Numerische und stochastische Grundlagen der Informatik

Peter Bastian
Universität Stuttgart, Institut für Parallele und Verteilte Systeme
Universitätsstraße 38, D-70569 Stuttgart
email: Peter.Bastian@ipvs.uni-stuttgart.de

Überarbeitet von *Thomas Ertl* und *Martin Falk*
Universität Stuttgart, Institut für Visualisierung und interaktive Systeme
Universitätsstraße 38, D-70569 Stuttgart
email: Thomas.Ertl@vis.uni-stuttgart.de

28. März 2009
$Id: numstoch-main.tex 1007 2009-02-05 14:38:48Z falkmn $

Inhaltsverzeichnis

Vorwort

Ziel dieser Vorlesung für Informatiker und Softwaretechniker im 3. Semester ist es eine Einführung in grundlegende Begriffe und Methoden der Numerik und der Stochastik zu geben. Besonderer Wert wird auch auf eine Begründung der Methoden gelegt, da nur so deren Grenzen erkannt werden können.

Erstmals steht im Wintersemester 2007/2008 ein Skript zur Vorlesung und ein Foliensatz zur Verfügung. Für die Erfassung des Textes in LaTeX danke ich Herrn Pascal Jäger recht herzlich. Alle verbleibenden Fehler gehen natürlich auf mein Konto.

Stuttgart, im Oktober 2007 *Peter Bastian*

1 Warum Numerik und Stochastik?

1.1 Modellbildung und Simulation

Die Wissenschaftliche Methode besteht aus den beiden Säulen Experiment und Theorie: Aus der Theorie werden Schlussfolgerungen gezogen und mit dem Experiment verglichen.

Die Theorie besteht in den „exakten“ Wissenschaften meist aus mathematischen Gleichungen (z. B. Differentialgleichungen).

Theorie und Experiment werden sukzessive verfeinert und verglichen, bis eine akzeptable Übereinstimmung vorliegt.

Man unterscheidet deterministische und stochastische Modelle:

- Deterministisch: Modell beschreibt eine Größe (z. B. Temperatur) in Abhängigkeit anderer Größen (z. B. Raum, Zeit) in eindeutiger Weise.
- Stochastisch: Modell beschreibt „Wahrscheinlichkeiten“ in Abhängigkeit von Parametern.

Oft können die Modellgleichungen nicht geschlossen (mit Papier und Bleistift oder Mathematica ...) gelöst werden. Dann führt man eine numerische Simulation durch.

Die Simulation (auch Wissenschaftliches Rechnen genannt) etabliert sich immer mehr als dritte Säule neben Theorie und Experiment. Vorteile sind:

- Undurchführbare Experimente werden möglich (z. B. Galaxienkollisionen).
- Teuere Experimente werden eingespart (z. B. Modelle im Windkanal).
- (Automatische) Optimierung von Prozessen.

Daher vielfältiger Einsatz auch in Industrie und Technik (etwa bei Strömungsberechnung, Festigkeit von Bauwerken).

Grundlage für alle diese Anwendungen sind numerische Algorithmen!

Diese Vorlesung ist auch wichtige Voraussetzung für die Visualisierung, Rechnerarchitektur, Grafische Ingenieursysteme, ...

Die prinzipielle Herangehensweise im Wissenschaftlichen Rechnen zeigt Abbildung 1. Die erfolgreiche Durchführung einer Simulation erfordert die interdisziplinäre Zusammenarbeit von Physikern oder Ingenieuren mit Mathematikern und Informatikern. Die Informatik leistet hier ihren Beitrag vor allem bei der Softwareentwicklung (auch Simulationsprogramme können sehr komplex sein), der Visualisierung und im (parallelen) Höchstleistungsrechnen.

In der Regel gibt es Unterschiede zwischen den simulierten und experimentell bestimmten Größen. Diese Unterschiede können verschiedene Gründe haben:

- Modellfehler: Ein relevanter Prozess wurde nicht oder ungenau modelliert (Temp. konstant, Luftwiderstand vernachlässigt, ...)
- Datenfehler: Messungen von Anfangsbedingungen, Randbedingungen, Werten für Parameter sind fehlerbehaftet.

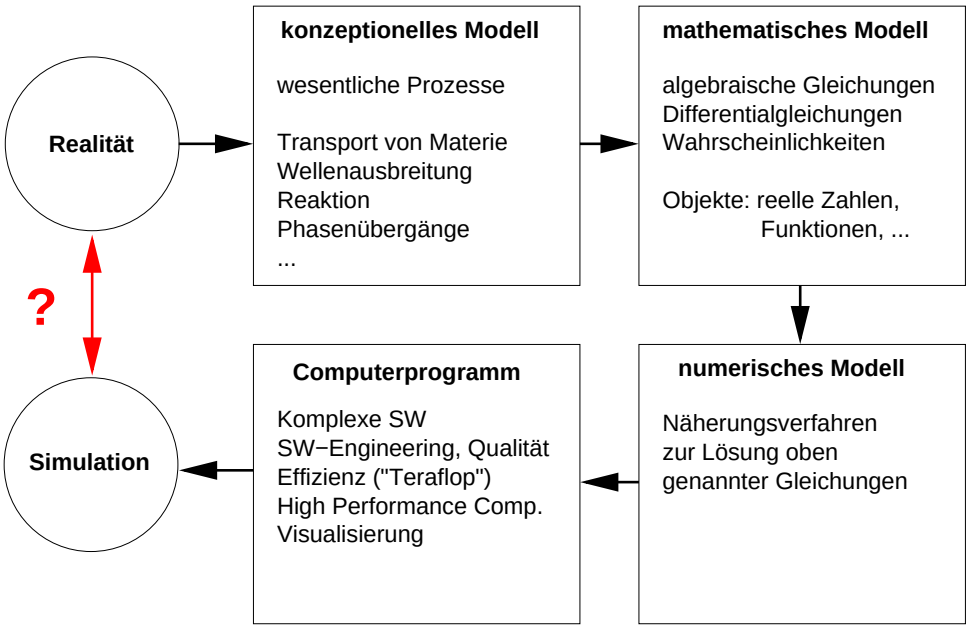


Abbildung 1: Prinzipielles Vorgehen im Wissenschaftlichen Rechnen.

- Rundungsfehler: Reelle Zahlen werden im Rechner genähert dargestellt.
- Diskretisierungsfehler: Funktionen müssen approximiert werden, z. B. durch (stückweise) Polynome, endliche Fourierreihe.
- Abbruchfehler: Reihenentwicklungen, Iterationen müssen irgendwann abgebrochen werden.

Sensibilisierung gegenüber diesen Fehlerquellen ist ein Hauptanliegen der Vorlesung!

1.2 Ein einfaches Beispiel: Das Fadenpendel

Pisa, 1582. Der Student Galileo Galilei sitzt in der Kirche und ihm ist langweilig. Er beobachtet den langsam über ihm pendelnden Kerzenleuchter und denkt: „Wie kann ich nur die Bewegung dieses Leuchters beschreiben?“.

Abbildung 2 zeigt das Fadenpendel, welches aus dem sogenannten konzeptionellen Modell resultiert.

Beim konzeptionellen Modell macht man sich Gedanken, welche Eigenschaften (physikalischen Prozesse) für die zu beantwortende Frage (Bewegung des Pendels) relevant sind (inklusive Genauigkeit)

Wir entscheiden uns für folgende Näherungen:

- Leuchter ist ein Massenpunkt mit der Masse m.
- Der Faden der Länge l wird als rigide und masselos angenommen.

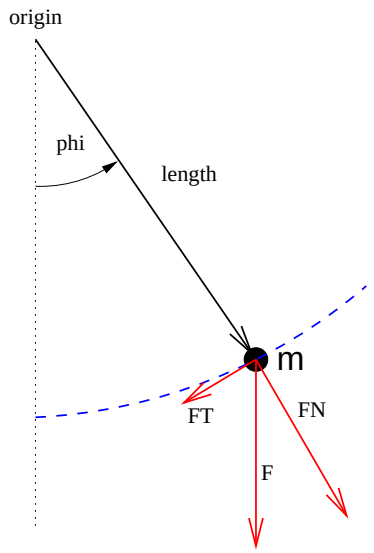


Abbildung 2: Das Fadenpendel.

- Der Luftwiderstand wird vernachlässigt.

