, zu einem Wert $p \in [0, 1]$ ist dasjenige x zu finden, für welches $F(x) = p$ gilt. Im Beispiel auf Seite 16 wurde gezeigt, wie die Fehlerfunktion

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

dazu verwendet werden kann, die Normalverteilungsfunktion

$$\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$

zu berechnen. In diesem Abschnitt soll untersucht werden, wie zu gegebenen Funktionswert p das zugehörige x bestimmt werden kann. Es soll also die Gleichung

$$\Phi(x) = p \qquad \Rightarrow \qquad f(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\Big(\frac{x}{\sqrt{2}}\Big) - p = 0$$

gelöst werden.

Sekantenverfahren

Die mit dem Sekantenverfahren gewonnen Iterationsfolge ist in der rechten Spalte in Tabelle 2.6 dargestellt. Da die erste Ableitung von f relativ langsam ändert, ist die Konvergenz zwar zunächst offenbar nur linear, beschleunigt sich dann aber auf fast quadratische Konvergenz, weil die Sekante die Tangente sehr gut approximiert, sich das Sekantenverfahren in ihrem Konvergenzverhalten also dem Newtonverfahren anzunähern beginnt. Wegen der unvermeidbaren Auslöschung bei der Berechnung der Sekantensteigung wird das Verfahren dann aber instabil, die Iteration bricht ab. Es können für den Typ `long double` nur etwa 18 signifikante Stellen gefunden werden, während das Newton-Verfahren noch drei weitere Stellen ermitteln kann.

Newton-Verfahren

Das Newton-Verfahren benötigt ausser dem Funktionswert auch noch die Ableitung

$$f'(x) = \frac{d}{dx}\frac{1}{\sqrt{\pi}}\int_0^{x/\sqrt{2}} e^{-t^2}\,dt = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Damit wird die Iterationsformel für das Newton-Verfahren:

$$x_{n+1} = x_n - \frac{\sqrt{\pi}}{2\sqrt{2}}e^{2x_n^2}\Big(\frac{1}{2} + \operatorname{erf}(x_n) - p\Big). \tag{2.12}$$

Wie erwartet konvergiert die Iterationsfolge quadratisch für geeignete Startwerte (siehe Tabelle 2.6). Der Startwert $x_0 = 0$ funktioniert für jedes beliebige p. Bei weiter von 0 entfernten Starwerte läuft man Gefahr, dass die Iteration zu betragsmässig grossen Werten x springt, was dann zu einem Überlauf führt.

2.3 Homotopie-Verfahren

Der Erfolg des Newton-Verfahrens hängt entscheidend von der Qualität der Anfangsschätzung x_0 ab. Allerdings ist es oft nicht einfach, eine solche Schätzung zu produzieren. Die folgende Idee kann dabei helfen.

Oft ist ein schwieriges Problem ein "deformierte" Variante eines weniger schwierigen Problems. Der Begriff der Homotopie

Definition 2.7. *Zwei Funktionen $f_0(x)$ und $f_1(x)$ heissen homotop, wenn es eine stetige Funktion*

$$F\colon \mathbb{R}\times I : (x,t) \mapsto F(x,t)$$

mit $I = [0.1]$ gibt derart, dass $f_0(x) = F(x,0)$ und $f_1(x) = F(x,1)$. Die partielle Funktion $x \mapsto F(x,t)$ für $t \in I$ wird auch mit f_t bezeichnet: $f_t(x) = F(x,t)$.

k	x_n nach Newtonverfahren	x_n nach Sekantenverfahren
0	0.00000000000000000000	0.00000000000000000000
1	1.12798272358394999736	1.00000000000000000000
2	1.50523868934241237280	1.31831529614238960517
3	1.63077255164383121513	1.53246084663801306026
4	1.64469272791792957300	1.62023672225157423321
5	1.64485360566334308020	1.64272141855681471658
6	1.64485362695147202343	1.64481100638619422225
7	1.64485362695147239662	1.64485355229014184382
8	1.64485362695147239662	1.64485362694885539842
9		1.64485362695147239597

Tabelle 2.6: Newton-Iteration und Sekantenverfahren zur Bestimmung der Inversen der Verteilungsfunktion $\Phi(x)$ der Normalverteilung, berechnet mit dem Typ `long double`. Das Sekantenverfahren ist auch sehr schnell, da die Sekante bereits sehr gut mit der vom Newton-Verfahren verwendeten Tangente übereinstimmt. Wegen Auslöschung kann es allerdings nicht die volle Genauigkeit erreichen.

Beispiel. Die Kepler-Gleichung ist

$$M = E - e \sin E,$$

wobei M gegeben und E gesucht ist. Dazu gehört die Funktion

$$f(E) = M - E + e \sin E.$$

Der Fall $e = 0$ ist ein trivial einfaches Problem, $E = M$ ist Nullstelle der Funktion

$$f_0(E) = M - E.$$

Eine Homotopie zwischen f_0 und $f_1 = f$ ist

$$F(x, t) = M - E + et \sin E.$$ ○

Eine Homotopie kann dazu verwendet werden, Startwerte für das Newtonverfahren zu liefern. Ist $x_0(t)$ eine Nullstelle der partiellen Funktion $x \mapsto F(x, t)$, dann kann $x_0(t)$ als Startwert zur Bestimmung einer Nullstelle von der partiellen Funktion $F(x, t')$ für $|t - t'| < \varepsilon$ dienen. Ist F differenzierbar bezüglich x, dann können einige Iterationen des Newton-Verfahrens aus dem Startwert $x_0(t)$ eine gute Lösung für $x_0(t')$ sein. Damit lässt sich der folgende Algorithmus konstruieren:

1. Starte mit der exakten Lösung $x_0 = x_0(0)$ und $t = 0$
2. Inkrementiere t um Δt
3. verbessere x_0 durch einige Iterationen des Newton-Verfahrens zu einer Nullstelle von $f_t(x)$.
4. Wiederhole Schritte 2 und 3 bis $t = 1$.

Auf diese Weise kann sichergestellt werden, dass jede Iteration des Newton-Verfahrens mit einem guten Schätzwert startet, wenn auch nur für eine immer bessere Approximation des eigentlichen Problems.

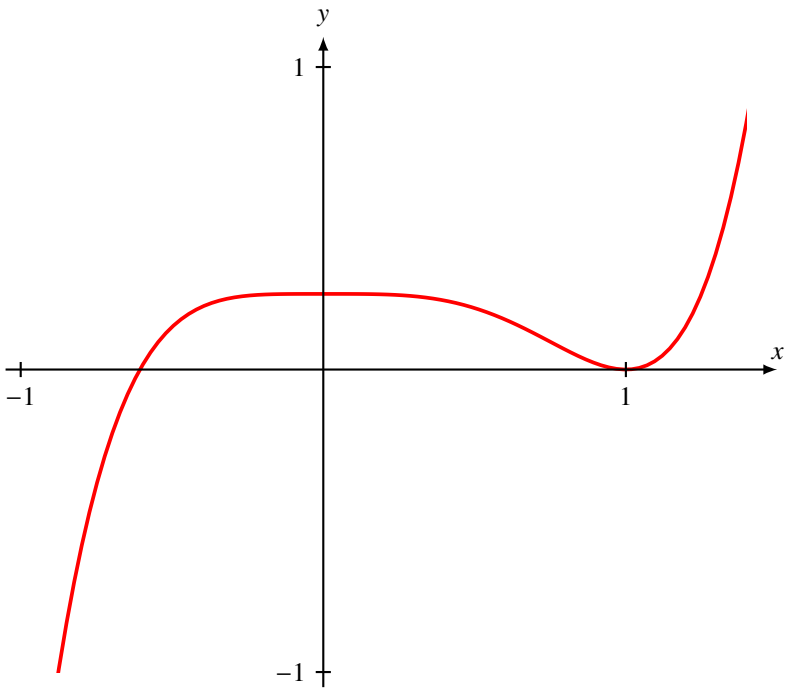


Abbildung 2.4: Graph zur Aufgabe 2.1, Nullstelle des Polynoms (2.13).

Übungsaufgaben

2.1. Finden Sie eine Lösung der Gleichung

$$x^5 - \frac{5}{4}x^4 + \frac{1}{4} = 0 \tag{2.13}$$

mit Hilfe des Newton-Verfahrens. Welche Konvergenzgeschwindigkeit stellen Sie fest?

Lösung. Die Iterationsformel für das Newton-Verfahren braucht die Ableitung von $f(x) = x^5 - \frac{5}{4}x^4 + \frac{1}{4}$, also

$$f'(x) = 5x^4 - 5x^3 = 5(x-1)x^3,$$

insbesondere hat $f'(x)$ eine Nullstelle bei $x = 1$, was in der Newton-Iteration zu Schwierigkeiten führen könnte. Ebenso ist 0 eine Nullstelle der Ableitung, so dass 0 und 1 ganz bestimmt keine guten Startwerte für die Newton-Iteration sind.

Die Iteration lautet jetzt

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^5 - \frac{5}{4}x_n^4 + \frac{1}{4}}{5(x_n-1)x_n^3}.$$

Um eine Lösung zu finden, braucht man jetzt noch einen guten Startwert. Der Graph in Abbilddung 2.4 zeigt, dass 1 eine doppelte Nullstelle ist, und dass es noch eine weitere Nullstelle in der Nähe von $x_0 = -0.5$ gibt.

Die Iteration liefert die folgenden Werte:

n	x_n
0	−0.500000000000000
1	−0.650000000000000
2	−0.610646335912608
3	−0.605893367985182
4	−0.605829597525286
5	−0.605829586188268
6	−0.605829586188268

Man hat also in 5 Iterationsschritten ein Resultate mit 15 Stellen Genaugikeit erhalten, quadratische Konvergenz ist klar sichtbar.

Die Nullstelle bei $x = 1$ macht dem Newton-Algorithmus dagegen etwas Mühe. Die Iteration mit Startwert $x_0 = 0.5$ liefert

n	x_n
0	0.500000000000000
1	1.15000000000000
2	1.08416125585600
3	1.04522243974522
4	1.02356853334768
5	1.01205249984377
6	1.00609759001936
7	1.00306721671745
8	1.00153829071860
9	1.00077032580427
10	1.00038545926066
11	1.00019280387677
12	1.00009642051980
13	1.00004821490799
14	1.00002410861546
15	1.00001205459851
16	1.00000602736919
17	1.00000301369634
18	1.00000150685215
19	1.00000075342128
20	1.00000037677627

Die Konvergenz ist wie erwartet nur linear.

Die Ableitung $f'(x) = 4x^3(x - 1)$ der Funktion $f(x)$ hat an der Stelle $x = 0$ eine dreifache Nullstelle, was auch daran erkennbar ist, dass $f(x)$ abgesehen von der Konstante von vierter Ordnung ist. Der Graph von $f(x)$ ist daher in einer Umgebung von 0 sehr flach. Startwerte x_0 in der Umgebung von 0 führen daher automatisch zu einem Wert x_1 sehr weit weg vom Nullpunkt. Zum Beispiel führt $x_0 = -10^{-4}$ auf $x_1 = -49995000499.9501$. Es braucht dann 114 Iterationen, bis x_n nahe genug bei der negativen Nullstelle liegt, dass man quadratische Konvergenz beobachten kann. ○

Kapitel 3

Interpolation

Der Satz von Stone-Weierstrass garantiert, dass jede stetige Funktione auf einem Interval beliebig genau durch Polynome approximiert werden kann. Polynome sind effizient berechenbar, es ist daher naheliegend, komplizierte Funktione durch Polynome zu approximieren, die möglichst viele für die vorliegende Anwendung relevante Eigenschaften mit der Funktion gemeinsam habe.

Leider sagt der Satz von Stone-Weierstrass nichts darüber, wie solche Polynome gefunden werden könnten. Ziel dieses Kapitels ist daher, einige Möglichkeiten zusammenzustellen, solche Approximationspolynome zu finden und insbesondere auch ihre Fehler abzuschätzen.

3.1 Lineare Interpolation und Polygonzüge

Oft sind von einer Funktion nur einzelne Werte bekannt, doch meist reicht dies, den ungefähren Verlauf ihres Graphen zu erahnen. Unter der Annahme, dass sich die Funktion zwischen den bekannten Werten nicht zu "wild" verhält, können Werte zwischen den bekannten Werten abgeschätzt werden. In diesem Abschnitt soll daher die folgende Aufgabe gelöst werden.

Aufgabe 3.1. *Gegeben $n+1$ Stützstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

im Intervall $[a, b]$ und bekannte Funktionswerte f_i, $0 \le i \le n$, finde eine stetige Funktion $f\colon [a, b] \to \mathbb{R}$ derart, dass $f(x_j) = f_j$ für $0 \le j \le n$.

3.1.1 Lineare Interpolation

Die Aufgabe (3.1) ist zu wenig präzise gestellt, es gibt unendlich viele Lösungen. Es müssen daher zusätzliche Bedingungen an die gesuchte Funktion f gestellt werden, damit die Lösung eindeutig bestimmt wird. Ein mögliche solche Bedingung ist, dass die Funktion f in jedem Teilintervall zwischen aufeinanderfolgenden Stützstellen linear ist. In diesem Fall haben die Werte ausserhalb des Intervalls offenbar keinen Einfluss auf den Verlauf im Inneren des Intervalls, es genügt also das Problem mit nur zwei Stützstellen zu lösen.

Satz 3.2. *Die einzige auf dem Intervall $[x_i, x_{i+1}]$ definierte lineare Funktion $f(x)$ mit Funktionswerten $f(x_i) = f_i$ und $f(x_{i+1}) = f_{i+1}$ hat die Werte*

$$f(x) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i) + f_i = \frac{x - x_{i+1}}{x_i - x_{i+1}} f_i + \frac{x_i - x}{x_i - x_{i+1}} f_{i+1} = f_i l_i(x) + f_{i+1} l_{i+1}(x) \quad x \in [x_i, x_{i+1}] \tag{3.1}$$

mit den linearen Funktionen

$$l_i(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} \qquad \text{und} \qquad l_{i+1}(x) = \frac{x_i - x}{x_i - x_{i+1}}.$$

Beweis von Satz 3.2. Die beiden linearen Funktion $l_i(x)$ und $l_{i+1}(x)$ haben an den Intervallenden die speziellen Werte

$$\begin{aligned} l_i(x_i) &= 1 & l_{i+1}(x_i) &= 0 \\ l_i(x_i) &= 0 & l_{i+1}(x_i) &= 1, \end{aligned}$$

wie man durch Einsetzen unmittelbar bestätigen kann. Die zweite Form in (3.1) ist daher als Linearkombination

$$f(x) = f_i l_i(x) + f_{i+1} l_{i+1}(x)$$

linearer Funktionen wieder linear, und dank der speziellen Werte von l_i und l_{i+1} folgt unmittelbar, dass

$$\begin{aligned} f(x_i) &= f_i l_i(x_i) + f_{i+1} l_{i+1}(x_i) = f_i \cdot 1 + f_{i+1} \cdot 0 = f_i \\ f(x_{i+1}) &= f_i l_i(x_{i+1}) + f_{i+1} l_{i+1}(x_{i+1}) = f_i \cdot 0 + f_{i+1} \cdot 1 = f_{i+1} \end{aligned}$$

gilt.

Der Bruch

$$m = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}$$

ist die Steigung der Geraden durch die Punkte (x_i, f_i) und $x_{i+1}, f_{i+1})$. Der erste Ausdruck auf der rechten Seite in (3.1) ist also $f(x) = m(x - x_i) + f_i$, dies ist die Gleichung einer Geraden mit Steigung m durch den Punkt (x_i, f_i), sie verläuft natürlich auch durch den Punkt (x_{i+1}, f_{i+1}). □

3.1.2 Polygonzüge

Wendet man das Resultate 3.2 auf jedes Teilintervall an, entsteht eine Interpolationsfunktion, deren Graph ein Polygonzug ist. Eine solche Funktion lässt sich einfacher beschreiben mit Hilfe der Interpolationsfunktionen mit den speziellen werten

$$l_i(x_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases} \tag{3.2}$$

Die Funktionen sind in Abbildung 3.1 für die Stützstellen $x_0 = 0$, $x_1 = 1, \ldots, x_6 = 6$ dargestellt. Die lineare Interpolationsfunktion kann jetzt als Linearkombination der Funktionen l_i geschrieben werden:

$$f(x) = \sum_{j=0}^{n} f_j l_j(x). \tag{3.3}$$

Diese Lösung des Interpolationsproblems kann für alle weiteren Interpolationsansätze in diesem Kapitel als Vorlage dienen. Es geht nämlich nicht darum, eine Interpolationsfunktion in geschlossener

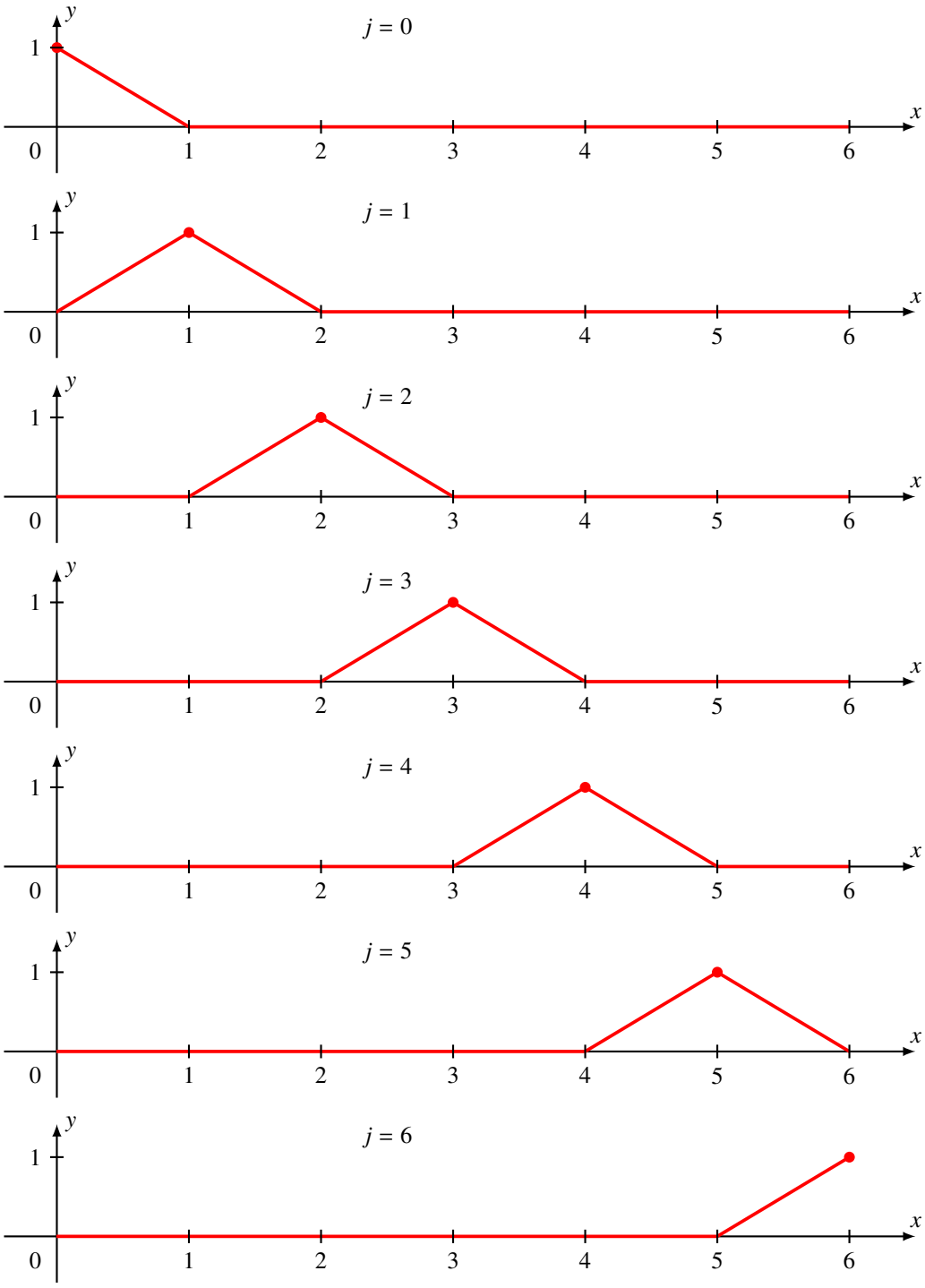


Abbildung 3.1: Basisfunktionen für die lineare Interpolation einer Funktion $f\colon [0,6] \to \mathbb{R}$ mit Stützstellen $0, 1, \ldots, 6$

Form hinzuschreiben, dies ist für die Funktionen $l_j(x)$ ohnehin nicht möglich. Es ist nur nötig, dass Funktionswerte $f(x)$ effizient berechnet werden können. Die letztere Aufgabe ist gelöst, wenn man die Funktionen $l_j(x)$ effizient berechnen kann. Es ist dann nur noch die Linearkombination (3.3) zu bilden.

Die Wahl der Funktionen $l_j(x)$, die natürlich die Bedingungen (3.2) erfüllen müssen, bestimmt die Eigenschaften der Interpolationsfunktion, die nach (3.3) gebildet wird. In diesem Abschnitt waren die l_j stückweise lineare Funktionen, also auch die Funktion $f(x)$. Im nächsten Abschnitt sollen die Funktionen Polynome sein, also wird die Interpolationsfunktion ebenfalls ein Polynom sein.

3.2 Interpolationspolynom

In diesem Abschnitt wird das folgende Problem gelöst.

Aufgabe 3.3 (Interplations-Polynom)**.** *Gegeben Stützstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

und Funktionswerte $f_i, 0 \leq i \leq n$, finde ein Polynome $l(x)$ mit der Eigenschaft $l(x_i) = f_i$ für alle $i = 0, 1, \ldots n$.

Gegeben sind also $n + 1$ Bedingungen, die das Polynom erfüllen muss. Abgesehen von trivialen Fällen wie dem Null-Polynom, muss ein Polynom im Allgemeinen mindestens den Grad n haben, damit alle Bedingungen durch geeignete Wahl der $n + 1$ Koeffizienten erfüllt werden können. Man könnte das Polynom nämlich in der Form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

ansetzen und die Stützstellen einsetzen. Lösung des Gleichungssystem

$$\begin{array}{ccccccc} a_N x_0^N + & a_{N-1} x_0^{N-1} + & \ldots + & a_1 x_0 + & a_0 x_0 = f_0 \\ a_N x_1^N + & a_{N-1} x_1^{N-1} + & \ldots + & a_1 x_1 + & a_0 x_1 = f_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \quad \vdots \\ a_N x_n^N + & a_{N-1} x_n^{N-1} + & \ldots + & a_1 x_n + & a_0 x_n = f_n \end{array} \tag{3.4}$$

liefert dann die gesuchten Koeffizienten. Dieser Weg ist allerdings sehr aufwendig, die Lösung eines linearen Gleichungssystems mit dem Gauss-Algorithmus benötigt $O(n^3)$ Operationen. Die sehr spezielle Struktur des Gleichungssystems sollte ermöglichen, das Polynom $l(x)$ auf direkterem Weg zu ermitteln.

3.2.1 Bestimmung des Interpolationspolynoms

Das allgemeine Interpolationsproblem kann leicht gelöst werden, wenn das folgende spezielle Interpolationsproblem gelöst ist.

Aufgabe 3.4 (Spezielle Interpolationspolynome)**.** *Gegeben die Stützstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

finde Polynome l_j vom Grad n derart, dass

$$l_j(x_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \textit{sonst.} \end{cases}$$

Jedes der Interpolationspolynome l_j hat Grad n, also hat auch eine beliebige Linearkombination den Grad höchstens n. Die Linearkombination

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

ist das gesuchte Interpolationspolynom, wie Einsetzen von x_i in

$$p(x_i) = \sum_{j=0}^{n} f_j l_j(x_i) = \sum_{j=0}^{n} f_j \delta_{ij} = f_i$$

bestätigt.

Beispiel. Ein besonders einfacher Fall ist $n = 1$. Gesucht ist eine lineare Funktion $l(x) = a_1 x + a_0$ derart, dass $l(x_0) = f_0$ und $l(x_1) = f_1$. Polynome l_0 und l_1 können leicht angegeben werden:

$$l_0(x) = \frac{x_1 - x}{x_1 - x_0} \qquad \text{und} \qquad l_1(x) = \frac{x - x_0}{x_1 - x_0}$$

haben die die geforderten Eigenschaften. Die gesuchte Interplationsfunktion ist daher

$$p(x) = \frac{x_1 - x}{x_1 - x_0} f_0 + \frac{x - x_0}{x_1 - x_0} f_1 = x \frac{f_1 - f_0}{x_1 - x_0} + \frac{x_1 f_0 - x_0 f_1}{x_1 - x_0}.$$

Der Koeffizient von x ist wie erwartet die Steigung der Geraden durch die Punkte (x_0, f_0) und (x_1, f_1). ◯

Ein Polynom vom Grad $n + 1$, welches in *allen* Stützstellen verschwindet, ist leicht zu finden, es ist

$$(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{n-1})(x - x_n).$$

Ein Polynom, welches nur an der Stützstelle x_j *nicht* verschwindet, ensteht, indem man den Faktor $(x - x_j)$ weglässt, es hat den Grad n. Wir führen dafür die Notation

$$(x - x_0)(x - x_1)(x - x_2) \cdots \widehat{(x - x_j)} \cdots (x - x_{n-1}(x - x_n),$$

der Hut bedeutet, dass dieser Faktor weggelassen werden soll. Allerdings hat dieses Polynom nicht den geforderten Wert 1, man muss es also noch mit einer geeigneten Konstante multiplizieren. Das gesuchte Polynom $l_j(x)$ hat daher die Form

$$l_j(x) = c_j(x - x_0)(x - x_1)(x - x_2) \cdots \widehat{(x - x_j)} \cdots (x - x_{n-1}(x - x_n).$$

Einesetzen von x_j ergibt

$$l_j(x_j) = 1 = c_j(x_j - x_0)(x_j - x_1)(x_j - x_2) \cdots \widehat{(x_j - x_j)} \cdots (x_j - x_{n-1})(x_j - x_n),$$

die Konstante c_j ist daher

$$c_j = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{1}{x_j - x_i}.$$

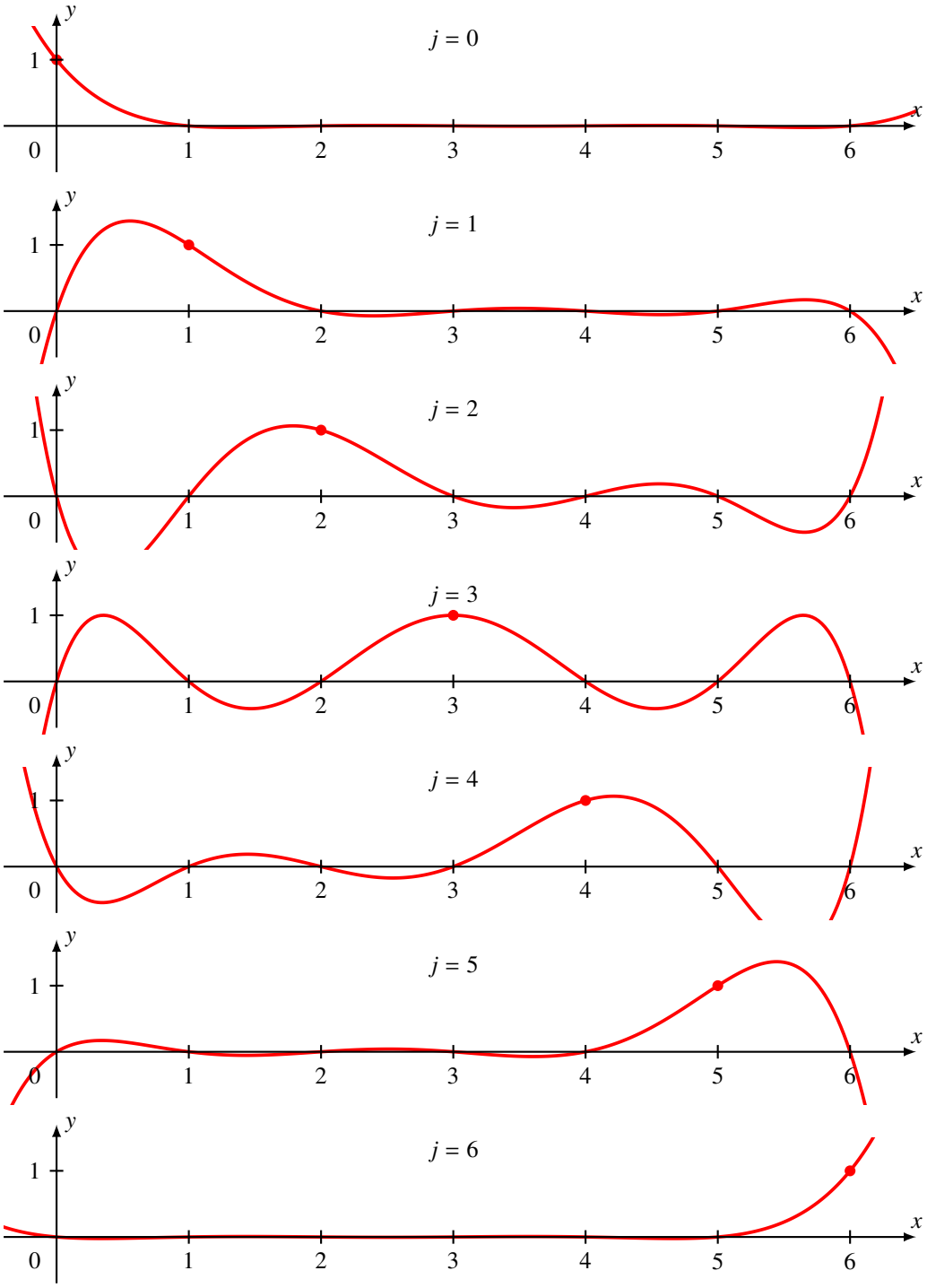


Abbildung 3.2: Polynome $l_j(x)$, welche des spezielle Interpolationsproblem 3.4 lösen.

Beispiel. Man finde ein Polynome, welches $l(0) = l(1) = 0$ und $l(\frac{1}{2}) = 1$ erfüllt. Wegen $f_0 = f_2 = 0$ ist nur das Polynome l_1 zu ermitteln, es ist

$$l(x) = l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{x(x-1)}{\frac{1}{2}(\frac{1}{2} - 1)} = \frac{1}{4}x(1 - x). \qquad \bigcirc$$

3.2.2 Fehler von Approximationspolynomen

Getreu der Maxime, dass wir zu jeder numerischen Lösungsformel auch Informationen über die zu erwartenden Fehler brauchen, entwickeln wir in diesem Abschnitt die Theorie des Fehlers der Approximationspolynome. Wir müssen zu diesem Zweck einen kleinen Ausflug in die Analysis unternehmen in einen Bereich, der im Unterricht manchmal etwas zu kurz kommt.

Wenn die Ableitung einer Funktion in einem Interval klein ist, dann werden auch die Funktionswerte im Inneren dieses Intervals nicht gross von den Werten am Rand abweichen können. Eine grosse Abweichung würde ja automatisch eine Steigung einer Sekanten und damit auch eine grosse Steigung einer Tangenten zur Folge haben. Dies ist die Idee, die den nachfolgend entwickelten Fehlerabschätzungen zu Grunde liegt.

Der Zwischenwertsatz

Der Ausgangspunkt aller nachfolgenden Überlegungen ist die intuitiv anschauliche Tatsache, dass eine stetige Funktion keine Sprünge macht.

Satz 3.5. *Eine auf dem Intervall $[a, b]$ stetige Funktion nimmt jeden Wert im Interval $[f(a), f(b)]$ an. Anders ausgedrückt, für jedes y zwischen $f(a)$ und $f(b)$ gibt es ein x zwischen a und b derart, dass $y = f(x)$.*

Dieser Satz war natürlich bereits die Grundlage des Verfahrens der Interval-Halbierung, mit welchem wir in Abschnitt 2.1.1 Gleichungen gelöst haben. Wenn die Funktion an den Intervallenden verschiedene Vorzeichen hat, dann muss es eine Nullstelle im Inneren des Intervalls geben. Die Intervallhalbierung hat in jedem Schritt ein neues Intervall konstruiert, das die Nullstelle enthielt.

Der Satz von Rolle

Der Satz von Rolle erweitert den Zwischenwertsatz auf die Ableitung einer differenzierbaren Funktion an (Abbildung 3.3).

Satz 3.6 (Rolle)**.** *Sei f eine auf dem Interval $[a, b]$ nicht konstante, stetig differenzierbare Funktion mit $f(a) = f(b)$, dann gibt es einen Punkt $\xi \in (a, b)]$ im Inneren des derart, dass $f'(\xi) = 0$.*

Der Satz von Rolle ist eine selbstverständlichkeit, wenn die Ableitung $f'(x)$ stetig ist, doch dies wird nicht vorausgesetzt, es wird nur verlangt, dass die Ableitung existiert. Ausserdem macht der Satz eine Aussage darüber, dass die Zwischenstelle ξ im Inneren des Intervals sei.

Beweis. Eine stetige Funktion hat auf dem kompakten Interval $[a, b]$ mindestens ein Maximum und ein Minimum. Da die Funktion nicht konstant ist, ist das Maximum oder das Minimum von $f(a)$ verschieden. Wir nehmen an $\xi \in [a, b]$ sei ein Maximum mit dieser Eigenschaft, das Argument für das Minimum ist völlig analog. Wegen $f(\xi) > f(a)$ ist ξ ein Punkt im Inneren des Intervals, also $\xi \in (a, b)$.

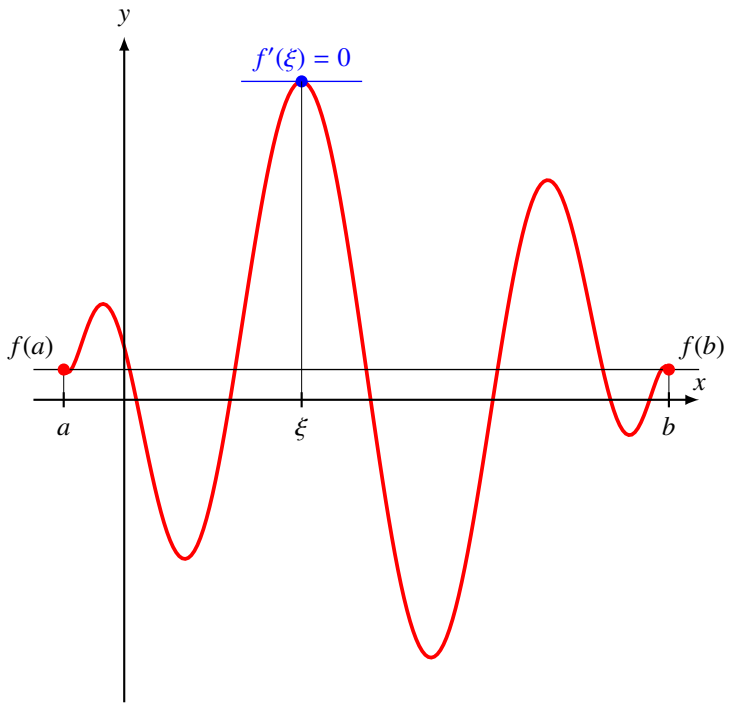


Abbildung 3.3: Satz von Rolle: eine nicht konstante differenzierbare Funktion, die an den Enden eines Intervals den gleichen Funktionswert hat, hat im Inneren des Intervals eine Stelle ξ mit Ableitung 0.

Wegen $f(\xi) \geq f(x) \forall x \in [a, b]$ folgt dann

$$f'(\xi) = \lim_{h \to 0+} \frac{f(\xi + h) - f(\xi)}{h} \leq 0$$
$$f'(\xi) = \lim_{h \to 0-} \frac{f(\xi + h) - f(\xi)}{h} \geq 0.$$

Da f differenzierbar ist, müssen diese beiden Grenzwerte übereinstimmen, also ist $f'(\xi) = 0$. □

Nullstellen und der Satz von Rolle

Satz 3.7. *Ist f eine differenzierbare Funktion auf dem Intervall $[a, b]$ mit Nullstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

die auf keinem Teilintervall $[x_i, x_{i+1}]$ konstant ist, dann hat f' im Inneren jedes Teilintervalls $[x_i, x_{i+1}]$ eine Nullstelle.

Das Polynom

$$l(x) = (x - x_0)(x - x_1) \ldots (x - x_{n-1})(x - x_n),$$

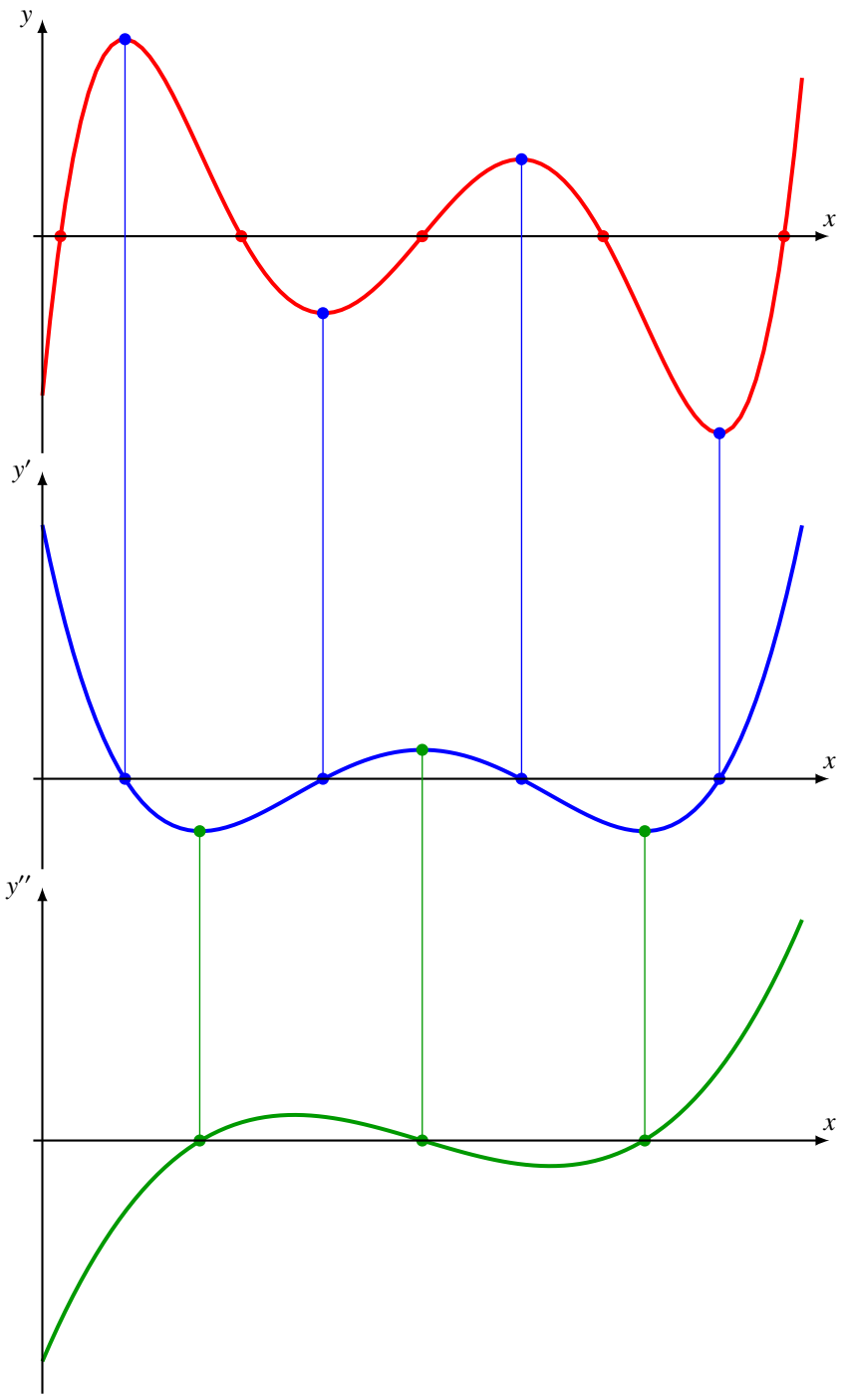


Abbildung 3.4: Schachtelung der Nullstellen von $f(x)$, $f'(x)$ und $f''(x)$. Der Satz von Rolle 3.6 impliziert, dass sich zwischen zwei Nullstellen von f immer eine Nullstelle von f' befindet, und ebenso zwischen zwei Nullstellen von f' eine von f''.

welches für die Konstruktion des Interpolationspolynoms verwendet wurde, hat genau die Nullstellen $x_0, x_1, \ldots, x_{n-1}, x_n$. Nach dem Satz 3.7 muss es zwischen je zwei aufeinanderfolgenden Nullstellen von l eine Nullstelle der Ableitung geben. Diese Situation ist in Abbildung 3.4 für den Fall $l(x) = (x+2)(x+1)x(x-1)(x-2)$ dargestellt.