Nun soll ein mathematisches Modell entwickelt werden. Wir beginnen mit der Frage, welche Kräfte auf den Körper wirken.

Der Körper wird auf eine Kreisbahn gezwungen; nur die *Tangentialkraft* ist relevant.

In Abhängigkeit der Auslenkung ϕ lautet diese:

$$\vec{F}_T(\phi) = -m\,g\sin(\phi)\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\end{pmatrix}.$$

Beispiel:

$$\vec{F}_T(0) = -mg\begin{pmatrix}0\\0\end{pmatrix}, \vec{F}_T(\pi/2) = -mg\begin{pmatrix}0\\1\end{pmatrix}.$$

Dies überlegt man sich so. Die Gewichtskraft zeigt immer nach unten, also

$$\vec{F}(\phi) = mg\begin{pmatrix}0\\-1\end{pmatrix}.$$

Die Normalkomponente zeigt immer in Richtung $\vec{n}(\phi) = (\sin\phi\,, -\cos\phi)^T$ und damit ist die Kraft in Normalenrichtung

$$\begin{aligned}\vec{F}_N(\phi) = (\vec{F}(\phi)\cdot\vec{n}(\phi))\vec{n} &= \left[mg\begin{pmatrix}0\\-1\end{pmatrix}\cdot\begin{pmatrix}\sin\phi\\-\cos\phi\end{pmatrix}\right]\begin{pmatrix}\sin\phi\\-\cos\phi\end{pmatrix}\\ &= mg\cos\phi\begin{pmatrix}\sin\phi\\-\cos\phi\end{pmatrix}.\end{aligned}$$

Damit rechnet man die Tangentialkraft aus $\vec{F}_T(\phi) + \vec{F}_N(\phi) = \vec{F}(\phi)$ aus:

$$\vec{F}_T(\phi) = \vec{F}(\phi) - \vec{F}_N(\phi) = mg \begin{pmatrix} 0 \\ -1 \end{pmatrix} - mg\cos\phi \begin{pmatrix} \sin\phi \\ -\cos\phi \end{pmatrix} = -mg \begin{pmatrix} \cos\phi\sin\phi \\ 1-\cos^2\phi \end{pmatrix}$$
$$= -mg\sin\phi \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}.$$

Beachte: Auslenkung entgegen Uhrzeigersinn ist positiv, sonst negativ. Auch eine Auslenkung größer π macht Sinn: rotierende Schiffschaukel.

Nach dem 2. Newton'schen Gesetz gilt nun

$$F(t) = m\,a(t)$$

(Kraft gleich Masse mal Beschleunigung).

Die Beschleunigung $a(t)$, Geschwindigkeit $v(t)$ und zurückgelegter Weg $s(t)$ hängen zusammen über

$$a(t) = \frac{dv(t)}{dt}, \qquad v(t) = \frac{ds(t)}{dt}.$$

Für unser Pendel gilt $s(t) = l\phi(t)$ (Setze z. B. $\phi = 2\pi$ ein) und damit

$$v(t) = \frac{d\,s(\phi(t))}{dt} = \frac{d\,l\phi(t)}{dt} = l\frac{d\phi(t)}{dt}$$

und entsprechend

$$a(t) = \frac{d\,v(\phi(t))}{dt} = l\frac{d^2\phi}{dt^2}(t).$$

Einsetzen in das 2. Newton'sche Gesetz liefert nun:

$$ml\frac{d^2\phi(t)}{dt^2} = -mg\sin(\phi(t)) \qquad \forall t > t_0.$$

Die Kraft ist hier skalar (vorzeichenbehafteter Betrag der Tangentialkraft), da wir nur den zurückgelegten Weg betrachten. Das Vorzeichen beschreibt die Richtung (rechts ist positiv).

Dies ist eine „gewöhnliche" Differentialgleichung 2. Ordnung für die Auslenkung ϕ in Abhängigkeit von der Zeit:

$$\frac{d^2\phi(t)}{dt^2} = -\frac{g}{l}\sin(\phi(t)) \qquad \forall t > t_0. \tag{1.1}$$

Um diese Gleichung eindeutig lösen zu können, benötigt man noch zwei Anfangsbedingungen (wegen der zweiten Ordnung):

$$\phi(0) = \phi_0, \qquad \frac{d\phi}{dt}(0) = u_0. \tag{1.2}$$

(Wir haben hier $t_0 = 0$ gesetzt).

Diese allgemeine Gleichung für das Pendel ist schwer „analytisch" zu lösen.

Für *kleine* Winkel ϕ gilt allerdings in guter Näherung

$$\sin(\phi) \approx \phi,$$

z.B. $\sin(0.1) = 0,099833417$.

Mit dieser *Näherung* reduziert sich die Gleichung zu

$$\frac{d^2\phi(t)}{dt^2} = -\frac{g}{l}\phi(t),$$

die man leicht lösen kann.

Der Ansatz $\phi(t) = A\cos(\omega t)$ liefert mit $\phi(0) = \phi_0$, $\frac{d\phi}{dt}(0) = 0$ dann die aus der Schule bekannte Formel

$$\phi(t) = \phi_0 \cos\left(\sqrt{\frac{g}{l}}t\right) \tag{1.3}$$

Die volle Gleichung wollen wir *numerisch* mit zwei verschiedenen Verfahren lösen.

Zunächst schreiben wir die eine Gleichung zweiter Ordnung in zwei Gleichungen erster Ordnung um (Das geht übrigens immer!):

$$\frac{d\phi(t)}{dt} = u(t), \qquad \frac{d^2\phi(t)}{dt^2} = \frac{du(t)}{dt} = -\frac{g}{l}\sin(\phi(t)).$$

Nun ersetzen wir die Ableitungen durch Differenzenquotienten:

$$\begin{aligned} \frac{\phi(t+\Delta t) - \phi(t)}{\Delta t} &\approx \frac{d\phi(t)}{dt} = u(t), \\ \frac{u(t+\Delta t) - u}{\Delta t} &\approx \frac{du(t)}{dt} = -\frac{g}{l}\sin(\phi(t)). \end{aligned}$$

Mit $\phi^n = \phi(n\Delta t)$, $u^n = u(n\Delta t)$ erhalten wir die Rekursion:

$$\phi^{n+1} = \phi^n + \Delta t\, u^n \qquad \phi^0 = \phi_0 \tag{1.4}$$

$$u^{n+1} = u^n - \Delta t\,(g/l)\,\sin(\phi^n) \qquad u^0 = u_0 \tag{1.5}$$

Dieses Verfahren ist nicht das einzig Mögliche.

Man kann auch eine Näherungsformel für die zweite Ableitung nutzen („Zentraler Differenzenquotient“):

$$\frac{\phi(t+\Delta t) - 2\phi(t) + \phi(t-\Delta t)}{\Delta t^2} \approx \frac{d^2\phi(t)}{dt^2} = -\frac{g}{l}\sin(\phi(t)).$$

Löst man nach $\phi(t+\Delta t)$ auf, so ergibt sich die Rekursionsformel ($n \geq 2$):

$$\phi^{n+1} = 2\phi^n - \phi^{n-1} - \Delta t^2\,(g/l)\,\sin(\phi^n) \tag{1.6}$$

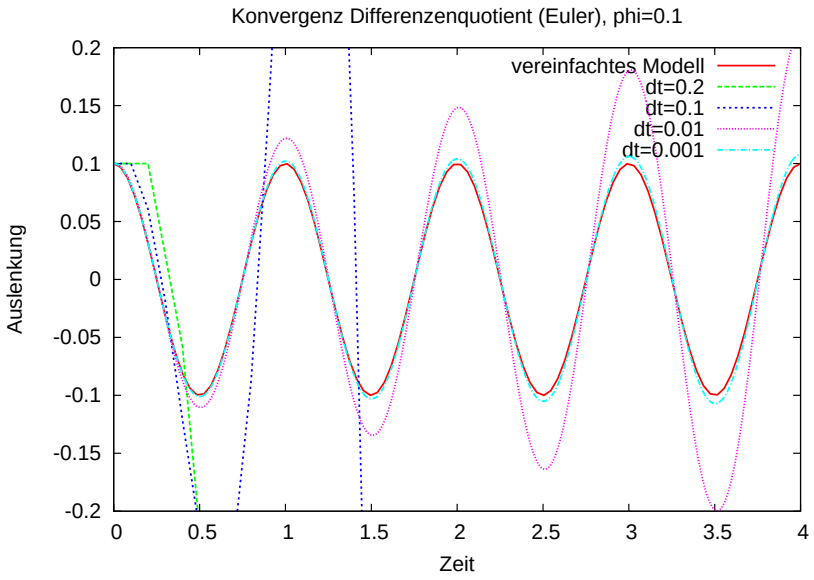


Abbildung 3: Simulation des Fadenpendels (volles Modell) bei $\phi_0 = 0.1 \approx 5.7°$ mit dem Eulerverfahren.

mit der Anfangsbedingung

$$\phi^0 = \phi_0, \qquad \phi^1 = \phi_0 + \Delta t\, u_0. \tag{1.7}$$

(Die zweite Bedingung kommt aus dem Eulerverfahren oben).

Nun auf zum Computer!

Abbildung 3 zeigt das Eulerverfahren in Aktion.

Für festen Zeitpunkt t und $\Delta t \to 0$ *konvergiert* das Verfahren.

Für festes Δt und $t \to \infty$ nimmt das Verfahren immer größere Werte an.

Abbildung 4 zeigt zum Vergleich das zentrale Verfahren für die gleiche Anfangsbedingung.

Im Unterschied zum expliziten Euler scheint das Verfahren bei festem Δt und $t \to \infty$ nicht unbeschränkt zu wachsen.

Nun können wir das volle Modell mit dem vereinfachten Modell vergleichen und sehen welche Auswirkungen die Annahme $\sin\phi \approx \phi$ auf das Ergebnis hat. Abbildung 5 zeigt die numerische Simulation.

Selbst bei 28.6° ist die Übereinstimmung noch einigermaßen passabel.

Für große Auslenkungen ist das vereinfachte Modell völlig unbrauchbar.

Die Form der Schwingung ist kein Kosinus mehr.

Das Pendel wird nahe π immer langsamer. Das ist die Schiffschaukel, die fast auf dem Kopf steht.

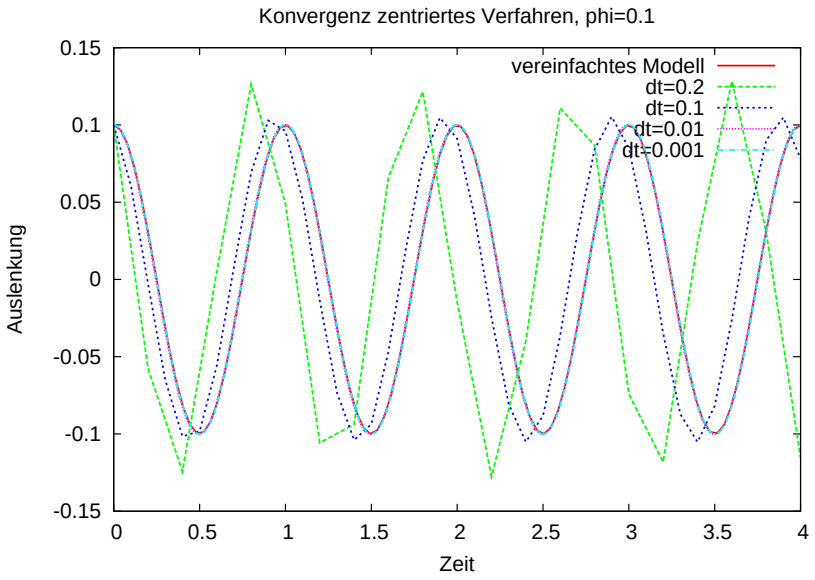


Abbildung 4: Simulation des Fadenpendels (volles Modell) bei $\phi_0 = 0.1 \approx 5.7°$ mit dem zentralen Verfahren.

Wie würde denn die Kurve bei einer umlaufenden Schiffschaukel aussehen?

1.3 Wo kommt jetzt die Stochastik ins Spiel?

Das Pendel ist ein klassisches Beispiel des *Determinismus* des 18. Jahrhunderts: Sind nur die Anfangsbedingungen bekannt, kann "alles" mittels mathematischer Gleichungen vorhergesagt werden.

Lotto (6 aus 49) ist ein sogenanntes Mehrkörpersystem, das auch durch mathematische Gleichungen und den Anfangszustand beschrieben werden kann. Warum rechnet dann niemand die nächsten Lottozahlen aus?

Dynamische Systeme: Es gibt Systeme, bei denen *winzigste* Unterschiede am Anfang nach endlicher Zeit sehr große Unterschiede im Zustand bewirken können ("Chaos"). Diese System sind *praktisch* nicht vorhersagbar.

Stochastische Modelle „beschreiben und untersuchen Vorgänge, die *zufällig* oder *vom Zufall beeinflusst* sind im Sinne von *nicht vorhersagbar*" [Hüb03].

Je nach Anwendung benutzt man stochastische oder deterministische Modelle (oder beides kombiniert), um ein System zu beschreiben.

1.4 Inhaltsübersicht der Vorlesung

Wie in jedem Wissensgebiet muss man auch hier bescheiden beginnen.

Wir werden in dieser Vorlesung die folgenden Themengebiete behandeln

- Gleitpunktzahlen, Gleitpunktarithmetik (2 Vorlesungen)
- Interpolation, Darstellung von Funktionen (4 Vorlesungen)
- Numerische Integration (2)
- Lösen linearer und nichtlinearer Gleichungen (5)
- Lösen gewöhnlicher Differentialgleichungen (2)
- Diskrete Wahrscheinlichkeitsräume (4)
- Kontinuierliche Wahrscheinlichkeitsräume (2)
- Statistik (1)

Die Zahl in Klammern gibt die Anzahl der Vorlesungen zu diesem Thema an.

1.5 Zusammenfassung

- Modellbildung und Simulation bzw. Wissenschaftliches Rechnen etabliert sich als dritte Säule in der Wissenschaftlichen Methode:
  - Man erhält Einsicht in komplexe Systeme, die nur mit Papier und Bleistift nicht möglich ist (im Sinne einer Ergänzung!).
  - Undurchführbare und/oder teuere Experimente können ersetzt werden.
  - Optimierung technischer Anlagen wird möglich.
- Dies hat vielfältige Anwendungen in Wissenschaft und Industrie.
- Informatiker tragen in diesem Umfeld z. B. in der Softwareentwicklung, Visualisierung und Höchstleistungsrechnen bei.
- Je nach Anwendungsfall werden stochastische und/oder deterministische Modelle verwendet.
- Mit dem Fadenpendel wurde das typische Vorgehen bei einer deterministische Modellierung und Simulation illustriert. Es wurden die zwei Fehlerarten Modellfehler und Diskretisierungsfehler demonstriert.