Die höheren Ableitungen $f^{(k)}$ haben ihre Nullstellen natürlich auch wieder zwischen den Nullstellen der Ableitung $f^{(k-1)}$. Die n-te Ableitung ist konstant und hat keine Nullstellen.

Der Mittelwertsatz der Differentialrechnung

Taylorreihe mit Restformel

Fehler des Lagrange-Interpolationspolynoms

Der folgende satz gibt vollständige Auskunft über den Fehler des Interpolationspolynoms.

Satz 3.8. *Sei p ein Polynom vom Grad n, welches mit der $n+1$-mal differenzierbaren Funktion f an den $n+1$ Stellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

übereinstimmt. Dann gibt es für jedes $x \in [a,b]$ ein $\xi_x \in [a,b]$ mit

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} l(x) \tag{3.5}$$

Beweis. An den Stützstellen x_i ist $f(x_i) - p(x_i) = 0$ und $l(x_i) = 0$, die Gleichung (3.5) ist also trivialerweise erfüllt.

Sei jetzt also $x \in [a,b]$ verschieden von allen x_i. Da $l(x) \neq 0$ ist, gibt es eine Zahl c derart, dass

$$f(x) - p(x) = cl(x) \qquad \Leftrightarrow \qquad f(x) - p(x) - cl(x) = 0. \tag{3.6}$$

Die Funktion $g(x) = f(x) - p(x) - cl(x)$ verschwindet in allen Stützstellen x_i und zusätzlich auch noch im Punkt x, sie hat also $n+1$ Nullstellen.

Nach dem Nullstellen-Schachtelungssatz 3.7 hat die $n+1$ Ableitung von g eine Nullstelle im Intervall. Es gibt also eine Zahl $\xi_x \in [a,b]$ mit $g^{(n+1)}(\xi_x) = 0$.

Da p Grad n hat, ist die $n+1$-te Ableitung 0. Das Polynom $l(x)$ hat die Form

$$l(x) = x^{n+1} - (x_0 + x_1 + \cdots + x_{n-1} + x_n)x^{n-1} + \cdots + (-1)^{n+1} x_0 x_1 \ldots x_{n-1} x_n,$$

seine $n+1$-Ableitung ist die Konstanten $(n+1)!$.

Die Folgerung $g^{(n+1)}(\xi_x) = 0$ wird damit zu

$$0 = f^{(n+1)}(\xi_x) - c(n+1)! \qquad \Rightarrow \qquad c = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}.$$

Einsetzen in (3.6) ergibt

$$f(x) - p(x) = cl(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} l(x),$$

wie behauptet. □

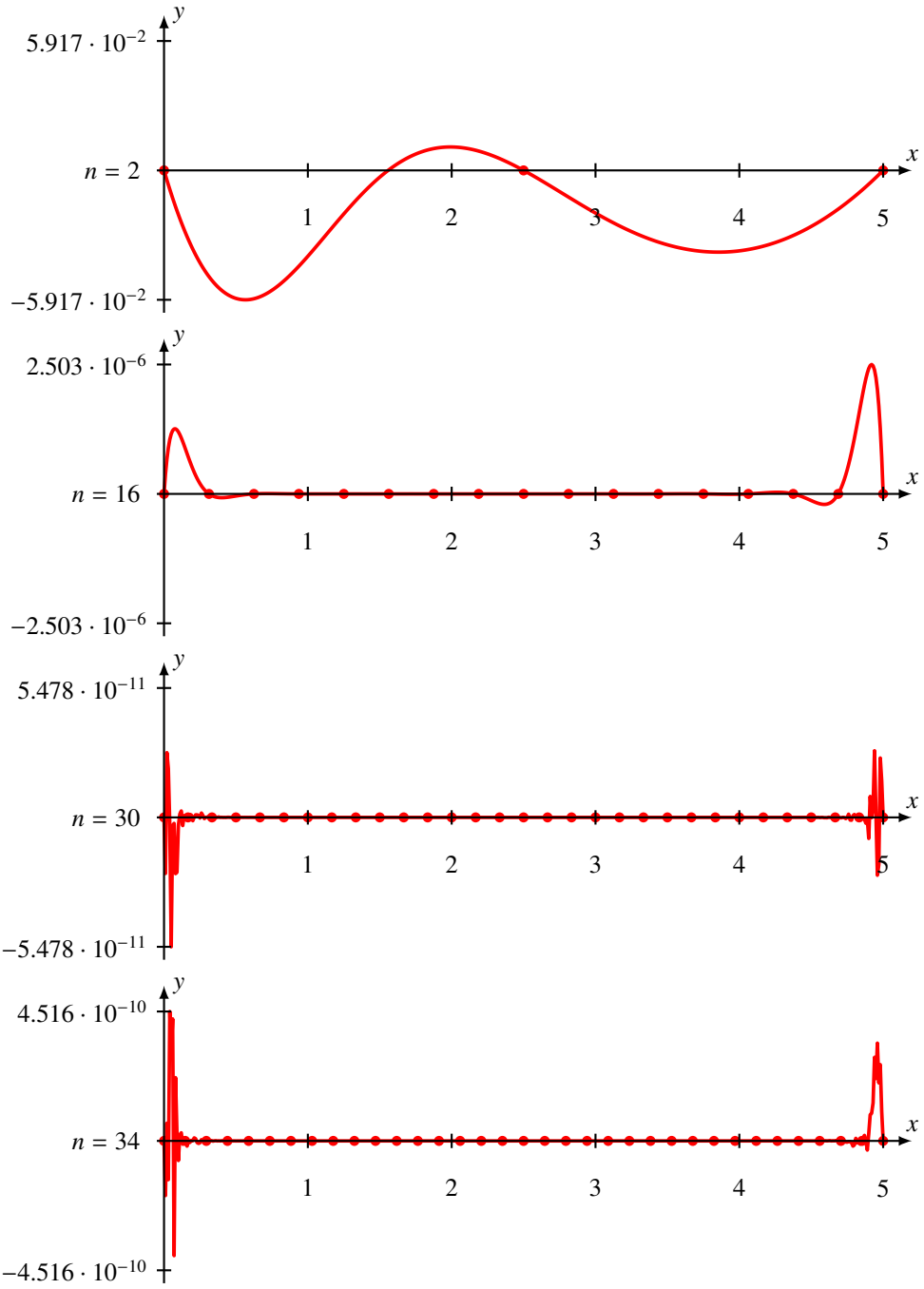


Abbildung 3.5: Fehler des Lagrange-Interpolationspolynoms für die Funktion $f(x) = e^{-x^2/2}/\sqrt{2\pi}$. Der Fehler nimmt mit der Anzahl der Stützstellen bis $n = 30$ ab, danach wird die Berechnung instabil und der Fehler nimmt wieder zu.

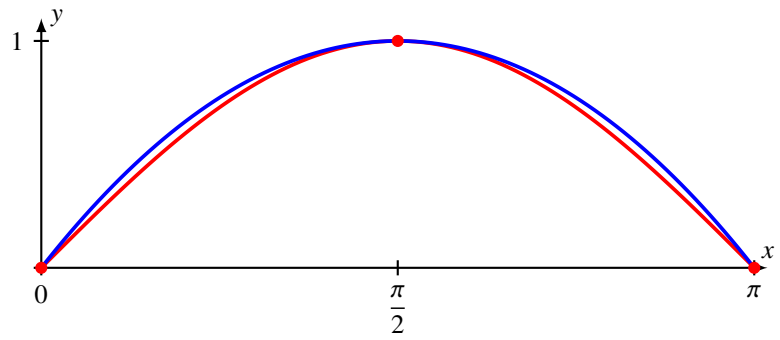


Abbildung 3.6: Interpolation der Funktion $f(x) = \sin x$ mit nur drei Stützstellen $x_0 = 0$, $x_1 = \frac{\pi}{2}$ und $x_2 = \pi$. Der Fehler ist deutlich kleiner als die Abschätzung mit Satz 3.8 erwarten lässt.

Dieser Satz erlaubt den Fehler eines Interpolationspolynoms abzuschätzen, wenn die $n+1$-te Ableitung der Funktion f bekannt ist. Wir bezeichnen mit

$$\|g\| = \sup_{a \le x \le b} |g(x)|$$

die *Supremun-Norm* der Funktion g im Intervall $[a, b]$.

Korollar 3.9. *Ist p ein Interpolationspolynom vom Grad n, welches mit der Funktion f in den Stellen $a = x_0 < x_1 < \cdots < x_{n+1} < x_n = b$ übereinstimmt, dann ist*

$$|f(x) - p(x)| \le \frac{\|f^{(n+1)}\|}{(n+1)!} |l(x)|.$$

Beispiel. Die Funktion $f(x) = \sin x$ soll mit den Stützstellen $x_0 = 0$, $x_1 = \frac{\pi}{2}$ und $x_2 = \pi$ interpoliert werden. Das Interpolationspolynom ist ein quadratisches Polynom mit Nullstellen x_0 und x_2, der Funktionswert bei x_1 muss 1 sein. Man kann sich davon überzeugen, dass das Polynom

$$p(x) = \frac{4}{\pi^2} x(\pi - x)$$

diese Eigenschaft hat. Wie gross ist der Fehler dieses Interpolationspolynoms?

Die dritten Ableitungen der Funktion $f(x) = \sin x$ sind, bekannt, es ist $f^{(3)}(x) = -\cos x$. Der Betrag von $f^{(3)}(x)$ wird also nie grösser als 1. Es folgt, dass

$$|f(x) - p(x)| \le \frac{1}{3!} l(x) = \frac{1}{6} |x(x - \tfrac{\pi}{2})(x - \pi)|$$

Die Ableitung des Polynoms auf der rechten Seite hat Nullstellen bei $\frac{\pi}{2} \pm \frac{\pi}{2\sqrt{3}}$, durch Einsetzen erhält man den maximalen Wert

$$\|f^{(3)}\| = \frac{\pi^3}{12\sqrt{3}} \simeq 1.49179.$$

Wir schliessen, dass das Interpolationspolynom niemals um mehr als 0.24863 vom Funktionswert abweichen kann. ○

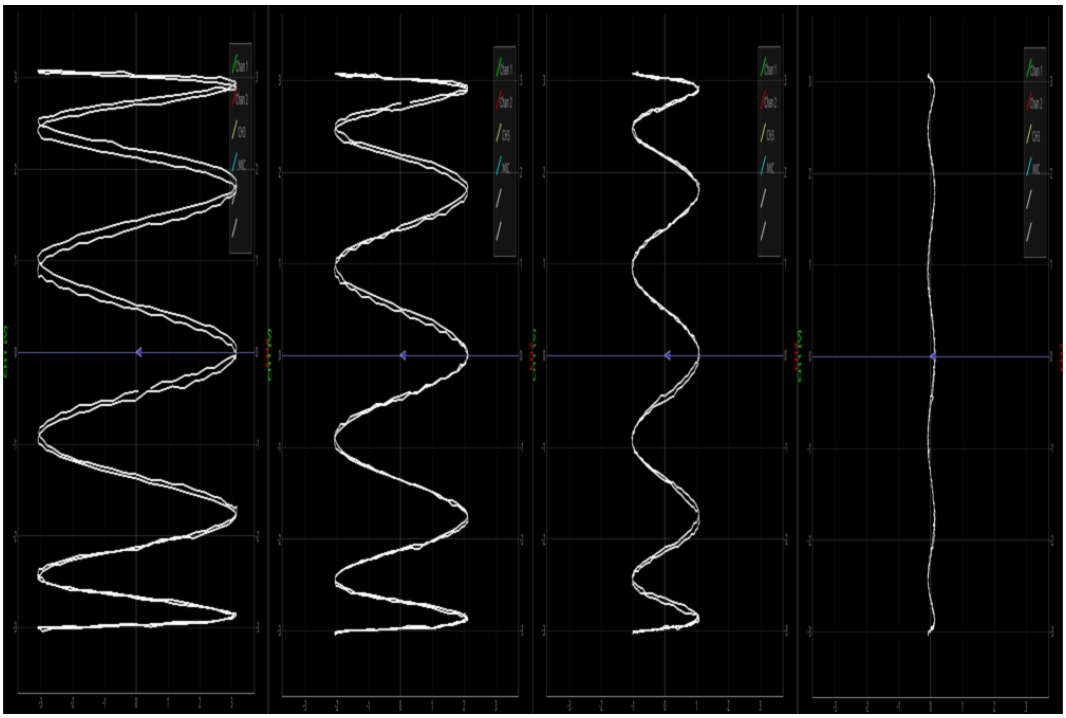


Abbildung 3.7: Lissajous-Figuren aus dem Artikel [1], die eine mögliche Lösung für ein Interpolationspolynom mit besonders kleinem Fehler suggerieren.

3.2.3 Wahl der Stützstellen und Tschebyscheff-Interpolationspolynom

Das Korollar 3.9 besagt, dass der Fehler des Interplationspolynom durch den Betrag von $l(x)$ begrenzt ist. Für äquidistante Stützstellen mit Abstand h kann man beobachten, dass die Oszillationen des Polynoms $l(x)$ gegen den Rand des Intervalls immer grösser werden. Für einen Punkte in der Mitte jedes Teilintervalls ist $h/2$ der kleinste mögliche Faktor in $l(x)$. Den grössten Faktor findet man für x im ersten oder letzten Teilinterval, er ist $b - a - h$. Ausserdem treten mehrere ähnlich grosse Faktoren auf. Für x in einem Intervall nahe $(a + b)/2$ sind die Faktoren dagegen nur halb so gross.

Die Oszillationen können verkleinert werden, wenn man dafür sorgt, dass mit grösserem Abstand von der Mitte des Intervalls der Abstand der Stützstellen ebenfalls verkleinert. Dies garantiert, dass neben den grossen Faktoren in der Nähe von $b - a$ auch wesentlich kleinere Faktoren auftreten, so dass die extremen Werte nahe den Intervallenden vergleichbar mit den Werten im Inneren des Intervalls werden.

Die beste Approximation durch ein Interpolationspolynom kann man also erwarten, wenn $l(x)$ im Interval $[a, b]$ keine besonders grossen Werte annimmt. Eine Funktion ähnlich wie $\sin x$ oder $\cos x$, die unendlich viele Extremewerte ± 1 haben, würde dieses Kriterium erfüllen, aber $\sin x$ und $\cos x$ sind keine Polynome. Sie sind auch auf auf einem viel grösseren Interval als nötig definiert, nämlich ganz $\mathbb{R}$. Ein verwandtes Beispiel sind Lissajous-Figuren, Abbildung 3.7 suggeriert, dass eine geeignete Lissajous-Figure als Graph für ein Interpolationspolynom mit sehr geringem Fehler dienen könnte, wenn man sie als Polynom darstellen kann. Eine solche Lissajous-Figur entsteht als

Kurve $t \mapsto (\cos t, \cos nt)$ oder eine anderes trigonometrisches Polynom als zweite Komponente. Es stellt sich also die Frage, ob $\cos nt$ als Polynom in $z = \cos t$ ausgedrückt werden kann.

Sei also

$$T_n(z) = T_n(\cos t) = \cos nt.$$

Für kleine n kann man unmittelbar verifizieren, dass $T_n(z)$ ein Polynom ist:

$$\begin{aligned} T_0(z) &= 1, \\ T_1(z) &= z, \\ T_2(z) &= \cos 2t = 2\cos^2 t - 1 = 2z^2 - 1 \quad \text{und} \\ T_3(z) &= \cos 3t = 4\cos^3 t - 3\cos t = 4z^3 - 3z. \end{aligned} \tag{3.7}$$

Aus den Additionstheoremen für den Kosinus folgt die Formel für die Summe von zwei Konsinus-Funktionen

$$\begin{aligned} \cos(n+1)t + \cos(n-1)t = 2\cos\frac{(n+1)t + (n-1)t}{2}\cos\frac{(n+1)t - (n-1)t}{2} &= 2\cos nt\cos t \\ T_{n+1}(z) + T_{n-1}(z) &= 2zT_n(z) \\ T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z). \end{aligned} \tag{3.8}$$

Wenn $T_n(z)$ und $T_{n-1}(z)$ Polynome sind, dann ist auch $T_{n+1}(z)$ ein Polynom. Aus den bereits gekannten Polynomen (3.7) und der Rekursionsformel (3.8) folgt jetzt mit vollständiger Induktion, dass alle $T_n(z)$ Polynome sind. Sie heissen *Tschebyscheff-Polynome*. Die Rekursionsformel kann dazu verwendet werden, die Polynome explizit zu berechnen. Zum Beispiel folgt für die nächsten paar Polynome

$$\begin{aligned} T_4(z) &= 8z^4 - 8z^2 + 1, \\ T_5(z) &= 16z^5 - 20z^3 + 5z \quad \text{und} \\ T_6(z) &= 32z^6 - 48z^4 + 18z^2 - 1. \end{aligned}$$

Da das Interpolationspolynom den führenden Koeffizienten 1 hat, muss $l(z) = 2^{1-n}T_n(z)$ gewählt werden.

Die Polynome $T_n(z)$ sind eigentlich nicht nötig, da für die Konstruktion des Interpolationspolynoms nur die Nullstellen nötig sind. Wegen $T_n(z) = \cos nt$ liegt eine Nullstelle genau dann vor, wenn $nt = \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$. Die zugehörigen Werte von z sind

$$z_k = \cos t = \cos\frac{\pi(2k+1)}{2n}.$$

In Abbildung 3.8 sind die Polynome $2^{n-1}l(x)$ vom Grad n oben für äquidistante Stützstellen und unten für Tschebyscheff-Stützstellen im Vergleich dargestellt. Wie in der Einleitung zu diesem Abschnitt angekündigt, oszillieren die Polynome für äquidistante Stütztstellen nahe den Intervallenden. Für Tschebyscheff-Stützstellen wird $2^{n-1}l(x)$ betragsmässig nie grösser als 1.

Abbildung 3.9 zeigt die Basispolynome vom Grad 7 $l_j(x)$ für Tschebscheff-Stütztstellen. Da bei Verwendung von Tschebyscheff-Stützstellen die Polynome $l(x)$ keine ausgeprägten Oszillationen an den Intervallenden aufweisen, sind auch die Basispolynome $l_j(x)$ vor allem in der Nähe der jeweiligen Stützstelle x_j wesentlich von 0 verschieden.

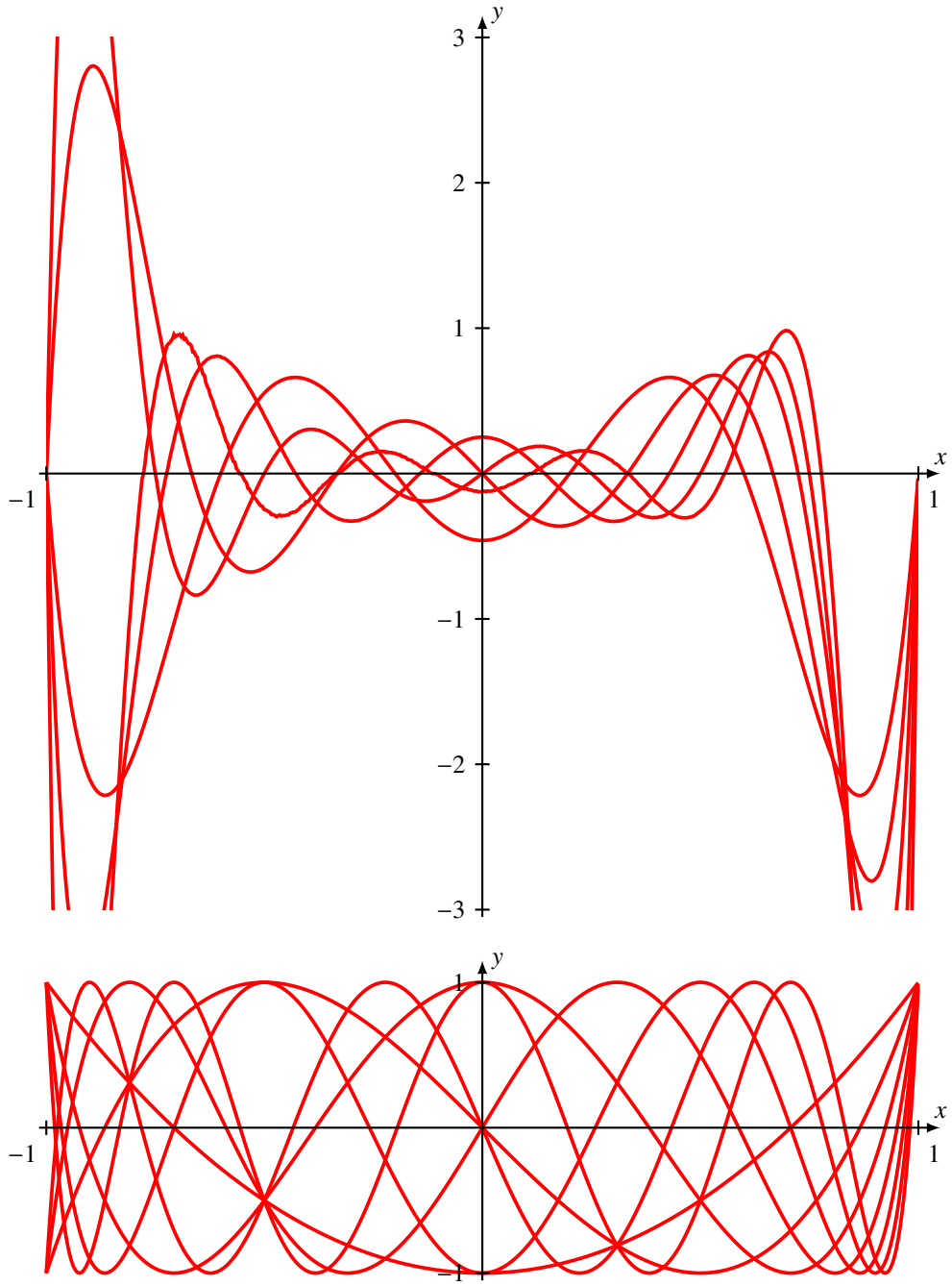


Abbildung 3.8: Vergleich der Oszillationen bei äquidistanten Stützstellen (oben) und bei Tschebyscheffstützstellen. Damit die Abweichungen sichtbar werden, sind die Polynome $l(x)$ vom Grad n mit dem Faktor 2^{n-1} skaliert. Bei Verwendung von Tschebyscheff-Stützstellen wächst $2^{n-1}l(x)$ nie über 1 an, während bei äquidistanten Stützstellen die in der Einleitung zu diesem Abschnitt diskutierten Oszillationen nahe den Intervallenden auftreten.

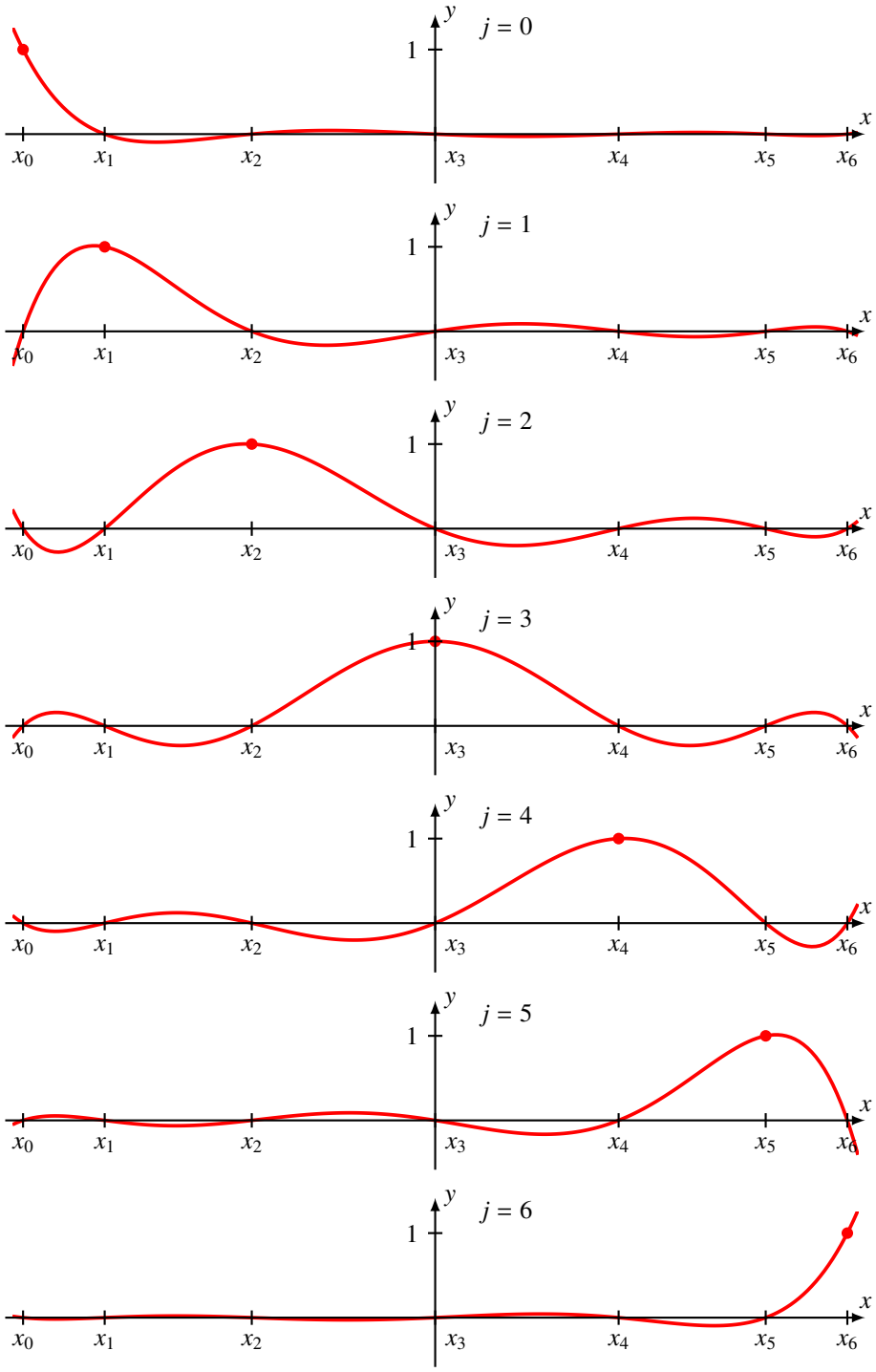


Abbildung 3.9: Basisinterpolationspolynome vom Grad 7 für Tschebyscheff-Stützstellen

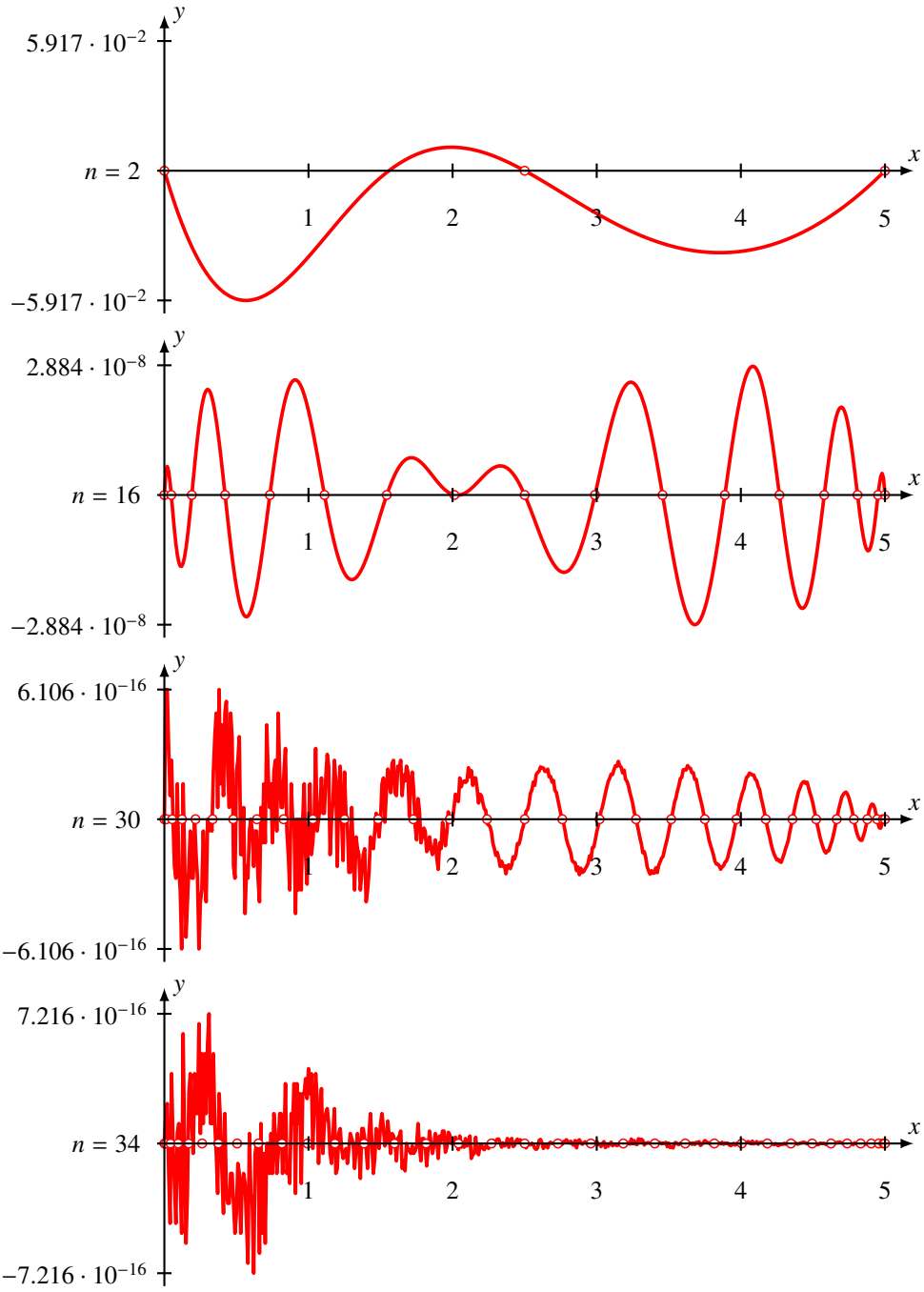


Abbildung 3.10: Fehler des Interpolationspolynomes für die Funktion $f(x) = e^{-x^2/2}/\sqrt{2\pi}$ mit Stützstellen nach Tschebyscheff. Der Fehler bleibt über das ganze Intervall gleichmässig. Für eine grosse Zahl von Stützstellen erreicht die Interpolation die Maschienengenauigkeit.

3.3 Hermite-Interpolation

Das Lagrange-Interpolationspolynom nimmt zwar in umittelbarer Nähe der Stützstellen zuverlässig Funktsionswerte nahe den gegeben Werten an, doch insbesondere gegen den Rand des Intervals können die oft beobachteten Oszillationen eine schlechte Approximation bewirken. Im Gegensatz zur Taylorreihe, deren Ableitung mindestens in der Nähe des Entwicklungspunktes auch mit der Ableitung der zu approximierenden Funktion übereinstimmt, gibt es für das Interpolationspolynom keine solche Garantie. Beide Schwierigkeiten könnten gemildert werden, indem gefordert wird, dass das Polynom nicht nur die gleichen Funktionswerte, sondern auch die gleichen Ableitungen bis zu einer bestimmten Ordnung haben soll. Dies ist die Idee der *Hermite-Interpolation*, die in diesem Abschnitt vorgestellt werden soll.

3.3.1 Aufgabenstellung

Das Hermite-Interpolationspolynom löst die folgende Approximationsaufgabe.

Aufgabe 3.10 (Hermite-Interpolationspolynom)**.** *Gegeben Stützstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

und Funktionswerte f_i, $0 \le i \le n$, und Werte $s_i^{(k)}$ der k-ten Ableitungen bis zur m-ten Ordnung, $1 \le k \le m$, finde ein Polynom h, mit

$$h(x_i) = f_i, \quad h^{(k)}(x_i) = s_i^{(k)}, \quad 0 \le i \le n, 1 \le k \le m. \tag{3.9}$$

Die Aufgabenstellung formuliert $N = (n+1)(k+1)$ Bedingungen für das Polynom h, es braucht also im Allgemeinen Polynom mindestens vom Grade $N = (n+1)(k+1)$, um alle diese Bedingungen erfüllen zu können. Ein elementarer Ansatz könnte sein, eine Polynom in der Form $a_N x^N + a_{N-1}x^{N-1} + \cdots + a_1 x + a_0$ anzusetzen, die Bedingungen (3.9) als lineare Gleichungen für die Koeffizienten auszuschreiben und das Gleichungssystem zu lössen. Dieses Vorgehen ist allerdings sehr aufwendig und numerisch nicht besonders stabil. Ein weg analog zur Bestimmung des Lagrange-Interpolationspolynomes in Abschnitt 3.2.1 ist daher angezeigt.

3.3.2 Bestimmung des Hermite-Interpolationspolynom

Wir führen die Konstruktion nur für den Fall $m = 1$ durch, also für Interpolationspolynome, die den Funktionswerten und ersten Ableitungen übereinstimmen. Wie in Abschnitt 3.2.1 suchen wir zunächst wieder eine Lösung des speziellen Interpolationsproblems.

Aufgabe 3.11 (Spezielles Hermite-Interpolationsproblem)**.** *Gegeben Stützstellen*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

finde Polynome h_j und h_j^1 vom Grad höchstens $2n+1$ derart, dass

$$\left.\begin{aligned} h_j(x_i) &= \delta_{ij} \\ h_j'(x_i) &= 0 \end{aligned}\right\} \forall i, j \qquad \textit{und} \qquad \left.\begin{aligned} h_j^1(x_i) &= 0 \\ h_j^{1\prime}(x_i) &= \delta_{ij} \end{aligned}\right\} \forall i, j$$

Lösung. Ein Polynom vom Grad $2n+2$, welches in allen Stützstellen eine doppelte Nullstelle hat, ist das Produkt

$$(x-x_0)^2(x-x_1)^2(x-x_2)^2\dots(x-x_{n-1})^2(x-x_n)^2.$$

Die Polynome h_j^1 haben in jeder Stützstelle ausser in x_j ein doppelte Nullstelle, die Nullstelle in x_j muss einfach sein. Ein solches Polynom kann man erhalten, indem man einen der Faktoren $(x-x_j)$ weglässt, oder zu

$$p_j(x) = (x-x_0)^2(x-x_1)^2\cdots\widehat{(x-x_k)^2}\dots(x-x_{n-1})^2(x-x_n)^2$$

einen solchen Faktor hinzufügt:

$$h_j^1(x) = c_j^2(x-x_j)p_j(x),$$

die Konstante c_j^2 muss passend gewählt werden, damit die Ableitung

$$h_j^{1\prime}(x) = c_j^2\underbrace{\frac{d}{dx}(x-x_j)}_{=1} + c_j^2(x-x_j)\frac{d}{dx}p_j(x)$$

den richtigen Wert bekommt. An der Stelle $x=x_j$ fällt der zweite Term weg und es bleibt

$$h_j^{1\prime}(x_j) = c_j^2 p_j(x_j).$$

und damit ist $c_j^2 = 1/p_j(x_j)$. Dies ist das Quadrat des entsprechenden Normierungsfaktors, der beim Lagrange-Interpolationspolynom zur Anwendung kam.

Die Polynome h_j haben in allen Stützstellen ausser x_j eine doppelte Nullstelle. Das Produkt $p_j(x)$ teilt diese Eigenschaft. Da es vom Grad $2n$ ist, haben wir nur die Freiheit, einen Linearfaktor der Form $(u_j(x-x_j)+v_j)$ hinzuzufügen, um h_j zu erhalten. Es müssen also u_j und v_j so gewählt werden, dass

$$h_j(x) = (u_j(x-x_j)+v_j)p_j(x) \qquad \Rightarrow \qquad \begin{cases} h_j(x_j) = v_j p_j(x_j) = 1 \\ h_j'(x_j) = u_j p_j(x_j) + v_j p_j'(x_j) = 0 \end{cases} \tag{3.10}$$

gilt. Aus der ersten Gleichung folgt $v_j = 1/p_j(x_j) = c_j^2$, aus der zweiten

$$u_j = -\frac{p_j'(x_j)}{p_j(x_j)^2} = -c_j^4 p_j'(x_j).$$

Einsetzen in die (3.10) ergibt

$$h_j(x) = \Big(-\frac{p_j'(x_j)}{p_j(x_j)^2}(x-x_j) + \frac{1}{p_j(x_j)}\Big)p_j(x) = \frac{p_j'(x_j)(x-x_j)+p_j(x_j)}{p_j(x_j)^2}p_j(x)$$

Andererseits ist $p_j(x)(x-x_j) = h_j^1(x)/c_j^1$, man kann also auch

$$h_j(x) = \frac{p_j(x)}{p_j(x_j)} - \frac{p_j'(x_j)}{c_j^1 p_j(x_j)^2}h_j^1(x) \tag{3.11}$$

schreiben. Der erste Term in (3.11) ist das Quadrat des Lagrange-Interpolationspolynoms $l_j(x)$. □

Mit der Lösung des speziellen Interpolations-Problems findet man jetzt auch eine Lösung für das allgemeine Problem. Das gesuchte Interplationspolynom ist

$$h(x) = \sum_{j=0}^{n} f_j h_j(x) + \sum_{j=0}^{n} s_j h_j^1(x).$$

3.3.3 Zwei Stützstellen

Der Fall zweier Stützstellen x_0 und x_1 ist von einiger praktischer Bedeutung. Er wird zum Beispiel im Abschnitt 3.5 zum Einsatz kommen. Die Polynome h und h^1 sollen daher für diesen Falll explizit berechnet werden.

Die Polynome p_0 und p_1 sind

$$p_0(x) = (x - x_1)^2 \qquad \text{und} \qquad p_1(x) = (x - x_0)^2,$$

woraus $c_0^2 = c_1^2 = (x_1 - x_0)^2$ folgt.