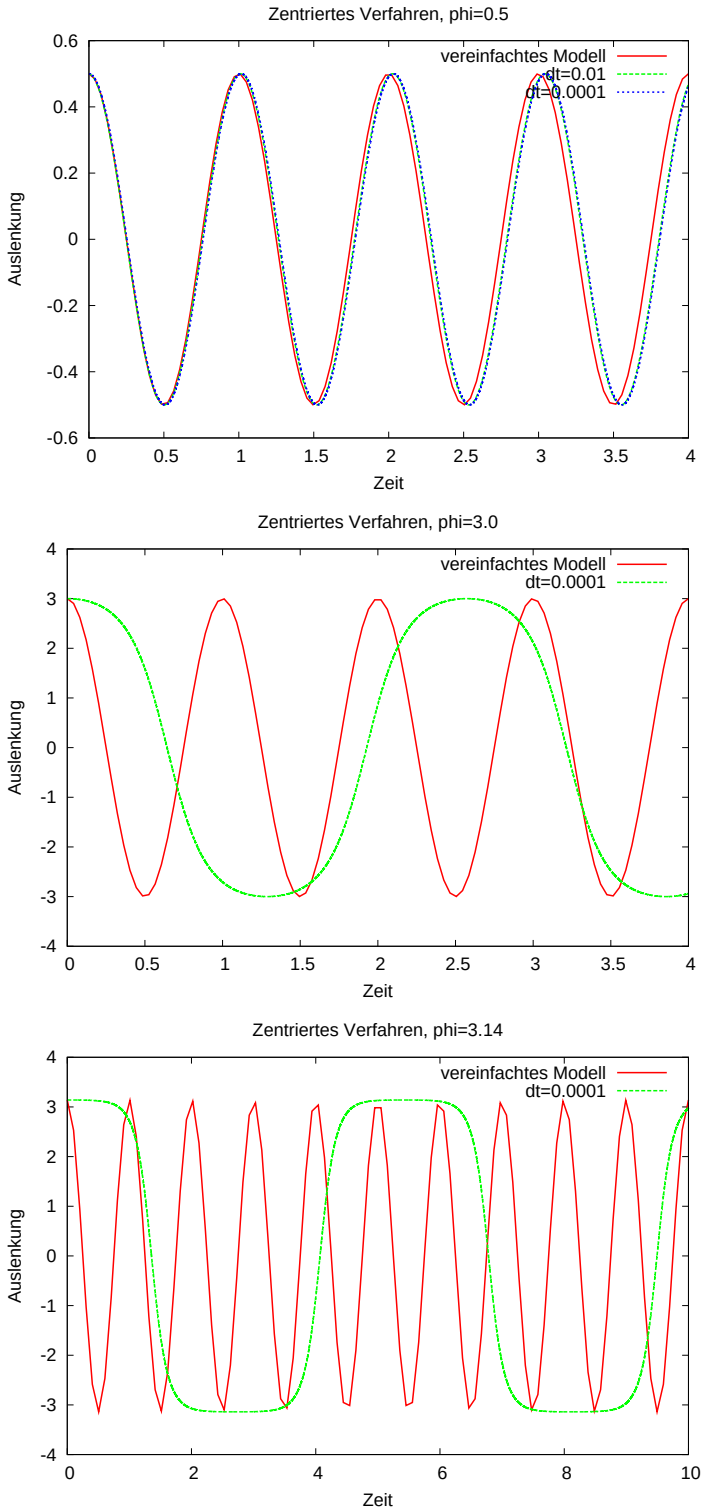


Abbildung 5: Vergleich von vollem und vereinfachtem Modell (jeweils in rot) bei den Winkeln $\phi = 0.5, 3.0, 3.14$ gerechnet mit dem zentralen Verfahren.

2 Fließkommazahlen und Fließkommaarithmetik

Alle Programmiersprachen stellen elementare Datentypen zur Repräsentation von Zahlen zur Verfügung. In C/C++ gibt es die folgenden:

unsigned int	$\mathbb{N}_0$
int	$\mathbb{Z}$
float	$\mathbb{R}$
double	$\mathbb{R}$
complex<double>	$\mathbb{C}$

Diese sind Idealisierungen der Zahlenmengen $\mathbb{N}_0, \mathbb{Z}, \mathbb{R}, \mathbb{C}$ aus der Mathematik.

Bei *unsigned int* und *int* besteht die Idealisierung darin, dass es eine größte (bzw. kleinste) darstellbare Zahl gibt. Ansonsten sind die Ergebnisse *exakt*.

Bei *float* und *double* kommt hinzu, dass die meisten innerhalb des erlaubten Bereichs liegenden Zahlen nur *näherungsweise* dargestellt werden können. Dies hat allerhand Auswirkungen, wenn man mit diesen Zahlen rechnet.

Beispiel 2.1 (Potenzreihe für e^x)**.** e^x lässt sich mit einer Potenzreihe berechnen:

$$e^x = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!} = 1 + \sum_{n=1}^{\infty} y_n.$$

Algorithmisch formulieren wir

$$y_1 = x; \qquad S_1 = 1 + y_1$$

und berechnen für $n = 2, 3, \ldots$

$$y_n = \frac{x}{n} y_{n-1}; \qquad S_n = S_{n-1} + y_n.$$

unter Nutzung verschiedener Genauigkeiten für die Fließkommaarithmetik.

Für $x = 1$ und `float`-Genauigkeit erhalten wir:

```
1.000000000000000e+00   2  2.500000000000000e+00
5.000000000000000e-01   3  2.666666746139526e+00
1.666666716337204e-01   4  2.708333492279053e+00
4.166666790843010e-02   5  2.716666936874390e+00
8.333333767950535e-03   6  2.718055725097656e+00
1.388888922519982e-03   7  2.718254089355469e+00
1.984127011382952e-04   8  2.718278884887695e+00
2.480158764228690e-05   9  2.718281745910645e+00
2.755731884462875e-06  10  2.718281984329224e+00
2.755731998149713e-07 ... 100 2.718281984329224e+00
0.000000000000000e+00  ex  2.718281828459045E0
```

... also 7 gültige Ziffern.

Für $x = 5$ ...

```
9.333108209830243e-06  ex  1.484131591025766E2
```

... dito.

Für $x = -1$ und `float`-Genauigkeit erhalten wir:

```
2.755731998149713e-07 11 3.678793907165527e-01
-2.505210972003624e-08 12 3.678793907165527e-01
2.087675810003020e-09 ex 3.678794411714423E-1
```

... 6 gültige Ziffern und für $x = -5$

```
-5.000000000000000e+00 2 8.500000000000000e+00
1.250000000000000e+01 3 -1.233333396911621e+01
-2.083333396911621e+01 4 1.370833396911621e+01
2.604166793823242e+01 ...  15 1.118892803788185e-03
-2.333729527890682e-02 16 8.411797694861889e-03
7.292904891073704e-03 ...  28 6.737461313605309e-03
1.221854423194557e-10 ...  100 6.737461313605309e-03
0.000000000000000e+00 ex 6.737946999085467E-3
```

nur noch 4 gültige Ziffern.

Für $x = -20$ und `float`-Genauigkeit sind ...

```
-2.000000000000000e+01 2 1.810000000000000e+02
2.000000000000000e+02 3 -1.152333374023438e+03
-1.333333374023438e+03 4 5.514333496093750e+03
6.666666992187500e+03 5 -2.115233398437500e+04
-2.666666796875000e+04 ...  31 -1.011914250000000e+06
-2.611609750000000e+06 32 6.203418750000000e+05
1.632256125000000e+06 33 -3.689042500000000e+05
-9.892461250000000e+05 34 2.130052500000000e+05
5.819095000000000e+05 35 -1.195144687500000e+05
-3.325197187500000e+05 36 6.521870312500000e+04
1.847331718750000e+05 ...  65 7.566840052604675e-01
-4.473213550681976e-07 66 7.566841244697571e-01
1.355519287926654e-07 67 7.566840648651123e-01
-4.046326296247571e-08 68 7.566840648651123e-01
1.190095932912527e-08 ex 2.061153622438557E-9
```

keine Ziffern mehr gültig. Das Ergebnis ist um 8 Größenordnungen daneben!

Für $x = -20$ und `double`-Genauigkeit erhält man

```
-1.232613988175268e+07 28 3.623690792934047e+06
8.804385629823344e+06 ...  94 6.147561828914626e-09
1.821561256740375e-24 95 6.147561828914626e-09
-3.834865803663947e-25 ex 2.061153622438557E-9
```

Immer noch um einen Faktor 3 daneben! Erst mit „vierfacher Genauigkeit“ erhält man

```
2.061153622438558339270045875294 7E-9
-4.185292933938207365036374157994 1E-41 118
2.061153622438558339270045875294 7E-9
7.093716837183402313620973149142 7E-42 ex
2.061153622438557827965940380155 8E-9
```

15 gültige Ziffern (bei ca 30 Ziffern „Rechengenauigkeit“). □

Dieses Beispiel wirft die folgenden Fragen auf:

- Was bedeutet überhaupt „Rechengenauigkeit“?
- Welche Genauigkeit können wir erwarten?
- Wo kommen diese Fehler her?
- Wie werden solche „Kommazahlen“ dargestellt und verarbeitet?

Bemerkung 2.2 (High-Precision Pakete)**.** Obige Berechnungen wurden mit den Paketen `qd` und `arprec` (beide `http://crd.lbl.gov/~dhbailey/mpdist/`) durchgeführt. `qd` erlaubt bis zu vierfache `double` Genauigkeit, `arprec` beliebige Genauigkeit.

Die GNU multiprecision library (`http://gmplib.org/`) ist eine Alternative.

2.1 Fließkommadarstellung von Zahlen

Zahlen werden in einem *Stellenwertsystem* (auch *polyadisches Zahlensystem*) folgendermaßen dargestellt;

$$x = \pm \ldots m_n \beta^n + \ldots + m_1 \beta^1 + m_0 + m_{-1}\beta^{-1} + \ldots + m_{-k}\beta^{-k} + \ldots \tag{2.1}$$

$\beta \in \mathbb{N}, \beta \geq 2$, heißt Basis.