Die Polynome h_0^1 und h_1^1 entstehen durch geeignete Normierung der Polynome $(x - x_0)p_0(x)$ und $(x - x_1)p_1(x)$, also

$$h_0^1(x) = \frac{(x - x_0)(x - x_1)^2}{(x_0 - x_1)^2} \qquad \text{beziehungsweise} \qquad h_1^1(x) = \frac{(x - x_0)^2(x - x_1)}{(x_0 - x_1)^2}.$$

Für die Polynome h_0^1 und h_1^1 sind die Konstaten u_0 und u_1 zu bestimmen. Die Ableitung der Polynome p_j sind

$$p_0'(x) = 2(x - x_1) \qquad \text{und} \qquad p_1'(x) = 2(x - x_0)$$

uund damit ist

$$u_0 = -\frac{p_0'(x_0)}{(x_0 - x_1)^4} = -2\frac{x_0 - x_1}{(x_0 - x_1)^4} = -\frac{2}{(x_0 - x_1)^3} \qquad \text{und} \qquad u_1 = -\frac{p_1'(x_1)}{(x_0 - x_1)^4} = \frac{2}{(x_0 - x_1)^3}$$

Aus (3.10) folgt jetzt

$$\begin{aligned} h_0(x) &= (u_0(x - x_0) + v_0)(x - x_1)^2 = -\frac{2}{(x_0 - x_1)^3}(x - x_0)(x - x_1)^2 + \frac{(x - x_1)^2}{(x_0 - x_1)^4} \\ h_1(x) &= (u_1(x - x_1) + v_1)(x - x_0)^2 = \frac{2}{(x_0 - x_1)^3}(x - x_1)(x - x_0)^2 + \frac{(x - x_0)^2}{(x_0 - x_1)^2} \end{aligned}$$

Der Spezialfall $x_0 = 0$

In diesem Fall schreiben wir $m = x_1$ für die Intervalllänge und erhalten die Polynome

$$\begin{aligned} h_0^1(x) &= \frac{x(x - m)^2}{m^2} & h_1^1(x) &= \frac{x^2(x - m)}{m^2} \\ h_0(x) &= \frac{2x(x - m)^2}{m^3} + \frac{(x - m)^2}{m^4} & h_1(x) &= -\frac{2x^2(x - m)}{m^3} + \frac{x^2}{m^4} \end{aligned} \tag{3.12}$$

Der Spezialfall $x_0 = 0$**,** $x_1 = 1$

Eine besonders einfache Form nehmen die Polynome h_j^0 und h_j^1 an, wenn man sie auf das Interval $[0, 1]$ spezialisiert. Wir bezeichnen diese Polynome mit grossen Buchstaben, sie sind

$$\begin{aligned} H_0^1(x) &= x(1 - x)^2 = x^3 - 2x^2 + x & H_1^1(x) &= (1 - x)x^2 = -x^3 + x^2 \\ H_0(x) &= (1 + 2x)(1 - x)^2 - 2x^3 + 5x^2 - 4x + 1 & H_1(x) &= (3 - 2x)x^2 = -2x^3 + 3x^2 \end{aligned} \tag{3.13}$$

Graphen dieser Polynome sind in Abbildung 3.11 dargestellt.

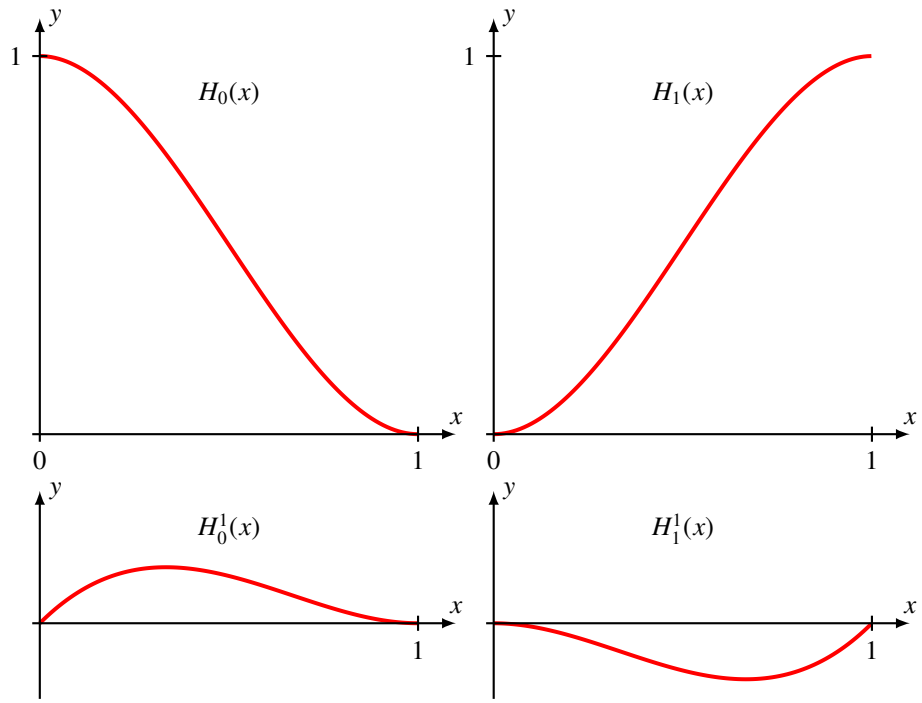


Abbildung 3.11: Hermite-Basispolynome für das Interval [0, 1] nach (3.13)

Diese Polynome können auch verwendet werden, die Polynome für ein beliebiges Intervall wieder zu gewinnen. Dazu setzen wir $x = (x - x_0)/m$ in die Polynome ein. Die Polynome $h_0(x) = H_0((x - x_0)/m)$ und $h_1(x) = H_1((x - x_0)/m)$ haben die Werte

$$h_0(x) = H_0((x-x_0)/m)\Big|_{x=x_0} = H_0(0) = 1 \quad \text{und} \quad h_0(x) = H_0((x-x_0)/m)\Big|_{x=x_1} = H_0(1) = 0$$
$$h_1(x) = H_1((x-x_0)/m)\Big|_{x=x_0} = H_1(0) = 0 \quad \text{und} \quad h_1(x) = H_1((x-x_0)/m)\Big|_{x=x_1} = H_1(1) = 1$$

und Ableitungen

$$h_i''(x_0) = \frac{d}{dx} H_i((x-x_0)/m)\Big|_{x=x_0} = H_i'((x-x_0)/m)\frac{1}{m}\Big|_{x=x_0} = \frac{H_i'(0)}{m} = 0$$

an den Intervalenden.

Tun wir dasselbe für die Polynome H_0^1 und H_1^1, erhalten wir

$$h_j^{1\prime}(x_i) = \frac{d}{dx} H_j^1((x-x_0)/m)\Big|_{x=x_i} = H_j^{1\prime}((x-x_0)/m)\Big|_{x=x_i}\frac{1}{m} = \frac{1}{m}H_j^{1\prime}(i) = \frac{1}{m}\delta_{ij},$$

dies ist bis auf den Faktor $1/m$ korrekt. Daraus lesen wir ab, dass wir die Polynome

$$h_j^1(x) = mH_j^1((x-x_0)/m)$$

für die Ableitungen verwenden müssen.

Zweite Ableitungen

Für die spätere Anwendung bei der Spline-Interpolation untersuchen wir auch noch die zweiten Ableitung des Hermite-Interpolationspolynoms im Fall zweier Stützstellen am Rande des Intervals. Wir tun dies für die Polynome (3.13) und kümmern uns später darum, was auf anderen Intervallen passiert.

$$\begin{aligned} H_0''(0) &= -6 & H_0''(1) &= 6 \\ H_1''(0) &= 6 & H_1''(1) &= -6 \\ H_0^{1\prime\prime}(0) &= -4 & H_0^{1\prime\prime}(1) &= 2 \\ H_1^{1\prime\prime}(0) &= -2 & H_1^{1\prime\prime}(1) &= 4 \end{aligned}$$

Unter Verwendung der Substition $x \to (x - x_0)/m$ können wir jetzt auch die Werte für die zweiten Ableitungen an den Intervallenden für bestimmen. Dazu berechnen wir erst die zweite Ableitung einer Funktion $f((x - x_0)/m)$:

$$\frac{d^2}{dx^2} f((x - x_0)/m) = \frac{d}{dx} f'((x - x_0)/m)\frac{1}{m} = f''((x - x_0)/m)\frac{1}{m^2}.$$

Angewendet auf die oben gefundenen Polynome bedeutet dies,

$$\begin{aligned} h_0''(x_0) &= -6/m^2 \quad \text{und} \quad h_0''(x_1) = 6/m^2 \\ h_1''(x_0) &= 6/m^2 \quad \text{und} \quad h_1''(x_1) = -6/m^2 \\ h_0^{1\prime\prime}(x_0) &= -4/m \quad \text{und} \quad h_0^{1\prime\prime}(x_1) = 2/m \\ h_1^{1\prime\prime}(x_0) &= -2/m \quad \text{und} \quad h_1^{1\prime\prime}(x_1) = 4/m \end{aligned}$$

3.4 Baryzentrische Formeln für Interpolationspolynome

Die Interpolationspolynome von Lagrange und Hermite haben in der bis jetzt gezeigten Form das folgende grundlegende Problem. Sie sind definiert über das Produkt

$$l(x) = (x - x_0)(x - x_1)\dots(x - x_{n-1})(x - x_n).$$

Ist die Zahl der Stützstellen gross und liegen erstrecken sich die Stützstellen über einen grossen Bereich, dann sind einzelne Faktoren $(x - x_i)$ immer gross. Zudem tritt bei der Berechnung eines Wertes in inmittelbarer Nähe der Stützstelle x_j in dem Faktor $(x - x_j)$ Auslöschung auf. Der grosse relative Fehler dieses Faktors wird durch die anderen Faktoren zu einem grossen absoluten Fehler aufgeblasen.

Andererseits ist klar, dass sich das Interpolationspolynom vor allem in der Nähe einer Stützstelle ändern sollte, wenn man den an der Stützstelle ändert. Die anderen Stützstellen sollten also nur einen geringen Einfluss auf den Wert des Interpolationspolynoms haben. Dies geht aus der bisherigen Form des Interpolationspolynoms ebenfalls nicht hervor.

Gesucht ist also eine Form des Interpolationspolynoms, welche einsichtig macht, dass Änderungen von Stützwerten sich vor allem in der nähe der betroffenen Stützstelle auswirken und die auch bei einer grossen Zahl von Stützstellen stabil sind.

Früher wurde gezeigt, dass das Interpolationspolynom für Funktionswerte f_j an den Stützstellen x_j durch die Linearkombination

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

gegeben ist. Für die Polynome $l_j(x)$ wurde

$$l_j(x) = \frac{(x-x_0)(x-x_1)\cdots\widehat{(x-x_j)}\cdots(x-x_n)}{(x_j-x_0)(x_j-x_1)\cdots\widehat{(x_j-x_j)}\cdots(x_j-x_n)}.$$

Schreibt man

$$w_j = \frac{1}{\prod\limits_{\substack{k=1\\k\neq j}}^{n}(x_j-x_k)},$$

dann kann man die Faktoren $l_j(x)$ auch als

$$l_j(x) = \frac{l(x)}{(x-x_j)} \cdot w_j$$

ausdrücken. Damit wird das Interpolationspolynom jetzt

$$p(x) = l(x)\sum_{j=0}^{n}\frac{w_j f_j}{x-x_j}. \tag{3.14}$$

Die Zahlen w_j hängen nur von den Stützstellen ab, nicht von den Funktionswerten f_j. Sie können also nach Festlegung der Stützstelle einmalig berechnet werden und verursachen danach keinen weiteren Berechnungsaufwand.

Das Interpolationspolynom wird besonders einfach, wenn alle Funktionswerte $f_j = 1$ sind. Da das konstante Polynom $p(x) = 1$ genau diese Werte annimmt, muss

$$1 = l(x)\sum_{j=0}^{n}\frac{w_j}{x-x_j}$$

gelten. Damit erhalten wir eine neue Darstellung für

$$l(x) = \frac{1}{\sum\limits_{j=0}^{n}\frac{w_j}{x-x_j}}. \tag{3.15}$$

In dieser Form wird vermieden, dass zur Berechnung von $l(x)$ eine grosse Anzahl Produkte mit potentiell grossen Faktoren gebildet werden muss. Sorgen bereiten vor allem die Faktoren $x - x_j$ für x weit entfernt von x_j. Stattdessen wird ein Summe von Summanden gebildet, die klein sind, wenn x weit von x_j entfernt ist.

Die vorteilhafte Formulierung (3.15) kann nun dazu verwendet werden, auch eine verbesserte Formulierung für das Interpolationspolynom aufzustellen. Dazu ersetzen wir den Faktor $l(x)$ in (3.14) durch (3.15) und erhalten

$$p(x) = \frac{\sum\limits_{j=0}^{n}\frac{w_j f_j}{x-x_j}}{\sum\limits_{j=0}^{n}\frac{w_j}{x-x_j}}.$$

Diese Form des Interpolationspolynoms ist ein gewichtetes Mittel der Werte f_j, gewichtet mit den Gewichten $w_j/(x - x_j)$. Diese Gewichte sind klein für x weit weg von x_j, die grössten Gewichte haben die Funktionswerte f_j nahe bei x.

3.5 Spline-Interpolation

Die Hermite-Interpolation ermöglicht Aporoximationspolynome zu finden, die sowohl Funktionswerte als auch Ableitungen an den Stützstellen mit der zu approximierenden Funktion gemeinsam haben. Dadurch wird der Fehler der Approximationspolynome zwar kleiner, aber es entsteht das zusätzliche Problem, dass die Ableitungen der Funktion bestimmt werden müssen.

Die Spline-Interpolation umgeht dieses Problem, indem sie an den Stützstellen nicht die gleichen Steigungen verlangt, sondern Steigungen, die zu einem möglichst "wenig gekrümmten" Graphen des Approximationspolynoms, welches natürlich immer noch in den Stützstellen die vorgegebenen Werte annehmen soll. Die Steigungen in den Stützstellen sind also Lösungen eines Optimierungsproblems, welches nicht die am besten passende, sonder die "schönste" Kurve durch die Stützstellen sucht.

3.5.1 Anforderungen and die interpolierende Funktion

Gegeben seien wie früher Punkte

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

auf und Funktionswerte f_i einer im übrigen unbekannten, aber ausreichend glatten Funktion $f\colon [a,b] \to \mathbb{R} : x \mapsto f(x)$, es ist also $f(x_i) = f_i$.

Gesucht ist eine stetige Funktion $g\colon [a,b] \to \mathbb{R} : x \mapsto g(x)$, die die folgenden natürliche Eigenschaften haben soll:

1. Die Funktion g nimmt in allen Stützstellen die Werte der Funktion f an, es ist also $g(x_i) = f_i \ \forall 0 \le i \le n$.

2. Die Funktion g ist stetig differenzierbar im ganze Interval. Insbesondere existiert die Ableitung $g'(x)$ in jedem Punkt x des Intervals $[a,b]$, der Graph von g kann also keine "Knicke" haben.

3. Im inneren jedes Teilintervalles $[x_i, x_{i+1}]$ ist die Funktion g beliebig oft stetig differenzierbar und die einseitigen Grenzwerte an den Enden der Teilintervalle existieren:

   $$\exists \lim_{x \to x_i +} g^{(k)}(x) \quad \forall 0 \le i < n \qquad \text{und} \qquad \exists \lim_{x \to x_i -} g^{(k)}(x) \quad \forall 0 < i \le n.$$

   Es wird nicht verlangt, dass die rechts- und linksseitigen Grenzwerte an den inneren Stützstellen $x1, \ldots, x_{n-1}$ übereinstimmen müssen.

4. Der Graph von g soll möglichst wenige gekrümmt sein. Da die zweite Ableitung einer Funktion ein Mass für die Krümmung des Graphen ist, kann dieses Kriterium dadurch realisiert werdenn, dass die Funktion g unter allen Funktionen, die die Bedingungen 1–3 erfüllen, das Integral

   $$J(g) = \int_a^b (g''(x))^2 \, dx$$

   minimiert.

Man beachte, dass nirgends verlangt wird, dass die Ableitungen von g an den Stützstellen irgendwie mit der Funktion f in Verbingung steht.

3.5.2 Das Optimierungsproblem

Zunächst ist nicht klar, ob das eben gestellt Optimierungsproblem überhaupt eine Lösung hat. In jedem Teilinterval $[x_i, x_{i+1}]$ geht es um ein Problem der folgenden Art. Gesucht ist eine Funktion, die an den Intervallenden die vorgegebene Werte $g(x_i) = f_i$ und $g(x_{i+1}) = f_{i+1}$ annimmt, im Inneren des Intervals beliebig oft stetig differenzierbar ist und zudem einen Integralausdruck

$$\int_{x_i}^{x_{i+1}} (g''(x))^2 \, dx$$

minimiert.

Diese Art von Problemen hat bereits Leonhard Euler in recht allgemeiner Form untersucht und zu diesem Zweck das Gebiet der Variationsrechnung geschaffen. Sie tauchen in der Physik zum Beispiel in der folgenden Form auf.

Beispiel. Ein Teilchen der Masse m bewegt sich entlang der y-Achse. Zur Zeit a befindet es sich bei f_0, zur Zeit b bei f_n. Auf das Teilchen wirkt ausserdem eine Kraft, die durch ihr Potential $V(y)$ beschrieben werden kann. Die Geschwindigkeit zur Zeit t ist $\dot{y}(t)$. Die Differenz von kinetischer und potentieller Energie ist die sogenannte Lagrange-Funktion

$$L(t, y, \dot{y}) = \frac{1}{2} m \dot{y}(t)^2 - V(y(t)). \tag{3.16}$$

In der Physik wird gezeigt, dass die Bewegung des Teilchens durch diejenige Funktion $y(t)$ beschrieben wird, welche das Integral

$$\int_a^b L(t, y(t), \dot{y}(t)) \, dt$$

minimiert. ○

Um zu zeigen, dass die Interpolationsfunktion g existiert, lösen wir daher das folgende, wesentlich allgemeinere Problem.

Satz 3.12. *Sei $L(x, y, y_1)$ eine in allen Argumenten beliebig oft stetig differenzierbare Funktion auf $[a, b]\times\mathbb{R}\times\mathbb{R}$. Es gibt eine glatte Funktion $y(x)$, die in den Intervalenden vorgegebene Werte $y(a) = y_a$ und $y(b) = y_b$ annimmt und ausserdem das Integral*

$$J(y) = \int_a^b L(x, y(x), y'(x)) \, dx$$

minimiert, sie ist Lösung der Euler-Lagrange-Differentialgleichung

$$\frac{d}{dx} \frac{\partial L}{\partial y_1}(x, y(x), y'(x)) - \frac{\partial L}{\partial y}(x, y(x), y'(x)) = 0. \tag{3.17}$$

Beweis. Wir gehen wie folgt vor: wir zeigen zunächst, dass eine solche Funktion eine Differentialgleichung erfüllen muss. Dann beziehen wir uns auf bekannte Sätze der Theorie der gewöhnlichen Differentialgleichungen, die besagen, dass die Gleichung eine glatte Lösung hat.

Sei jetzt also $y(x)$ eine Funktion mit $y(a) = y_a$ und $y(b) = y_b$, die das Integral $J(y)$ minimiert. Ändern wir die Funktion ein klein wenig, dann muss der Wert von J zunehmen. Wir vollziehen die Änderung, indem wir eine Funktion $h(x)$ wählen mit $h(a) = 0$ und $h(b) = 0$. Die Funktionen $y_\varepsilon = y + \varepsilon h$ erfüllen dann alle die Bedingung $y_\varepsilon(a) = y_a$ und $y_\varepsilon(b) = y_b$, insbesondere müssen sie alle

einen Wert $J(g+\varepsilon h)$ ergeben, der grösser ist als $J(g)$. Insbesondere muss die Ableitung von $J(y+\varepsilon h)$ nach ε an der Stelle $\varepsilon = 0$ verschwinden.

Wir berechnen die Ableitung von $J(y + \varepsilon h)$ nach ε:

$$\begin{aligned} 0 = \frac{d}{d\varepsilon} J(y+\varepsilon h)\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon} \int_a^b L(x, y(x) + \varepsilon h(x), y'(x) + \varepsilon h'(x))\, dx \Big|_{\varepsilon=0} \\ &= \int_a^b \frac{\partial L}{\partial y}(x, y(x), y'(x))\, h(x) + \frac{\partial L}{\partial y_1} L(x, y(x), y'(x)\, h'(x)\, dx \\ &= \int_a^b \frac{\partial L}{\partial y}(x, y(x), y'(x))\, h(x)\, dx + \int_a^b \frac{\partial L}{\partial y_1} L(x, y(x), y'(x)\, h'(x)\, dx \end{aligned} \tag{3.18}$$

Das zweite Integral enthält die Ableitung $h'(x)$, über die wir nicht viel wissen. Wir können diese aber durch partielle Integration los werden:

$$\int_a^b \frac{\partial L}{\partial y_1} L(x, y(x), y'(x)\, h'(x)\, dx = \Big[\frac{\partial L}{\partial y_1} L(x, y(x), y'(x))\, h(x) \Big]_a^b - \int_a^b \frac{d}{dx} \frac{\partial L}{\partial y_1} L(x, y(x), y'(x))\, h(x)\, dx$$

h war so gewählt, dass die Werte, an den Intervalenden verschwinden, also $h(a) = h(b) = 0$. Der erste Terme verschwindet daher und es bleibt

$$= - \int_a^b \frac{d}{dx} \frac{\partial L}{\partial y_1} L(x, y(x), y'(x))\, h(x)\, dx.$$

Einsetzen in (3.18) ergibt die Gleichung

$$0 = - \int_a^b \Big(\frac{d}{dx} \frac{\partial L}{\partial y_1}(x, y(x), y'(x)) - \frac{\partial L}{\partial y}(x, y(x), y'(x)) \Big) h(x)\, dx. \tag{3.19}$$

Gleichung (3.19) muss für jede beliebige Funktion $h(x)$ gelten. Wir möchten zeigen, dass das nur möglich ist, wenn die grosse Klammer im Integral verschwindet.

Nehmen wir an, die grosse Klammer sei an einer Stelle im Intervall von 0 verschieden. Dann wird sie wegen der Stetigkeit auch in einer kleinen Umgebung dieser Stelle immer noch das gleiche Vorzeichen haben. Wir wählen eine Funktion h, die in der gleichen kleinen Umgebung positiv ist und sonst überall verschwindet. Das Integral muss dann nur noch über diese kleine Umgebung erstreckt werden und die Funktion, die integriert wird, hat in der ganzen Umgebung das gleiche Vorzeichen. Insbesondere kann das Integral nicht verschwinden. Somit ist gezeigt, dass die grosse Klammer verschwinden muss, oder dass die Gleichung

$$\frac{d}{dx} \frac{\partial L}{\partial y_1}(x, y(x), y'(x)) - \frac{\partial L}{\partial y} L(x, y(x), y'(x)). \tag{3.20}$$

gelten muss. □

Beispiel. Wir wenden die Euler-Lagrange-Gleichung auf die Lagrange-Funktion (3.16) an, dabei erhalten wir

$$\left.\begin{aligned} \frac{\partial L}{\partial y} &= -V'(y) \\ \frac{\partial L}{\partial \dot{y}} &= m\dot{y} \end{aligned}\right\} \quad \Rightarrow \quad \frac{d}{dt} \frac{\partial L}{\partial \dot{y}} - \frac{\partial L}{\partial y} = \frac{d}{dt} m\dot{y} + V'(y) = 0 \quad \Rightarrow \quad m\ddot{y} = -V'(y).$$

Dies ist das 2. Newtonsche Gesetz. ○

3.5.3 Lösung des Optimierungsproblems

Leider lässt sich der Satz 3.17 nicht direkt auf das Interpolationsproblem anwenden, weil im Ausdruck $J(g)$ die zweite Ableitung von g vorkommt. Wir führen daher die Rechnung, die auf die Euler-Lagrange-Differentialgleichung geführt hat, nochmals in diesem Spezialfall durch. Wieder sei h eine Funktion, die in jeder Stützstelle verschwindet. Die Minimalitätsbedingung ist dann

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon} \int_{x_i}^{x_{i+1}} (g''(x) + \varepsilon h''(x))^2 \, dx \Big|_{\varepsilon=0} \\ &= \int_{x_i}^{x_{i+1}} 2g''(x)h''(x) + 2\varepsilon h''(x)^2 \, dx \Big|_{\varepsilon=0} \\ &= \int_{x_i}^{x_{i+1}} 2g''(x)h''(x) \, dx. \end{aligned} \tag{3.21}$$

Wie bei der Euler-Lagrange-Gleichung können wir durch partielles Integrieren die zweite Ableitung der Funktion h los werden:

$$\begin{aligned} 0 &= \Big[g''(x)h'(x) \Big]_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} g'''(x)h'(x) \, dx \\ &= \Big[g''(x)h'(x) \Big]_{x_i}^{x_{i+1}} - \Big[g'''(x)h(x) \Big]_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} g^{(4)}(x)h(x) \, dx. \end{aligned} \tag{3.22}$$

Auf Grund der Definition von h verschwindet der mittlere Term.

Bedingungen im Inneren der Teilintervalle

Jetzt nutzen wir wieder die freie Wahlmöglichkeit der Funktion h aus. Wir können die Funktion so wählen, dass $h(x_i) = h(x_{i+1}) = h'(x_i) = h'(x_{i+1}) = 0$ ist, dann verschwinden die ersten beiden Terme. Das Integral verschwindet nur dann immer, wenn der Integrand verschwindet, wenn also $g^{(4)}(x) = 0$ im Inneren jedes Teilintervals $[x_i, x_{i+1}]$. Es folgt, dass in jedem Teilinterval die Funktion g ein kubisches Polynom sein muss.

Bedingungen an den Stützstellen

Aus dem verbleibenden ersten Term von Gleichung (3.22) lässt sich noch mehr über die zweiten Ableitungen der Funktion g schliessen. Die Summe dieser Terme muss ja ebenfalls 0 ergeben, also

$$\begin{aligned} 0 &= \sum_{i=0}^{n-1} \Big[g''(x)h'(x) \Big]_{x_i}^{x_{i+1}} = \sum_{i=0}^{n-1} (g''(x_{i+1}-)h'(x_{i+1}) - g''(x_i+)h'(x_i)) \\ &= -g''(x_0+)h'(x_0) + \sum_{i=1}^{n-1} h'(x_i)(g''(x_i-) - g''(x_i+)) + g''(x_n-)h'(x_n). \end{aligned}$$

Indem man für h eine Funktion wählt, die an allen Stützstellen verschwindet und in genau einer Stützstelle Ableitung 1 hat, was mit einem Hermite-Interpolationspolynom sicher möglich ist, schliesst man

$$g''(x_i-) = g''(x_i+) \quad \forall 1 \leq i < n. \tag{3.23}$$

Die Funktion ist also zweimal stetig differenzierbar. Schliesslich müssen auch die Terme an den Enden der Summe verschwinden. Eine Funktion h, die in allen Stützstellen zusammen mit der ersten

Ableitung in den inneren Stützstellen verschwindet und deren erste Ableitung in genau einem der Endpunkte 1 ist zeigt, dass ausserdem

$$g''(x_0+) = g''(x_n-) = 0 \tag{3.24}$$

sein muss.

Ein Gleichungssystem für die Steigungen

Zur Lösung des eingangs gestellten Interpolationsproblems ist jetzt also für jedes Teilinterval $[x_i, x_{i+1}]$ ein kubisches Polynom $g_i(x)$ zu finden, mit folgenden Eigenschaften:

$g_i(x_i) = f_i$	$g_i(x_{i+1}) = f_{i+1}$	$0 \le i \le n$	$2n+2$ Bedingungen
$g'_{i-1}(x_i) = g'_i(x_i)$	$g''_{i-1}(x_i) = g''_i(x_i)$	$1 \le i < n$	$2n$ Bedingungen
$g''_0(x_0) = 0$	$g''_n(x_n) = 0$		2 Bedingungen

Dies sind $4n+4$ lineare Bedingungen für $n+1$ Polynome, die je 4 Koeffizienten haben. Es sollte sich also ein lineares Gleichungssystem finden lassen, welches diese Koeffizienten findet.

Aus Abschnitt 3.3 ist bekannt, dass die kubischen Polynome $g_i(x)$ durch die bereits bekannten Funktionswerte f_i und die noch zu findenen Steigungen in den Stützstellen bestimmt sind. Wir schreiben daher $s_i = g'_i(x_i)$ für die Steigungen und machen es uns zum Ziel ein Gleichungssystem für die s_i zu finden.

In Abschnitt 3.3.3 haben wir Hermite-Interopationspolynome für zwei Stützstellen zusammengestellt. Wir haben dort die Polynome H_i und H_i^1 konstruiert, aus denen sich mit der Substition $x \to (x - x_0)/m$ die Hermite-Interpolationspolynome für das Interval $[x_0, x_0 + m]$ bilden liess. Wir bezeichnen die Länge des Intervalls $[x_i, x_{i+}]$ mit $m_i = x_{i+} - x_i$.

Die gesuchte Funktion im Interval ist daher

$$g_i(x) = f_i H_0((x-x_i)/m_i) + f_{i+1} H_1((x-x_i)/m_i) + s_i m_i H_0^1((x-x_i=)/m_i) + s_{i+1} m_i H_1^1((x-x_i)/m_i). \tag{3.25}$$

Diese Funktion hat die richtigen Funktionswerte und Ableitungen an den Intervallenden.

Die Steigungen s_i in (3.25) ist noch nicht bekannt, aber die Bedingung an die zweiten Ableitungen wurde noch nicht ausgenutzt. Die zweiten Ableitungen

$$\begin{aligned}
i = 0 \qquad & 0 = g_0''(x_0) = -\frac{6f_0}{m_0^2} + \frac{6f_1}{m_0^2} - \frac{4s_0}{m_0} + \frac{2s_1}{m_0} \\
i = 1 \qquad & g_0''(x_1) = \frac{6f_0}{m_0^2} - \frac{6f_1}{m_0^2} + \frac{2s_0}{m_0} - \frac{4s_1}{m_0} \\
& = g_1''(x_1) = -\frac{6f_1}{m_1^2} + \frac{6f_2}{m_1^2} - \frac{4s_1}{m_1} + \frac{2s_2}{m_1} \\
i = 2 \qquad & g_1''(x_2) = \frac{6f_1}{m_1^2} - \frac{6f_2}{m_1^2} + \frac{2s_1}{m_1} - \frac{4s_2}{m_1} \\
& = g_2''(x_2) = -\frac{6f_2}{m_2^2} + \frac{6f_3}{m_2^2} - \frac{4s_2}{m_2} + \frac{2s_3}{m_2} \\
& \vdots \\
i = n \qquad & 0 = g_n''(x_n) = \frac{6f_{n-1}}{m_n^2} - \frac{6f_n}{m_n^2} + \frac{2s_{n-1}}{m_n} - \frac{4s_n}{m_n}
\end{aligned}$$

In allen Gleichungen kommt der Faktor 2 vor, den wir herausdividieren können. Schaffen wir die Terme in f_i auf die rechte Seite und sammeln die Terme mit s_i auf der linken Seite, erhalten wir das Gleichungssystem

$$\begin{aligned}
\frac{2}{m_0}s_0 + \quad \frac{1}{m_0}s_1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &= 3\frac{f_1-f_0}{m_0^2}\\
\frac{1}{m_0}s_0 + \quad \left(\frac{2}{m_0}+\frac{2}{m_1}\right)s_1 + \quad \frac{1}{m_1}s_2 \qquad\qquad\qquad\qquad &= 3\frac{f_2-f_1}{m_1^2}\\
\frac{1}{m_1}s_1 + \quad \left(\frac{2}{m_1}+\frac{2}{m_2}\right)s_2 + \quad \frac{1}{m_2}s_3 \qquad\qquad &= 3\frac{f_3-f_2}{m_2^2}\\
\ddots \qquad\qquad \ddots \qquad\qquad &\vdots\\
\frac{1}{m_{n-2}}s_{n-1} + \quad \frac{2}{m_{n-1}}s_n &= 3\frac{f_n-f_{n-1}}{m_{n-1}^2}
\end{aligned} \tag{3.26}$$

Die Koeffizientenmatrix und die rechte Seite dieses Gleichungsssytems sind

$$A = \begin{pmatrix}
\frac{2}{m_0} & \frac{1}{m_0} & & & & & \\
\frac{1}{m_0} & \frac{2}{m_0}+\frac{2}{m_1} & \frac{2}{m_1} & & & & \\
 & \frac{1}{m_1} & \frac{2}{m_1}+\frac{2}{m_2} & \frac{1}{m_2} & & & \\
 & & \frac{1}{m_2} & \ddots & \ddots & & \\
 & & & \ddots & \ddots & \frac{1}{m_{n-2}} & \\
 & & & & \frac{1}{m_{n-2}} & \frac{2}{m_{n-1}}
\end{pmatrix} \quad \text{und} \quad b = \begin{pmatrix}
3\frac{f_1-f_0}{m_0^2}\\
3\frac{f_2-f_1}{m_1^2}\\
3\frac{f_3-f_2}{m_2^2}\\
\vdots\\
3\frac{f_{n-1}-f_{n-2}}{m_{n-2}^2}\\
3\frac{f_n-f_{n-1}}{m_{n-1}^2}
\end{pmatrix}.$$

Die Gleichungen werden besonders einfach, wenn alle Abstände gleich sind, zum Beispiel $m = m_0 = \ldots m_{n-1}$. Dann kann man die Gleichungen mit m multiplizieren und bekommt für die Koeffizientenmatrix und die rechte Seite

$$A = \begin{pmatrix}
2 & 1 & & & & \\
1 & 2 & 1 & & & \\
 & 1 & 2 & 1 & & \\
 & & 1 & \ddots & \ddots & \\
 & & & \ddots & \ddots & 1\\
 & & & & 1 & 2
\end{pmatrix} \quad \text{und} \quad b = \frac{3}{m}\begin{pmatrix}
f_1-f_0\\
f_2-f_1\\
f_3-f_2\\
\vdots\\
f_{n-1}-f_{n-2}\\
f_n-f_{n-1}
\end{pmatrix}$$

3.5.4 Bézier-Kurven und Splines in der Ebene

Übungsaufgaben

3.1. Das Polynom $p(x)$ soll die Funktionswerte der Sinusfunktion an den Stellen $k\frac{\pi}{2}$ für ganzzahliges

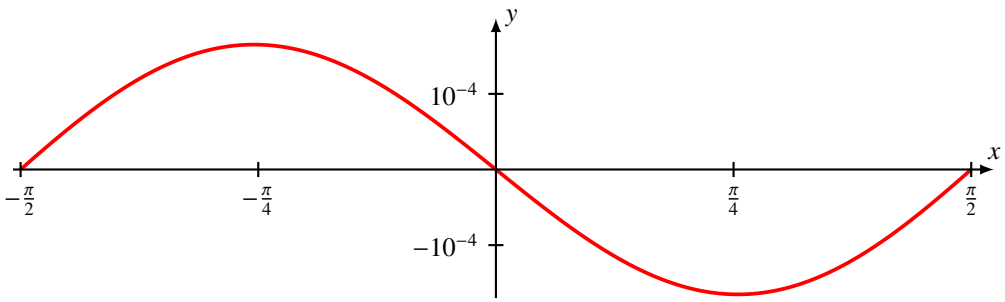


Abbildung 3.12: Fehler des Interpolationspolynoms für $\sin x$ mit Stützstellen $\frac{\pi}{2}k$ mit ganzzaligen k mit $-10 \le k \le 10$.

k mit $-10 \le k \le 10$ interpolieren. Wie gross ist der Fehler des Interpolationspolynoms für $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

Lösung. Zunächst halten wir fest, dass 21 Stützstellen verwendet werden, dass also $n = 20$ ist. Nach der Fehlerformel für das Interpolationspolynom gilt

$$|f(x) - p(x)| \le \frac{|l(x)|}{21!}|f^{(21)}(x)|$$

mit $f(x) = \sin x$. Die Ableitungen von f sind wieder trigonometrische Funktionen, d. h. $|f^{(21)}(x)| \le 1$. Der Fehler ist daher

$$|f(x) - p(x)| \le \frac{|l(x)|}{21!}.$$

Um $l(x)$ abzuschätzen verwendet man

$$\begin{aligned} |l(x)| &= |x - x_0| \cdot |x - x_1| \cdots |x - x_n| \\ &\le (11\frac{\pi}{2} \cdot 10\frac{\pi}{2} \cdot \cdots \cdot 2\frac{\pi}{2})^2 \frac{\pi}{2} \\ &= \left(\frac{\pi}{2}\right)^{21} (11!)^2 \end{aligned}$$

Damit kann man jetzt den Fehler abschätzen:

$$|\sin x - p(x)| \le \left(\frac{\pi}{2}\right)^{21} \frac{(11!)^2}{21!} = 22 \cdot \left(\frac{\pi}{2}\right)^{21} \binom{22}{11}^{-1} = 0.40972.$$

In der Tat ist der Fehler viel kleiner als diese Schranke vermuten lässt. In Abbildung 3.13 kann man erkennen, dass das Interpolationspolynom die Funktion in der Mitte des Intervalls sehr genau wiedergeben kann. Nur am Rande weicht es wegen des Runge-Phänomens stark ab. In Abbildung 3.12 ist der Fehler $\sin x - p(x)$ für $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ dargestellt, es zeigt sich, dass der Betrag des Fehlers kleiner als $1.66 \cdot 10^{-4}$ ist. ○

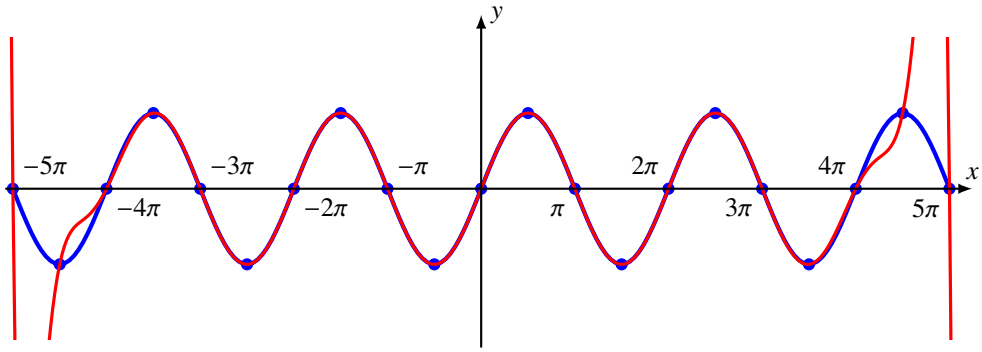


Abbildung 3.13: Graph des Interpolationspolynoms für $f(x) = \sin x$ mit Stützstellen $\frac{\pi}{2}k$ für ganzzahlige k mit $-10 \leq k \leq 10$. Trotz des grossen Abstandes und der sehr speziellen Wahl der Stützstellen folgt das Interpolationspolynom (rot) der Funktion (blau) in der Mitte des Definitionsbereichs sehr genau.

Kapitel 4

Integration

Die Wahrscheinlichkeitsdichte der Standardnormalverteilung führt auf das Integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

für die Verteilungsfunktion. Man kann beweisen, dass es für diesen Integranden keine Stammfunktion in analytischer Form als algebraischer Ausdruck bekannter Funktionen geben kann. Es bleibt also nur die numerische Berechnung. Dieses Kapitel stellt einige Methoden zur Berechnung von Integralen zusammen.

4.1 Riemann-Integral und Trapezregel

Die Stammfunktion

$$F(x) = \int_{a}^{x} f(t)\, dt$$

einer Funktion $f(x)$ ist die Lösung der besonders einfachen Differentialgleichung $y' = f(x)$, man könnte also dieses Kapitel einfach überspringen und für die Berechnung von Integralen auf die Methoden zur Lösung gewöhnlicher Differentialgleichungen verweisen. Dies ist aber nicht unbedingt sinnvoll. Das bestimmte Integral einer Funktion zwischen zwei Grenzen ist ein wesentlich einfacheres Konzept als die Lösung einer Differentialgleichung, man darf daher davon ausgehen, dass es auch einfachere Berechnungsverfahren dafür geben dürfte. Die relativ komplizierten Lösungsverfahren für gewöhnliche Differentialgleichungen dürften viel zu viel rechnen für das gestellte Problem.