Die $m_i \in \mathbb{N}_0, 0 \leq m_i < \beta$ heißen Ziffern.

Alternativ sind *Additionssysteme* (z. B. römische Zahlen) möglich.

Die Darstellung von Zahlen hat eine sehr interessante Geschichte, siehe [Knu98, p.194] für Details.

Die Babylonier nutzten 1750 v. Chr $\beta = 60$ (deswegen 60 Sekunden). Die Basis 10 hat sich in Europa ab ca. 1585 durchgesetzt. Pascal[1] erkannte 1658, dass man jedes $\beta \geq 2$ verwenden kann.

Im Rechner legen technische Gründe (Digitaltechnik)

$$\beta = 2, m_i \in \{0, 1\}$$

nahe. m_i ist dann ein Bit.

Bei Festkommazahlen wählt man $n, k \in \mathbb{N}$ fest und hat dann

$$x = \sum_{i=-k}^{n} m_i \beta^i.$$

β^{-k} ist dann die „Auflösung“ (kleinster Abstand zweier Festkommazahlen).

Bei wissenschaftliche Anwendungen kommen Zahlen sehr unterschiedlicher Größe vor, etwa in den physikalischen Konstanten

Plancksches Wirkungsquantum:	$6.6260693 * 10^{-34}\ Js$
Ruhemasse Elektron:	$9.11 * 10^{-28}\ g$
Avogadro Konstante:	$6.021415 * 10^{23}\ mol^{-1}$

Für Zahlen sehr unterschiedlicher Größe werden Festkommazahlen ineffizient.

Die sogenannten Fließkommazahlen (auch Fließpunkt, Gleitpunkt, engl. floating point numbers) erlauben dann eine effizientere Darstellung.

Definition 2.3 (normierte Fließkommazahlen)**.** $\mathbb{F}(\beta, r, s) \subset \mathbb{R}$ besteht aus den Zahlen mit folgenden Eigenschaften:

[1] Blaise Pascal, 1623-1662, frz. Mathematiker und Philosoph.

- $\forall x \in \mathbb{F}(\beta, r, s)$ gilt $x = m\beta^e$ mit

$$m = \pm \sum_{i=1}^{r} m_i \beta^{-i}, \qquad e = \pm \sum_{j=0}^{s-1} e_j \beta^j$$

  m heißt *Mantisse* (engl. *mantissa* oder *fraction*) und e *Exponent*.
- $\forall x \in \mathbb{F}(\beta, r, s)$ gilt $x = 0 \vee m_1 \neq 0$ (Normierung), d. h.

$$|x| = 0 \ \vee \ \beta^{-1} \leq |m| < 1.$$

Sind β, r, s klar (oder egal) so schreiben wir einfach $\mathbb{F}$. □

Beispiel 2.4. $\underline{\mathbb{F}(10,3,1)}$ besteht aus Zahlen der Form

$$x = \pm(m_1 \cdot 0.1 + m_2 \cdot 0.01 + m_3 \cdot 0.001) \cdot 10^{\pm e_0}$$

$(m_1 \neq 0 \vee m_1 = m_2 = m_3 = 0)$, z. B. $0.999 \cdot 10^1, 0.123 \cdot 10^{-1}, 0$.

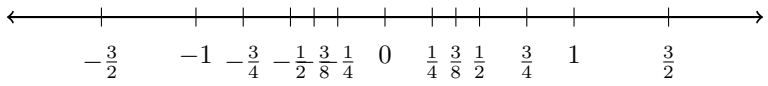
$0.014 \in \mathbb{F}(10,3,1)$ da $0.014 = 0.140 \cdot 10^{-1}$, aber

$0.000000000014 \notin \mathbb{F}(10,3,1)$ da $0.000000000014 = 0.14 \cdot 10^{-10}$

$\underline{\mathbb{F}(2,2,1)}$ besteht aus Zahlen der Form $(x = 0 \vee m_1 \neq 0)$

$$x = \pm \left(m_1 \frac{1}{2} + m_2 \cdot \frac{1}{4} \right) \cdot 2^{\pm e_0}.$$

Somit also $\mathbb{F}(2,2,1) = \{-\frac{3}{2}, -1, -\frac{3}{4}, -\frac{1}{2}, -\frac{3}{8}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}\}$, bzw. graphisch:



Dies überlegt man so:

- 0 ist klar.
- $m_1 = 1, m_2 = 0$ gibt $1/2$, $m_1 = 1, m_2 = 1$ gibt $3/4$.
- Multiplikation mit $2, 1, 1/2$ $(e_0 = 1, 0, -1)$ liefert $\mathbb{F}(2,2,1)$.

Beachte den größeren Abstand bei der Null wegen Normierung! □

Die größte bzw. kleinste darstellbare Zahl in $\mathbb{F}(\beta, r, s)$ ist:

$$\begin{aligned} X_{+/-} &= \pm \underbrace{\overbrace{(\beta-1)\{\beta^{-1} + \cdots + \beta^{-r}\}}^{1-\beta^{-r}}}_{m_i = \beta - 1} \cdot \beta \underbrace{\overbrace{(\beta-1)\{\beta^{s-1} + \cdots + \beta^0\}}^{\beta^s - 1}}_{e_i = \beta - 1} \\ &= \pm(1 - \beta^{-r})\beta^{\beta^s - 1} \end{aligned}$$

Die kleinste positive bzw. größte negative Zahl in $\mathbb{F}(\beta, r, s)$ ist:

$$x_{+/-} = \pm \underbrace{\beta^{-1}}_{\substack{\text{kleinste Mantisse} \\ \text{bei Normierung}}} \cdot \beta^{\overbrace{-(\beta-1)\{\beta^{s-1}+\cdots+\beta^0\}}^{-(\beta^s-1)}} = \pm\beta^{-\beta^s}$$

Damit gilt $\mathbb{F}(\beta, r, s) \subset D(\beta, r, s) = [X_-, x_-] \cup \{0\} \cup [x_+, X_+] \subset \mathbb{R}$.

2.2 Runden und Rundungsfehler

Sind beliebige Zahlen $x, y \in \mathbb{R}$ gegeben, so sind diese erst in Fließkommazahlen zu verwandeln.

Wir benötigen eine Abbildung $\mathrm{rd} : D \to \mathbb{F}$, (Für $x \notin D$ muss man sein Problem umformulieren oder $\mathbb{F}$ größer machen).

Sinnvollerweise fordert man

$$|x - \mathrm{rd}(x)| \leq \min_{y\in\mathbb{F}} |x - y| \quad \forall x \in D. \tag{2.2}$$

Mit

$$\mathrm{left}(x) = \max\{y \in \mathbb{F} \,|\, y \leq x\}, \quad \mathrm{right}(x) = \min\{y \in \mathbb{F} \,|\, y \geq x\}$$

gilt dann

$$\mathrm{rd}(x) = \begin{cases} \mathrm{left}(x) & \text{falls} \quad |x - \mathrm{left}(x)| < |x - \mathrm{right}(x)| \\ \mathrm{right}(x) & \text{falls} \quad |x - \mathrm{right}(x)| < |x - \mathrm{left}(x)| \\ ? & \quad x = \frac{\mathrm{left}(x)+\mathrm{right}(x)}{2} \end{cases}$$

Für die im letzten Fall erforderliche Rundung gibt es verschiedene Möglichkeiten.

Sei $x = \mathrm{sign}(x)(\sum_{i=1}^{\infty} m_i\beta^{-i})\beta^e$ die normierte Darstellung von $x \in D \subset \mathbb{R}$.

Aufrunden, natürliche Rundung:

$$\mathrm{rd}(x) = \begin{cases} \mathrm{left}(x) = \mathrm{sign}(x)(\sum_{i=1}^{r} m_i\beta^{-i})\beta^e & \text{falls } 0 \leq m_{r+1} < \beta/2 \\ \mathrm{right}(x) = \mathrm{left}(x) + \beta^{e-r} & \text{falls } \beta/2 \leq m_{r+1} < \beta \end{cases}$$

Gerade Rundung (β sei gerade):

$$\mathrm{rd}(x) = \begin{cases} \mathrm{left}(x) & |x - \mathrm{left}(x)| < |x - \mathrm{right}(x)| \vee \\ & (|x - \mathrm{left}(x)| = |x - \mathrm{right}(x)| \wedge m_r \text{ gerade}) \\ \mathrm{right}(x) & \text{sonst} \end{cases}$$

Mit dieser Wahl gilt, dass m_r immer gerade ist wenn gerundet werden musste.