Wir erwarten daher, dass spezialisierte Verfahren zur Berechnung von Integralen folgende Eigeschaften haben:

1. Einfache Anwendung: Der Code zur Berechnung eines Integrals sollte sehr viel einfacher ausfallen als der Code zur Lösung einer Differentialgleichung.

2. Allgemein anwendbar: Das Verfahren sollte für eine grosse Klasse von Integranden anwendbar sein, auch für solche, für die Lösungsverfahren für gewöhnliche Differentialgleichungen schwierig zu konstruieren sind. Zum Beispiel verlangen die Eindeutigkeitssätze für gewöhnliche Differentialgleichungen typischerweise eine Lifshitz-Eigenschaft. Integrale sollten aber auch von Funktionen berechnet werden können, die eine solche Eigenschaft nicht haben.

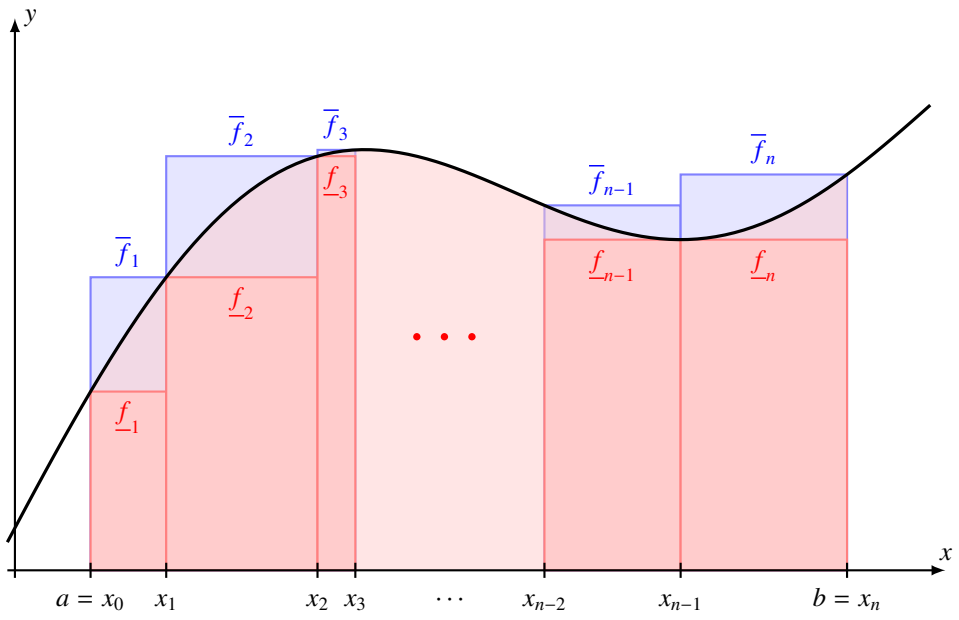


Abbildung 4.1: Definition der oberen und unteren Riemann-Summe und des Riemann-Integrals

3. Schnelle Konvergenz für "gute" Integranden. Glatte Integranden sollten nur an wenigen Stellen ausgewertet werden müssen und bereits gute Resultate ergeben.

Der besser geeignete Ausgangspunkt für die Konstruktion eines Integrationsverfahrens ist daer die usrprüngliche Definition des Riemann-Integrals, welche wir in Abschnitt 4.1.1 rekapitulieren. Daraus ergibt sich dann die Mittelpunktsformel in Abschnitt 4.1.2, deren Fehlerverhalten in Abschnitt 4.1.4 untersucht wird.

4.1.1 Das Riemann-Integral

Im Analysisunterricht wird das Riemann-Integral

$$I = \int_a^b f(x)\,dx \tag{4.1}$$

einer Funktion $f(x)$ zwischen den Grenzen a und b üblicherweise wie folgt definiert. Zunächst wird das Interval $[a, b]$ mit Hilfe einer Menge $D = \{x_0, x_1, x_2, \ldots, x_{n-1}, x_n\}$ von Zwischenpunkten mit der Eigenschaft

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

unterteilt. Das *Korn* der Unterteilung ist die Länge des längsten Teilintervals

$$\delta(D) = \max_{0 \le i < n} |x_{i+1} - x_i|.$$

Als Approximationen des Integrals (4.1) werden dann die obere und untere Riemann-Summe

$$\overline{I}(D) = \sum_{i=1}^{n} \overline{f}_i (x_i - x_{i-1}) \qquad \text{mit} \qquad \overline{f}_i = \max_{\xi \in [x_{i-1}, x_i]} f(\xi)$$

$$\underline{I}(D) = \sum_{i=1}^{n} \underline{f}_i (x_i - x_{i-1}) \qquad \text{mit} \qquad \underline{f}_i = \min_{\xi \in [x_{i-1}, x_i]} f(\xi)$$

gebildet. Aufgrund der Konstruktion ist klar, dass $\underline{I} \le I \le \overline{I}$ sein muss. Für einigermassen "glatte" Funtionen wird der Unterschied zwischen $\overline{I}(D)$ und $\underline{I}(D)$ kleiner werden, wenn man die Unterteilung verfeinert. Falls die obere und die untere Riemann-Summe bei Verfeinerung gegen den gleichen Grenzwert streben, sagt man, das *Riemann-Integral* existiere und habe den Wert

$$I = \int_a^b = \lim_{\delta(D)\to 0} \underline{I}(D) = \lim_{\delta(D)\to 0} \overline{I}(D).$$

Wenn das Integral existert, kann man offenbar auch jeden beliebigen anderen anderen Wert der Funktion in den Teilintervallen $[x_{i-1}, x_i]$ verwenden. Eine Summe der Form

$$\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) \qquad \text{mit} \qquad \xi_i \in [x_{i-1}, x_i],$$

auch *Riemann-Summe* genannt, ist für jede beliebige Wahl der Unterteilung D und der Zwischenpunkte ξ_i ein Approximation des Integrals I von (4.1).

Die Riemann-Summe liefert also bereits eine direkte Berechnungsmöglichkeit für ein beliebiges Integral, und sie liefert uns auch bereits ein Möglichkeit, die Grössenordnung des zu erwartenden Fehlers abzuschätzen: Die Differenz $\overline{I}(D) - \underline{I}(D)$ ist der grösste mögliche Fehler. Der Berechnungsaufwand lässt sich ebenfalls sehr gut abschätzen. Es muss eine Summe von n Termen gebildet werden, in jeder Summe wird eine Differenz aufeinanderfolgender Teilpunkte gebildet und ein Produkt mit dem Funktionswert. Es wird offensichtlich, dass ausser für sehr einfache Funktionen, für die man das Integral auch analytisch ausrechnen könnte, der Hauptteil der Arbeit in der Berechnung der Funktionswert $f(\xi_i)$ steckt.

Die Riemann-Summe beinhaltet einige Wahlmöglichkeiten, mit der die Berechnung optimiert werden kann. Wir können die Teilpunkte D wählen und zum Beispiel die Unterteilung dort feiner wählen, wo die Funktion sich schnell verändert. Wir können auch die Zwischenpunkte wählen. Für beides ist jedoch eine detaillierte Kenntnis der Funktion notwendig, welche nur durch die Berechnung zusätzlicher Funktionswerte gewonnen werden kann.

4.1.2 Mittelpunktsregel

Die einfachste Art der Unterteilung des Intervalls ist Teilinterval konstanter Länge zu verwenden. Wir schreiben

$$h = \frac{b-a}{n} \qquad \text{und} \qquad x_i = a + ih$$

für die Intervalllänge und die Teilpunkte. Werten wir die Funktion im Mittelpunkt eines Teilintervals aus, also in $\xi = (x_{i-1} + x_i)/2 = x_{i-1} + \frac{1}{2}h = x_i - \frac{1}{2}h$, erhalten wir als Approximation für das Integral (4.1) die *Mittelpunktsregel*

$$M(h) = \sum_{i=1}^{n} f\left(x_i - \frac{h}{2}\right) \cdot h = h \sum_{i=1}^{n} f(a + (i - \tfrac{1}{2})h). \tag{4.2}$$

Die Approximation wird in Abbildung 4.2 veranschaulicht. Für die Berechnung der Summe (4.2) sind genau n Auswertungen der Funktion $f(x)$ notwendig.

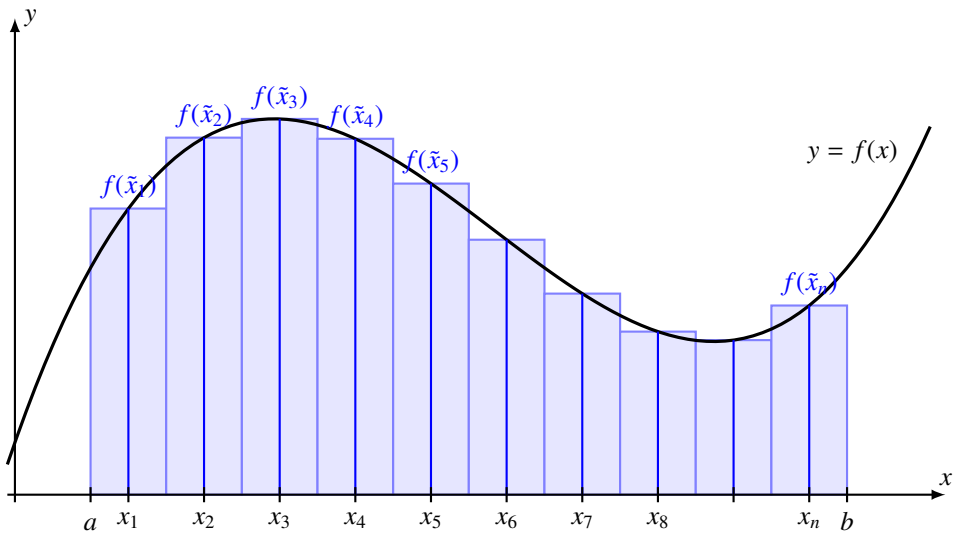


Abbildung 4.2: Approximation eines Integrals mit Hilfe der Mittelpunktsregel

Beispiel. Als Beispiel approximieren wir das Integral

$$I = \int_1^9 x\,dx = \left[\frac{x^2}{2}\right]_1^9 = \frac{9^2 - 1^2}{2} = \frac{81-1}{2} = 40$$

mit Schritten der Schrittweite $h = 2$. Die Zwischenpunkte sind $x_1 = 2$, $x_2 = 4$, $x_3 = 6$, $x_4 = 8$ oder $x_n = 2n$. Die Mittelpunktregel liefert die Approximation

$$M(2) = \sum_{i=1}^{n} f(2i)\cdot 2 = \sum_{i=1}^{n} 4i = 4\sum_{i=1}^{n} i = 4\cdot\frac{n(n+1)}{2} = 4\cdot\frac{4\cdot 5}{2} = 4\cdot 2\cdot 5 = 40.$$

In diesem Beispiel liefert die Mittelpunktformel also den exakten Wert des Integrals. Grund dafür ist natürlich, dass der Graph eine Gerade ist, und die "fehlenden" Dreieck in der Summe $M(h)$ links und rechts vom Funktionswert genau gleich gross sind, sich diese Fehler also wegheben. ○

Das Beispiel illustriert, dass die Approximation gar nicht so schlecht ist, wie sich auf Grund der deutlich hervortretenden Stufen in Abbildung 4.2 vielleicht vermuten lässt.

4.1.3 Trapezregel

Statt die Fläche unter dem Graphen von $f(x)$ in jedem Teilinterval durch den Wert im Mittelpunkt des Intervalls zu berechnen wie bei der Mittelpunktsregel, könnte sie auch approximiert werden durch ein Trapez mit Eckpunkten $(x_{i-1}, 0)$, $(x_i, 0)$, $(x_{i-1}, f(x_{i-1}))$ und $(x_i, f(x_i))$. Die Mittellinie eines solchen Trapezes hat die Länge $\frac{1}{2}(f(x_{i-1}) + f(x_i))$, die Höhe ist h. Der Flächeninhalt eines einzelnen Trapezes ist daher $\frac{h}{2}(f(x_{i-1} + f(x_i))$ und es ergibt sich die Approximation

$$\int_a^b f(x)\,dx \simeq T(h) = \frac{h}{2}\sum_{i=1}^{n}\big(f(x_{i-1} + f(x_i)\big)$$

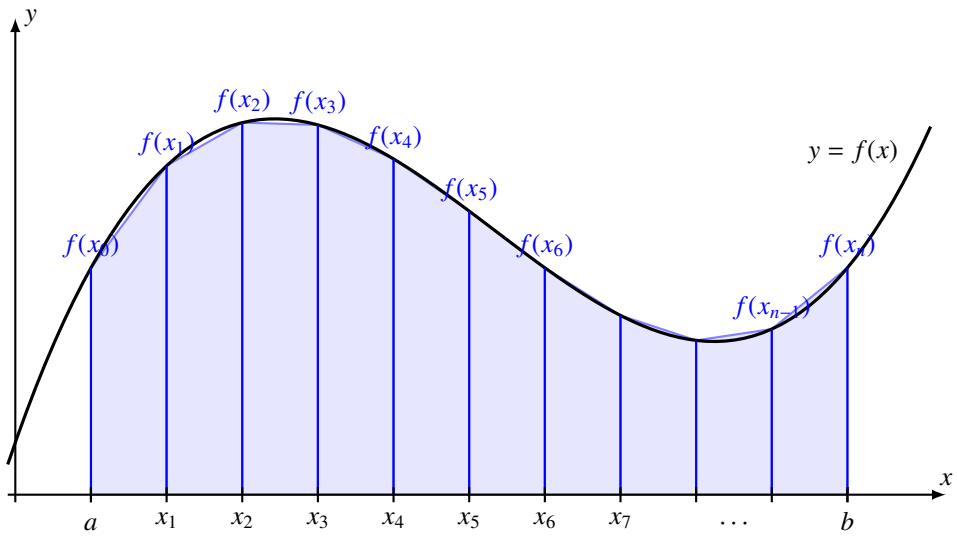


Abbildung 4.3: Approximation eines Integrals mit Hilfe der Trapezregel

$$= (\tfrac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-2}) + f(x_{n-1}) + \tfrac{1}{2}f(x_n)) \cdot h \qquad (4.3)$$

für das Integral. In Abbildung 4.3 ist die Summe der Trapezflächen dargestellt. Die Approximation $T(h)$ in (4.3) heisst *Trapezregel*. Die Trapezregel erfordert genau $n + 1$ Funktionsauswertungen.

Verfeinerung

Die Trapezregel $T(h)$ benötigt die Funktionswerte an den Stellen $x_i = a + hi$. Halbiert man die Schrittweite, müssen zusätzlich die Funktionswerte an den Zwischenpunkten

$$\frac{x_{i-1} + x_i}{2} = a + h(i - \tfrac{1}{2})$$

für $i = 1, \ldots, n$ bestimmt werden. Das sind genau die Funktionswerte, die in der Mittelpunktformel summiert wurden. Somit kann man die Trapezsumme $T(\frac{h}{2})$ durch Umordnen der Terme in der Summe durch die Trapezsummen $T(h)$ und die Mittelpunktformel $M(h)$ ausdrücken:

$$\begin{aligned}
T(\tfrac{h}{2}) &= \frac{h}{2} \cdot (\tfrac{1}{2}f(a) + f(a + 1 \cdot \tfrac{h}{2}) + f(a + 2 \cdot \tfrac{h}{2}) + \ldots f(a + (2n-1) \cdot \tfrac{h}{2}) + \tfrac{1}{2}f(a + 2n \cdot \tfrac{h}{2})) \\
&= \frac{h}{2} \cdot (\text{gerade Terme}) + \frac{h}{2} \cdot (\text{ungerade Terme}) \\
&= \frac{h}{2} \left(\tfrac{1}{2}f(a) + f(a + 2 \cdot \tfrac{h}{2}) + f(a + 4 \cdot \tfrac{h}{2}) + \cdots + \tfrac{1}{2}f(a + (2n) \cdot \tfrac{h}{2})\right) \\
&\quad + \frac{h}{2} \left(f(a + 1 \cdot \tfrac{h}{2}) + f(a + 3 \cdot \tfrac{h}{2}) + \cdots + f(a + (2n-1) \cdot \tfrac{h}{2})\right) \\
&= \frac{1}{2} \cdot h \cdot \left(\tfrac{1}{2}f(a) + f(a + h) + f(a + 2h) + \cdots + \tfrac{1}{2}f(a + nh)\right) \\
&\quad + \frac{1}{2} \cdot h \cdot \left(f(a + (1 - \tfrac{1}{2})h) + f(a + (2 - \tfrac{1}{2})h) + \cdots + f(a + (n - \tfrac{1}{2})h)\right)
\end{aligned}$$

$$= \frac{1}{2}T(h) + \frac{1}{2}M(h). \tag{4.4}$$

Halbierung der Schrittweite bedeutet daher nur, dass man die Mittelpunktregel $M(h)$ zusätzlich auswerten muss.

Für die Berechnung von $T(\frac{h}{2})$ sind $2n + 1$ Funktionsauswertungen notwendig, davon entfallen $n + 1$ auf die Berechnung von $T(h)$ und n auf die Berechnung von $M(h)$.

Beispiel. Als Beispiel berechnen wir das Integral

$$I = \int_0^1 \sqrt{1 - x^2}\, dx. \tag{4.5}$$

Es berechnet die Fläche des im ersten Quadranten liegenden Teils des Einheitskreises, hat also den Wert $I = \frac{\pi}{4} \simeq 0.78539816$.

Das folgende Octave-Programm kann zur Berechnung verwendet werden:

```
1  a = 0;
2  b = 1;
3  h = b-a;
4
5  function y = f(x)
6      y = sqrt(1 - x*x);
7  endfunction
8
9  function summe = M(h, a, b, f)
10     x = a + h/2;
11     s = 0;
12     while (x < b)
13         s = s + f(x);
14         x = x + h;
15     end
16     summe = h * s;
17 endfunction
18
19 T = h * ( (1/2) * f(a) + (1/2) * f(b) )
20
21 for k = (2:20)
22     T = (1/2) * T + (1/2) * M(h, a, b, @f);
23     h = h/2;
24 endfor
```

In Zeile 5 wird der Integrand als Funktion definiert. In Zeile 9 wird die Funktion `M(h,a,b,f)` definiert, die die Mittelpunktsumme $M(h)$ der Funktion f berechnet. Es ist nicht nötig, eine Funktion für $T(h)$ zu definieren, da sich der Wert $T(\frac{h}{2})$ aus $T(h)$ und $M(h)$ ergibt. In Zeile 19 wird der erste Wert $T(1)$ berechnet, In der nachfolgenden Schleife zwischen Zeilen 21 und 24 wird die Rekursion (4.4) angewendet. Die sich ergebenden Zahlenwerte sind in Tabelle 4.1 zusammengestellt.

Für den Wert von $T(h)$ in Zeile k der Tabelle sind $2^{k-1} + 1$ Funktionsauswertungen nötig, auf der letzen Zeile sind dies 524289 Funktionsauswertungen. Es fällt auf, dass die Konvergenz ziemlich langsam ist. Der Fehler von $T(h)$ nimmt ungefähr so schnell ab, wie h kleiner wird. Sollte sich diese Systematik bestätigen lassen, könnte man daraus ein Verfahren zur Beschleunigung der Konvergenz ableiten. ○

k	$T(h)$	h
1	0.5000000000000000	1.0000000000000000
2	0.6830127018922193	0.5000000000000000
3	0.7489272670256102	0.2500000000000000
4	0.7724547860892934	0.1250000000000000
5	0.7808132594569352	0.0625000000000000
6	0.7837756057192828	0.0312500000000000
7	0.7848242281949210	0.0156250000000000
8	0.7851951980991537	0.0078125000000000
9	0.7853263957393075	0.0039062500000000
10	0.7853727881799138	0.0019531250000000
11	0.7853891916347545	0.0009765625000000
12	0.7853949913528619	0.0004882812500000
13	0.7853970419019385	0.0002441406250000
14	0.7853977668874246	0.0001220703125000
15	0.7853980232097235	0.0000610351562500
16	0.7853981138335553	0.0000305175781250
17	0.7853981458739578	0.0000152587890625
18	0.7853981572019572	0.0000076293945312
19	0.7853981612070116	0.0000038146972656
20	0.7853981626230124	0.0000019073486328
	0.7853981633974483	

Tabelle 4.1: Approximation des Integrals (4.5) mit Hilfe der Trapezregel. Unterstrichen die Stellen des Wertes von $T(h)$, die bereits korrekt sind.

4.1.4 Fehler von Trapez- und Mittelpunktsregel

Im Beispiel im vorangenangenen Abschnitt hat sich in Tabelle 4.1 eine Gesetzmässigkeit in Entwicklung der Fehler von $T(h)$ gezeigt. Es schien, als wäre $I - T(h) \sim h$. Ziel dieses Abschnittes ist, dies nachzuweisen. Als Werkzeug dafür benötigen wir die Euler-Maclaurinsche Summenformel.

Partielle Integration

Der Fehler der Trapezregel dürfte umso grösser werden, je stärker der Graph der Funktion $f(x)$ gekrümmt ist. Wir vermuten daher einen Zusammenhang zwischen dem Fehler und der zweiten Ableitung von $f(x)$. Um diesen Zusammenhang zu finden, betrachten wir das Integral

$$I = \int_0^1 g(t)\,dt. \tag{4.6}$$

Den Integranden können wir auch als Produkt $1 \cdot g(t)$ der konstanten Funktion 1 mit $g(t)$ schreiben und das Integral partiell ausführen:

$$I = \int_0^1 1 \cdot g(t)\,dt = \Big[p_1(t) \cdot g(t)\Big]_0^1 - \int_0^1 p_1(t) \cdot g'(t)\,dt$$

wobei wir für $p_1(t)$ ein beleibiges Polynom wählen können, dessen Ableitung $p_1'(t) = 1$ ist. Das zweite Integral können wir nochmals partiell integrieren und erhalten

$$= \Big[p_1(t)\cdot g(t)\Big]_0^1 - \Big[p_2(t)\cdot g'(t)\Big]_0^1 + \int_0^1 p_2(t)\cdot g''(t)\,dt$$

wobei wieder $p_2'(t) = p_1(t)$ sein muss. Dieser Prozess lässt sich natürlich wiederholen und ergibt nach k Schritten

$$\begin{aligned} = &\Big[p_1(t)\cdot g(t)\Big]_0^1 - \Big[p_2(t)\cdot g'(t)\Big]_0^1 + \Big[p_3(t)\cdot g''(t)\Big]_0^1 - \Big[p_4(t)\cdot g'''(t)\Big]_0^1 + \dots \\ &- (-1)^k\Big[p_k(t)\cdot g^{(k-1)}(t)\Big]_0^1 + (-1)^k\int_0^1 p_k(t)g^{(k)}(t)\,dt \end{aligned} \tag{4.7}$$

Die Polynome $p_k(t)$ sind bis jetzt noch nicht eindeutig festgelegt, es ist nur die Rekursionsformel

$$p_k'(t) = p_{k-1}(t) \qquad \text{für alle } k > 0 \text{ und} \qquad p_0(t) = 1$$

gegeben. Durch eine geschickte Wahl der Polynome soll jetzt aus der Entwicklung möglichst viel Information über das Integral extrahiert werden.

Wir betrachten zunächst den Fall, dass $g(t)$ selbst ein Polynom ist. Genügend hohe Ableitungen eines Polynoms verschwinden, ist die Ordnung k der Ableitung grösser als der Grad des Polynoms, dann ist $g^{(k)}(t) = 0$. Dies bedeutet, dass in der Summe (4.7) der Integralterm verschwindet, sobald k grösser als der Grad von g ist. Aber auch für $k = \deg g$ kann man den letzten Term zum verschwinden bringen, wenn man $p_k(t)$ geeignet wählt. In diesem Fall ist $g^{(k)}(t)$ konstant, der Integralterm verschwindet also genau dann, wenn das Integral von $p_k(t)$ verschwindet. Wir legen daher die Polynome $p_k(t)$ durch die Forderung fest, dass

$$\int_0^1 p_k(t)\,dt = 0 \qquad \forall k \tag{4.8}$$

sein soll.

Die Polynome $p_k(t)$

Wir berechnen die ersten paar Polynome der Folge $p_k(t)$. Zunächst ist $p_1(t) = t + C$, wir müssen also C so wählen, dass das Integral

$$\int_0^1 p_1(t)\,dt = \int_0^1 t + C\,dt = \Big[\frac{1}{2}t^2 + Ct\Big]_0^1 = \frac{1}{2} + C$$

verschwindet. Es folgt $C = -\frac{1}{2}$ und $p_1(t) = t - \frac{1}{2}$.

Für $p_2(t)$ folgt zunächst durch Integration $p_2(t) = \frac{1}{2}t^2 - \frac{1}{2}t + C$ und dann aus der Integralbedingung

$$\int_0^1 p_2(t)\,dt = \Big[\frac{1}{6}t^3 - \frac{1}{4}t^2 + Ct\Big]_0^1 = \frac{1}{6} - \frac{1}{4} + C = -\frac{1}{12} + C,$$

woraus $C = \frac{1}{12}$ folgt und damit $p_2(t) = \frac{1}{2}t^2 - \frac{1}{2}t - \frac{1}{12}$. Auf dies Art kann man schrittweise die folgenden Polynome finden:

$$p_1(t) = t - \frac{1}{2}$$

$$
\begin{aligned}
p_2(t) &= \frac{1}{2}t^2 - \frac{1}{2}t + \frac{1}{12} \\
p_3(t) &= \frac{1}{6}t^3 - \frac{1}{4}t^2 + \frac{1}{12}t \\
p_4(t) &= \frac{1}{24}t^4 - \frac{1}{2}t^3 + \frac{1}{24}t^2 - \frac{1}{720} \\
p_5(t) &= \frac{1}{120}t^5 - \frac{1}{240}t^4 + \frac{1}{360}t^3 - \frac{1}{3600}t
\end{aligned}
$$

Der führende Koeffizient des Polynoms $p_k(t)$ ist $1/k!$, indem man diesen ausklammert, kann man etwas einfachere Polynome bekommen:

$$
\begin{aligned}
& & B_0(t) &= 1 \\
p_1(t) &= 1 \cdot B_1(t) & B_1(t) &= t - \frac{1}{2} \\
p_2(t) &= \frac{1}{2!} \cdot B_2(t) & B_2(t) &= t^2 - t + \frac{1}{6} \\
p_3(t) &= \frac{1}{3!} \cdot B_3(t) & B_3(t) &= t^3 - \frac{3}{2}t^2 + \frac{1}{2}t \\
p_4(t) &= \frac{1}{4!} \cdot B_4(t) & B_4(t) &= t^4 - 2t^3 + t^2 - \frac{1}{30} \\
p_5(t) &= \frac{1}{5!} \cdot B_5(t) & B_5(t) &= t^5 - \frac{5}{2}t^4 - \frac{5}{3}t^3 - \frac{1}{6}t
\end{aligned}
$$

Die Polynome $B_k(t)$ heissen *Bernoulli-Polynome*.

Eigenschaften der Polynome $p_k(t)$

Auch ohne die expliziten Formeln des vorangenangen Abschnittes lassen sich nützliche Symmetrieeigenschaften der Polynome $p_k(t)$ bezüglich des Punktes $t = \frac{1}{2}$ ableiten.

Die Symmetrieeigenschaften sind etwas verschleiert durch die Wahl des Intervals $[0, 1]$ als Definitionsbereich. Wir untersuchen die Symmetrie der Polynome bezüglich des Punktes $t_0 = \frac{1}{2}$. Wir schreiben $\tilde{p}_k(t)$ für die Stammfunktion von $p_{k-1}(t)$ mit Integrationskonstante $C_k = 0$, es ist also

$$
p_k(t) = \tilde{p}_k(t) + C_k \qquad \text{und} \qquad \tilde{p}_k(0) = 0.
$$

Die Integrationskonstante C_k wurde im vorangegangenen Abschnitt jeweils bestimmt.

Zunächst halten wir fest, dass die Polynome $p_k(t)$ alle entweder gerade oder ungerade sind bezüglich t_0.

Lemma 4.1. *Die Polynome $p_{2k}(t)$ sind gerade bezüglich t_0, die Polynome $p_{2k+1}(t)$ sind ungerade bezüglich t_0. Insbesondere ist $p_{2k}(0) = p_{2k}(1)$ und $p_{2k+1}(1) = -p_{2k+1}(0)$.*

Beweis. Die ersten zwei Polynome $p_0(t) = 1$ und $p_1(t) = t - \frac{1}{2}$ haben tatsächlich die genannten Symmetrieeigenschaften. Um die Aussage zu beweisen muss jetzt also nur untersucht werden, dass $p_k(t)$ die verlangten Eigenschaften hat, wenn alle vorangegangenen Polynome die "richtigen" Symmetrien hat.

Nehmen wir an, dass $p_{k-1}(t)$ gerade ist bezüglich t_0, wir müssen zeigen, dass $p_k(t)$ ungerade bezüglich t_0 ist. Zunächst ist klar, dass $p_k(t) - p_k(t_0)$ ungerade ist bezüglich t_0. Folglich verschwindet

das Integral von $p_k(t) - p_k(t_0)$ oder

$$0 = \int_0^1 p_k(t) - p_k(t_0)\,dt = \int_0^1 p_k(t)\,dt - p_k(t_0) = -p_k(t_0),$$

d. h. auch $p_k(t_0)$ verschwindet und $p_k(t)$ ist bereits ungerade bezüglicch t_0.

Nehmen wir jetzt an, dass $p_{k-1}(t)$ ungerade ist bezüglich t_0, wir müssen zeigen, dass $p_k(t)$ gerade ist. Es ist

$$\begin{aligned}
p_k(t_0+\tau) - p_k(t_0) &= \int_{t_0}^{t_0+\tau} p_{k-1}(t)\,dt = \int_0^{\tau} p_{k-1}(t_0+s)\,ds \\
&= -\int_0^{\tau} p_{k-1}(t_0-s)\,ds = \int_0^{-\tau} p_{k-1}(t_0+\tilde{s})\,d\tilde{s} \qquad \text{mit} \qquad \tilde{s} = -s \\
&= \int_{t_0}^{t_0-\tau} p_{k-1}(\tilde{t})\,d\tilde{t} \qquad \text{mit} \qquad \tilde{t} = t_0 + \tilde{s} \\
&= p_k(t_0-\tau) - p_k(t_0) \\
\Rightarrow \qquad p_k(t_0+\tau) &= p_k(t_0-\tau),
\end{aligned}$$

das Polynom $p_k(t)$ ist also gerade bezüglich t_0. □

Die spezielle Wahl der Polynome $p_k(t)$ führt jetzt dazu, dass die Formel für das Integral (4.6) von $g(t)$ über das Interval $[0, 1]$ weiter vereinfacht werden kann. Die Werte von $p_k(t)$ an den Intervallgrenzen haben gemäss Lemma 4.1 gleiche Werte für gerade k und entgegengesetzte Werte für ungerade k. Es gilt

$$\begin{aligned}
I = \int_0^1 g(t)\,dt = p_1(1)(g(1)+g(0)) - p_2(1)(g'(1)-g'(0)) + p_3(1)(g''(1)+g''(0)) \\
- \cdots + (-1)^k \int_0^1 p_k(t) g^{(k)}(t)\,dt
\end{aligned} \tag{4.9}$$

Für $k = 2$ ergibt sich die einfachere Formel

$$\int_0^1 g(t)\,dt = \underbrace{\frac{1}{2}(g(1)+g(0))}_{=T(1)} + p_2(1)(g'(0)-g'(1)) + \int_0^1 p_2(t)\,g''(t)\,dt, \tag{4.10}$$

welche einen Zusammenhang zwischen dem Integral und der einfachsten Form der Trapezsumme herstellt.

Euler-Maclaurin Summenformel

Unser Interesse gilt den Summen $T(h)$ und $M(h)$, wir möchten gerne den Fehler solche Summen abschätzen. Die Formel (4.10) vergleicht ein einzelnes Trapez (den Wert $T(1)$ auf der rechten Seite) mit dem gesuchten Integral. und beschreibt eine Formel für den Unterschied. Um den Fehler der Trapezformel zu bestimmen, müssen wir Sumen von vielen Trapezen berechnen.

Wir bestimmen jetzt das Integral

$$I_n = \int_0^n g(t)\,dt.$$

Wir zerlegen das Integral in viele einzelne Teilintegrale auf Intervallen der Länge 1:

$$\int_0^n g(t)\,dt = \int_0^1 g(t)\,dt + \int_1^2 g(t)\,dt + \int_2^3 g(t)\,dt + \cdots + \int_{n-1}^n g(t)\,dt$$

und wenden die Formel (4.9) auf jeden einzelnen Term an

$$\begin{aligned}
&= p_1(1)\sum_{i=1}^{n}(g(i)+g(i-1)) + p_2(1)\sum_{i=1}^{n}(g'(i)-g'(i-1)) \\
&\quad + p_3(1)\sum_{i=1}^{n}(g''(i)+g''(i-1)) + p_4(1)\sum_{i=1}^{n}(g'''(i)-g'''(i-1)) \\
&\quad + \int_0^1 p_k(t)\sum_{i=1}^{n} g^{(4)}(k+t)\,dt.
\end{aligned}$$

In den Summen über Differenzen heben sich die inneren Terme jeweils weg:

$$\begin{aligned}
\sum_{i=1}^{n}(g'(i)-g'(i-1)) &= g'(1)-g'(0)+g'(2)-g'(1)+g'(3)-g'(2)+\cdots+g'(n)-g'(n-1) \\
&= g'(n)-g'(0) \\
\sum_{i=1}^{n}(g'''(i)-g'''(i-1)) &= g'''(1)-g'''(0)+g'''(2)-g'(1)+\cdots+g'''(n)-g'''(n-1) \\
&= g'''(n)-g'''(0).
\end{aligned}$$

Damit kann man die Formel für das Integral weiter vereinfachen und erhält

$$\begin{aligned}
\int_0^n g(t)\,dt &= p_1(1)\sum_{i=1}^{t}(g(i)+g(i-1)) + p_2(1)(g'(n)-g'(0)) \\
&\quad + p_3(1)\sum_{i=1}^{n}(g''(i)+g''(i-1)) + p_4(1)(g'''(n)-g'''(0)) + \int_0^1 p_4(t)\sum_{i=1}^{n} g^{(4)}(i+t)\,dt.
\end{aligned} \tag{4.11}$$

Dies ist ein Spezialfall der *Euler-Maclaurinschen Summenformel*.

Fehler der Trapezformel

Die Euler-Maclaurinsche Summenformel (4.11) kann noch etwas vereinfacht werden. Wir wissen, dass $p_3(t)$ eine bezüglich $t_0 = \frac{1}{2}$ ungerade Funktion ist, also ist $p_3(1) = -p_3(0) = 0$, wie man aus der früher hergeleiteten expliziten Formel für $p_3(t)$ berechnen kann. Damit vereinfacht sich die Formel zu

$$\begin{aligned}
\int_0^n g(t)\,dt &= \sum_{i=1}^{n}\frac{1}{2}(g(i-1)+g(i)) + p_2(1)(g'(n)-g'(0)) + p_4(1)(g'''(n)-g'''(0)) \\
&\quad + \int_0^1 p_4(t)\sum_{i=1}^{n} g^{(4)}(i+t)\,dt
\end{aligned}$$

Der erste Term auf der rechten Seite sieht aus wie eine Trapezsumme, nur ist es die falsche Funktion und es sind die falschen Teilpunkte.

Durch eine geeignete Variablentransformation können wir das Intervall $[a, b]$ auf das Intervall $[0, n]$ transformieren. Dazu schreiben wir

$$x = a + \frac{b-a}{n} \cdot t = a + ht \qquad \text{mit} \qquad h = \frac{b-a}{n}.$$

Die Funktion

$$g(t) = f(x) = f(a + ht)$$

kann dann zur Berechnung des Integrals herangezogen werden. Zunächst folgt aus der Variablentransformation

$$\begin{aligned} \int_0^n g(t)\,dt &= \int_0^n f(\underbrace{a+ht}_{=x})\,dt \qquad \text{mit} \qquad t = \frac{x-a}{h} \\ &= \frac{1}{h}\int_a^b f(x)\,dx \\ \Rightarrow \qquad \int_a^b f(x)\,dx &= h\int_0^n g(t)\,dt. \end{aligned}$$

In der Euler-Maclaurinschen Summenformel brauchen wir die Ableitungen von $g(t)$ die wir natürlich wieder durch Ableitungen von $f(x)$ ausdrücken möchten:

$$\begin{aligned} g'(t) &= hf'(a+ht) = hf'(x) \\ g''(t) &= h^2 f''(x) \\ &\vdots \quad \vdots \\ g^{(k)}(t) &= h^k f^{(k)}(x) \end{aligned}$$

Setzen wir dies alles in die Summenformel ein, erhalten wir

$$\begin{aligned} \int_a^b f(x)\,dx &= \sum_{i=1}^n \frac{h}{2}(f(x_{i-1}) + f(x_i)) + \frac{h^2}{12}(f'(b) - f'(a)) - \frac{h^4}{720}(f'''(b) - f'''(a)) \\ &\quad + h\int_0^1 p_4(t) \sum_{i=1}^n g^{(4)}(i+t)\,dt \\ &= T(h) + \frac{h^2}{12}(f'(b) - f'(a)) - \frac{h^4}{720}(f'''(b) - f'''(a)) \\ &\quad + h\sum_{i=1}^n \int_0^1 p_4(t) g^{(4)}(i+t)\,dt. \end{aligned}$$

Die Integrale im letzten Term können wir wieder auf die Variable x transformieren. Dazu verwenden wir für den Summanden i die Variablentransformation

$$x = x_{i-1} + ht \quad \Rightarrow \quad t = \frac{1}{h}(x - x_{i-1}), \quad dt = \frac{1}{h}\,dx$$

und erhalten

$$\begin{aligned}\int_0^1 p_4(t)g^{(4)}(i+t)\,dt &= \int_{x_{i-1}}^{x_i} p_4(\tfrac{1}{h}(x-x_{i-1}))h^4 f^{(4)}(x)\,\frac{1}{h}\,dx \\ &= h^3\int_{x_{i-1}}^{x_i} p_4(\tfrac{1}{h}(x-x_{i-1}))f^{(4)}(x)\,dx\end{aligned}$$

Die Funktion p_4 im Integranden kann mit Hilfe der früher abgeleiteten expliziten Formel für $p_4(t)$ abgeschätze werden, sei K eine Konstante so dass $|p_4(t)| \le K$ für $0 \le t \le 1$. Dann ist können wir das Integral auf der rechten Seite abschätzen durch

$$\left|\int_0^1 p_4(t)g^{(4)}(i+t)\,dt\right| \le h^3\int_{x_{i-1}}^{x_i} K|f^{(4)}(x)|\,dx$$

Summieren wir über i erhalten wir

$$\left|\sum_{i=1}^{n}\int_0^1 p_4(t)g^{(4)}(i+t)\,dt\right| \le h^3\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} K|f^{(4)}(x)|\,dx = h^3K\int_a^b f^{(4)}(x)\,dx = h^3K_1$$

mit einer neuen Konstanten K_1.

Nach dieser langen Rechnung können wir jetzt den Fehler für die Trapezformel vollständig hinschreiben

$$\begin{aligned}\int_a^b f(x)\,dx &= T(h) + h^2\cdot\frac{f'(b)-f'(a)}{12} - h^4\frac{f'''(b)-f'''(a)}{720} + h^4K_1 \\ &= T(h) + Ch^2 + O(h^4). \qquad (4.12)\end{aligned}$$

Die dominante Komponente des Fehlers ist also der Term Ch^2, der verbleibende Fehler ist von der Grössenordnung $O(h^4)$, also im allgemeinen viel kleiner.