Dies Wahl vermeidet eine Drift, die bei Aufrunden auftreten kann (siehe Übungsaufgabe).

Wir wollen nun den bei der Rundung entstehenden Fehler analysieren.

Zunächst eine allgemeine Definition zum Fehlerbegriff

Definition 2.5 (Absoluter und relativer Fehler)**.** Sei $x' \in \mathbb{R}$ eine Näherung von $x \in \mathbb{R}$ dann heißt

$$\Delta x = x' - x \tag{2.3}$$

absoluter Fehler und für $x \neq 0$ heißt

$$\varepsilon_{x'} = \frac{\Delta x}{x} = \frac{x' - x}{x} \tag{2.4}$$

relativer Fehler. Oft nutzen wir die Form

$$x' \overset{(2.3)}{=} x + \Delta x = x\left(1 + \frac{\Delta x}{x}\right) \overset{(2.4)}{=} x(1 + \varepsilon_{x'})$$

□

Motivation zum relativen Fehler.

Bei der Entfernung Erde-Sonne ($\approx 1,5 * 10^8 km$) sind $100km$ ein relativ kleiner Fehler

$$\varepsilon_{x'} = 6.6 \cdot 10^{-7}.$$

Bei der Entfernung Stuttgart-Paris ($\approx 600km$) dagegen schon

$$\varepsilon_{x'} = 0.1\overline{6}.$$

Damit gilt für den Rundungsfehler das

Lemma 2.6 (Rundungsfehler)**.** Der absolute Rundungsfehler bei Rundung von $x \in D(\beta, r, s)$ nach $\mathbb{F}(\beta, r, s)$ ist höchstens

$$|x - \mathrm{rd}(x)| \leq \frac{1}{2}\beta^{e-r}. \tag{2.5}$$

Der relative Rundungsfehler kann abgeschätzt werden durch

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq \frac{1}{2}\beta^{1-r} \tag{2.6}$$

Die Größe $eps := \frac{1}{2}\beta^{1-r}$ heißt Maschinengenauigkeit, in der englischen Literatur heißt β^{1-r} oft *ulp* (*units last place*).

Beweis: (2.5) gilt sofort wegen (2.2). Für (2.6) zeigt man:

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq \frac{1}{2}\frac{\beta^{e-r}}{|m|\beta^e} \overset{\substack{\text{Normierung} \\ |m| \geq \beta^{-1}}}{\leq} \frac{1}{2}\frac{\beta^{-r}}{\beta^{-1}} = \frac{1}{2}\beta^{1-r}$$

2.3 Fließkommaarithmetik

Auf dem Körper $\mathbb{R}$ sind die Operationen $* \in \{+, -, \cdot, /\}$ definiert.

Wir benötigen auch entsprechende *Maschinenoperationen* $\circledast : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ für $\circledast \in \{\oplus, \ominus, \odot, \oslash\}$.

Dabei soll $\oplus$ dem $+$ Operator entsprechen, $\ominus$ dem $-$, usw.

Wenn $x, y \in \mathbb{F}$ folgt daraus *nicht*, dass $x * y \in \mathbb{F}$, sondern es ist eventuell eine Rundung erforderlich.

Man fordert für die Maschinenoperationen folgende Eigenschaft:

$$x \circledast y = \mathrm{rd}(x * y) \quad \forall x, y \in \mathbb{F}.$$

Man sagt $\circledast$ ist *exakt gerundet*. Dass dies effizient möglich ist, motivieren wir durch ein Beispiel.

Beispiel 2.7 (Guard digit)**.** Sei $\mathbb{F} = \mathbb{F}(10, 3, 1)$ und betrachte $\ominus$. Sei weiter $x = 0.215 \cdot 10^8$, $y = 0.125 \cdot 10^{-5}$.

Naive Realisierung von $x \ominus y = \mathrm{rd}(x - y)$ erfordert schieben von y auf den größeren Exponenten $y = 0.125 \cdot 10^{-13} \cdot 10^8$ und subtrahieren der Mantissen:

$$\begin{array}{rcl} x & = & 0.2150000000000000 \cdot 10^8 \\ y & = & 0.0000000000000125 \cdot 10^8 \\ \hline x - y & = & 0.2149999999999875 \cdot 10^8 \end{array}$$

Runden auf drei Stellen liefert dann $x \ominus y = 0.215 \cdot 10^8$. Dies erfordert einen Addierer mit $2\beta^s$ Stellen!

Das Ergebnis hätten wir auch durch die Abfolge *Schieben*, *Runde* y, *Rechne* bekommen.

Im Allgemeinen ist das aber nicht gut wie folgendes Beispiel zeigt:

$$\begin{array}{rcl} x & = & 0.101 \cdot 10^1 \\ y & = & 0.993 \cdot 10^0 \end{array} \quad \to \quad \begin{array}{rcll} x & = & 0.101 & \cdot 10^1 \\ y & = & 0.0993 & \cdot 10^1 \\ \hline x \ominus y & = & 0.002 & \cdot 10^1 \end{array}$$

für den relativen Fehler im Ergebnis gilt dann

$$\frac{(x \ominus y) - (x - y)}{(x - y)} = \frac{0.02 - 0.017}{0.017} \approx 0.176 \approx 35 eps$$

bei

$$eps = \frac{1}{2} 10^{-3+1} = 0.005$$

Nun spendieren wir eine Stelle mehr, d.h. wir nutzen einen $r + 1$-stelligen Addierer:

$$\begin{array}{rcll} x & = & 0.1010 & \cdot 10^1 \\ y & = & 0.0993 & \cdot 10^1 \\ \hline x - y & = & 0.0017 & \cdot 10^1 \end{array}$$

Das Ergebnis $x \ominus y = 1,7 \cdot 10^{-2}$ ist exakt!

Allgemein kann man zeigen: Mit *einer* zusätzlichen Stelle (sog. guard digit) gilt

$$\frac{(x \ominus y) - (x - y)}{x - y} \leq 2eps.$$

Mit noch einer Stelle mehr erreicht man die exakte Rundung! □

Die Fließkommaarithmetik hat allerdings noch ein paar Überraschungen parat ...

Bemerkung 2.8. Assoziativ- und Distributivgesetz gelten in $\mathbb{F}$ im allgemeinen nicht, d.h. es ist für $x, y, z, \in \mathbb{F}$:

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z), \qquad (x \oplus y) \odot z \neq (x \odot z) \oplus (y \odot z)$$

Insbesondere gilt

$$x \oplus y = x \qquad \forall |y| \leq \underbrace{|x| \frac{\beta^{-r}}{2}}_{\substack{x \text{ um } r \text{ Stellen nach rechts,} \\ \frac{1}{2} \text{damit } m_{r+1} \leq \frac{\beta}{2}}} = |x| \frac{\beta}{\beta} \frac{\beta^{-r}}{2} = \frac{eps}{\beta} |x|$$

Allerdings gilt das Kommutativgesetz $x \oplus y = y \oplus x; x \odot y = y \odot x$.

Es gelten noch ein paar weitere einfache Gesetze, wie etwa (keine vollständige Liste)

$$\begin{gathered} (-x) \odot y = -(x \odot y), \qquad 1 \odot x = x; \\ x \odot y = 0 \qquad \text{genau dann, wenn } x = 0 \text{ oder } y = 0; \\ (-x)/y = x \oslash (-y) = -(x \oslash y); \\ x \odot z \leq y \odot z \qquad \text{falls } x \leq y \text{ und } z > 0. \end{gathered}$$

Es gibt auch bemerkenswertere Resultate wie: Sind u, v normalisierte Fließkommazahlen und

$$\begin{aligned} u' &= (u \oplus v) \ominus v, & v' &= (u \oplus v) \ominus u, \\ u'' &= (u \oplus v) \ominus v', & v'' &= (u \oplus v) \ominus u', \end{aligned}$$

so gilt

$$u + v = (u \oplus v) + \left((u \ominus u') \oplus (v \ominus v'')\right).$$

Dies erlaubt eine Berechnung des Fehlers mittels Fließkommaarithmetik.

Siehe [Knu98, 4.2.2, Theorem B]. □

2.4 Der IEEE-754 Standard

Bis in die 1980er Jahre waren viele verschiedene Fließkommazahlen in Gebrauch. Die Eigenschaften von $\circledast$ waren nicht genormt (z.B. $\ominus$ exakt gerundet oder nur ein guard digit?).

Ziel des 1985 verabschiedeten IEEE-754 Standards: Portabilität von Programmen!

IEEE-754 legt $\beta = 2$ fest und definiert vier Genauigkeitsstufen: *single, single-extended, double, double-extended.*

Diese haben folgende Parameter:

Parameter	Format			
	single	single-ext	double	double-ext
e_{max}	$+127$	1023	$+1023$	> 16383
e_{min}	-126	≤ -1022	-1022	≤ -16382
Bits für exp	8	≤ 11	11	15
Bits für alles	32	43	64	79

$\oplus, \ominus, \odot, \oslash, \sqrt{}$ sind exakt gerundet.

Betrachte double Genauigkeit genauer:

- Formatbreite : 64 Bit
- davon 11 Bit für Exponent
- bleiben 53 für Mantisse
- davon 1 Bit Vorzeichen bleiben 52 Bit Mantisse.

Da $x \in \mathbb{F}$ normiert dargestellt wird und $\beta = 2$, gilt *immer* $m_1 = 1$ es sei denn $x = 0$. Kodiert man die Null anders so muss m_1 nicht gespeichert werden (sog. hidden bit).

Der Exponent wird *vorzeichenlos* mittels

$$e = c - 1023 \text{ für } \underbrace{c = c_0 2^0 + \cdots + c_{10} 2^{10}}_{11 \text{ Bits}} \in [\mathbf{1}, \mathbf{2046}]$$

dargestellt.

$c = 0, m = 0$ kodiert den Fall $x = 0$, $c = 2047, m \neq 0$ den Fall NaN (not a number) und $c = 2047, m = 0$ kodiert den Fall ∞ (Überlauf).

Im IEEE Format wird dann *nicht* abgebrochen, sondern z. B. mit der Definition $x \circledast \text{NaN} = \text{NaN}$ weiter gerechnet.

IEEE-754 kennt auch vier verschiedene Rundungsarten, die man umschalten kann. *Default* ist *round to nearest*, es gibt noch *round to zero* (d.h. abschneiden, immer näher zur Null hin), *round to* ∞ (macht nie kleiner), *round to* $-\infty$ (macht nie größer). Dies ist wichtig im Zusammenhang mit Intervallarithmetik.

IEEE-754 definiert Grundrechenarten und Wurzel als exakt gerundet. Über Funktionen wie sin oder exp wird nichts gesagt.

Dabei tritt das *Tabellenmacher-Dilemma* auf: Angenommen exp soll auf vier Stellen genau berechnet werden. Man findet bei 5 Stellen

$$\exp(1.626) = 5.0835.$$

Soll nun ab- oder aufgerundet werden. Genauere Rechnung (Reihe!) liefert

$$\exp(1.626) = 5.0835000$$

und man ist nicht schlauer. Problem: Bei einer transzendenten Funktion kann es beliebig lange dauern bis man $\exp(1.626) < 5.0835$ oder $\exp(1.626) > 5.0835$ findet.

`http://lipforge.ens-lyon.fr/www/crlibm/` ist eine freie Bibliothek korrekt gerundeter mathematischer Funktionen auf Basis des IEEE-754.

Vorsicht beim x86: x86 Register verwenden das double-extended Format, im Speicher wird nur double verwendet. Werden Variablen im Register gehalten (Optimierung!), entstehen so andere Resultate, als wenn diese im Speicher gehalten werden.

2.5 Zusammenfassung

Wichtiges in dieser Vorlesung:

- Stellenwertsystem und Definition normierter Fließkommazahlen.
- Relativer und absoluter Fehler.
- Rundung und Rundungsfehler.
- Exakt gerundete Fließkommaoperationen.
- IEEE-754 Standard für Fließkommazahlen und Fließkommaoperationen.

Eine ausführliche Darstellung zur Fließkommazahlen findet man in dem Artikel von

David Goldberg: *What Every Computer Scientist Should Know About Floating Point Arithmetic, Computing Surveys, 1991* [Gol91].

3 Fehleranalyse

3.1 Auslöschung

Auslöschung ist ein wichtiges Phänomen bei der Subtraktion von Fließkommazahlen.

Bereits in Beispiel 2.7 haben wir in anderem Zusammenhang gesehen, dass bei der Subtraktion in etwa gleich großer Zahlen große relative Fehler entstehen können.

Beobachtung 3.1. (a) Es seien $x, y \in \mathbb{F}$. Die Operation $\ominus : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ sei exakt gerundet, d.h. $x \ominus y = \mathrm{rd}(x - y)$. Dann gilt für den relativen Fehler im Ergebnis:

$$\frac{(x \ominus y) - (x - y)}{(x - y)} = \frac{\mathrm{rd}(x - y) - (x - y)}{(x - y)} \underset{\text{Lemma 2.6}}{\leq} eps \qquad (x - y \neq 0)$$

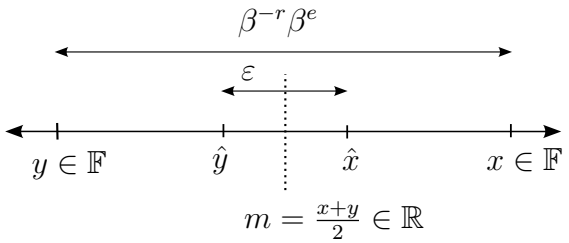
Also kein Problem.

(b) Nun seien $x, y \in \mathbb{F}$ *gerundete Eingaben*, d.h. es gibt $\hat{x}, \hat{y} \in \mathbb{R}$ so dass $x = \mathrm{rd}(\hat{x})$ und $y = \mathrm{rd}(\hat{y})$. Für den relativen Fehler bezüglich des exakten Ergebnisses gilt dann:

$$\frac{(x \ominus y) - (\hat{x} - \hat{y})}{(\hat{x} - \hat{y})} \sim \frac{1}{\epsilon} \qquad \text{für bestimmte } |\hat{x} - \hat{y}| = \epsilon$$

Beweis: Wir betrachten $\mathbb{F}(\beta, r, s)$. Wähle $x - y = \beta^{-r}\beta^e$, $m = (x + y)/2$ sowie $\hat{y} = m - \epsilon/2, \hat{x} = m + \epsilon/2$. Dann gilt:

$$\frac{(x \ominus y) - (\hat{x} - \hat{y})}{(\hat{x} - \hat{y})} = \frac{\beta^{-r}\beta^e - \epsilon}{\epsilon} = \frac{\beta^{-r}\beta^e}{\epsilon} - 1.$$



□

Beispiel 3.2 (Zur Auslöschung)**.** Sei $\mathbb{F} = \mathbb{F}(10, 4, 1)$

$$\begin{array}{rcl|rcl} \hat{x} & = & 0,11258762 \cdot 10^2 & \to \quad x = \mathrm{rd}(\hat{x}) & = & 0,1126 \cdot 10^2 \\ \hat{y} & = & 0,11244891 \cdot 10^2 & \to \quad y = \mathrm{rd}(\hat{y}) & = & 0,1124 \cdot 10^2 \\ \hline \hat{x} - \hat{y} & = & 0,13871 \cdot 10^{-1} & x - y & = & 0,200 \cdot 10^{-1} \end{array}$$

und damit

$$\frac{0,2 \cdot 10^{-1} - 0,13871 \cdot 10^{-1}}{0,13871 \cdot 10^{-1}} \approx 0.44 \approx 883 eps \quad !$$

bei $eps = 0.0005$. □

Nochmal: Die Beobachtung sagt, dass der Fehler in der Subtraktion bei *gerundeten Eingaben* beliebig groß werden kann.

3.2 Rundungsfehleranalyse

Wir beschäftigen uns nun damit wie man die auftretenden Fehler im allgemeinen analysieren kann. Dazu sehen wir uns erst mal an wie eine numerische Berechnung eigentlich abläuft.

Eine numerische Berechnung im Computer verarbeitet Eingaben

$$x_1, \ldots, x_m \qquad x_i \in \mathbb{F}$$

und produziert mittels eines Algorithmus die Ausgaben

$$y_1, \ldots, y_n, \qquad y_i \in \mathbb{F} \quad .$$

Der Algorithmus bestehe dabei nur aus den Maschinenoperationen $\oplus, \ominus, \odot, \oslash$ (später auch $\sqrt[\circ]{}$).