Man beachte, dass diese Abschätzungen nur funktionieren, wenn die dritten Ableitungen von $f(x)$ an den Grenzen und die vierten Ableitungen von $f(x)$ im inneren des Intervals $[a,b]$ nicht übermässig gross werden.

Wie muss h gewählt werden?

Wie gross h genau gewählt werden muss hängt natürlich von der Konstanten C ab. Aus

$$C = \frac{f'(b)-f'(a)}{12}$$

kann man gegebenenfalls ein Abschätzung gewinnen. Zum Beispiel weiss man, dass die Ableitung der Funktion $\sin x$ betragsmässig niemals grösser als 1 werden kann. Für das Integral der Sinusfunktion kann man also davon ausgehen, dass höchstens $C = 1/6$ ist. Einen Fehler der Grösse ε erhält man also, wenn man h so wählt, dass $h^2/6 < \varepsilon$ ist, oder $h < \sqrt{6\varepsilon}$. Für 10 Nachkommastellen bedeutet dies $h < \sqrt{6\cdot 10^{-10}}$. Um das Integral über das Interval $[0,\pi]$ mit dieser Schrittweite zu berechnen, sind über 128000 Funktionsauswertungen notwendig.

Man kann die Frage auch umgekehrt stellen: Welche Genauigkeit kann man sinnvollerweise von der Trapezregel erwarten? Dabei kann man ein in vernünftiger Zeit durchführebare Rechnung von etwa 10^3 bis 10^4 Funktionsauswertungen zu Grunde legen. Dies führt im genannte Beispield es Integrals der Sinusfunktion auf eine Schrittweite zwischen $h = 0.003$ und $h = 0.0003$. Für diese h

wird der Fehler zwischen 0.00001 und 0.0000001 gross. Viel mehr als etwa sechs Nachkommastellen lassen sich mit der Trapezregel alleine also nicht gewinnen. Dies wird auch durch die Rechnungen im Beispiel im nächsten Abschnitt bestätigt.

Das Beispiel auf Seite 98 zeigt auch, dass die Konvergenz noch schlechter sein kann. Da die Ableitung des Integranden in diesem Bespiel an der Stelle $x = 1$ divergiert, wird der Koeffiziente C in der Fehlerformel sehr gross.

4.2 Romberg-Algorithmus

Die bisher betrachteten Methoden zur Berechnung von Integralen zeichnen sich dadurch aus, dass die Fehler bei einer gegebenen Schrittweite sehr genau bekannt sind. Es sollte daher möglich sein, die Ideen von Abschnitt **??** anzuwenden und die Konvergenz des Verfahrens zu beschleunigen, ohne dass weitere Funktionswerte berechnet werden müssen.

Elmination des Fehlers

Wir kehren zurück zur Berechnung des Integrals

$$I = \int_a^b f(x)\,dx$$

mit Hilfe der Trapezregel und bezeichnen wieder mit $T(h)$ die Approximation mit der Trapezregel mit Schrittweite h. In (4.12) haben wir den Fehler von $T(h)$ bestimmt. Wir nehmen an, dass der Fehler vierter Ordnung im Vergleich zum Term Ch^2 vernachlässigbar ist. Wenn wir die Schrittweite immer wieder halbieren, erhalten wir immer bessere Approximationen

$$\begin{aligned} I &= T(h) + Ch^2 \\ &= T\Big(\frac{h}{2}\Big) + \frac{Ch^2}{4}. \end{aligned}$$

Nehmen wir für den Moment an, dass dies exakte Gleichungen sind, dann können wir die Unbekannte C eliminieren und nach der Unbekannten I auflösen. Dazu multiplizieren wir die zweite Gleichung mit 4 und subtrahieren die erste Gleichung. Wir erhalten

$$3I = 4T\Big(\frac{h}{2}\Big) - T(h) \qquad \Rightarrow \qquad I = \frac{4T(h/2) - T(h)}{3}.$$

Natürlich ist dies auch noch nicht der exakte Wert des Integrals, wir haben die die Terme vierter Ordnung vernachlässigt. Wir können aber daraus schliessen, dass

$$T^*(h/2) = \frac{4T(h/2) - T(h)}{3}$$

eine Approximation für das Integral ist, deren Fehler nur noch von vierter Ordnung $O(h^4)$ sein wird.

Da wir für die Approximationen $T^*(h)$ auch wieder eine Approximation für die Fehler haben, können wir das gleiche Prinzip nochmals anwenden und auch den Fehler vierter Ordnung eliminieren. Wir multiplizieren die zweite Gleichung in

$$I = T^*(h) + Ch^4$$

k	$T(\pi 2^{-k})$	$T^*(\pi 2^{-k})$	$T^{**}(\pi 2^{-k})$
0	0.0000000000000002		
1	1.5707963267948966	2.0943951023931953	
2	1.8961188979370398	2.0045597549844207	1.9985707318238357
3	1.9742316019455508	2.0002691699483877	1.9999831309459855
4	1.9935703437723393	2.0000165910479355	1.9999997524545718
5	1.9983933609701441	2.0000010333694123	1.9999999961908441
6	1.9995983886400366	2.0000000645300009	1.9999999999407068
7	1.9998996001842024	2.0000000040322576	1.9999999999990750
8	1.9999749002350549	2.0000000002520060	1.9999999999999891
9	1.9999937250705797	2.0000000000157545	2.0000000000000044
10	1.9999984312683776	2.0000000000009770	1.9999999999999920
11	1.9999996078171345	2.0000000000000533	1.9999999999999918

Tabelle 4.2: Berechung des Integrals $\int_0^\pi \sin x\,dx$ mit Hilfe der Trapezregel (Spalte $T(\pi 2^{-k})$) und Beschleunigung der Konvergenz mit Hilfe des Romberg-Algorithmus. Für die Resultate auf der Zeile $k > 0$ sind $2^k + 1$ Funktionsauswertungen notwendig.

$$= T^*(h/2) + C\frac{h^4}{16}$$

mit 16 und subtrahieren die erste, so erhalten wir

$$15I = 16T^*(h/2) - T(h) \qquad \Rightarrow \qquad I = \frac{16T^*(h/2) - T^*(h)}{15}.$$

Dieses Verfahren der fortlaufenden Verbesserung der Konvergenz durch Elimination des Fehlers ist bekannt als *Romberg-Algorithmus*.

Ein Beispiel

Wir berechnen das Integral

$$I = \int_0^\pi \sin x\,dx = 2$$

mit Hilfe der Trapezregel und wenden das Beschleunigungsschema von Romberg auf die erhaltenen Werte an.

Die gefundenen Resultate sind in Tabelle 4.2 dargestellt. Man kann erkennen, dass die Maschinengenauigkeit von 15 Nachkommastellen mit der Trapezregel auch bei $k = 11$ nicht erreicht wird. Für die erste Romberg-Beschleunigung T^* wird die Genauigkeit bei $k = 11$ erreicht, sie kann aber durch Erhöhung von k nicht mehr verbessert werden, da die aufsummierten Rundungsfehler wichtiger werden. In beiden Fällen muss der Integrand 2049 mal ausgewertet werden.

Die zweite Romberg-Beschleunigung erreicht Maschinengenauigkeit bereits bei $k = 9$, als mit nur 513 Funktionsauswertungen. Die Romberg-Beschleunigung ist in diesem Beispiel also ausserordentlich erfolgreich.

Allgemeine Form des Romberg-Verfahrens

In noch allgemeinerer Form kann man die Formeln des Romberg-Algorithmus wie folgt schreiben. Die Werte $T_{k,0} = T(2^{-k}h)$ mit $k = 0, 1, \ldots$ sind die Werte der Trapezsumme mit $h = 2^{-k}(b-a)$. Die Beschleunigungen sind

$$\begin{aligned} T_{k,1} &= \frac{4T_{k,0} - T_{k-1,0}}{4-1} \\ T_{k,2} &= \frac{4^2T_{k,1} - T_{k-1,1}}{4^2-1} \\ &\vdots \quad \vdots \\ T_{k,l} &= \frac{4^lT_{k,l-1} - T_{k-1,l-1}}{4^l-1} \end{aligned}$$

Dieses Verfahren funktioniert und beschleunigt die Konvergenz, sofern die Ableitungen von $f(x)$ bis zur Ordnung $2l$ ausreichend glatt und beschränkt sind.

Wieviele Stellen Genauigkeit lassen sich erreichen?

Das Romberg-Verfahren basiert darauf, dass man die Trapezsummen für sukzessive halbiert Schrittlänge berechnet. In jeder Iteration wird die Schrittweite also halbiert. Die Fehler von $T_{k,l}$ sind von der Ordnung h^{2l}, also

$$T_{k,l} = I + O(2^{-2(l+1)})$$

Auf jeder Zeile wird der Fehler also um den Faktor $2^{-2l} = \frac{1}{4^l}$ kleiner, dies entspricht einem Gewinn von zwei Binärstellen oder $2(l+1)\log_{10} 2 = 0.60206(l+1)$ Dezimalstellen. Man kann dies auch in der Tabelle 4.2 verfolgen. In der ersten Spalte mit $l = 0$ braucht es etwa 10 Zeilen, um 6 Dezimalstellen zu gewinnen, in der zweiten Spalte mit $l = 1$ sind 6 Dezimalstellen bereits nach 5 Zeilen erreicht, 12 Dezimalstellen nach 10 Zeilen. In der dritten Spalte mit $l = 2$ erreicht man 12 Stellen bereits ungefähr nach 5 Zeilen.

Kapitel 5

Gewöhnliche Differentialgleichungen

Gewöhnliche Differentialgleichungen gehören zu den wichtigsten Problemen, die vor allem mit numerischen Methoden gelöst werden. Nur selten lassen sich Probleme der Praxis in geschlossener Form lösen. In diesem Kapitel wird nach einer Rekapitulation der Problemstellung in Abschnitt 5.1 zunächst das Grundprinzip der numerischen Lösungsverfahren vorgestellt. In Abschnitt 5.4 wird das Runge-Kutta-Verfahren entwickelt, welches einen guten Mittelweg zwischen Genauigkeit und Komplexität darstellt. Später im Kapitel kommen Mehrschrittverfahren und das Randwertproblem zur Rede.

5.1 Problemstellung

Eine gewöhnliche Differentialgleichung für eine reellwertige Funktion $y(x)$ stellt einen Zusammenhang her zwischen der Funktion und ihren Ableitungen. Wir schreiben die Ableitungen als y', y'', y''' und $y^{(n)}$ für die n-te Ableitung. Wir lassen oft das Argument der Funktion weg. Beispiele von Differentialgleichungen sind

$$\begin{aligned} y' &= -Ny && \text{Ordnung: 1} \\ y'' &= -\omega^2 y && \text{Ordnung: 2} \\ x^2y'' + xy' + (x^2 - n^2)y &= 0 && \text{Ordnung: 2} \end{aligned}$$

Die Abhängigkeit kann in expliziter Form als

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \tag{5.1}$$

oder in impliziter Form

$$F(x, y, y', \ldots, y^{(n)}) = 0$$

gegeben sein. Die Ordnung einer Differentialgleichung ist die höchste vorkommende Ableitung.

Differentialgleichungen erster Ordnung lassen sich mit Hilfe eines Richtungsfeldes visualisieren, wie in Abbildung 5.1 dargestellt. In jedem Punkt (x, y) der x-y-Ebene wird die Steigung $y' = f(x, y)$ eingezeichnet. Eine Lösung der Differentialgleichung hat in diesem Bild als Graph eine Kurve in der x-y-Ebene, die an jeder Stelle als Tangente das Richtungsfeld haben.

Insbesondere in Anwendungen in der Physik ist die Zeit die unabhängige Variable. Die abhängige Variable ist dann zum Beispiel die Ortskoordinate $x(t)$ und wir bezeichnen ihre Ableitungen mit

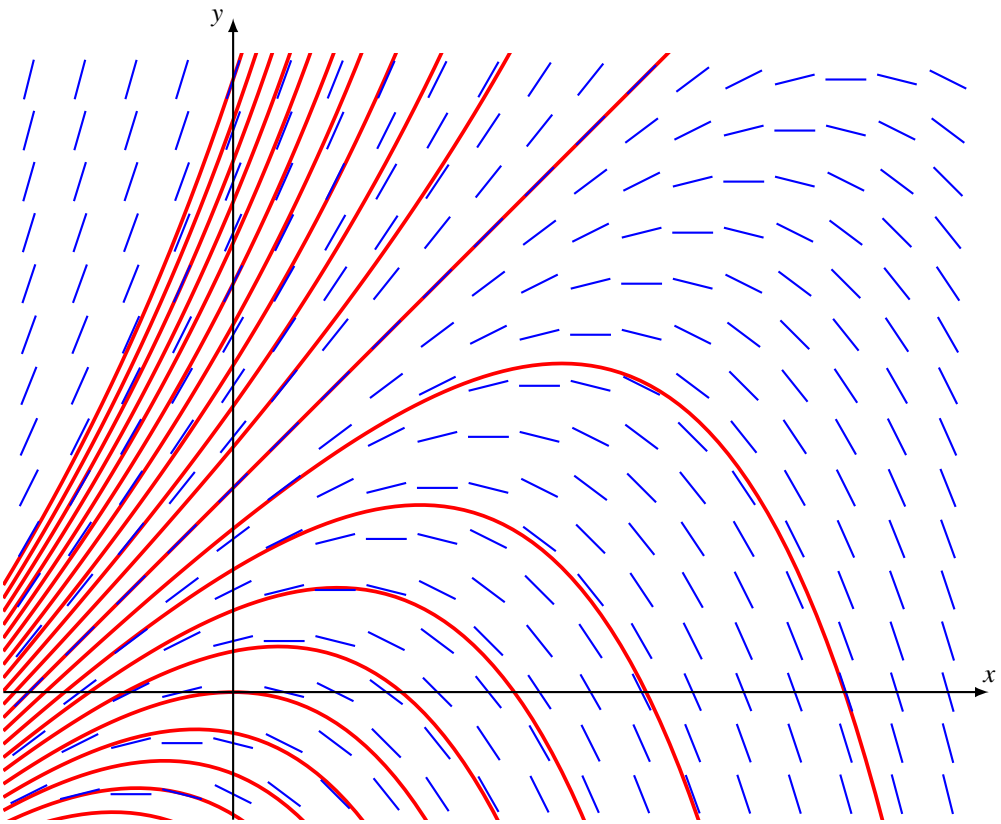


Abbildung 5.1: Richtungsfeld der Differentialgleichung $y' = y - x$ mit einzelnen Lösungskurven.

$\dot{x}(t)$ für die Geschwindigkeit, $\ddot{x}(t)$ für die Beschleunigung. Dieses Beispiel suggeriert auch, dass die abhängige Variable ein Vektor sein kann, den man als den Ortsvektor eines Teilchens interpretieren kann. Auch die Funktion $f(t, x, \dots, x^{(n-1)})$ muss dann vektorwertig sein, und ebenso alle Argumente ausser dem ersten von f.

5.1.1 Reduktion der Ordnung

Eine Differentialgleichung n-ter Ordnung für eine skalare Funktion kann in eine Vektor-Differentialgleichung erster Ordnung für eine n-dimensionale vektorwertige Funktion umgewandelt werden. Ist $y(x)$ die gesuchte Funktion in der Differentialgleichung (5.1), dann kann man den Vektor

$$u(x) = \begin{pmatrix} y(x) \\ y'(x) \\ \vdots \\ y^{(n-1)}(x) \end{pmatrix} \in \mathbb{R}^n$$

bilden. Er erfüllt die Differentialgleichung

$$\frac{d}{dx}\begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix} = \begin{pmatrix} y' \\ y'' \\ \vdots \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} y' \\ y'' \\ \vdots \\ f(x, y, y', \dots, y^{(n-1)}) \end{pmatrix}. \tag{5.2}$$

Der Vektor auf der rechten Seite hängt nur von x, der Funktion y und ihren Ableitungen bis zur $n-1$-ten Ordnung ab, also von u, man kann (5.2) daher als

$$\frac{d}{dx}u = \tilde{f}(x, u) \tag{5.3}$$

schreiben. Im Folgenden werden wir fast ausschliesslich Differentialgleichungen erster Ordnung der Form $y' = f(x, y)$ betrachten, und dabei stillschweigend zulassen, dass y ein Vektor ist.

5.1.2 Anfangswertprobleme

Die Differentialgleichung $y' = f(x, y)$ alleine kann eine Lösungsfunktion $y(x)$ nicht festlegen, sie codiert nur, wie sich die Lösung verändern wird. Es ist also zusätzlich die Angabe eines Punktes der Lösungskurve notwendig. Man nennt das Problem, eine Funktion $y(x)$ zu finden, welche

$$y' = f(x, y) \qquad \text{und} \qquad y(0) = y_0$$

erfüllt, ein *Anfangswertproblem*. Ein Anfangswertproblem verlangt für die gewöhnliche Differentialgleichung n-ter Ordnung verlangt also die Angabe der Werte von $y(0), y'(0), \dots, y^{(n-1)}(0)$

Existenz und Eindeutigkeit von Lösungen

Die Existenz und Eindeutigkeit einer Lösung ist aus den Beispielen und graphischen Darstellungen intuitiv verständlich, für einen exakten Beweis sind jedoch zusätzliche Voraussetzungen nötig.

Definition 5.1. *Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ heisst global* Lipschitz-stetig*, wenn es eine Zahl L gibt*

$$|f(x_2) - f(x_1)| \le L\,|x_2 - x_1| \tag{5.4}$$

für alle Vektoren $x_1, x_2 \in \mathbb{R}^n$. Eine Funktion heisst lokal Lipschitz-stetig im Punkt x_0, wenn die Bedingung (5.4) *für x_i in einer Umgebung von x_0 erfüllt ist.*

Eine Funktion ist insbesondere dann lokal Lipschitz-stetig, wenn sie stetig differenzierbar ist. In diesem Fall ist die Ableitung $f'(x)$ in einer Umgebung von x_0 beschränkt, also $|f(x)| < M$, und der Mittelwertsatz der Differentialgleichung sagt, dass

$$|f(x_2) - f(x_1)| \le M|x_2 - x_2|$$

ist, f ist also lokal Lipschitz-stetig.

Satz 5.2 (Picard-Lindelöf)**.** *Ist die Funktion $f(x, y)$ lokal Lipschitz-stetig bezüglich der Variablen y für $x \in [x_0, b]$ und $|y - y_0| < R$. Dann hat das Anfangswertproblem*

$$y'(x) = f(x, y) \qquad \textit{und} \qquad y(x_0) = y_0$$

ein eindeutige Lösung, die in einem Intervall $[x_0, x_0 + \varepsilon)$ definiert ist.

In diesem Buch werden die Funktionen f der Differentialgleichungen meistens stetig differenzierbar sein, so dass der Satz 5.2 in unseren Anwendungen die lokale Existenz und Eindeutigkeit einer Lösung garantiert.

5.1.3 Randwertprobleme

Wenn man einen Ball wirft, wird seine Bewegung durch die Vektordifferentialgleichung zweiter Ordnung

$$\frac{d^2}{dt^2}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ -\dfrac{g}{m} \end{pmatrix}$$

beschrieben. Die Bahn ist ausserdem bestimmt durch die Anfangsbedingungen, d. h. den Anfangspunkt und die Anfangsgeschwindigkeit der Bahn. Praktischer Ballwurf verlangt aber, dass ein Ziel getroffen wird. Die Aufgabenstellung ist daher eine Bahnkurve $\gamma(t)$ zu finden, welche sowohl durch den Anfangspunkt als auch den Zielpunkt verläuft.

Die Lösung einer Differentialgleichungen erster Ordnung für eine unbekannte reellwertige Funktion $y(x)$ ist vollständig durch einen einzigen Anfangswert bestimmt. Eine Differentialgleichung zweiter Ordnung für eine unbekannte rellwertige Funktion $y(x)$ verlangt dagegen zwei Anfangswerte, nämlich für $y(0)$ und $y'(0)$. In Analogie zum Problem des Ballwurfs könnte die Lösungsfunktion auch festgelegt werden durch den Wert für $x = 0$ und $x = 1$. Gesucht ist also eine Funktion $y(x)$ auf dem Intervall $[0, 1]$, die

$$y'' = f(x, y, y') \qquad \text{mit} \qquad y(0) = y_0, \qquad y(1) = y_1 \tag{5.5}$$

erfüllt. Die Lösungsfunktion muss also bestimmte Werte am Rand des Definitionsbereichs annehmen, man spricht von einem *Randwertproblem*.

Beispiel. Wir lösen die Differentialgleichung $y'' = -y$ mit den Randwerten $y(0) = 1$ und $y(1) = 2$. Die homogene Differentialgleichung hat die Funktionen

$$y(x) = A\cos x + B\sin x$$

als allgemeine Lösung. Die Konstanten A und B müssen so gewählt werden, dass die Randwerte korrekt sind. Setzt man $x = 0$ und $x = 1$ ein, erhält man die linearen Gleichungen

$$\begin{aligned} a &= y(0) = A\cos 0 + B\sin 0 = A \\ b &= y(1) = A\cos 1 + B\sin 1 \qquad \Rightarrow \qquad \frac{b - a\cos 1}{\sin 1}. \end{aligned}$$

Die Lösung des Randwertproblems ist daher die Funktion

$$y(x) = a\cos x + \frac{b - a\cos 1}{\sin 1}\sin x,$$

wie man auch durch Einsetzen von $x = 0$ und $x = 1$ verifizieren kann. ○

5.1.4 Höhere Ableitungen

Die Differentialgleichung $y' = f(x, y)$ erlaubt nicht nur die erste Ableitung einer Funktion zu bestimmen. Durch Ableitung nach x können wir auch die höheren Ableitungen bestimmen, die eine Lösung der Differentialgleichung haben muss, die durch den Punkt (x, y).

$$\begin{aligned} &y'(x) = f(x, y), \\ &y''(x) = \frac{dy'(x)}{dx} = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}y'(x) = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}f(x,y), \end{aligned} \tag{5.6}$$

$$
\begin{aligned}
y'''(x) &= \frac{dy''(x)}{dx} \\
&= \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y\,\partial x} y'(x) + \Big(\frac{\partial^2 f(x,y)}{\partial x\,\partial y} + \frac{\partial^2 f(x,y)}{\partial y^2} y'(x)\Big) f(x,y) \\
&\quad + \frac{\partial f(x,y)}{\partial y}\Big(\frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y} f(x,y)\Big), \\
&= \frac{\partial^2 f(x,y)}{\partial x^2} + 2\frac{\partial^2 f(x,y)}{\partial x\,\partial y} f(x,y) + \frac{\partial^2 f(x,y)}{\partial y^2} f(x,y)^2 + \Big(\frac{\partial f(x,y)}{\partial y}\Big)^2 f(x,y) + \frac{\partial f(x,y)}{\partial y}\frac{\partial f(x,y)}{\partial x}.
\end{aligned}
\tag{5.7}
$$

Die Terme werden offensichtlich schnell kompliziert.

5.1.5 Ableitung nach der Anfangsbedingung

Verändert man die Anfangsbedingung, ändert sich auch die Lösung, die Komponente y_i ist also eine Funktion von x und von allen Anfangswerten y_{j0}:

$$y_i(x, y_{10}, \ldots, y_{n0}).$$

Wenn man untersuchen will, wie empfindlich die y_i auf Änderungen der Anfangswerte reagieren, dann sucht man die Ableitungen der y_i nach den Anfangswerten, also die sogenannte Jacobi-Matrix

$$J(x) = \begin{pmatrix} \frac{\partial y_1}{\partial y_{10}} & \cdots & \frac{\partial y_1}{\partial y_{n0}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial y_{10}} & \cdots & \frac{\partial y_n}{\partial y_{n0}} \end{pmatrix}$$

Wir schreiben abkürzend auch

$$J(x) = \frac{\partial y}{\partial y_0}.$$

Für $x = 0$ ist $y(x) = y(0) = y_0$, die Ableitung der y-Werte nach den Anfangswerten ist daher die Einheitsmatrix:

$$J(0) = E,$$

und es liegt nahe, dass auch $J(x)$ eine Differentialgleichung erfüllen muss.

Wie ändert sich $J(x)$ zwischen x und $x + \Delta x$? Die Werte von y ändern sich um

$$\frac{dy}{dx} = f(x,y), \tag{5.8}$$

aber y hängt von den Anfangsbedingungen ab. Leiten wir (5.8) nach y_0 ab, finden wir

$$\frac{\partial}{\partial y_0}\frac{dy}{dx} = \frac{\partial}{\partial y_0} f(x,y). \tag{5.9}$$

Die rechte Seite wir mit der Kettenregel zu

$$\frac{\partial}{\partial y_{j0}} f_i(x,y). = \sum_{k=1}^{n} \frac{\partial f_i(x,y)}{\partial y_k} \underbrace{\frac{\partial y_k}{\partial y_{j0}}}_{J_{ij}(x)}.$$

Schreiben wir die Ableitungen von f nach y in die Matrix

$$\frac{\partial f(x,y)}{\partial y} = \begin{pmatrix} \frac{\partial f_1(x,y)}{\partial y_1} & \cdots & \frac{\partial f_1(x,y)}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x,y)}{\partial y_1} & \cdots & \frac{\partial f_n(x,y)}{\partial y_n} \end{pmatrix} = F(x,y),$$

kann die Gleichung (5.9) mit dem Matrizenprodukt als

$$\frac{\partial}{\partial y_0}\frac{dy}{dx} = F(x,y)J(x)$$

sehr viel kompakter geschrieben werden. Die Ableitungen auf der linken Seite vertauschen, und wir erhalten

Satz 5.3. *Ist $y(x, y_0)$ die Lösung der Differentialgleichung*

$$y' = f(x,y) \qquad \textit{mit Anfangsbedingung} \qquad y(0) = y_0,$$

dann erfüllt die Jacobi-Matrix

$$\frac{\partial y}{\partial y_0} = J(x)$$

die Differentialgleichung

$$\frac{d}{dx}\frac{\partial y}{\partial x} = \frac{d}{dx}J(x) = F(x,y)J(x) \qquad \textit{mit Anfangsbedingung} \qquad J(x) = E, \tag{5.10}$$

wobei F die Ableitung von f nach y ist,

$$F(x,y) = \frac{\partial f(x,y)}{\partial y}.$$

Vor allem bei der numerischen Lösung mit dem Computer lässt sich die Jacobi-Matrix also gleich mit bestimmen, indem man die gefundene Lösung y als Input für die Gleichung (5.10) verwendet.

Beispiel. Wir betrachten wieder die Schwingungsdifferentialgleichung $y'' = -y$, oder vielmehr die Form erster Ordnung

$$\frac{d}{dx}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -y_1. \end{pmatrix}.$$

Die Ableitungsmatrix $F(x, y)$ ist

$$F(x,y) = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Die Jacobi-Matrix erfüllt also die Differentialgleichung

$$\frac{d}{dx}J(x) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} J(x).$$

Da wir die Differentialgleichung bereits vollständig gelöst und die Lösung

$$\begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix} = \begin{pmatrix} y_{10}\cos x + y_{20}\sin x \\ -y_{10}\sin x + y_{20}\cos x \end{pmatrix} = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} \begin{pmatrix} y_{10} \\ y_{20} \end{pmatrix}$$

gefunden haben, können wir die Jacobi-Matrix auch aus der Lösung berechnen, und verifizieren, ob sie die Differentialgleichung erfüllt. Die Jacobi-Matrix ist

$$J(x) = \frac{\partial y}{\partial y_0} = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}.$$

Die Ableitung davon ist

$$\frac{d}{dx} J(x) = \begin{pmatrix} -\sin x & \cos x \\ -\cos x & -\sin x \end{pmatrix}.$$

Setzen wir dies in die Differentialgleichung (5.10) ein finden wir

$$F(x,y)J(x) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} = \begin{pmatrix} -\sin x & \cos x - \cos x & -\sin x \end{pmatrix} = \frac{d}{dx} J(x),$$

die Differentialgleichung ist also tatsächlich erfüllt. ○

Die Möglichkeit, die Jacobi-Matrix zu berechnen, wird sich im Abschnitt 5.7.2 bei der numerischen Lösung von Randwertproblemen als besonders nützlich erweisen.

Für die numerische Berechnung von $y(x)$ und $J(x)$ müssen die beiden Differentialgleichungen in eine einzige zusammengefasst werden. Wir fassen dazu die Vektoren $y(x)$ und die Matrix $J(x)$ in einen einzigen Vektor zusammen. Wir können diese Zusammenfassung schreiben als

$$Y(x) = \begin{pmatrix} y(x) \\ J(x) \end{pmatrix},$$

wobei wir irgend eine Methode verwenden, eine Matrix als Vektor anzuordnen. Welche Methode dazu verwendet wird ist egal, solange es immer die Gleiche ist. Zum Beispiel können die Matrixelemente zeilenweise in den Spaltenvektor Y abgefüllt werden, oder spaltenweise. In dieser Notation wird die zusammengefasste Differentialgleichung

$$\frac{d}{dx} \begin{pmatrix} y \\ J \end{pmatrix} = \begin{pmatrix} f(x,y) \\ F(x,y)J \end{pmatrix} \qquad \text{mit der Anfangsbedingung} \qquad Y(0) = \begin{pmatrix} y_0 \\ E \end{pmatrix}, \tag{5.11}$$

wobei wir die unbekannten Funktionen rot hervorgehoben haben, um die Abhängigkeiten deutlicher zu machen.

Wir schreiben dies noch aus für den Fall $n = 2$. In diesem Fall sind insgesamt sechs Funktionen zu bestimmen, die zwei Komponenten von y, $y_1(x)$ und $y_2(x)$, und die vier Komponenten von $J(x)$. Um daraus eine einzige Differentialgleichung zu machen, packen wir die sechs Funktionen in einen Vektor $Y(x)$

$$Y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \\ J_{11}(x) \\ J_{12}(x) \\ J_{21}(x) \\ J_{22}(x) \end{pmatrix}.$$

Für die Ableitung der ersten zwei Komponenten verwenden wir die Differentialgleichung (5.8), für die J-Komponenten aber die Differentialgleichung (5.10). So erhalten wir die kombinierte Differentialgleichung

$$\frac{d}{dx}Y(x) = \frac{d}{dx}\begin{pmatrix} y_1(x) \\ y_2(x) \\ J_{11}(x) \\ J_{12}(x) \\ J_{21}(x) \\ J_{22}(x) \end{pmatrix} = \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \\ \dfrac{\partial f_1(x,y)}{\partial x_1} J_{11}(x) + \dfrac{\partial f_1(x,y)}{\partial x_2} J_{21}(x) \\ \dfrac{\partial f_1(x,y)}{\partial x_1} J_{12}(x) + \dfrac{\partial f_1(x,y)}{\partial x_2} J_{22}(x) \\ \dfrac{\partial f_2(x,y)}{\partial x_1} J_{11}(x) + \dfrac{\partial f_2(x,y)}{\partial x_2} J_{21}(x) \\ \dfrac{\partial f_2(x,y)}{\partial x_1} J_{12}(x) + \dfrac{\partial f_2(x,y)}{\partial x_2} J_{22}(x) \end{pmatrix}. \tag{5.12}$$

5.1.6 Abhängigkeit von Parametern

Im Allgemeinen wird eine Differentialgleichung von Parametern abhängen, so dass die Lösung nicht nur von der Anfangsbedingung, sondern auch von den Werten dieser Parameter abhängen wird. Um dies explizit zu machen, fassen wir die Parameter in einen Vektor c zusammen, und machen die Abhängigkeit der Differentialgleichung von c durch ein zusätzliches Argument der Funktion f explizit:

$$\frac{d}{dx}y = f(x, y, c).$$

Die Lösung $y(x)$ hängt dann zusätzlich von der Wahl der Parameter ab, wir können das durch die Schreibweise $y(x, c)$ sichtbar machen.

Oft sucht man dann geeignete Parameter, für die die Lösung der Differentialgleichung bestimmte Eigenschaften hat, zum Beispiel durch einen bestimmten Punkt verläuft, d. h. wir suchen Werte c derart, dass $y(x_1, c) = y_1$. Ein besonders effizientes Verfahren zur Bestimmung solcher Parameterwerte ist das Newton-Verfahren, welches aber die Ableitungen

$$\frac{\partial y(x,c)}{\partial c} = \begin{pmatrix} \dfrac{\partial y_1(x,c)}{\partial c_1} & \cdots & \dfrac{\partial y_1(x,c)}{\partial c_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_n(x,c)}{\partial c_1} & \cdots & \dfrac{\partial y_n(x,c)}{\partial c_n} \end{pmatrix} = C(x)$$

von $y(x, c)$ nach c benötigt.

Um die Abhängigkeit von c zu berechnen, leiten wir die Differentialgleichung nach c ab, wir erhalten

$$\frac{\partial}{\partial c}\frac{d}{dx}y(x,c) = \frac{\partial f(x,y,c)}{\partial y}\frac{\partial y}{\partial c} + \frac{\partial f(x,y,c)}{\partial c}.$$

Wieder können auf der linken Seite die beiden Ableitungen vertauscht werden, so dass eine Differentialgleichung für $C(x)$ entsteht:

$$\begin{aligned} \frac{d}{dx}\frac{\partial y(x,c)}{\partial c} &= \frac{\partial f(x,y,c)}{\partial y}\frac{\partial y(x,c)}{\partial c} + \frac{\partial f(x,y,c)}{\partial c} \\ \frac{d}{dx}C(x) &= \frac{\partial f(x,y,c)}{\partial y}C(x) + \frac{\partial f(x,y,c)}{\partial c}. \end{aligned}$$

Dies ist eine lineare inhomogene Differentialgleichung erster Ordnung für die Matrix $C(x)$.

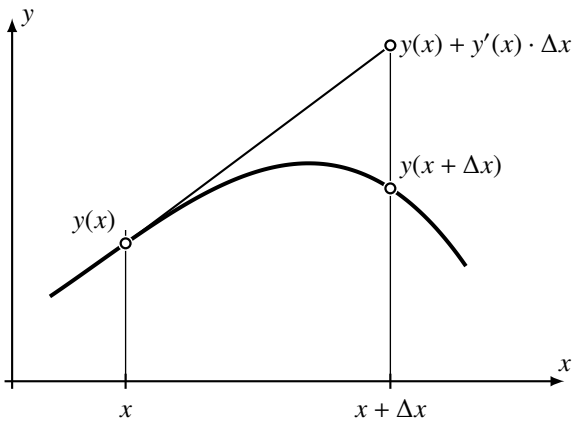


Abbildung 5.2: Lineare Approximation von $y(x + \Delta x)$ durch Information, die am Punkt x verfügbar ist.

5.2 Grundprinzip numerischer Lösungsverfahren

Wir versuchen die Differentialgleichung

$$y' = -\alpha y, \qquad y(0) = y_0 \tag{5.13}$$

numerisch zu lösen. Die Lösung ist natürlich bekannt, es ist

$$y(x) = y_0 e^{-\alpha x}. \tag{5.14}$$

Dazu unterteilen wir die x-Achse in diskrete Abschnitte der Länge h, genannt die *Schrittweite*, und bezeichnen die Teilpunkte mit $x_k = kh$. Das Ziel ist jetzt, $y(x_k)$ näherungsweise zu berechnen. Wir schreiben y_k für die Näherungswerte von $y(x_k)$. Die Ableitung liefert eine lineare Approximation für $y(x)$, nämlich

$$y(x + \Delta x) \simeq y(x) + y'(x) \cdot \Delta x$$

(Abbildung 5.2). Für die Punkte x_k bedeutet das

$$y(x_{k+1}) \simeq y(x_k) + y'(x_k) \cdot h.$$

Die Differentialgleichung liefert Werte für $y'(x_k)$ aus x_k und $y(x_k)$, damit können wir aus dieser Approximation ein allgemeines Näherungsverfahren für die Lösung einer Differentialgleichung konstruieren.

Satz 5.4 (Euler-Verfahren)**.** *Die Differentialgleichung*

$$y' = f(x, y), \qquad y(0) = y_0 \tag{5.15}$$

und die Schrittweite h definieren eine Folge

$$y_k = y_{k-1} + h \cdot f(x_{k-1}, y_{k-1}), \quad k > 0,$$

mit $x_k = kh$, die eine Näherung für die Funktionswerte $y(x_k)$ der Lösung $y(x)$ der Differentialgleichung (5.15) *ist.*

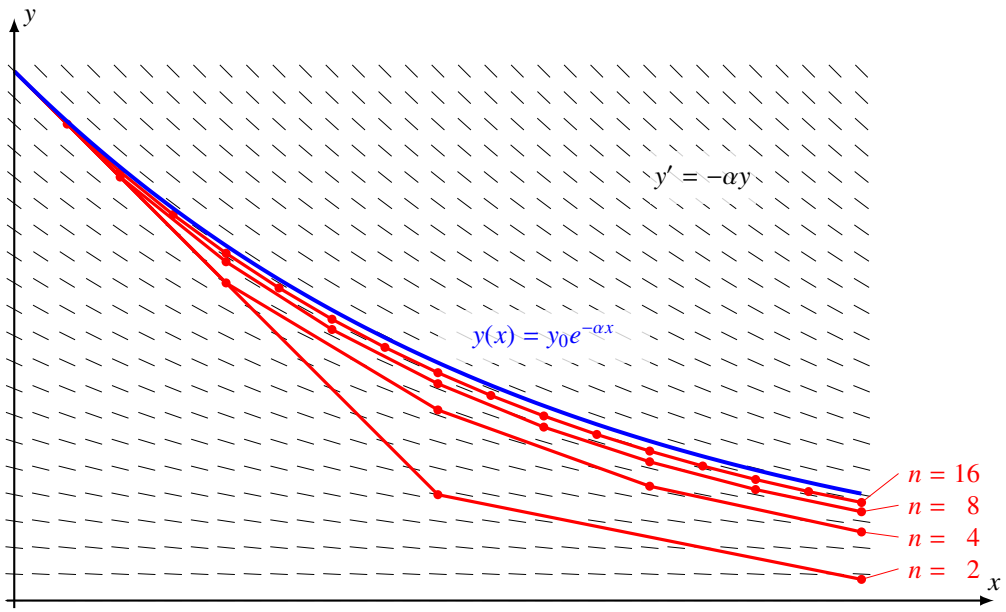


Abbildung 5.3: Approximationen der Lösung der Differentialgleichung $y' = -\alpha y$ mit verschiedener Anzahl Schritte (rot) nähern sich für wachsendes n der exakten Lösung (blau).

Dieses Verfahren ist nicht besonders gut, wie wir im Folgenden zeigen wollen. Die Diskussion soll vor allem zeigen, worauf bei der Weiterentwicklung des Verfahrens geachtet werden muss.

Im vorliegenden Beispiel liefert die Differentialgleichung (5.13) den Wert $y'(x_k) = -\alpha y(x_k)$ für die Ableitung, woraus wir die Rekursionsformel

$$y_{k+1} = y_k - \alpha y_k \dot{h}.$$

gewinnen. Die Rekursionsgleichung kann in diesem Fall exakt gelöst werden, und wir finden

$$y(x_{k+1}) = y(x_k) - \alpha y(x_k)h = (1 - \alpha h)y(x_k) = \cdots = (1 - \alpha h)^{k+1} y_0 \tag{5.16}$$

für die Näherung y_k der Funktionswerte $y(x_k)$.

Wir möchten $y(x)$ für einen ganz bestimmten x-Wert berechnen. Dazu unterteilen wir das Intervall $[0, x]$ in n Teilschritte der Breite x/n, und wenden die Formel (5.16) an:

$$y(x) = y(x_n) = (1 - \alpha h)^n y_0 = \left(1 + \frac{-\alpha x}{n}\right)^n y_0.$$

Für eine grosse Zahl von Teilschritten erhalten wir so tatsächlich die korrekte Lösung:

$$\lim_{n\to\infty} y_0 \left(1 + \frac{-\alpha x}{n}\right)^n = y_0 e^{-\alpha x}.$$

Abbildung 5.3 zeigt, wie die durch (5.16) gegebenen Approximationen mit zunehmendem n der exakten Lösung $y(x) = e^{-\alpha x}$ näher kommen.

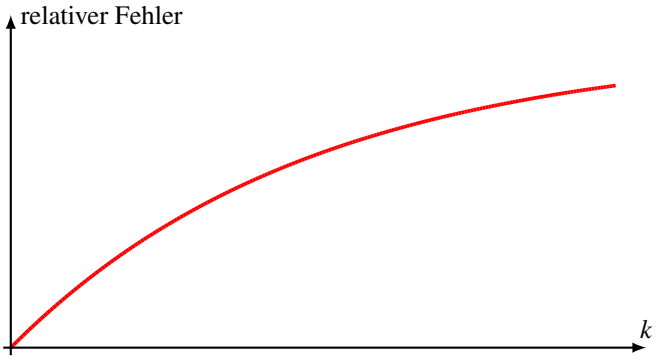


Abbildung 5.4: Relativer Fehler des Euler-Verfahrens für die Differentialgleichung (5.13) in Abhängigkeit von der Anzahl k der Schritte.

Wir können auch den Fehler des numerischen Verfahrens berechnen. Bei der Schrittweite h ist der Fehler von y_k die Differenz

$$y(x_k) - y_k = y_0 e^{-\alpha k h} - y_0(1-\alpha h)^k = y_0((e^{-\alpha h})^k - (1-\alpha h)^k) = y_0 e^{-\alpha h k}\left(1 - \left(\frac{1-\alpha h}{e^{-\alpha h}}\right)^k\right).$$

Man beachte, dass der Zähler $1 - \alpha h$ die Approximation y_1 ist, als eine Approximation von $e^{-\alpha h}$, dem Nenner. Schreiben wir

$$q = \frac{1-\alpha h}{e^{-\alpha h}},$$

für den Quotienten zwischen der Approximation und dem korrekten Wert, dann ist sicher immer $q < 1$. Den Fehler können wir jetzt schreiben

$$y(x_k) - y_k = y_0 e^{-\alpha h k}(1 - q^k) = y(x_k)(1 - q^k).$$

Der relative Fehler des Verfahrens ist also

$$\frac{y(x_k) - y_k}{y(x_k)} = (1 - q^k).$$

Ganz unabhängig von der Schrittweite h wird der relative Fehler des Verfahrens immer gegen 1 streben, der Fehler wird also von der gleichen Grössenordnung wie die berechneten Resultate.

Die Abbildung 5.4 zeigt, dass zu Beginn des Verfahrens der relative Fehler ungefähr linear mit der Anzahl der Schritt zunimmt. Um eine angemessene Genauigkeit über einen grösseren Bereich zu erreichen, muss das Euler-Verfahren also sehr viel kleinere Schritte und eine entsprechend grössere Anzahl von Schritten ausführen, die entsprechend viel Rechenzeit benötigen.

Ein praktisch nützliches Verfahren muss also anstreben, mit einer sehr viel kleineren Anzahl von Schritten eine viel grössere Genaugikeit der Approximation zu erreichen.

5.3 Fehler-Entwicklung numerischer Lösungen

Wir betrachten wieder die Differentialgleichung (5.15) und versuchen, den Fehler eines Näherungsverfahrens zu bestimmen, welches Schritte der Grösse h durchführt, um den Wert $y(x)$ zu approximieren.

Das Euler-Verfahren verwendet Schritte der Form

$$y_{k+1} = y_k + hf(x_k, y_k).$$

In jedem einzelnen Schritt entsteht ein Fehler, dessen Grösse wir aus der Taylor-Entwicklung

$$y(x + \Delta x) = y(x) + y'(x) \cdot \Delta x + R(x)\Delta x^2$$

abschätzen können. Die Funktion $R(x)$ ist beschränkt und beschreibt den verbleibenden Fehler. Um $y(x)$ zu approximieren, müssen $n = x/h$ Schritte der Schrittweite h durchgeführt werden, von denen jeder einen Fehler von der Grössenordnung $R(x)h^2$ hat. Der Gesamtfehler ist daher von der Grössenordnung

$$y(x) - y_n = O\left(R(x)h^2 \frac{x}{h}\right) = O(h),$$

er ist also von erster Ordnung in h. Um eine zusätzliche Stelle Genauigkeit zu erhalten, muss man also zehnmal so viele Schritte von zehnmal kleinerer Grösse durchführen, wodurch auch wieder Rundungsfehler eingeführt werden.

Könnte man den Fehler des Einzelschrittes wesentlich verkleinern, würde auch die Abhängigkeit des Fehlers des Verfahrens vorteilhafter. Wäre der Fehler des Einzelschrittes $O(h^k)$ statt $O(h^2)$, dann wäre der Gesamtfehler des Verfahrens nur noch $O(h^{k-1})$. Für $k = 3$ bedeutet dies, dass eine Halbierung der Schrittweite zwar doppelt so viele Schritte braucht, aber auch, dass in jedem Schritt nur ein Achtel des Fehlers auftritt. Der Gesamtfehler ist also nur ein Viertel. Mit zehnmal mehr Arbeit kann man also nicht nur eine Stelle an Genauigkeit gewinnen, sondern gleich deren zwei.

Man nennt ein Verfahren, bei dem der Gesamt-Fehler von der Grössenordnung $O(h^k)$ ist, von einem Verfahren k-ter Ordnung. Das Euler-Verfahren ist also ein Verfahren erster Ordnung oder ein lineares Verfahren. In der Praxis werden Verfahren bis zu vierter und fünfter Ordnung verwendet, so dass eine zehnmal kleinere Schrittweite zu gleich vier Stellen Genauigkeitsgewinn führen. Das Ziel der kommenden Abschnitte muss daher sein, einfach berechnebare Approximationen der Funktion $y(x)$ mit möglichst geringen Einzelschrittfehlern zu finden.

5.4 Einschritt-Verfahren

Die relativ geringe Genauigkeit des Eulerschrittes beruht darauf, dass die zu Beginn des Schrittes berechnete Ableitung $f(x_k, y_k)$ nur für das linke Ende des Intervalls $[x_k, x_k + h]$ zutrifft, weiter rechts im Intervall wird die Abweichung immer grösser. Eine mögliche Lösung des Problems könnte darin bestehen, statt nur einer linearen Näherung zusätzliche Glieder der Taylorreihe

$$y(x + \Delta x) = y(x) + y'(x) \cdot \Delta x + \frac{1}{2}y''(x) \cdot \Delta x^2 + \frac{1}{6}y'''(x) \cdot \Delta x^3 + o(\Delta x^3) \tag{5.17}$$

zu verwenden. In (5.17) werden höhere Ableitungen von $y(x)$ benötigt, während die Differentialgleichung nur die erste Ableitung liefert. Die höheren Ableitungen wurden aber bereits im Abschnitt 5.1.4 berechnet.

Wir untersuchen, wie sich das Verfahren für die Beispiel-Gleichung (5.13) anwenden lässt. Dort gilt

$$\begin{aligned} & y'(x) = f(x, y) = -\alpha y \\ \Rightarrow \qquad & \frac{\partial f}{\partial x} = 0 \qquad\qquad \frac{\partial f}{\partial y} = -\alpha. \end{aligned}$$

i	x	$e^{-\alpha x}$	Euler	kubisch
1	0.1	0.95122942	0.95000000	0.95122917
2	0.2	0.90483742	0.90250000	0.90483693
3	0.3	0.86070798	0.85737500	0.86070728
4	0.4	0.81873075	0.81450625	0.81872987
5	0.5	0.77880078	0.77378094	0.77879973
6	0.6	0.74081822	0.73509189	0.74081702
7	0.7	0.70468809	0.69833730	0.70468675
8	0.8	0.67032005	0.66342043	0.67031859
9	0.9	0.63762815	0.63024941	0.63762660
10	1.0	0.60653066	0.59873694	0.60652902

Tabelle 5.1: Näherungswerte für die Lösung $e^{-\alpha x}$ der Beispieldifferentialgleichung (5.13) nach dem Euler-Verfahren und nach dem kubischen Verfahren (5.18) mit einer Schrittweite von 0.1. Unterstrichen ist jeweils die Stellen, die nach Rundung auf die angegebene Anzahl stellen mit dem exakten Wert übereinstimmt.

Alle zweiten Ableitungen verschwinden. Die Gleichungen werden damit einfach:

$$y''(x) = -\alpha f(x,y) = \alpha^2 y$$
$$y'''(x) = \alpha^2 f(x,y) = -\alpha^3 y.$$

Statt der linearen Approximation sollte daher die kubische Approximation

$$y_{k+1} = y_k - \alpha h y_k + \frac{1}{2}\alpha^2 h^2 y_k - \frac{1}{6}\alpha^3 h^3 y_k = y_k \underbrace{\left(1 - \alpha h + \frac{1}{2}\alpha^2 h^2 - \frac{1}{6}\alpha^3 h^3\right)}_{\simeq e^{-\alpha h}} \tag{5.18}$$

verwendet werden. Dass man hier mit einer grösseren Genauigkeit rechnen darf ist schon daran erkennbar, dass der Klammerausdruck auf der rechten Seite eine viel bessere Approximation von $e^{-\alpha x}$ ist also der Faktor $(1-\alpha h)$ im Euler-Verfahren. Genauer erwarten wir, dass wir hier ein kubisches Verfahren konstruiert haben.

In Tabelle 5.1 werden die Resultate des kubischen Verfahrens denen des Euler-Verfahrens gegenübergestellt. Im ersten Schritt ist der Fehler des Euler-Verfahrens kleiner als 10^{-2}, was einer Einheit in der zweiten Nachkommastelle entspricht. Der Fehler des kubischen Verfahrens ist kleiner als 10^{-6}, eine Einheit in der sechsten Nachkommastelle, ungefähr die von einem kubischen Verfahren zu erwartende Verbesserung. Nach zehn Rechenschritten liefert das Euler-Verfahren dank Rundung gerade noch eine korrekte Stelle, während das kubische Verfahren immer noch gerundet fünf korrekte Stellen gibt.

Es wurde bereits darauf hingewiesen, dass die Terme für die Ableitungen sehr kompliziert werden. noch viel gravierender ist allerdings, dass auch die partiellen Ableitungen von f nach x und y bekannt sein müssen. Es ist zwar im Prinzip möglich, diese zu berechnen, der Rechenaufwand dafür kann aber so erheblich sein, dass er den Genauigkeitsgewinn leicht wieder zunichte machen kann. Praktisch nützliche Verfahren müssen daher danach streben, die höheren Ableitungen von $y(x)$ ausschliesslich aus Funktionswerten von $f(x,y)$ zu berechnen.

Wir möchten aber weiterhin nur y_{k+1} ausschliesslich aus x_k und y_k berechnen, also in einem einzelnen Schritt der Form

$$y_{k+1} = y_k + h\,F(x_k, y_k, h).$$

Die Funktion $F(x, y, h)$ heisst die *Inkrement-Funktion* des Verfahrens. Für das Euler-Verfahren ist $F(x, y, h) = f(x, y)$. Es soll also eine Inkrement-Funktion gefunden werden, bei der $y(x + \Delta x)$ durch $y(x) + \Delta x \cdot F(x, y, \Delta x)$ bis auf Terme höherer Ordnung approximiert werden kann.

5.4.1 Quadratische Verfahren

Ein quadratisches Verfahren verwendet eine Inkrement-Funktion $F(x, y, h)$, welche

$$y(x + h) = y(x) + hF(x, y, h) + O(h^3)$$

erfüllt. Aus den einleitenden Bemerkungen von **??** folgt, dass dieses Ziel möglicherweise dadurch erreicht werden kann, dass man Werte von f für verschiedene x geeignet miteinander kombiniert. Ein denkbarer Ansatz dafür ist

$$F(x, y, h) = af(x, y) + bf(x + \alpha h, y + \beta hf(x, y)),$$

oder anders ausgedrückt: Man führt zuerst etwas ähnliches wie einen Eulerschritt durch, um zum Punkt $(x+\alpha h, y+\beta hf(x, y))$ zu gelangen. Dort berechnet man den Wert von f, und bildet dann einen geeigneten Mittelwert davon mit $f(x, y)$. Durch geeignete Wahl von a, b, α und β sollte es möglich sein, dass die Inkrement-Funktion einen Fehler höchstens dritter Ordnung hat, womit wir dann ein Integrationsverfahren zweiter Ordnung gewonnen hätten.

Wir müssen jetzt die Parameter a, b, α und β bestimmen. Da wir mit dem übereinstimmen der ersten zwei Ableitungen nur zwei Bedingungen haben, können wir nicht erwarten, dass wir eine eindeutige Lösung finden werden. Vielmehr werden einzelne Parameter frei wählbar sein, es wird eine ganze Familie von quadratischen Lösungsverfahren entstehen, parametrisiert durch eine der Variablen a, b, α und β.

Wir berechnen nun $F(x, y, h)$ bis zur zweiten Ordnung, damit wird $y(x+h)$ bis zur dritten Ordnung ausdrücken können.

$$\begin{aligned} f(x + \alpha h, y + \beta hf(x, y)) &= f(x, y) + \alpha h\frac{\partial f(x, y)}{\partial x} + \beta h\frac{\partial f(x, y)}{\partial y} + O(h^2) \\ F(x, y, h) &= af(x, y) + bf(x + \alpha h, y + \beta hf(x, y)) \\ &= (a + b)f(x, y) + \left(\alpha b\frac{\partial f(x, y)}{\partial x} + \beta b\frac{\partial f(x, y)}{\partial y}f(x, y))\right)h + O(h^2). \end{aligned} \quad (5.19)$$

Damit dies bis zur zweiten Ordnung mit dem Inkrement zwischen x und $x + h$ übereinstimmt, muss (5.19) mit der Taylorreihe von $y(x)$ übereinstimmen, also mit

$$\frac{y(x + h) - y(x)}{h} = y'(x) + \frac{1}{2}y''(x)h + O(h^2) = f(x, y) + \frac{1}{2}\frac{\partial f(x, y)}{\partial x} + \frac{1}{2}\frac{\partial f(x, y)}{\partial y}f(x, y) + O(h^2), \quad (5.20)$$

wobei wir für $y''(x)$ die Gleichung (5.6) verwendet haben. Durch Koeffizientenvergleich finden wir die Bedingungen

$$a + b = 1, \quad \alpha b = \frac{1}{2}, \quad \beta b = \frac{1}{2}.$$

Einzig b kommt in allen drei Gleichungen vor, und bestimmt den Wert der jeweiligen anderen Variablen:

$$a = 1 - b, \quad \alpha = \beta = \frac{1}{2b}.$$

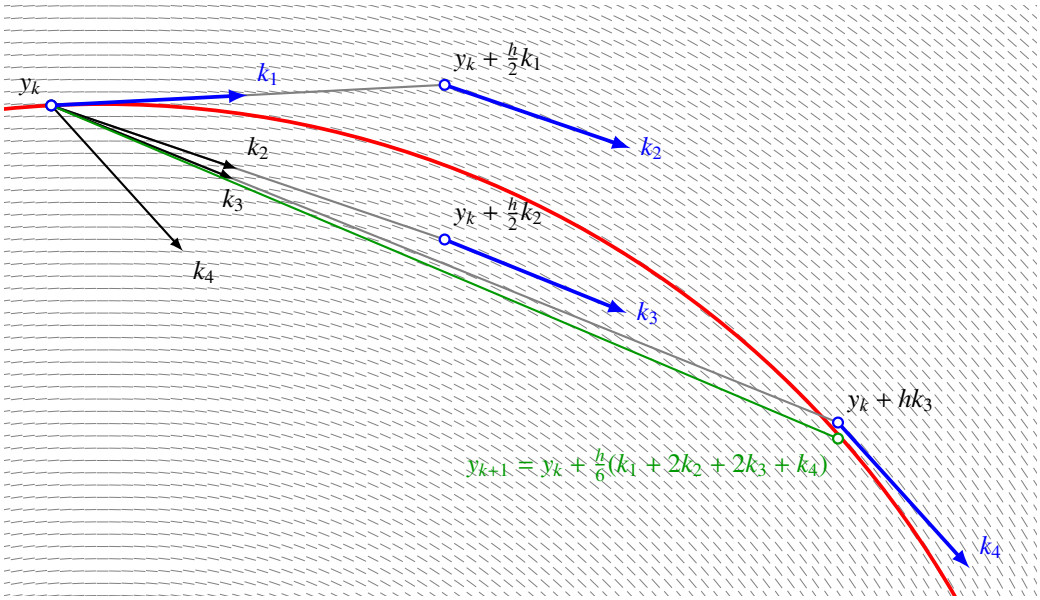


Abbildung 5.5: Zusammenspiel der Richtungen k_1 bis k_4 bei einem Einzelschritt des Runge-Kutta-Verfahrens vierter Ordnung. Der nächste Punkt y_{k+1} wird gemäss Formel (5.23) berechnet.

Jeder Wert von b zwischen 0 und 1 liefert ein Verfahren mit quadratischer Genauigkeit.

Der Parameterwert $b = 1$ führt auf $\alpha = \beta = 1$ und $a = 0$, die Rekursionsformel ist in diesem Falle

$$y_{k+1} = y_k + hf\Big(x_k + \frac{h}{2}, y_k + \frac{h}{2}f(x_k, y_k)\Big). \tag{5.21}$$

Das Verfahren führt also erst einen halben Eulerschritt zum Punkt $(x_k + \frac{1}{2}h, y_k + \frac{h}{2}f(x_k, y_k))$ durch, berechnet dort mit Hilfe von f die Steigung, die dann für einen Eulerschritt der Länge h verwendet wird. Daher heisst dieses Verfahren auch das *verbesserte Euler-Verfahren*.

Verwendet man $b = \frac{1}{2}$, folgt zunächst $a = \frac{1}{2}$ und $\alpha = \beta = 1$. Daraus erhält man die Rekursionsformel

$$y_{k+1} = y_k + \frac{h}{2}\Big(f(x_k, y_k) + f(x_k + h, y_k + hf(x_k, y_k))\Big). \tag{5.22}$$

In diesem Verfahren führt man also zuerst einen Eulerschritt der Länge h durch, mit dem man zum Punkt $(x_k + h, y_k + hf(x_k, y_k))$ gelangt. Dort berechnet mit mit Hilfe von f die Steigung. Das arithmetische Mittel dieser Steigung mit der im Euler-Verfahren verwendeten Steigung $f(x_k, y_k)$ im Punkt x_k wird dann als Steigung für einen Eulerschritt verwendet. Statt eines einzigen Steigungswertes werden hier also zwei Steigungswerte von den Enden des Intervalls $[x_k, x_k + 1]$ gemittelt. Wegen der Ähnlichkeit dieses Vorgehens mit dem später zu besprechenden Runge-Kutte-Verfahren heisst diese Verfahren auch das *vereinfachte Runge-Kutta-Verfahren*.

5.4.2 Runge-Kutta-Verfahren

Das *Runge-Kutta-Verfahren* erweitert die Inkrement-Funktion derart, dass der Einzelschritt bis zur fünften Ordnung mit der Taylorreihe von $y(x)$ übereinstimmt. So entsteht ein Verfahren vierter Ordnung, es stellt einen guten Kompromiss zwischen Genauigkeit und Rechenaufwand dar.

i	x	$y(x) = e^{-\alpha x}$	Euler	verbessert	vereinfacht	Runge-Kutta
0	0.0	1.00000000	1.000	1.00000000	1.00000000	1.0000000000
1	0.1	0.95122942	0.950	0.95125000	0.95125000	0.9512294271
2	0.2	0.90483742	0.902	0.90487656	0.90487656	0.9048374229
3	0.3	0.86070798	0.857	0.86076383	0.86076383	0.8607079834
4	0.4	0.81873075	0.814	0.81880159	0.81880159	0.8187307620
5	0.5	0.77880078	0.773	0.77888502	0.77888502	0.7788007936
6	0.6	0.74081822	0.735	0.74091437	0.74091437	0.7408182327
7	0.7	0.70468809	0.698	0.70479480	0.70479480	0.7046881031
8	0.8	0.67032005	0.663	0.67043605	0.67043605	0.6703200606
9	0.9	0.63762815	0.630	0.63775229	0.63775229	0.6376281672
10	1.0	0.60653066	0.598	0.60666187	0.60666187	0.6065306762

Tabelle 5.2: Vergleich der Genauigkeit der verbesserten numerischen Verfahren. Unterstrichen jeweils die nach Rundung korrekten Stellen der Lösung.

Da vier Ableitungen korrekt dargestellt werden müssen, ist zu erwarten, dass vier verschiedene Werte von f an verschiedenen Punkten (x, y) ausgewertet und geeignet miteinander kombiniert werden müssen. Genauer: Man bestimmt zuerst die Werte

$$\begin{aligned} k_1 &= f(x_k, y_k) \\ k_2 &= f\Big(x_k + \frac{h}{2}, y_k + \frac{h}{2}k_1\Big) \\ k_3 &= f\Big(x_k + \frac{h}{2}, y_k + \frac{h}{2}k_2\Big) \\ k_4 &= f(x_k + h, y_k + hk_3) \end{aligned}$$

und setzt diese dann zusammen, um den nächsten Wert y_{k+1} zu berechnen:

$$y_{k+1} = y_k + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{5.23}$$

(Abbildung 5.5). Man kann die Formeln wie folgt interpretieren. Zuerst wird ein halber Eulerschritt mit der Steigung $k_1 = f(x_k, y_k)$, durchgeführt, und am Zielpunkt die Steigung k_2 ermittelt. Mit dieser Steigung wird dann erneut ein halber Schritt von (x_k, y_k) aus durchgeführt, und am Zielpunkt erneut die Steigung k_3 ermittelt. Damit führt man einen ganzen Schritt aus, an dessen Zielpunkt man die Steigung k_4 findet. Diese vier Steigungen werden jetzt gewichtet gemittelt, wobei k_2 und k_3 doppeltes Gewicht erhalten, und mit dieser Steigung wird ein ganzer Schritt vorgenommen.

Die Formeln für die k_i sowie (5.23) können ganz ähnlich wie das verbesserte Euler-Verfahren bzw. das vereinfachte Runge-Kutta-Verfahren begründet werden. Der Aufwand dafür ist aber beträchtlich, so dass wir auf die detaillierte Darstellung dieser Herleitung verzichten wollen.

Die Tabelle 5.2 demonstriert die überragende Genauigkeit des Runge-Kutta-Verfahrens. Trotz der relativ grossen Schrittweite von $h = 0.1$ erreicht das Verfahren nach zehn Schritten eine Genauigkeit von sieben signifikanten Stellen. Da in jedem Schritt die Funktion f viermal ausgewertet werden muss, ist der Rechenaufwand mit dem Runge-Kutta-Verfahren viermal grösser als im Euler-Verfahren, letzteres kann aber mit nur einer signifikanten Stelle kaum als brauchbar bezeichnet werden. Passt man in jedem Verfahren die Schrittweite so an, dass für die Berechnung der Näherung für $y(1)$ immer gleich viele Auswertungen der Funktion $f(x, y)$ nötig sind, ergeben sich die Resultate in

Verfahren	h	Schritte	y_n	Fehler
Euler-Verfahren	0.025	40	0.60462232	0.00190834
verbessertes Euler-Verfahren	0.05	20	0.60656285	−0.00003219
vereinfachtes Runge-Kutta-Verfahren	0.05	20	0.60656285	−0.00003219
Runge-Kutta-Verfahren	0.1	10	0.60653067	−0.00000001

Tabelle 5.3: Vergleich der verschiedenen Verfahren bei gleichbleibendem Rechenaufwand. Die Schrittweite wurde jeweils so angepasst, dass in allen Verfahren bis zum Wert $x = 1$ die gleiche Anzahl von Auswertungen der Funktion f notwendig wurde.

Tabelle 5.3. Bei gleichem Rechenaufwand ist das Runge-Kutta-Verfahren um viele Grössenordungen präziser. Es gibt daher eigentlich keinen praktischen Grund, überhaupt je etwas anderes als das Runge-Kutta-Verfahren zu verwenden.

5.5 Mehrschritt-Verfahren

In den Einschritt-Verfahren wurde wiederholt die Funktion f ausgewertet, um die Inkrement-Funktion für einen einzigen Schritt zu bestimmen. Das Ziel dabei war, $y(x + h)$ in Übereinstimmung mit der Taylorreihe bis zu möglichst hoher Ordnung zu bestimmen. Im Runge-Kutta-Verfahren wurden dabei halbe Eulerschritte durchgeführt, man hat also eigentlich die Auflösung nochmals halbiert, um die Inkrement-Funktion zu ermitteln. Diese Zwischenwerte geben dem Verfahren die Information über die höheren Ableitungen der Funktionen.

Sobald einige Werte der Lösung berechnet sind, lässt sich die Krümmung der Lösungskurve auch aus diesen Werten ablesen. Es sollte daher auch möglich sein, aus mehreren bereits ermittelten Werten $y_n, y_{n+1}, \dots, y_{n+s-1}$ den nächsten Wert y_{n+s} mit der verlangten Genauigkeit zu berechnen. Der Vorteil eines solchen Vorgehens ist, dass für jeden Schritt nur eine einzige Auswertung der Funktion f nötig ist, nicht mehrere wie bei den besprochenen Einschritt-Verfahren.

Als Beispiel versuchen wir daher ein Verfahren aufzubauen, welches y_{n+2} aus den bereits berechneten Werten y_n und y_{n+1} berechnet. Wir nehmen dabei an, dass y_n und y_{n+1} exakt sind. Der neue Datenpunkt soll mit Hilfe eines Ausdrucks der Form

$$y_{n+2} = y_{n+1} + h(af(x_{n+1}, y_{n+1}) + bf(x_n, y_n)) \tag{5.24}$$

gefunden werden. Die Näherung kann wieder mit Hilfe der Ableitungen alleine durch Werte bei x_{n+1} ausgedrückt werden:

$$\begin{aligned}
y_{n+2} &= y_{n+1} + h(af(x_{n+1}, y_{n+1}) + bf(x_{n+1} - h, y_n)) \\
&= y_{n+1} + haf(x_{n+1}, y_{n+1}) + hbf(x_{n+1} - h, y_{n+1} - hf(x_{n+1}, y_{n+1}) + O(h^2)) \\
&= y_{n+1} + haf(x_{n+1}, y_{n+1}) + hb\Big(f(x_{n+1}, y_{n+1}) - h\frac{\partial f(x_{n+1}, y_{n+1})}{\partial x} \\
&\qquad - h\frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} f(x_{n+1}, y_{n+1}) + \frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} O(h^2)\Big) \\
&= y_{n+1} + (a + b)hf(x_{n+1}, y_{n+1}) - bh^2\Big(\frac{\partial f(x_{n+1}, y_{n+1})}{\partial x} + \frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} f(x_{n+1}, y_{n+1}) + O(h^3)\Big).
\end{aligned}$$

i	x	$y(x) = e^{-\alpha x}$	Euler	Adams-Bashforth	Runge-Kutta
0	0.0	1.00000000	1.00000000	1.00000000	1.0000000000
1	0.1	0.95122942	0.95000000	0.95128178	0.9512294271
2	0.2	0.90483742	0.90250000	0.90493564	0.9048374229
3	0.3	0.86070798	0.85737500	0.86084752	0.8607079834
4	0.4	0.81873075	0.81450625	0.81890734	0.8187307620
5	0.5	0.77880078	0.77378094	0.77901048	0.7788007936
6	0.6	0.74081822	0.73509189	0.74105738	0.7408182327
7	0.7	0.70468809	0.69833730	0.70495334	0.7046881031
8	0.8	0.67032005	0.66342043	0.67060827	0.6703200606
9	0.9	0.63762815	0.63024941	0.63793648	0.6376281672
10	1.0	0.60653066	0.59873694	0.60685645	0.6065306762

Tabelle 5.4: Vergleich der Genauigkeit der Verfahren von Euler, Adams-Bashforth und Runge-Kutta. Als Startwerte für das Adams-Bashforth-Verfahren wurden die Werte $y(-h) = e^{-\alpha h}$ und $y(0) = 1$ verwendet, um keine zusätzlichen Fehler aus der Durchführung des ersten Schrittes hinzuzufügen.

Sie muss bis zur zweiten Ordnung mit der Taylorreihe übereinstimmen:

$$\begin{aligned} y(x_{n+2}) &= y_{n+1} + hy'(x_{n+1}) + \frac{1}{2}h^2y''(x_{n+1}) + O(h^3) \\ &= y_{n+1} + hf(x_{n+1}, y_{n+1}) + \frac{1}{2}h^2\Big(\frac{\partial f(x_{n+1}, y_{n+1})}{\partial x} + \frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} f(x_{n+1}, y_{n+1})\Big). \end{aligned}$$

Vergleicht man Koeffizienten, findet man

$$a + b = 1 \quad -b = \frac{1}{2} \quad \Rightarrow \quad a = \frac{3}{2}.$$

Aus der Formel (5.24) wird somit die Iterationsformel

$$y_{n+2} = y_{n+1} + h\Big(\frac{3}{2}f(x_{n+1}, y_{n+1}) - \frac{1}{2}f(x_n, y_n)\Big). \tag{5.25}$$

Diese Rekursionsformel definiert ein quadratisches Verfahren, das *Adams-Bashforth-Verfahren* mit $s = 2$.

Das Verfahren kann ähnlich wie das Runge-Kutta-Verfahren auf höhere Ordnung erweitert werden. Man findet nach einiger Rechnung

$$\begin{aligned} s = 1: &\quad y_{n+1} = y_n + hf(x_n, y_n) \\ s = 2: &\quad y_{n+2} = y_{n+1} + h\Big(\frac{3}{2}f(x_{n+1}, y_{n+1}) - \frac{1}{2}f(x_n, y_n)\Big) \\ s = 3: &\quad y_{n+3} = y_{n+2} + h\Big(\frac{23}{12}f(x_{n+2}, y_{n+2}) - \frac{4}{3}f(x_{n+1}, y_{n+1}) + \frac{5}{12}f(x_n, y_n)\Big) \\ s = 4: &\quad y_{n+4} = y_{n+3} + h\Big(\frac{55}{24}f(x_{n+3}, y_{n+3}) - \frac{59}{24}f(x_{n+2}, y_{n+2}) + \frac{37}{24}f(x_{n+1}, y_{n+1}) - \frac{3}{8}f(x_n, y_n)\Big) \end{aligned}$$

Es ist also möglich, ausgehend von dieser Idee Verfahren beliebig hoher Ordnung zu produzieren.

In der Tabelle 5.4 wird das Adams-Bashforth-Verfahren verglichen mit dem lineare Euler-Verfahren und dem Verfahren vierter Ordnung von Runge-Kutta. Die Verbesserung der Genauigkeit des

Adams-Bashforth-Verfahrens gegenüber dem Euler-Verfahren ist konsistent damit, dass das Adams-Bashforth-Verfahren ein quadratisches Verfahren ist.

Nachteilig an den Mehrschritt-Verfahren ist die Notwendigkeit, genügend viele Werte $y_n, \ldots, y_{n+s-1}$ mit ausreichend hoher Genauigkeit zu bestimmen, bevor das Mehrschritt-Verfahren seine Schritte der Ordnung s beginnen kann. Solange diese Werte nicht zur Verfügung stehen, kann ein Mehrschritt-Verfahren nur Schritte niedrigerer Ordnung als s durchführen.

Bei einem Einschritt-Verfahren kann in jedem Schritt die Schrittweite h verändert werden, zum Beispiel für Bereiche von x-Werten, in denen die Steigung von $y(x)$ sehr rasch ändert.

Für die Beispiel-Differentialgleichung (5.13) können wir das Adams-Bashforth-Verfahren zweiter Ordnung ($s = 2$) vollständig analysieren. Die Rekursionsformel wird zu

$$y_{n+2} = y_{n+1} + h\Big(\frac{3}{2}(-\alpha y_{n+1}) - \frac{1}{2}(-\alpha y_n)\Big) = \Big(1 - \frac{3}{2}\alpha h\Big)y_{n+1} + \frac{\alpha h}{2}y_n.$$

Dies ist eine Differenzengleichung mit konstanten Koeffizienten, man kann sie mit Hilfe eines Potenzansatzes lösen. Wir nehmen also an, dass $y_n = \lambda^n$, und setzen dies in die Rekursionsformel ein. Ausserdem kürzen wir $\alpha h/2 = \delta$ ab. Wir erhalten

$$\lambda^{n+2} - (1 - 3\delta)\lambda^{n+1} - \delta\lambda^n = 0.$$

Nach Division durch λ^n erhalten wir die quadratische Gleichung

$$\lambda^2 - (1 - 3\delta)\lambda - \delta = 0$$

für λ mit den Lösungen

$$\lambda_\pm = \frac{1}{2}(1 - 3\delta) \pm \frac{1}{2}\sqrt{(1 - 3\delta)^2 + 4\delta}.$$

Da δ klein ist, wird λ_- ebenfalls klein sein, während λ_+ näher bei 1 sein wird. Der dominante Einfluss auf die Lösung rührt also von λ_+ her. Um diesen Unterschied genauer zu verstehen, verwenden wir eine lineare Approximation der Wurzel auf der rechten Seite von $\lambda_\pm$:

$$\begin{aligned}
\sqrt{1+x} &= 1 + \frac{x}{2} - \frac{x^2}{4} + \frac{3x^3}{8} - \ldots \\
\sqrt{x} &= \sqrt{x_0 + x - x_0} = \sqrt{x_0}\sqrt{1 + \frac{x - x_0}{x_0}} = \sqrt{x_0}\Big(1 + \frac{1}{2}\frac{x - x_0}{x_0} - \frac{1}{4}\frac{(x - x_0)^2}{x_0^2} + \ldots\Big) \\
&= \sqrt{x_0} + \frac{1}{2}\frac{x - x_0}{\sqrt{x_0}} - \frac{1}{4}\frac{(x - x_0)^2}{\sqrt{x_0}^3} + \ldots
\end{aligned}$$

Wir verwenden diese Approximation mit $x_0 = (1 - 3\delta)^2$ und $x - x_0 = -4\delta$

$$\begin{aligned}
\sqrt{(1 - 3\delta)^2 + 4\delta} &= (1 - 3\delta)\Big(1 + \frac{1}{2}\frac{4\delta}{(1 - 3\delta)^2} - \frac{1}{4}\frac{16\delta^2}{(1 - 3\delta)^4} + \ldots\Big) \\
&= (1 - 3\delta) + \frac{1}{2}\frac{4\delta}{1 - 3\delta} - \frac{1}{4}\frac{16\delta^2}{(1 - 3\delta)^3} + \ldots \\
&= 1 - 3\delta + 2\delta(1 + 3\delta) - 4\delta^2 + O(\delta^3) \\
&= 1 - 3\delta + 2\delta + 2\delta^2 + O(\delta^3) \\
&= 1 - 3\delta + 2\delta + \frac{1}{2}(2\delta)^2 + O(\delta^3).
\end{aligned}$$

Damit können wir jetzt λ_+ bis zur zweiten Ordnung berechnen:

$$\begin{aligned}\lambda_+ &= \frac{1}{2}\Big((1-3\delta) + (1-3\delta) + 2\delta + \frac{1}{2}(2\delta)^2\Big) + O(\delta^3)\\ &= 1 - 2\delta + \frac{1}{2}(2\delta)^2 + O(\delta^3)\\ &= e^{-2\delta} + O(\delta^3).\end{aligned}$$

Die exakte Lösung erfüllt $y_{n+1} = e^{-2\delta} y_n$, der Faktor λ_+ stimmt bis auf Terme mindestens dritter Ordnung mit $e^{-2\delta}$ überein. Damit ist erneut bestätigt, dass wir es mit einem quadratischen Verfahren zu tun haben.

Wir können auch λ_- berechnen, und erhalten

$$\lambda_- = -\delta - 2\delta^2 + O(\delta^3).$$

Da δ klein ist, ist eine Komponente der Lösung bereits nach drei Schritten kleiner als $O(\delta^3)$, und spielt daher im Vergleich zu den von λ_+ herrührenden Lösungen in dritter Ordnung keine Rolle.

5.6 Software

Die im letzten Abschnitt entwickelten numerischen Verfahren zur Lösung einer Differentialgleichung kommen ausschliesslich mit Auswertungen der Funktion f aus, die Ableitungen der Funktion f müssen nicht bekannt sein. Es sollte also ein Leichtes sein, eine Softwarebibliothek zur Verfügung zu stellen, mit der eine beliebige gewöhnliche Differentialgleichung gelöst werden kann. Als Input braucht es nur die Funktion f und die Anfangsbedingungen.

Als Beispiel wollen wir in diesem Abschnitt die Differentialgleichung

$$y'' + y = \sin\frac{x}{10}, \qquad y(0) = y'(0) = 0$$

in verschiedenen Programmierumgebungen lösen. Als erstes bringen wir die Differentialgleichung wieder in die Standardform einer Vektordifferentialgleichung erste Ordnung:

$$\frac{d}{dt}Y = \frac{d}{dx}\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \begin{pmatrix}y_2\\ -y_1 + \sin\dfrac{x}{10}\end{pmatrix} = f(x, Y). \tag{5.26}$$

Ein numerisches Verfahren braucht also als Input eine Anfangsbedingung sowie die Funktion f. Ausserdem muss es Möglichkeiten bereitstellen, wie man den Gang der Rechnung beeinflussen kann, z. B. um die x-Werte anzugeben, für die die $Y(x)$ bestimmt werden sollen, oder um Genauigkeitsziele zu erreichen.

5.6.1 Octave

In Octave steht eine einzige Funktion `lsode` zur Verfügung, welche auf zuverlässige Art Differentialgleichungen löst. Der Anwender muss eine Implementation der Funktion f zur Verfügung stellen, allerdings werden die Argument in der umgekehrten Reihenfolge zu der erwarte, die wir in diesem Skript bisher verwendet haben. Für die Beispieldifferentialgleichung (5.26) kann man sie zum Beispiel so definieren:

x	$y_{\text{numerisch}}(x)$	$y_{\text{exakt}}(x)$	Fehler
0	0.00000000	0.00000000	0.00000000
1	0.00158525	0.00158528	0.00000003
2	0.01090682	0.01090678	0.00000003
3	0.02858723	0.02858716	0.00000007
4	0.04756207	0.04756212	0.00000004
5	0.05957416	0.05957437	0.00000020
6	0.06276426	0.06276444	0.00000018
7	0.06337942	0.06337932	0.00000010
8	0.07002849	0.07002811	0.00000037
9	0.08576626	0.08576594	0.00000032
10	0.10528405	0.10528416	0.00000010
100	0.84661503	0.84661930	0.00000427
1000	−0.55228836	−0.55234514	0.00005678
2000	0.90392063	0.90373523	0.00018540
3000	−0.99018339	−0.99032256	0.00013917
4000	0.75185982	0.75202340	0.00016358
5000	−0.25298074	−0.25252044	0.00046030
6000	−0.30093757	−0.30056348	0.00037408
7000	0.76905079	0.76889872	0.00015207
8000	−1.00327790	−1.00396748	0.00068958
9000	0.88860437	0.88793099	0.00067338
10000	−0.50337856	−0.50335983	0.00001873

Tabelle 5.5: Exakte und numerische Lösung der Beispieldifferentialgleichung berechnet mit der Funktion `lsode` von Octave.

```
global omega = 0.1;

function yprime = f(y, x)
        global omega;
        yprime = [ y(2); -y(1) + sin(omega * x) ];
endfunction
```

Beim Aufruf der Funktion `lsode` muss man den *Namen* der Funktion, die Anfangsbedingung, sowie einen Vektoren mit x-Werten, für die man die Lösung ausgegeben haben möchte, als Argumente übergeben. Der erste Wert im x-Vektor muss der x-Wert für die Anfangsbedingung sein, in unserem Fall also 0. Um die Werte von $y(x)$ für ganzzahlige Werte von x zu erhalten, muss man also die Befehle

```
y0 = zeros(2,1);
x = (0:10) * deltax;
lsode("f", y0, x)
```

ausführen. Als Rückgabewert erhält man eine Matrix, die in jeder Zeile die Werte von $y(x)$ und $y'(x)$ zum entsprechenden Wert von x aus dem `x`-Argument enthält. Die Resultate sind zusammen mit den Werten der exakten Lösung (5.14) in der dritten Spalte in der Tabelle 5.5 zusammengestellt. Es ist gut erkennbar, wie der Fehler anfänglich langsam ansteigt, dann aber unter Kontrolle bleibt. Die Dokumentation der Funktion `lsode` beschreibt, wie man mit Hilfe von Optionen ihr Verhalten und insbesondere die Grösse der Fehler weiter beeinflussen kann.

5.6.2 GNU Scientific Library

Während Octave dem Benutzer die Wahl eines geeigneten Verfahrens abnimmt und ihm überhaupt wenig Kontrolle über den Gang der Rechnung gibt, kann ein C-Programmierer durch den Einsatz der GNU Scientific Library (GSL) die volle Kontrolle über alle Aspekte der Iteration erhalten. Der Preis ist eine wesentlich höhere Komplexität. Ziel dieses Abschnitts ist, ein einfaches Beispielprogramm zu zeigen, welches als Basis eigener Programme dienen kann. Es verwendet eine Runge-Kutta-Verfahren achter Ordnung.

Die Funktionen zum Lösen von gewöhnlichen Differentialgleichungen der GSL haben alle das Präfix `gsl_odeiv2_`. Zunächst braucht es natürlich wieder eine Implementation der Funktion f. Die GSL übergibt zwei Arrays, im einen findet die Funktion die aktuellen Y-Werte, im anderen soll sie die Werte der Ableitung zurückgeben. Für die Beispiel-Differentialgleichung (5.26) sieht der Code wie folgt aus:

```
int f(double x, const double y[], double f[], void *params) {
        double  omega = *(double *)params;
        f[0] = y[1];
        f[1] = -y[0] + sin(omega * x);
        return GSL_SUCCESS;
}
```

Der Parameter `params` dient dazu, der Funktion zusätzliche Parameter zu übergeben. In unserem Fall ist das nur die Zahl ω. Da `params` ein `void`-Pointer ist, kann eine beliebige Struktur zur Parameterübergabe verwendet werden.

Die Differentialgleichung wird beschrieben durch eine Struktur vom Typ `gsl_odeiv2_system`, welche ausser Zeigern auf die Funktion und die Parameter-Struktur auch noch die Dimension der Vektoren enthält. Es kann auch noch ein Funktionszeiger für eine Funktion übergeben werden, die die Jacobi-Matrix berechnet, in unserem Beispiel wird dies jedoch nicht benötigt.

Die eigentliche Berechnung wird von einer “driver”-Funktion durchgeführt. Diese sorgt im wesentlichen für die Wahl der Schrittweite, verwaltet Datenstrukturen, und ruft die Funktionen auf, die die einzelnen Schritte durchführen. Die Treiber-Funktion führt die einzelnen Schritte (im Sinne der in Abschnitt **??** besprochenen Einschritt-Verfahren) mit Hilfe der Schritt-Funktionen durch, von denen die Bibliothek eine ganze Reihe bereitstellt. Die Funktion `gsl_odeiv2_step_rk4` ist das klassische Runge-Kutta-Verfahren vierter Ordnung, welches in Abschnitt 5.4.2 beschrieben wurde. Im Beispielverfahren verwenden wir `gsl_odeiv2_step_rk8pd`, das Runge-Kutta Prince-Dormand Verfahren achter Ordnung. Für Aufgaben allgemeiner Art ebenfalls sehr gut geeignet ist das Runge-Kutta-Fehlberg-Verfahren fünfter Ordnung mit dem Namen `gsl_odeiv2_step_rkf45`. Diese Datenstrukturen werden mit dem Code

```
double  omega = 0.1;    /* Parameter fuer f */
gsl_odeiv2_system       system = { f, NULL, 2, &omega }; /* DG-System */
gsl_odeiv2_driver       *driver
        = gsl_odeiv2_driver_alloc_y_new(&system, gsl_odeiv2_step_rk8pd,
                1e-6, 1e-6, 0.0);
```

initialisiert. Durch Austausch des zweiten Arguments der Driver-Allozierungs-Funktion kann man leicht das Verfahren wechseln und so Zeitaufwand und Genauigkeit für verschiedene Lösungsverfahren vergleichen.

Um die Rechnung durchzuführen, muss jetzt die Driver-Funktion so oft angewendet werden, wie man Punkt der Lösungskurve ausgeben will. Dazu dient die Funktion `gsl_odeiv2_driver_apply`. An Argumenten braucht sie den eben initialisierten Driver, den aktuellen x-Wert, den x_{next}-Wert, für

den der nächste Punkt ausgegeben werden soll, sowie einen Vektor, in dem der aktuelle Anfangswert für $Y(x)$ übergeben und $Y(x_{\text{next}})$ zurückgegeben wird. x wird als Referenz übergeben, wenn die Funktion zurückkehrt, findet man dort den neuen aktuellen Wert von x, also im Erfolgsfall x_{next}. In unserem Fall brauchen wir $X(x)$ für ganzzahlige x, die folgende Schleife bewerkstelligt dies:

```
double  x = 0.0;
double  y[2] = { 0.0, 0.0 };
long    lastcounter = evalcounter;
for (int i = 1; i <= 10000; i++) {
        double  xnext = i;
        int     status = gsl_odeiv2_driver_apply(driver, &x, xnext, y);
        if (status != GSL_SUCCESS) {
                fprintf(stderr, "error: return value = %d\n", status);
                return EXIT_FAILURE;
        }
        /* Output */
        ...
}
```

Man kann die Funktion f im Programm natürlich auch mit einem Zähler ausstatten und damit herausfinden, wie viele Aufrufe der Funktion für die numerische Lösung benötigt werden. Es stellt sich heraus, dass für das erste Intervall von 0 bis 1 die Funktion f 131 mal aufgerufen wird, hier versucht die Bibliothek die optimale Schrittweite h zu bestimmen. In allen folgenden Intervallen der Länge 1 von n bis $n+1$ werden nur noch jeweils 13 Aufrufe der Funktion benötigt. Verwendet man stattdessen das Runge-Kutta-Fehlberg-Verfahren, werden pro Intervall 18 Auswertungen der Funktion f benötigt, und die Genauigkeit sinkt auf zwei Stellen nach dem Komma.

5.7 Randwertprobleme

Die bisher beschriebenen Verfahren gehen von einer Anfangsbedingung aus, und berechnen die dadurch eindeutig festgelegte Lösungskurve. Randwertproblem, beschrieben in Abschnitt 5.1.3, verknüpfen dagegen Werte von einzelnen Komponenten von Y an den Rändern eines Intervalls.

5.7.1 Einführende Beispiele

Wir betrachten zwei prototypische Randwertprobleme, die das allgemeine Lösungsverfahren motivieren sollen. In der ersten Aufgabe sind wir dank einer expliziten Form der Lösung nach Einsetzen der Randwertbedingung die Parameter durch Lösen von Gleichungen zu bestimmen.

Aufgabe 5.5. *Mit einem nur der Schwerkraft unterworfenen Ball, der im Ursprung des Koordinatensystems geworfen wird, soll ein Ziel im Punkt P getroffen werden. In welcher Richtung und mit welcher Anfangsgeschwindigkeit muss er geworfen werden?*

Um das Problem einfach zu halten, modellieren wir diese Aufgabe wie folgt. Der Ball der Masse m bewegt sich in der x-y-Ebene, wobei die Schwerkraft in negativer y-Richtung zeigt. Das Newtonsche Gesetz liefert die Differentialgleichung zweiter Ordnung

$$m\frac{d^2}{dt^2}\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}0\\-mg\end{pmatrix}. \tag{5.27}$$

Die Masse m kann herausgekürzt werden. Gesucht ist eine Lösung so, dass die Bahn durch die Punkte $(0,0)$ und $P=(p,0)$ geht.

x	$y_{\text{numerisch}}(x)$	$y_{\text{exakt}}(x)$	Fehler
1	0.01584477	0.01584477	−0.00000000
2	0.10882786	0.10882787	−0.00000000
3	0.28425071	0.28425071	−0.00000001
4	0.46979656	0.46979656	−0.00000000
5	0.58112927	0.58112926	0.00000001
6	0.59856974	0.59856972	0.00000002
7	0.58436266	0.58436265	0.00000001
8	0.62466692	0.62466694	−0.00000001
9	0.74961115	0.74961117	−0.00000003
10	0.90492231	0.90492232	−0.00000001
20	0.82626555	0.82626556	−0.00000000
30	0.24234669	0.24234664	0.00000005
40	−0.83971103	−0.83971092	−0.00000011
50	−0.94210772	−0.94210787	0.00000015
60	−0.25144906	−0.25144893	−0.00000013
70	0.58545210	0.58545205	0.00000005
80	1.09974465	1.09974456	0.00000009
90	0.32597838	0.32597860	−0.00000023
100	−0.49836792	−0.49836823	0.00000031
200	1.01037929	1.01037877	0.00000052
300	−0.89702589	−0.89702630	0.00000041
400	0.83859090	0.83859101	−0.00000010
500	−0.21777633	−0.21777543	−0.00000090
600	−0.31235410	−0.31235239	−0.00000171
700	0.72675906	0.72676124	−0.00000219
800	−1.09422992	−1.09422790	−0.00000202
900	0.80223761	0.80223872	−0.00000111
1000	−0.59500324	−0.59500363	0.00000040
2000	−0.97606658	−0.97606187	−0.00000471
3000	−1.03200392	−1.03199479	−0.00000914
4000	−0.79047800	−0.79047372	−0.00000428
5000	−0.37269297	−0.37270218	0.00000921
6000	0.08785158	0.08783273	0.00001886
7000	0.49827647	0.49826463	0.00001184
8000	0.80219750	0.80220742	−0.00000992
9000	0.94568799	0.94571577	−0.00002778
10000	0.86607968	0.86610200	−0.00002232

Tabelle 5.6: Lösungen der Beispieldifferentialgleichung (5.26) mit Hilfe der GNU Scientific Library (GSL).

Genau genommen ist dies nicht ein Randwertproblem wie in Abschnitt 5.1.3, denn es wird nicht verlangt, dass der Ball zu einer bestimmten Zeit t beim Punkt P eintrifft. Die Differentialgleichung bedeutet aber, dass die Horizontalgeschwindigkeit des Balls konstant ist (die horizontale Beschleunigung ist immer 0). Ist v_x die Horizontalgeschwindigkeit, dann erreicht der Ball zur Zeit $t_1 = p/v_x$ die x-Koordinate des Ziels. Gesucht ist also die anfängliche Vertikalgeschwindigkeit, die man dem Ball geben muss, dass zur Zeit p/v_x die y-Komponente der Lösung den Wert 0 hat. In dieser Form liegt ein Randwertproblem wie in Abschnitt 5.1.3 vor.

Die Lösungen der Differentialgleichung 5.27 sind aus dem Physik-Unterricht bekannt: es gilt

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} v_x t \\ v_y t - \frac{1}{2} g t^2 \end{pmatrix} \tag{5.28}$$

Damit lässt sich auch das Randwertproblem lösen. Für $t = v_x/p$ muss $y(t) = 0$ sein, also

$$\begin{aligned} y(t) = y\Big(\frac{p}{v_x}\Big) = v_y \frac{p}{v_x} - \frac{1}{2} g \Big(\frac{p}{v_x}\Big)^2 = 0 \\ \Rightarrow \qquad v_y = \frac{v_x}{p} \frac{1}{2} g \frac{p^2}{v_x^2} = \frac{gp}{2v_x}. \end{aligned} \tag{5.29}$$

Offenbar gibt es zu jedem v_x einen passenden Wert von v_y, mit dem das Ziel getroffen wird.

Die Differentialgleichung (5.27) ist nicht in einer Form, die der numerischen Lösung zugänglich ist. Wir schreiben Sie daher als Differentialgleichung erster Ordnung für vierdimensionale Vektoren:

$$\frac{d}{dt} Y = \frac{d}{dt} \begin{pmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \dot{x} \\ \dot{y} \\ 0 \\ -g \end{pmatrix}. \tag{5.30}$$

Gesucht ist eine Lösung, die die Randbedingungen

$$Y(0) = \begin{pmatrix} 0 \\ 0 \\ v_x \\ v_y \end{pmatrix}, \qquad Y\Big(\frac{p}{v_x}\Big) = \begin{pmatrix} p \\ 0 \\ ? \\ ? \end{pmatrix} \tag{5.31}$$

erfüllt. Darin stehen die roten Einträge für Werte, die nicht vorgegeben sind. Aus der Symmetrie des Problems kann man natürlich auch die Endgeschwindigkeit ablesen. Zu bestimmen ist also v_y so, dass die Lösungskurve durch den Punkt $(p, 0)$ geht.

Wird statt der Horizontalkomponenten der Anfangsgeschwindigkeit die gesamte Anfangsgeschwindigkeit v_0 vorgegeben, dann muss der Winkel gefunden werden, unter dem der Ball geworfen werden muss, um das Ziel zu trefen. Bei der Elevation α sind die Komponenten der Anfangsgeschwindigkeit $v_x = v_0 \cos\alpha$ und $v_y = v_0 \sin\alpha$. Setzt man dies in die Bedingung (5.29) ein, findet man

$$\begin{aligned} v_0 \sin\alpha &= \frac{gp}{2v_0 \cos\alpha} \\ 2 \sin\alpha \cos\alpha &= \frac{gp}{v_0^2} \\ \sin 2\alpha &= \frac{gp}{v_0^2} \end{aligned}$$

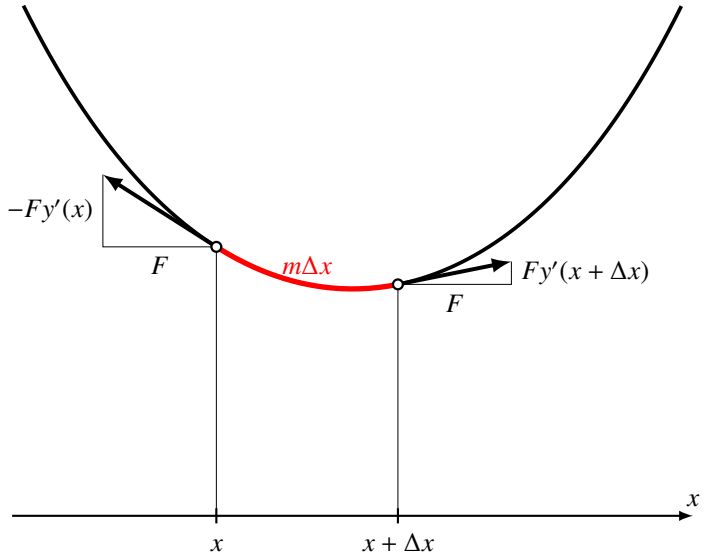


Abbildung 5.6: Herleitung der Differentialgleichung der Kettenlinie

$$\alpha = \frac{1}{2} \arcsin \frac{gp}{v_0^2}.$$

Im Nenner rechts steht im wesentlichen die kinetische Energie, je mehr kinetische Energie der Ball zu Beginn hat, desto kleiner ist der Winkel, man trifft das Ziel mit einer sehr flachen Bahn. Kleine Winkel reichen auch für geringe Schwerkraft (g klein) und kurze Distanzen (p klein). Die maximale Distanz wird erreicht, wenn das Argument des Arcussinus den Wert 1 erreicht, grösser darf p nicht werden, weil es sonst keine Lösung mehr für α gibt. Die Maximaldistanz ist daher

$$p_{\max} = \frac{v_0^2}{g}.$$

Aufgabe 5.6. *Ein Seil ist zwischen zwei Punkten aufgehängt, welche Form nimmt es allein unter der Wirkung seines Eigengewichtes an?*

Die Lösungskurve dieses Problems heisst die *Kettenlinie*. Auch in diesem Fall können wir wieder eine Lösungsfunktion $y(x)$ finden, aber die Bestimmung der Parameter ist jetzt nicht mehr in geschlossener Form möglich. Wir behelfen uns mit einer numerischen Lösung.

Lösung. Zunächst brauchen wir für eine Differentialgleichung, deren Lösung die gesuchte Kurve beschreibt. Zur Herleitung dient die Abbildung 5.6. Die Masse des Seils zwischen den beiden Punkten x und $x+\Delta x$ wird von den beiden eingezeichneten Kräften getragen. Die horizontalen Komponenten tragen nicht dazu bei, das Seil zu tragen, sie haben daher entlang des ganzen Seils immer die gleiche Grösse F. Die Masse des Seilstücks ist proportional zu seiner Länge, der Proportionalitätsfaktor ist die lineare Massendichte μ. Nach dem dritten Newtonschen Gesetz sind die vertikalen Kraftkomponenten gleich gross wie die Gewichtskraft des Seilstücks:

$$F(y'(x+\Delta x) - y'(x)) = \mu g \sqrt{(y(x+\Delta x) - y(x))^2 + \Delta x^2}$$

oder nach Divison durch $F\Delta x$:

$$\frac{y'(x+\Delta x)-y'(x)}{\Delta x} = \frac{\mu g}{F}\sqrt{1+\left(\frac{y(x+\Delta x)-y(x)}{\Delta x}\right)^2}.$$

Schreibt man $a = \mu g/F$ und geht zur Grenze $\Delta x \to 0$ über, erhält man die Differentialgleichung

$$y''(x) = a\sqrt{1+y'(x)^2}. \tag{5.32}$$

Diese Differentialgleichung hat die Funktion

$$y(x) = \frac{1}{a}\cosh ax + C$$

als Lösung, wie man durch Nachrechnen einsehen kann[1]. Die Ableitungen von $y(x)$ sind

$$y'(x) = \sinh ax$$
$$y''(x) = a\cosh ax.$$

Eingesetzt in die Differentialgleichung erhält man

$$a\cosh ax = a\sqrt{1+\sinh^2 ax},$$

was sich zu

$$\cosh^2 ax - \sinh^2 ax = 1$$

umformen lässt, diese Gleichung ist für hyperbolische Funktionen immer erfüllt.

Die Differntialgleichung (5.32) ist autonom, also sind auch verschobene Funktionen Lösungen. Die allgemeine Lösung des Problems ist daher die Funktion

$$y(x) = \frac{1}{a}\cosh a(x-x_0) + C. \tag{5.33}$$

Die Bedeutung der Konstante C ist leichter verständlich, wenn man sie als $C = y_0 - 1/a$ schreibt, also

$$y(x) = \frac{1}{a}\cosh a(x-x_0) - \frac{1}{a} + y_0.$$

Setzt man $x = x_0$ ein, findet man $y(x_0) = y_0$, d. h. der Punkt (x_0, y_0) ist der Scheitelpunkt des Graphen von $y(x)$.

[1] Man kann die Gleichung (5.32) natürlich auch direkt lösen. Dazu bestimmt man zuerst die Funktion $z(x) = y'(x)$, welche die Differentialgleichung

$$z'(x) = a\sqrt{1+z(x)^2}$$

erfüllt. Diese Gleichung lässt sich durch Separation lösen:

$$\frac{dz}{dx} = a\sqrt{1+z^2} \quad\Rightarrow\quad \frac{1}{a}\int\frac{dz}{\sqrt{1+z^2}} = \int dx \quad\Rightarrow\quad \frac{1}{a}\operatorname{asinh} z = x + C_1 \quad\Rightarrow\quad z(x) = \sinh a(x+C_1).$$

Die Funktion $y(x)$ bekommt man jetzt durch Integration

$$y(x) = \int z(x)\,dx = \frac{1}{a}\cosh a(x+C_1) + C_2,$$

mit $C_1 = -x_0$ und $C_2 = C$ genau die vorgeschlagene Lösung.

Damit kann jetzt das Anfangswertproblem gelöst werden. Wir verlangen, dass die im Punkt $x = x_1$ der Funktionswert $y(x_1) = y_1$ sein soll, und die Steigung $y'(x_1) = m$. Setzt man die Lösung (5.33) in die Bedingung für die Steigung ein, erhält man

$$m = \sinh a(x_1 - x_0) \qquad \Rightarrow \qquad \operatorname{asinh} m = a(x_1 - x_0) \qquad \Rightarrow \qquad x_0 = x_1 - \frac{1}{a} \operatorname{asinh} m.$$

Die Anfangsbedingung für $y(x_1)$ liefert

$$y_1 = y(x_1) = \frac{1}{a} \cosh a(x_1 - x_0) + y_0 - \frac{1}{a} = \frac{1}{a} \cosh \operatorname{asinh}(m) + y_0 - \frac{1}{a} = \frac{1}{a} \sqrt{1 + m^2} + y_0 - \frac{1}{a}$$

$$\Rightarrow \quad y_0 = y_1 + \frac{1}{a} - \frac{1}{a} \sqrt{1 + m^2}.$$

Damit ist das Anfangswertproblem vollständig gelöst,

$$y(x) = \frac{1}{a} \cosh(a(x - x_1) + \operatorname{asinh} m) + y_1 - \frac{1}{a} \sqrt{1 + m^2} \tag{5.34}$$

ist die allgemeine Lösung zur Anfangsbedingung $y(x_1) = y_1$, $y'(x_1) = m$.

Mit dieser Lösung kann jetzt auch das Randwertproblem gelöst werden. Es muss ein Wert m gefunden werden, so dass $y(x_2) = y_2$, man muss also die Gleichung

$$y_2 = \frac{1}{a} \cosh(a(x_2 - x_1) + \operatorname{asinh} m) + y_1 - \frac{1}{a} \sqrt{1 + m^2} =: f(m)$$

nach m auflösen. Das ist leider nicht so einfach, weil m in der transzendenten Funktion asinh auftritt.

Wir können die Gleichung jedoch iterativ mit dem Newton-Verfahren nach Satz **??** lösen. Dazu brauchen wir die Ableitung von f, sie ist

$$f'(m) = \frac{\sinh(x_2 - x_1 + \operatorname{asinh} m) - m}{\sqrt{m^2 + 1}}.$$

Damit können wir die Iterationsformel

$$m_{\text{neu}} = m - \frac{f(m) - y_2}{f'(m)}$$

für den korrekten Wert m finden. Sie kann leicht in Octave implementiert werden, wie Listing 5.1 zeigt.

In Tabelle 5.7 ist der Gang der Berechnung für den Fall $x_1 = -1$, $x_2 = 2$, $y_1 = 1$ und $y_2 = 2$ mit dem Startwert $m = 0$ dargestellt. Wie erwartet ist die Konvergenz quadratisch, in jedem Schritt verdoppelt sich die Anzahl korrekter Stellen. In Abbildung 5.7 sind die Graphen von $y(x)$ zu den im Newton-Algorithmus ermittelten Werten von $m = y'(x_1)$ blau dargestellt, die Lösungskurve ist rot. ○

Listing 5.1: Octave-Programm zur Bestimmung der Anfangssteigung m im Kettenlinien-Problem

```
1 global x1 = -1
2 global x2 =  2
3 global y1 =  1
4 global y2 =  2
5
```

	m	$f(m)$
1	0.0000000000000000	8.0676619957777653
2	−0.8053266839760436	2.5748454835378896
3	−1.3999183852140562	0.5757860863682827
4	−1.6180720459804230	0.0383682292517731
5	−1.6347219547661627	0.0001812733120350
6	−1.6348013659846743	0.0000000040625958
7	−1.6348013677644737	0.0000000000000004
8	−1.6348013677644739	−0.0000000000000002

Tabelle 5.7: Numerische Lösung des Randwertproblems für die Kettenlinie mit Randbedingungen $x_1 = -1$, $x_2 = 2$, $y(x_1) = y_1$ und $y(x_2) = y_2$.

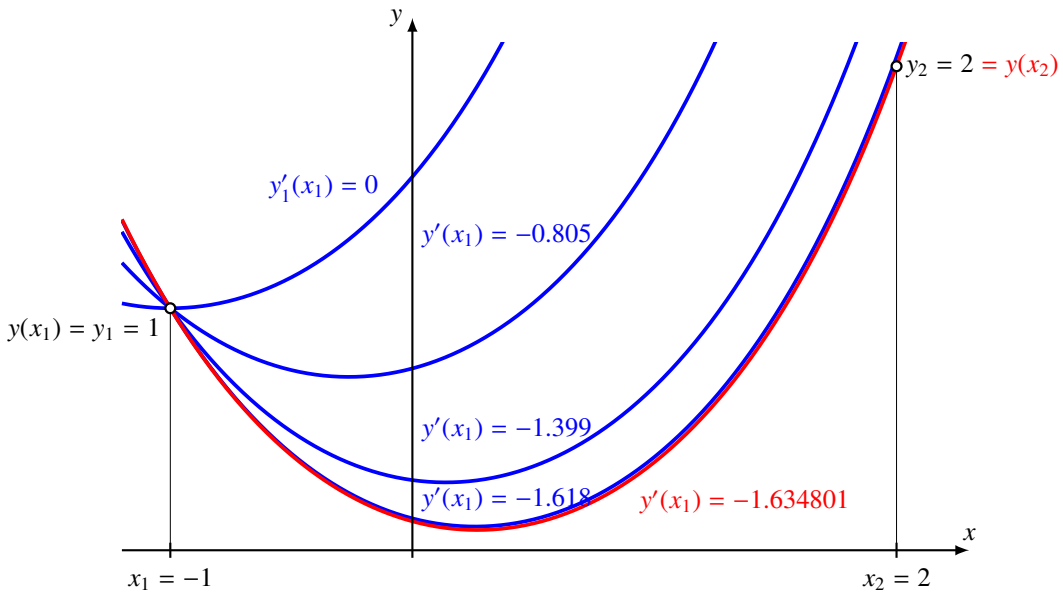


Abbildung 5.7: Iterative Bestimmung der Kettenlinie zwischen den Punkten $(-1, 2)$ und $(2, 2)$ mit Hilfe des Newton-Algorithmus.

```
function v = f(m)
        global x1 x2 y1 y2
        v = cosh(x2 - x1 + asinh(m)) + y1 - y2 - sqrt(1 + m^2);
endfunction

function v = fprime(m)
        global x1 x2 y1 y2
        v = (sinh(x2 - x1 + asinh(m)) - m)/sqrt(1 + m^2);
endfunction

m = 0;

for i = (1:10)
        printf("%2d   %20.16f %20.16f\n", i, m, f(m));
        m = m - f(m)/fprime(m);
endfor
```

5.7.2 Schiess-Verfahren

Wenn man experimentell versucht, ein Ziel zu treffen, dann wird man in wiederholten Versuchen die Richtung anpassen, so dass man dem Ziel immer näher kommt. Genau dies haben wir in der Aufgabe 5.6 gemacht, wo wir iterativ die noch unbekannen Anfangsbedingung $y(x_1) = m$ bestimmt haben, mit der die Lösung durch den rechten Randpunkt $y(x_2) = y_2$ geht.

Auch in der Aufgaben 5.5 konnten wir die Parameter direkt bestimmen. Doch auch dort läuft die Lösung auf die Bestimmung einer geeigneten Anfangsbedingung hinaus. Der y-Wert zur Zeit p/v_x hängt von der Vertikalgeschwindigkeit ab, wir bezeichnen ihn mit $h(v_y)$. Man verändert also v_y, bis die Gleichung $h(v_y) = 0$ erfüllt ist. Um das Randwertproblem zu lösen, muss man also die Gleichung $h(v_y) = 0$ numerisch lösen.

Man kann dies z. B. dadurch machen, dass man nach zwei Werten von v_y sucht, so dass die zum einen gehörige Bahn unter dem Punkt P durchgeht, während der Ball im anderen Fall darüber hinwegfliegt. Durch wiederholte Halbierung des Intervalls kann man dann den korrekten Wert für v_y immer genauer eingrenzen[2]. Der Nachteil dieses Verfahrens ist, dass mit jedem Schritt die Genauigkeit nur um in Bit ansteigt, es sind also sehr viele Iterationen notwendig.

Schnellere Konvergenz kann mit dem Newton-Verfahren erreicht werden, welches in Anhang **??** beschrieben wird. Für die Anwendung des Newton-Verfahrens auf das Randwert-Problem ist die Bestimmung der Steigung der Funktion nötig, die die Abweichung der Kurve von der Randbedingung am rechten Rand angibt. Wir müssen also berechnen, wie schnell sich $y(p/v_x)$ ändert, wenn v_y verändert wird. Dies ist die Ableitung

$$h'(v_y) = \frac{\partial y}{\partial v_y},$$

ein Eintrag der Jacobi-Matrix. In Abschnitt 5.1.5 wurde gezeigt, wie man auch für die Jacobi-Matrix eine Differentialgleichung aufstellen kann, die man natürlich ebenfalls mit den früher beschriebenen numerischen Bibliotheken lösen kann.

Das Randwertproblem kann daher mit folgendem Algorithmus numerisch gelöst werden.

1. Beginne mit einer Schätzung für v_y

[2]Tatsächlich wird dieses Verfahren in der Artillerie verwendet. Der Schiesskommandant beobachtet die einschlagenden Granaten und kommandiert Änderungen der Anfangs-Elevation an die Geschützbatterien. Dabei sucht er Einschläge, die aus seiner Perspektive vor bzw. hinter dem Ziel liegen, und halbiert dann das Intervall, bis die Einschläge dem Ziel genügend nahe kommen.

Abbildung 5.8: Lösungen des Anfangswertproblems (5.30) und (5.31). Das Newton-Verfahren korrigiert v_y derart, dass $h(v_y) = 0$ wird. So wird die Lösung des Randwertproblems (rot) gefunden.

n	v_y	t	$x(t)$	$y(x)$	$\frac{\partial y}{\partial v_y}$	$v_{y,\text{new}}$	Δ
0	7.0000	2.5	20.0	−13.156250	2.5	12.26250000	−5.2625000000
1	12.2625	2.5	20.0	−0.000004	2.5	12.26250145	−0.0000014458
2	12.2625	2.5	20.0	0.000000	2.5	12.26250143	0.0000000204

Tabelle 5.8: Newton-Algorithmus für das Ball-Problem, Resultate der numerischen Rechnung. v_y wird in drei Schritten mit einer Genauigkeit von mehr als 10 Stellen gefunden.

2. Finde numerisch die Lösung des Anfangswertproblems mit v_y als anfängliche Vertikalgeschwindigkeit. Berechnet dabei auch die Jacobi-Matrix

3. Lese die $h(v_y)$ aus der Lösung zur Zeit p/v_x ab, und $h'(v_y)$ aus der Jacobi-Matrix und verwendet den Newton-Algorithmus (Satz **??**), um eine verbesserte Schätzung von v_y zu bekommen.

4. Wiederhole Schritte 2 und 3 bis die Randbedingung für $t = p/v_x$ genügend genau erfüllt ist.

5. Die Lösung des Anfangswertproblems mit diesem v_y ist die Lösung des gestellten Randwertproblems.

Beispiel. Wir führen den eben skizzierten Algorithmus für das Ball-Problem durch. Um die Jacobi-Matrix zu berechnen, müssen wir die Ableitung von f berechnen:

$$\frac{\partial f(x,y)}{\partial y} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.35}$$

Da die rechte Seite nicht von y abhängt, können wir die Gleichung für die Jacobi-Matrix ganz unabhängig von y lösen. Da F so einfach ist, kann man das Matrizenprodukt direkt ausrechnen, so wird die Differentialgleichung für J

$$\begin{pmatrix} J'_{11} & J'_{12} & J'_{13} & J'_{14} \\ J'_{21} & J'_{22} & J'_{23} & J'_{24} \\ J'_{31} & J'_{32} & J'_{33} & J'_{34} \\ J'_{41} & J'_{42} & J'_{43} & J'_{44} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} J_{11} & J_{12} & J_{13} & J_{14} \\ J_{21} & J_{22} & J_{23} & J_{24} \\ J_{31} & J_{32} & J_{33} & J_{34} \\ J_{41} & J_{42} & J_{43} & J_{44} \end{pmatrix} = \begin{pmatrix} J_{31} & J_{32} & J_{33} & J_{34} \\ J_{41} & J_{42} & J_{43} & J_{44} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.36}$$

Daraus kann man ablesen, dass die Elemente J_{3j} und J_{4j} sich nicht ändern, sie bleiben also konstant. Aber auch in den ersten zwei Zeilen können sich nur die Elemente J_{13} und J_{24} ändern, die Differentialgleichungen für diese Elemente sind

$$J'_{13} = 1$$
$$J'_{24} = 1$$

oder in Matrixform:

$$J(x) = \begin{pmatrix} 1 & 0 & x & 0 \\ 0 & 1 & 0 & x \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.37}$$

Die Lösung $J_{13}(x) = x$ und $J_{24} = x$. Damit haben wir die nötige Information, um den Newton-Algorithmus durchzuführen. In Tabelle 5.8 sind die Resultate der numerischen Rechnung zusammengestellt. Es zeigt sich, dass der korrekte Wert für v_y in drei Iterationen mit 10 Stellen Genauigkeit gefunden werden kann. Damit ist das Randwertproblem numerisch gelöst. ○

Kapitel 6

Lineare Gleichungssysteme

Die lineare Algebra ist fundamental in vielen Bereichen der angewandeten Mathematik. Eine grosse Zahl von Methoden zur Lösung linearer Gleichungssysteme, zur Zerlegung von Matrizen und zur Lösung des Eigenwertproblems sind über die Jahrhunderte entwickelt worden und werden zum Teil bereits in Anfängervorlesungen unterrichtet. Insbesondere der Gausssche Elminitationsalgorithmus gehört zu den grundlegenden Techniken der numerischen linearen Algebra, wird hier aber als bekannt vorausgesetzt.

Die meisten Techniken gehen von relativ kleinen Gleichungssystemen aus. Sie sind aber schlicht nicht leistungsfähig genug oder stabil genug für grosse Systeme, wie sie zum Beispiel bei der Lösung von partiellen Differentialgleichungen oder Simulationen komplexer Systeme auftreten. Auch ist die Laufzeit für die exakte Lösung oft zu lang. Zum Beispiel hat der Gauss-Algorithmus für n Unbekannte Laufzeitkomplexität $O(n^3)$, was für Gleichungsyssteme mit $n > 10^5$ zu prohitiv grossem Aufwand führt. Kompromisse zwischen exakter Lösung und Durchführbarkeit in vernünftiger Zeit sind daher unumgänglich. Bereits Gauss hat daher iterative numerische Methoden entwickelt.

Dieses Kapitel präsentiert einige wenige Algorithmen des überaus weiten Feldes der numerischen linearen Algebra, welche die Vielfalt dieses Gebietes illustrieren sollen. Eine vertiefte Darstellung kann gefunden werden in [4]. In Abschnitt 6.1 wird gezeigt, wie sich Gleichungssysteme unter gewissen Voraussetzungen auch iterativ lösen lassen. Die QR-Zerlegung ist eine andere Formulierung des Problems, eine Basis zu orthonormalisieren, welches schon vom Gram-Schmidt-Algorithmus gelöst wurde, welcher allerdings gewisse Stabilitätsprobleme hat. Der in Abschnitt 6.2 vorgestellte, auf Spiegelungen basierende Algorithmus ist effizienter und stabiler. Das Eigenwertproblem für symmetrische Matrizen ist von grundlegender Bedeutung für die Anwendungen, die Lösung über das charakteristische Polynom, welches man oft in den Grundlagenvorlesungen lernt, ist jedoch nur für sehr kleine Matrizen praktikabel. Abschnitt 6.3 stellt das Jacobi-Verfahren zur Diagonalisierung symmetrischer Matrizen vor, welches ebenfalls iterativ arbeitet.

Weitere Verfahren der numerischen linearen Algebra werden in einzelnen Artikeln des zweiten Teiles vorgetellt.

6.1 Iterative Gleichungslösung nach Gauss-Seidel

Gegeben ist eine lineares Gleichungssystem von n Gleichungen mit n Unbekannten, welches wir als

$$\begin{array}{ccccccc} a_{11}x_1 + & a_{12}x_2 + & \dots & + & a_{1n}x_n = & b_1 \\ a_{21}x_1 + & a_{22}x_2 + & \dots & + & a_{2n}x_n = & b_2 \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ a_{21}x_1 + & a_{n2}x_2 + & \dots & + & a_{nn}x_n = & b_n \end{array} \tag{6.1}$$

Abgekürzt wird das Gleichungssystem auch $Ax = b$ geschrieben, wobei $A = (a_{ij})$ die Koeffizientenmatrix ist, $x = (x_k)$ der Vektor der Unbekannten und $b = (b_k)$ der Vektor der rechten Seiten.

6.1.1 Iterative Lösung nach Gauss-Seidel

Jede der Gleichungen (6.1) kann nach Variablen aufgelöst werden, sofern der zugehörige Koeffizient von 0 verschieden ist. Gleichung k in (6.1) ist

$$a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kk}x_k + \dots a_{kn}x_n = b_k$$

Aufgelöst nach x_k ist dies

$$x_k = \frac{1}{a_{kk}}(b_k - a_{k1}x_1 - a_{k2}x_2 - \cdots - a_{kn}x_n),$$

sofern $a_{kk} \neq 0$. Diese Gleichung kann dazu verwendet werden, die Werte für die Unbekannten zu verbessern.

Wir verwenden daher die Notation $x^{(m)}$ für die m-te Approximation der Lösung. Mit dieser Notation können wir die Iterations

Satz 6.1 (Gauss-Seidel-Iteration)**.** *Unter geeigneten Voraussetzungen konvergiert die Folge $x^{(m)}$ definiert durch*

$$x_k^{(m)} = b_k - a_{k1}x_1^{(m)} - \cdots - a_{k,k-1}x_{k-1}^{(m)} - a_{k,k+1}x_{k+1}^{(m-1)} - \cdots - a_{kn}x_n^{(m-1)} \tag{6.2}$$

mit Startwert $x^{(0)} = 0$ konvergiert gegen die Lösung x des Gleichungssystems $Ax = b$.

In Abschnitt 6.1.3 werden die Bedingungen genauer untersucht, die Konvergenz des Verfahrens gegen die Lösung garantieren können.

Beispiel. Sei das Gleichungssystem gegeben durch

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix} \qquad \text{und} \qquad b = \begin{pmatrix} 7 \\ 6 \\ 5 \end{pmatrix}. \tag{6.3}$$

Die Berechnung der Folge $x^{(m)}$ nach (6.2) liefert die Werte in Tabelle 6.1. Die Konvergenz scheint linear zu sein. ○

m	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$
0	0.0000000	0.0000000	0.0000000
1	3.5000000	0.8333333	0.1666667
2	3.0000000	0.9444444	0.2638889
3	2.8958333	0.9467593	0.2893519
4	2.8819444	0.9429012	0.2937886
5	2.8816552	0.9415187	0.2942065
6	2.8821373	0.9412187	0.2941610
7	2.8823102	0.9411762	0.2941284
8	2.8823476	0.9411747	0.2941194
9	2.8823531	0.9411759	0.2941177
10	2.8823531	0.9411764	0.2941176
∞	2.8823529	0.9411764	0.2941176

Tabelle 6.1: Lösung des Gleichungssystems mit Koeffizientenmatrix A und rechter Seite b aus (6.3) mit Hilfe des Gauss-Seidel-Algorithmus. In der letzten Zeile die exakten Resultate, erhalten mit dem Gauss-Algorithmus.

6.1.2 Matrixformulierung

Die Iterationsformel (6.2) verknüpft bei der Berechnung von $x^{(m)}$ Komponenten von $x^{(m-1)}$ und $x^{(m)}$, was es etwas schwieriger macht, die Iteration als Fixpunktiteration der Form $x^{(m)} = Fx^{(m-1)}$ zu schreiben mit einer $n \times n$-Matrix F. Um dies zu erreichen zerlegen wir die Matrix A in drei Summanden $A = L + D + U$, wobei L eine untere Dreiecksmatrix mit Nullen auf der Diagonalen sein soll, D eine Diagonalmatrix und U eine obere Dreiecksmatrix mit Nullen auf der Diagonalen, also

$$L = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & 0 & \dots & 0 & 0 \\ a_{31} & a_{3,2} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \dots & 0 & 0 \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{n,n-1} & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1,n-2} & a_{1,n-1} & a_{1n} \\ 0 & 0 & \dots & a_{1,n-2} & a_{2,n-1} & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & a_{n-2,n-1} & a_{n-2,n} \\ 0 & 0 & \dots & 0 & 0 & a_{n-1,n} \\ 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

$$D = \operatorname{diag}(a_{11}, a_{22}, \dots, a_{n-1,n-1}, a_{nn})$$

Die Iterationsformel (6.2) lässt sich mit diesen Matrizen schreiben als

$$Dx^{(m)} = b - Lx^{(m)} - Ux^{(m-1)}.$$

Auflösen nach $x^{(m)}$ führt auf

$$x^{(m)} = (D + L)^{-1}(b - Ux^{(m-1)}). \tag{6.4}$$

Die Form (6.4) für das Gauss-Seidel-Iterationsverfahren ist jetzt die einer Fixpunkt-Iteration.

6.1.3 Konvergenzbedingung

In Kapitel 1 haben wir gelernt, dass eine Fixpunktiteration konvergiert, wenn der Betrag der Ableitung < 1 ist. Hier liegt jedoch eine Matrix-Iteration mit der Abbildung

$$F(x) = \underbrace{(D + L)^{-1}b}_{= c} - (D + L)^{-1}Ux = c - (D + L)^{-1}Ux$$

vor. Die Ableitung ist daher ebenfalls eine Matrix, nämlich

$$D_x F = (D + L)^{-1} U,$$

und der Fehler der Iteration m ist

$$\delta_m = (D + L)^{-1} U \delta_{m-1}. \tag{6.5}$$

Konvergenz kann also nur vorliegen, wenn dieser Vektor im Laufe der Iteration immer kleiner wird. Dies ist zum Beispiel dann der Fall, wenn die *Norm* der Matrix kleiner als 1 ist:

Definition 6.2. *Die* Norm *einer Matrix M ist*

$$\|M\| = \max\{|Mx| \mid x \in \mathbb{R}^n \wedge |x| = 1\}.$$

Für einen Vektor $x \in \mathbb{R}^n$ gilt $|Mx| \leq \|M\| \cdot |x|$.

Die Bedingung (6.5) bedeutet jedoch nicht, dass die Norm der Ableitung < 1 sein muss, es genügt, wenn genügend hohe Potenzen der Ableitung eine Norm < 1 haben.

Beispiel. Die Matrix

$$M = \begin{pmatrix} 0 & 2 \\ \frac{1}{3} & 0 \end{pmatrix}$$

hat Norm

$$\|M\| = \max_{|x|=1} |Mx| = \max_{t \in \mathbb{R}} \sqrt{2^2 \cos^2 t + \frac{1}{3^2} \sin^2 t} \geq 2.$$

Da aber

$$M^2 = \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \qquad \Rightarrow \qquad \|M^2\| = \frac{2}{3}$$

ist, wird eine Iteration mit Ableitungsmatrix M trotzdem konvergieren, weil der Fehler nach jedem zweiten Schritt um den Faktor $\frac{2}{3}$ kleiner geworden ist. ○

Dies führt uns auf das Konzept des Spektralradius

Definition 6.3. *Der* Spektralradius *der Matrix M ist*

$$\varrho(M) = \limsup_{n \to \infty} \|A^n\|^{\frac{1}{n}}.$$

Das Gauss-Seidel-Iterationsverfahren ist also genau dann für alle Startwerte x_0 linear konvergent, wenn der Spektralradius

$$\varrho((L + D)^{-1} U) < 1$$

ist.

6.2 QR-Zerlegung mit Spiegelungen

Das Orthonormalisierungsproblem verlangt, dass zu gegebenen Vektoren $a_1, \ldots, a_n \in \mathbb{R}^l$ orthonormierte Vektoren $q_1, \ldots, q_n$ gefunden werden derart, dass

$$a_k = \langle q_1, \ldots, q_k \rangle \qquad \forall 1 \leq k \leq n. \tag{6.6}$$

Die Bedingung (6.6) bedeutet, dass es Zahlen r_{ik} mit $i \leq k$ derart, dass

$$a_k = q_1 r_{11} + q_2 r_{12} + \cdots + q_k r_{1k} \tag{6.7}$$

Schreiben wir die Komponenten des Vektors a_k als $l \times n$-Matrix A mit Einträgen a_{ik} und analog für die Vektoren q_k, dann kann (6.7) in Matrixform geschrieben werden als

$$A = QR, \qquad \text{mit} \quad R = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ 0 & r_{22} & \ldots & r_{2n} \\ 0 & 0 & \ldots & r_{3n} \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

Darin ist Q eine $l \times n$-Matrix und R ist eine $n \times n$-Matrix. Die Aufgabe, eine Menge von Vektoren $\{a_1, \ldots, a_n\}$ zu orthonormieren, ist also gleichbedeutend damit, für die Matrix A eine Zerlegung $A = QR$ zu finden, wobei Q orthogonal und R eine obere Dreiecksmatrix sein soll.

6.2.1 Gram-Schmidt-Orthonormalisierung

Das einfachste und anschaulichste Orthonormalisierungsverfahren, der Gram-Schmidt-Prozess verwendet die Formeln

$$\begin{aligned} q_1 &= \frac{a_1}{|a_1|} \\ q_2 &= \frac{a_2 - (q_1 \cdot a_2) q_1}{|a_2 - (q_1 \cdot a_2) q_1|} \\ q_3 &= \frac{a_3 - (q_1 \cdot a_3) q_1 - (q_2 \cdot a_3) q_2}{|a_3 - (q_1 \cdot a_3) q_1 - (q_2 \cdot a_3) q_2|} \\ &\vdots \\ q_k &= \frac{a_k - (q_1 \cdot a_k) q_1 - \cdots - (q_{k-1} \cdot a_k) q_{k-1}}{|a_k - (q_1 \cdot a_k) q_1 - \cdots - (q_{k-1} \cdot a_k) q_{k-1}|}, \end{aligned}$$

um die orthonormierten Vektoren q_k zu finden. Wegen der Differenzen im Zähler und Nenner besteht die Gefahr von Auslöschung, falls der Winkel zwischen a_k und $\langle a_1, \ldots, a_{k-1} \rangle$ sehr klein ist. Der Gram-Schmidt-Prozess ist daher in seiner Grundform nicht immer stabil.

6.2.2 Spiegelungen

Mit dem Skalarprodukt kann zu jedem Paar a, b von Vektoren gleicher Länge eine Matrix gefunden werden, welche eine Spiegelung beschreibt, die a auf b abbildet, alle Vektoren orthogonal zu a und b aber unverändert lässt.

Sei $n = (b - a)/|b - a|$ der Einheitsvektor in Richtung $b - a$. Es ist $(a - b) \cdot (a + b) = |a|^2 - |b|^2$, also auch $n \cdot (a + b) = 0$.

Die Abbildung

$$s : x \mapsto x - 2n(n \cdot x)$$

hält auf n orthogonale Vektoren x wegen $n \cdot x = 0$ fest. Für die Vektoren a und b ist $(b-a) \cdot a = -(b-a) \cdot b$, so dass

$$\begin{aligned} a &= \underbrace{\frac{1}{2}(a+b)}_{=u} + \underbrace{\frac{1}{2}(a-b)}_{=v} = u + v \\ b &= \frac{1}{2}(a+b) - \frac{1}{2}(a-b) = u - v \end{aligned}$$

gilt mit orthogonalen Vektoren $u = \frac{1}{2}(a+b)$ und $v = \frac{1}{2}(a-b)$. Für die beiden Summanden gilt

$$\begin{aligned} s(u) &= u, \\ s(v) &= -v \end{aligned}$$

und daher

$$\begin{aligned} s(a) &= s(u+v) = s(u) + s(v) = u - v = b, \\ s(b) &= s(u-v) = s(u) - s(v) = u + v = a. \end{aligned}$$

Die Abbildung s vertauscht also die beiden Vektoren, sie ist eine Spiegelung in der von a und b aufgespannten Ebene, an einer Geraden mit der Normalen n.

Der Vektor n ermöglicht auch, die lineare Abbildung s mit Hilfe einer Matrix s zu schreiben. Das Skalarprodukt $n \cdot x$ ist das Matrixprodukt $n^t x$, der Vektor $n(n \cdot x)$ als Matrixprodukt $nn^t x$ berechnet werden. Damit wird

$$s(x) = x - 2n(n \cdot x) = Ex - 2(nn^t)x = (E - 2nn^t)x \qquad \Rightarrow \qquad S = E - 2nn^t$$

die Matrix der Abbildung s. Wir haben damit den folgenden Satz bewiesen.

Satz 6.4. *Zu gegebenen Vektoren a und b gleicher, nicht verschwindender Länge gibt es eine orthogonale Matrix*

$$S_{a,b} = E - 2nn^t \qquad \textit{mit} \quad n = (b-a)/|b-a|,$$

die zu einer Spiegelung gehört, die a auf b abbildet und umgekehrt und alle Vektoren senkrecht auf a und b fest lässt.

6.2.3 QR-Zerlegung mit Spiegelungen

Aus der QR-Zerlegung $A = QR$ erhält man durch Multiplikation mit Q^t die Gleichung $Q^t A = R$. Die Aufgabe, die QR-Zerlegung zu finden, ist also gleichbedeutend damit, eine Matrix durch Multiplizieren mit orthogonalen Matrizen auf Dreiecksform zu bringen. Das Produkt der orthogonalen Matrizen ist Q^t. In diesem Abschnitt wird ein Algorithmus basierend auf den Matrizen $S_{a,b}$ vorgestellt, der genau dies leistet.

Sei $a \in \mathbb{R}^l$ ein nicht verschwindender Vektor. Dann ist der Vektor $b = (|a|, 0, \ldots, 0)^t$ ein Vektor gleicher Länge. Nach Satz 6.4 ist die Matrix $S_{a,b}$ eine Spiegelung, die a auf b abbildet.

Sei jetzt ein Vektor $a \in \mathbb{R}^l$ gegeben und eine Zahl $k < l$. Wir zerlegen den Vektor a in zwei Teile

$$a = \begin{pmatrix} \tilde{a} \\ \bar{a} \end{pmatrix} \quad \text{mit} \quad \tilde{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} \bar{a} = \begin{pmatrix} a_k \\ \vdots \\ a_l \end{pmatrix}.$$

Die Matrix $Q_{\bar{a}}$ bildet $\bar{a}$ auf einen Vektor ab, von dem nur die erste Komponente von 0 verschieden ist. Daraus wird die Matrix

$$Q_a = \begin{pmatrix} E & 0 \\ 0 & Q_{\bar{a}} \end{pmatrix}, \tag{6.8}$$

die den Vektor

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ a_k \\ \vdots \\ a_n \end{pmatrix} \qquad \text{auf} \qquad Q_a a = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ |\bar{a}| \\ \vdots \\ 0 \end{pmatrix}$$

abbildet. Q_a lässt Vektoren x, für die $\bar{x} = 0$ ist, unverändert.

Daraus lässt sich jetzt ein QR-Algorithmus konstruieren.

Satz 6.5 (QR-Zerlegung mit Spiegelungen)**.** *Gegeben ist die $l \times n$-Matrix A mit $n \leq l$. Dann gibt es orthogonale Matrizen $Q_1, \ldots, Q_n$ derart, dass in jeder Matrix der Folge*

$$A_0 = A, \quad A_i = Q_i A_{i-1} \quad \textit{für } 0 < i \leq n$$

die ersten i Spalten die Form einer oberen Dreicksmatrix haben. Mit $Q = Q_1^t \ldots Q_n^t$ und $R = A_n = Q_n \ldots Q_1 A$ gilt $A = QR$.

Beweis. Wir müssen die Matrix Q_i aus A_{i-1} konstruieren. Sei a die i-te Matrix von A_{i-1}, dann setzen wir $Q_i = Q_a$ mit der Matrix Q_a nach (6.8). Die Multiplikation mit Q_i bringt die Spalte i der Matrix A_{i-1} in die für A_i verlangte Form, ohne die Spalten mit Spaltenindex $< i$ zu verändern, da deren Komponenten dort, wo Q_a etwas bewirkt, alle verschwinden. □

Beispiel. Die Vektoren

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad a_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \qquad a_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

sollen orthonormalisiert werden. Die Matrix Q_1 spiegelt den Vektor

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{auf} \quad \begin{pmatrix} \sqrt{3} \\ 0 \\ 0 \end{pmatrix}$$

ab. Die zugehörige Matrix ist

$$Q_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{2} - \frac{1}{2\sqrt{3}} & -\frac{1}{2} - \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{2} - \frac{1}{2\sqrt{3}} & \frac{1}{2} - \frac{1}{2\sqrt{3}} \end{pmatrix} \qquad A_1 = Q_1 A = \begin{pmatrix} \sqrt{3} & \frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{1}{\sqrt{3}} & -\frac{1}{2} - \frac{1}{2\sqrt{3}} \\ 0 & -\frac{1}{\sqrt{3}} & \frac{1}{2} - \frac{1}{2\sqrt{3}} \end{pmatrix}.$$

Die zweite Matrix Q_2 ist

$$Q_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad R = A_2 = Q_2 A_1 = \begin{pmatrix} \sqrt{3} & \frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}$$

Wegen $R = (Q_2 Q_1)A$ folgt $A = (Q_2 Q_1)^t R$ und damit

$$Q = (Q_2 Q_1)^t = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\sqrt{\frac{2}{3}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Damit ist die QR-Zerlegung der Matrix A gefinden. Tatsächlich bilden die Spalten von Q eine orthonormierte Basis von $\mathbb{R}^3$, sie stimmt mit der Basis überein, die der Gram-Schmidt-Prozess aus den Spalten der Matrix A ableitet. ○

Der Nachteil dieser Methode ist, dass die Matrizen $S_{a,b}$ typischerweise sehr viele Einträge haben und dass damit die Matrixprodukte mit $S_{a,b}$ aufwendig zu berechnen sind.

6.3 Diagonalisierung mit dem Jacobi-Verfahren

Die Diskretisierung linearer partieller Differentialgleichungen wie der Wellengleichung führen immer auf symmetrische Eigenwertprobleme, also auf Gleichungen der From $Av = \lambda v$ mit $A = A^t$. Aus der linearen Algebra ist bekannt, dass es in diesem Fall eine orthogonale Matrix O gibt mit

$$\begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} = O^t A O$$

Die Matrix $O^t A O$ hat also die Basisvektoren $e_i = (0, \dots, 1, \dots, 0)$ als Eigenvektoren. O bildet e_i auf den i-ten Eigenvektor von A ab, in der i-ten Spalten von O steht der i-te Eigenvektor von A. Findet man O, kann man daraus die Eigenvektoren ablesen.

Das Jacobi-Verfahren versucht, O als Zusammensetzung von Drehungen in jeweils zwei Dimensionen aufzubauen. Gleichzeitig wird die Matrix A "in place" auf Diagonalform reduziert, so dass man dort die Eigenwerte ablesen kann.

6.3.1 Jacobi-Verfahren in zwei Dimensionen

In zwei Dimensionen hat eine orthogonale Matrix O immer die Form

$$O = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

Die symmetrische Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

soll damit auf Diagonalform gebracht werden. In

$$\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}\begin{pmatrix}\cos\alpha & \sin\alpha\\ -\sin\alpha & \cos\alpha\end{pmatrix}$$

müssen die Elemente ausserhalb der Diagonalen zu 0 werden, dann stehen auf der Diagonalen die gesuchten Eigenwerte. Durch Nachrechnen findet man für α die Bedingung

$$\begin{aligned}0 &= a_{11}\sin\alpha\cos\alpha + a_{12}(\cos^2\alpha - \sin^2\alpha) - a_{22}\sin\alpha\cos\alpha\\ &= (a_{11}-a_{22})\frac{1}{2}\sin 2\alpha + a_{12}\cos 2\alpha\\ \cot 2\alpha &= \frac{a_{22}-a_{11}}{2a_{12}}.\end{aligned}$$

Mit Hilfe der goniometrischen Beziehung

$$\vartheta = \cot 2\alpha = \frac{1-\tan^2\alpha}{2\tan\alpha}$$

kann man sie als quadratische Gleichung für $\tan\alpha$ betrachten, nämlich

$$\tan^2\alpha + 2\vartheta\tan\alpha - 1 = 0,$$

welche die Lösungen

$$\tan\alpha = \vartheta \pm \sqrt{\vartheta^2+1}$$

hat. In der Matrix O werden nur Sinus und Cosinus benötigt, diese kann man durch algebraische Ausdrücke berechnen:

$$\cos\alpha = \frac{1}{\sqrt{1+\tan^2\alpha}} \tag{6.9}$$

$$\sin\alpha = \frac{\tan\alpha}{\sqrt{1+\tan^2\alpha}} \tag{6.10}$$

Insbesondere sind keine numerische aufwendigen trigonometrischen Operationen notwendig, ausser der Wurzel in (6.9) und (6.10) sind alle Schritte mit den Grundoperationen durchführbar.

6.3.2 Beliebige Dimension

Für $n > 2$ lässt sich die Reduktion auf Diagonalform nicht mehr in einem Schritt durchführen. Der folgende Algorithmus führt jedoch zum Erfolg.

1. Initialisiere die Matrix O als Einheitsmatrix: $O = I$
2. Für jedes Paar von Indizes (p, q) mit $q > p$ führe die folgenden zwei Schritte aus.
3. Finde eine Matrix

   $$O_{pq}\begin{pmatrix}\cos\alpha & \sin\alpha\\ -\sin\alpha & \cos\alpha\end{pmatrix}$$

   wie in Abschnitt 6.3.1, welche die Teilmatrix

   $$A_{pq} = \begin{pmatrix}a_{pp} & a_{pq}\\ a_{qp} & a_{qq}\end{pmatrix}$$

   auf Diagonalform bringt: $O_{pq}^t A_{pq} O_{pq}$.

4. Bilde die Matrix

$$O' = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & & 0 \\ & \ddots & \vdots & & \vdots & & \vdots \\ 0 & & \cos\alpha & \dots & \sin\alpha & \dots & 0 \\ & & \vdots & \ddots & \vdots & & \vdots \\ 0 & & -\sin\alpha & \dots & \cos\alpha & \dots & 0 \\ & & \vdots & & \vdots & \ddots & \\ 0 & \dots & 0 & & 0 & & 1 \end{pmatrix}.$$

5. Setzte $A := O'^t A O'$ und $O := OO'$.

6. Wiederhole das Verfahren ab Schritt 2, falls noch Indexpaare (p, q) mit $|a_{pq}| > \varepsilon$ vorkommen.

Am Ende dieses Verfahrens steht in A die Diagonalmatrix mit den Eigenwerten, in O steht die orthogonale Matrix, die A auf Diagonalform gebracht, sie enthält die Eigenvektoren in den Spalten.

Kapitel 7

Partielle Differentialgleichungen

Kapitel 8

Periodische Funktionen

Literatur

[1] CloudyPadmal. *Makeing Shapes with PSLab Oscilloscope*. März 2018. URL: https://blog.fossasia.org/making-shapes-with-pslab-oscilloscope/.

[2] *Kahan summation algorithm*. 29. Feb. 2020. URL: https://en.wikipedia.org/wiki/Kahan_summation_algorithm.

[3] Andreas Müller. *Source Code Repository*. 2020. URL: https://github.com/AndreasFMueller/SeminarNumerik.git.

[4] David S. Watkins. *Fundamentals of Matrix Computations*. 3. Aufl. John Wiley und Sons, Inc., 2010.

Teil II

Anwendungen und weiterführende Themen

Übersicht

Im zweiten Teil kommen die Teilnehmer des Seminars selbst zu Wort. Die im ersten Teil dargelegten mathematischen Methoden und grundlegenden Modelle werden dabei verfeinert, verallgemeinert und auch numerisch überprüft.

Kapitel 9

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

9.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

9.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{9.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

9.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 9.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

9.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

9.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

9.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

9.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 10

Van der Pol-Differentialgleichung

Manuel Cattaneo und Niccolo Galliani

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

10.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

10.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{10.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

10.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 10.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

10.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

10.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

10.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

10.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 11

Iteration der logistischen Gleichung

Michael Schneeberger

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

11.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

11.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{11.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

11.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 11.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

11.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

11.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

11.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

11.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 12

Kettenbrüche

Benjamin Bouhafs-Keller

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

12.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

12.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{12.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

12.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 12.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

12.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

12.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

12.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

12.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 13

Taylor-Reihe und Differentialgleichungen

Fabio Marti

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

13.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

13.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{13.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 13.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

13.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga.

Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

13.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 14

Finite Elemente in der Ebene

Joël Rechsteiner

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

14.1 Einleitung

Das Kochrezept für das Vorgehen bei der Finiten Element Methode in der Ebene ist analog zur FEM im 1 Dimensionalen Raum. Jedoch müssen zusätzliche Parameter beachtet werden, die vorher einfach gegeben oder geschenkt wurden. Der weitere Schritt zur FEM im 3 Dimensionalen Raum ist fast identisch, wird jedoch in diesem "Buch/Kapitel"nicht näher erläutert. Zu erwähnen ist noch, dass das Kochrezept nicht einfach anwenden heisst von Formeln im Falle der Finiten Elemente. Es müssen mehrere Überlegungen gemacht werden, die jedoch der gleichen Idee folgen.

Das Konzept oder die Idee der Finiten Elemente ist wie folgend:

1. Differentialgleichungssystem aufstellen und in eine abgeschwächte Form bringen wie z.B. DGL in ein äqualentes minimalproblem umsetzen.
2. Ansatzfunktionen für Lösung finden bzw. die Lösungsfunktion approximieren

3. Minimalprinzip auf die Approximation anwenden durch Multiplikation mit den Wichtungsfunktionen
4. Lineare Gleichungen für die Koeffizienten aufstellen durch Integration <-evt. nur Gallerkin
5. Das Gleichungssystem lösen und dadurch die Koeffizienten der Approximation bestimmen

14.2 Problemstellung

Anforderung an $h(x)$:

- 1. stetig Differenzierbar (Steigung)
- 2. stetig Differenzierbar (Krümmung)
- an Knoten muss $h(x) = 0$ sein

Poisson-Gleichung:

$$\frac{\partial^2 u(x,y)}{\partial x^2} \frac{\partial^2 u(x,y)}{\partial y^2} = -u(x,y) \in \Omega \tag{14.1}$$

Randbedingungen:

$$u = 0, (x,y) \in \Omega \tag{14.2}$$

$$u = U(x,y) \in \Omega \tag{14.3}$$

$$\frac{\partial u}{n} = 0, (x,y) \in \Omega \tag{14.4}$$

$\Rightarrow$ 4 Parameter $\Rightarrow$ Polynom 3. Grades

$$\iint_{\Omega} (u_2^2 + u_y^2)) \, dxdy \tag{14.5}$$

$$\iint_{\Omega} u^2 \, dxdy \tag{14.6}$$

$$\iint_{\Omega} u \, dxdy \tag{14.7}$$

Und daraus müsste dann die Lösung gefunden werden um die Koeffizienten der folgenden Gleichung zu finden.

$$\int g(x)dx = f_i \int h_0 dx + f_{i+1} \int h_1 dx + s_i \int h_0^1 dx + s_{i+1} \int h_1^1 dx \tag{14.8}$$

14.2.1 De finibus bonorum et malorum

$$\int_a^b x^2 \, dx = \left[\frac{1}{3} x^3 \right]_a^b = \frac{b^3 - a^3}{3}. \tag{14.9}$$

14.3 Lösung

Zuerst muss die DGL in ein äqualentes Minimalprobelm übersetzt werden in der Form

$$\int_{\Omega} (\nabla u)^2 + \lambda u^2 dx \tag{14.10}$$

Approximation der unbekannten Funktion mit

$$u(x,y) = \sum_{k=0} N_k^e(x,y) \cdot u(x,y)_k^e \tag{14.11}$$

N = Formfunktion des Elementes e und Knoten k
Zu lösendes GL- System

$$Ax = b \tag{14.12}$$

wobei A eine Matrix, x der Vektor der Unbekannten und b der Ansatzfunktionenvektor ist.

14.3.1 linearer Ansatzfunktion

$$u(x,y) = c_1 + c_2 x + c_3 y \tag{14.13}$$

Gemäss Abschnitt 2.2.2 Buch von Schwarz sind die Werte entlang einer Seite eines Dreiecks-Elements gleich der angrenzenden Linie eines anderen Dreiecks, wenn die Eckpunktwerte gleich sind. Die Frage Warum soll dass so sein? Weil Funktion $h(x)$ bei jedem Element gleich ist sowie es sich bei $h(x)$ hierbei um eine Lineare Funktion handelt.

14.3.2 Quadratischer Ansatz

Abschnitt

14.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

14.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est

eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Kapitel 15

Die Gleichung von Burgers

Michael Schmid

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

15.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

15.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{15.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 15.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

15.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

15.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 16

Padé-Approximation

Cédric Renda

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

16.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

16.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{16.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 16.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

16.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

16.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 17

QR-Zerlegung mit Givens-Rotationen

Manuel Tischhauser

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

17.1 Einleitung

Wie in Abschnitt 6.2 schon genauer betrachtet, kann eine Matrix A mit l Zeilen und n linear unabhängigen spalten ($l \geq n$) in zwei Matrizen zerlegt werden:

$$A = QR.$$

Q ist dabei wieder ein $l \times n$-Matrix mit, diesmal aber orthonormierten Spalten. R ist eine ober Dreiecksmatrix der Grösse $n \times n$.

17.1.1 Anwedungsbeispiel Least-Squares

Die *QR*-Zerlegung kann unter anderem beim lösen von Gleichungssystemen angewendet werden. Ein in den Parametern lineares Gleichungssystem

$$Ax = b \tag{17.1}$$

hat in praktischen Fällen keine Lösung, wenn die $l \times n$ Matrix mehr Zeilen als Spalten hat, bzw. überdefiniert ist. Mit der Methode der kleinsten Quadrate kann aber ein Vektor $\hat{x}$ gefunden werden welcher die L^2-Norm

$$||A\hat{x} - b||_2$$

minimiert. Dieses Problem kann geometrisch gelöst werden, indem man es wie folgt betrachtet:

$$A\hat{x} = b - b_\perp,$$

wobei $b_\perp$ die zu den Spalten in A orthogonale Komponente von b ist. Multipliziert am beide Seiten mit A^T erhält man

$$A^TA\hat{x} = A^Tb - \underbrace{A^Tb_\perp}_{=0} \quad \Leftrightarrow \quad A^TA\hat{x} = A^Tb \tag{17.2}$$

Wobei der letzte Term verschwindet, da alle Spalten von A und $b_\perp$ orthogonal zueinander stehen und somit alle Skalarprodukte verschwinden. Das Matrixprodukt A^TA ist unter Umständen aufwändig zu berechnen. Geht man nun aber den Umweg über die *QR*-Zerlegung kann die Gleichung 17.2 umgeschrieben werden:

$$(QR)^TQR\hat{x} = (QR)^Tb$$

was sich zu

$$R\hat{x} = Q^Tb \tag{17.3}$$

vereinfacht.

17.2 Problemstellung

In Abschnitt 6.2.1 wird das Gram-Schmidt-Orthonormalisierungsverfahren vorgestellt. Es liefert zwar eine sehr anschauliche Methode zur *QR*-Zerlegung, ist aber numerisch nicht stabil wenn die die Spaltenvektoren von A nahezu parallel sind.

Wie in [1] schon als Beispiel benutzt, soll dies nun Anhand der Matrix

$$A = \begin{pmatrix} 1+\epsilon & 1 & \\ 1 & 1+\epsilon & 1 \\ 1 & 1 & 1+\epsilon \end{pmatrix}$$

genauer untersucht werden. ϵ ist dabei eine so „kleine“ Zahl, sodass ϵ^2 im Rechner nicht mehr darstellbar ist bzw. $\epsilon^2 \approx 0$. Nach Gram-Schmidt wird die erste Komponente von Q berechnet als

$$q_1 = \frac{a_1}{|a_1|} = \frac{1}{\sqrt{3+2\epsilon+\epsilon^2}} \begin{pmatrix} 1+\epsilon \\ 1 \\ 1 \end{pmatrix} \approx \frac{1}{\sqrt{3+2\epsilon}} \begin{pmatrix} 1+\epsilon \\ 1 \\ 1 \end{pmatrix}.$$

Schon hier kommt es zu einem kleinen Rundungsfehler.. Die zweite Spalte von Q berechnet man als

$$q_2 = \frac{a_2 - (q_1^T a_2)q_1}{|a_2 - (q_1^T a_2)q_1|} = \frac{z_1}{|z_1|}.$$

Dies wird im Rechner in zwei Schritten ausgeführt: Zuerst wird z ausgerechnet und dann nach dessen Betrag normiert um auf q_2 zu kommen. Im ersten Schritt also:

$$z_1 = \begin{pmatrix} 1 \\ 1+\epsilon \\ 1 \end{pmatrix} - \underbrace{\frac{1}{\sqrt{3+2\epsilon}}(3+2\epsilon)\frac{1}{\sqrt{3+2\epsilon}}}_{=1} \begin{pmatrix} 1+\epsilon \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\epsilon \\ \epsilon \\ 0 \end{pmatrix}$$

und

$$|z_1| = \sqrt{(-\epsilon)^2 + \epsilon^2 + 0^2} = \sqrt{2\epsilon^2} = \sqrt{2}\epsilon.$$

Im zweiten Schritt ergibt sich somit

$$q_2 = \frac{z_1}{|z_1|} = \frac{1}{\sqrt{2}\epsilon}\begin{pmatrix} -\epsilon \\ \epsilon \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

Die dritte Spalte von Q wird berechnet als

$$q_3 = \frac{a_3 - (q_1^T a_3)q_1 - (q_2^T a_3)q_2}{|a_3 - (q_1^T a_3)q_1 - (q_2^T a_3)q_2|} = \frac{z_2}{|z_2|}$$

Dies ergibt

$$z_2 = \begin{pmatrix} 1 \\ 1 \\ 1+\epsilon \end{pmatrix} - - \underbrace{\frac{1}{\sqrt{3+2\epsilon}}(3+2\epsilon)\frac{1}{\sqrt{3+2\epsilon}}}_{=1} \begin{pmatrix} 1+\epsilon \\ 1 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{2}}\underbrace{(-1+1+0)}_{=0}\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \epsilon \\ 0 \\ -\epsilon \end{pmatrix}$$

und

$$|z_2| = \sqrt{\epsilon^2 + 0^2 + (-\epsilon)^2} = \sqrt{2}\epsilon.$$

Wiederum im zweiten Schritt ergibt sich

$$q_3 = \frac{z_2}{|z_2|} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Nach diesem Vorgehen kommt man also auf

$$Q = \begin{pmatrix} q_1 & q_2 & q_3 \end{pmatrix} = \begin{pmatrix} \frac{1+\epsilon}{\sqrt{3+2\epsilon}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3+2\epsilon}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3+2\epsilon}} & 0 & -\frac{1}{\sqrt{2}} \end{pmatrix} \tag{17.4}$$

Q ist nach Definition orthonormiert. Die Spalten müssen also ein Betrag von jeweils 1 haben. Dies stimmt, wenn man ϵ in 17.4 vernachlässigt. Betrachtet man wieder unter Vernachlässigung von ϵ die Winkel zwischen den Spalten (angegeben im jeweiligen Index), kommt man auf

$$\theta_{12} = \frac{\pi}{2}, \quad \theta_{13} = \frac{\pi}{2}, \quad \theta_{23} = \frac{2\pi}{3}.$$

Q ist also nicht orthogonal und somit ist auch $Q^{-1} \neq Q^T$.

17.3 Lösung

todo:
Givens-Rotation

17.4 Folgerungen

todo:
Vergleich der beiden Implementationen.
Plot mit von Winkelabweichung gegenüber $\pi/2$ („Orthogonalitätsfehler “)mit verschieden ϵ als Qualitätskontrolle der Zerlegung

Literatur

[1] Tin-Yau Tam. *QR decompostion: History and its Applications*. 10. Dez. 2010.

Kapitel 18

Numerische Laplace-Inversion

Severin Weiss

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

18.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

18.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{18.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 18.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

18.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

18.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 19

Störungstheorie

Daniel Bucher und Thomas Kistler

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

19.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

19.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{19.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

19.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 19.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

19.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

19.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

19.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

19.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 20

Gauss-Quadratur

Mike Schmid

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

20.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

20.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{20.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

20.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 20.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

20.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

20.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

20.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

20.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 21

Schrittlängensteuerung

Reto Fritsche

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

21.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

21.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{21.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 21.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

21.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

21.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 22

Francis-Algorithmus

Tobias Grab

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

22.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

22.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{22.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

22.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 22.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

22.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

22.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

22.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

22.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 23

Stabile Berechnung von Legendre Polynomen

Patrick Elsener

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

23.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

23.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{23.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

23.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 23.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

23.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

23.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga.

Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

23.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

23.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 24

Die Methode der konjugierten Gradienten

Raphael Unterer

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

24.1 Herleitung des Algorithmus

Erste, grobe Herleitung des Algorithmus, ohne viel Text...

Der Algorithmus versucht ein (grosses) lineares Gleichungssystem der Form

$$Ax = b \quad x \in \mathbb{R}^N, \tag{24.1}$$

zu lösen. Wir definieren folgende Voraussetzungen für die Matrix A:

- A ist positiv definit, d.h. $x^T Ax > 0$
- A ist Symmetrisch

Die Stärke des CG-Verfahren ist es, in N schritten zur exakten Lösung zu finden. Das kann man sich gut vorstellen als steepest descent eines Problems, welches schön auf die Koordinatenachsen ausgerichtet ist.

24.1.1 Minimierungsproblem

Eine Lösung für x kann durch das Minimierungsproblem

$$\min_x \Phi(x) = \frac{1}{2} x^T A x - x^T b \tag{24.2}$$

gefunden werden. Der folgende Beweis zeigt, wieso dies ein sinnvoller Ansatz ist.

Wir definieren eine zweite Variable $z = x + \lambda y$, was uns erlaubt die folgende Differenz auszurechnen

$$\Phi(z) - \Phi(x) = \frac{1}{2} (x + \lambda y)^T A (x + \lambda y) - (x + \lambda y)^T b - \frac{1}{2} x^T A x + x^T b \tag{24.3}$$

$$= \frac{1}{2} \left(x^T A x + x^T A \lambda y + \lambda y^T A x + \lambda y^T A \lambda y \right) - x^T b - \lambda y^T b - \frac{1}{2} x^T A x + x^T b. \tag{24.4}$$

Da A symmetrisch ist, können die Terme $x^T A \lambda y$ und $\lambda y^T A x$ zusammengefasst werden (analog zur binomischen Formel)

$$\Phi(z) - \Phi(x) = \frac{1}{2} \cancel{x^T A x} + x^T A \lambda y + \frac{1}{2} \lambda y^T A \lambda y - \cancel{x^T b} - \lambda y^T b - \frac{1}{2} \cancel{x^T A x} + \cancel{x^T b} \tag{24.5}$$

$$= \lambda x^T A y + \frac{1}{2} \lambda^2 y^T A y - \lambda y^T b \tag{24.6}$$

$$= \frac{\lambda^2}{2} y^T A y + \lambda y^T (A x - b). \tag{24.7}$$

Nun können wir den Beweis führen, indem wir $\Phi(z) \geq \Phi(x)$ setzen (da $\Phi(x)$ ja minimiert wird)

$$\Phi(z) \geq \Phi(x) \tag{24.8}$$

$$\Phi(z) = \frac{\lambda^2}{2} y^T A y + \lambda y^T (A x - b) + \Phi(x) \tag{24.9}$$

$$0 \leq \frac{\lambda^2}{2} y^T A y + \lambda y^T (A x - b) \quad \forall \quad y \in \mathbb{R}^N. \tag{24.10}$$

Der erste Term ist dabei quadratisch in λ, A ist positiv definit und somit ist immer $\frac{\lambda^2}{2} y^T A y \geq 0$. Beim zweiten Term ist diese Bedingung nur erfüllt für alle y, wenn $Ax - b = 0$. Damit Bewiesen, dass eine Lösung für die Gleichung $Ax = b$ durch Minimierung von $\Phi(x)$ gefunden wird. □

24.1.2 Optimale Schrittweite

Gegeben:

- Aktueller Index k
- Suchrichtung d_k
- Startpunkt x_k

Wir suchen nun die optimale Schrittweite α, um möglichst nahe an die Lösung zu kommen in der gegebenen Suchrichtung. Dazu stellen wir wieder das Minimierungsproblem auf

$$\min_\alpha \Phi(x_k + \alpha d_k) = \frac{1}{2} x_k^T A x_k + \alpha x_k^T A d_k + \frac{1}{2} \alpha^2 d_k^T A d_k - x_k^T b - \alpha d_k^T b. \tag{24.11}$$

In α haben wir hier eine quadratische Gleichung, welche einer nach oben geöffneten Parabel entspricht (da $d_k^T A d_k \geq 0$). Somit ist es möglich ein klares Minimum zu finden, indem wir die Gleichung nach α ableiten und null setzen

$$\frac{\partial \Phi(x_k + \alpha d_k)}{\partial \alpha} = x_k^T A d_k + \alpha d_k^T A d_k - d_k^T b = 0. \tag{24.12}$$

Dies ergibt für α

$$\alpha = \frac{d_k^T b - x_k^T A d_k}{d_k^T A d_k} = \frac{d_k^T (b - A x_k)}{d_k^T A d_k}. \tag{24.13}$$

Wenn wir nun den Fehler der momentanen Approximation als Residuum $r_k = b - A x_k$ bezeichnen erhalten wir

$$\alpha = \frac{\langle d_k, r_k \rangle}{\langle d_k, d_k \rangle_A}, \tag{24.14}$$

wobei $\langle d_k, d_k \rangle_A = d_k^T A d_k$ das verallgemeinerte Skalarprodukt zu A darstellt. Somit haben wir nun die optimale Schrittlänge gefunden.

24.1.3 Optimale Suchrichtung

Nun muss nur noch die optimale nächste Suchrichtung d_{k+1} gefunden werden. Diese ist Orthogonal im Sinne von A auf d_k in Richtung des Residuums r_k und kann mithilfe des Gram-Schmidt-Orthogonalisierungsverfahren gefunden werden.

$$d_{k+1} = r_{k+1} - \frac{\langle d_k, r_{k+1} \rangle_A}{\langle d_k, d_k \rangle_A} d_k \tag{24.15}$$

24.2 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

24.3 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2 \, dx = \left[\frac{1}{3} x^3 \right]_a^b = \frac{b^3 - a^3}{3}. \tag{24.16}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

24.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 24.4. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

24.4 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

24.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

24.5 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

24.5.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 25

Störungstheorie für das Eigenwertproblem

Nicolas Tobler

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

25.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

25.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{25.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

25.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 25.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

25.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

25.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga.

Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

25.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

25.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: `https://de.wikipedia.org/wiki/BibTeX`.

Kapitel 26

Numerische Ableitung

Martin Stypinski

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

26.1 Einleitung

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

26.2 Problemstellung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{26.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

26.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio 26.3. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

26.3 Lösung

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

26.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

26.4 Folgerungen

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

26.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Index