Die Berechnung der einzelnen y_i können wir als Funktionen ausdrücken:

$$\forall i \in \{1, \ldots, n\} : y_i = f_i(x_1, \ldots, x_m) \text{ mit } f_i : \mathbb{F}^m \to \mathbb{F} \quad .$$

Also etwa

$$\begin{aligned} f_1(x_1, x_2) &= x_1 \odot x_1 \ominus x_2 \odot x_2. \\ \text{oder } f_2(x_1, x_2) &= (x_1 \ominus x_2) \odot (x_1 \oplus x_2) \quad . \end{aligned}$$

Natürlich können wir alles auch kompakt in vektorieller Form schreiben

$$x = (x_1, \ldots, x_m)^T, \quad y = (y_1, \ldots, y_n)^T, \quad f = (f_1, \ldots, f_n)^T, \quad y = f(x).$$

Zu der Abbildung $f : \mathbb{F}^m \to \mathbb{F}^n$ können wir eine entsprechende Abbildung $\hat{f} : \mathbb{R}^m \to \mathbb{R}^n$ definieren bei der alle Maschinenoperationen durch die exakten mathematischen Operationen $+, -, \cdot, /$ (und $\sqrt{}$) ersetzt sind.

Schließlich können wir auch Eingaben $\hat{x}_1, \ldots \hat{x}_m \in \mathbb{R}$ betrachten und

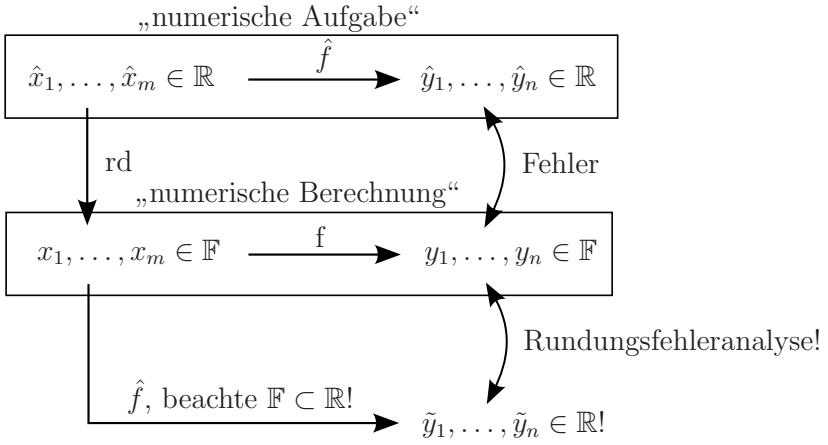
$$x_i = \mathrm{rd}(\hat{x}_i)$$

als gerundet auffassen. Die Ausgaben $\hat{y}_j = \hat{f}_j(\hat{x})$ würden wir dann als das exakte Ergebnis auffassen.

Also für f_1, f_2 von oben:

$$\hat{f}_1(x_1, x_2) = x_1^2 - x_2^2 = (x_1 - x_2) \cdot (x_1 + x_2) = \hat{f}_2(x_1, x_2)$$

$\hat{f}_1, \hat{f}_2$ sind gleich, f_1, f_2 jedoch nicht!



Wir setzen $\tilde{y}_j = \hat{f}_j(x_1, \ldots, x_m)$.

Im allgemeinen können die $\hat{x}_i$ noch mit einem Datenfehler behaftet sein.

Definition 3.3 (Rundungsfehleranalyse)**.** Die Rundungsfehleranalyse untersucht den Fehler in der numerischen Berechnung unter der Annahme, dass die *Eingaben* Maschinenzahlen sind. Also mit den Bezeichnung aus der Abbildung:

$$\frac{y_i - \tilde{y}_i}{\tilde{y}_i} = \frac{f_i(\overbrace{x_1, \ldots, x_m}^{\in \mathbb{F}}) - \hat{f}_i(x_1, \ldots, x_m)}{\hat{f}_i(x_1, \ldots, x_m)} \qquad \text{(wie immer } \tilde{y}_i \neq 0) \quad .$$

Die Rundung der Eingaben bleibt ausser acht. □

Bemerkung 3.4. Wir schreiben hier $x_i = \mathrm{rd}(\hat{x}_i)$ d.h. $x_i \in \mathbb{F}, \hat{x}_i \in \mathbb{R}$. Später verwenden wir auch $\hat{x}_i = \mathrm{rd}(x_i)$ mit $\hat{x}_i \in \mathbb{F}, x_i \in \mathbb{R}$ um Schreibarbeit zu sparen. Also immer auf den Kontext achten! □

Ausgangspunkt der Rundungsfehleranalyse ist immer die Annahme exakt gerundeter Operationen, d.h.

$$x_1 \circledast x_2 = \mathrm{rd}(x_1 * x_2) = (x_1 * x_2)(1 + \epsilon_*)$$

mit $|\epsilon_*| \leq eps$. Beachte jedoch, dass $\epsilon_* = \epsilon_*(x_1, x_2)$ für jede Operation und Eingabe potentiell verschieden ist.

Beispiel 3.5 (Zur Rundungsfehleranalyse)**.** (a) $f_1(x_1, x_2) = x_1 \odot x_1 \ominus x_2 \odot x_2$, also

$$\begin{aligned}
u &= x_1 \odot x_1 = x_1^2(1+\epsilon_1) \\
v &= x_2 \odot x_2 = x_2^2(1+\epsilon_2) \\
y = f_1(x_1, x_2) &= u \ominus v = (u - v)(1+\epsilon_3) \\
&= (x_1^2(1+\epsilon_1) - x_2^2(1+\epsilon_2))(1+\epsilon_3) \\
&= x_1^2(1+\epsilon_1)(1+\epsilon_3) - x_2^2(1+\epsilon_2)(1+\epsilon_3) \\
&= x_1^2 - x_2^2 + x_1^2(\epsilon_1+\epsilon_3) - x_2^2(\epsilon_2+\epsilon_3) + \underbrace{x_1^2\epsilon_1\epsilon_3 - x_2^2\epsilon_2\epsilon_3}_{\text{da } \epsilon_i \text{ klein lässt man die weg}} \\
&\doteq x_1^2 - x_2^2 + x_1^2(\epsilon_1+\epsilon_3) - x_2^2(\epsilon_2+\epsilon_3)
\end{aligned}$$

$\doteq$ bedeutet „in erster Näherung“.

Für den relativen Fehler erhalten wir

$$\begin{array}{rcccc} \frac{f_1(x_1,x_2)-\overbrace{\hat{f}_1(x_1,x_2)}^{=x_1^2-x_2^2}}{\hat{f}_1(x_1,x_2)} & = & \frac{x_1^2}{x_1^2-x_2^2}(\epsilon_1+\epsilon_3) & + & \frac{x_2^2}{x_2^2-x_1^2}(\epsilon_2+\epsilon_3) \\ & = & \underbrace{\frac{1}{1-(\frac{x_2}{x_1})^2}}_{k_1}(\epsilon_1+\epsilon_3) & + & \underbrace{\frac{1}{1-(\frac{x_1}{x_2})^2}}_{k_2}(\epsilon_2+\epsilon_3) \end{array}$$

Die Faktoren k_1, k_2 heißen *Fehlerverstärkungsfaktoren*: Sie messen wie sich der Rundungsfehler einer Maschinenoperation im späteren Ergebnis auswirkt. Wir sehen:

- Für $x_1 \approx x_2$ wird k_1, k_2 sehr groß.
- Für $x_1 \ll x_2$ oder $x_1 \gg x_2$ geht einer gegen 0 und einer gegen 1.

(b) $f_2(x_1, x_2) = (x_1 \ominus x_2) \odot (x_1 \oplus x_2)$. Hier erhalten wir

$$\begin{array}{rcccl} u & = & x_1 \ominus x_2 & = & (x_1 - x_2)(1+\epsilon_1) \\ v & = & x_1 \oplus x_2 & = & (x_1 + x_2)(1+\epsilon_2) \\ y = f_2(x_1,x_2) & = & u \odot v & = & (u \cdot v)(1+\epsilon_3) \\ & = & & & ((x_1 - x_2)(1+\epsilon_1)(x_1+x_2)(1+\epsilon_2))(1+\epsilon_3) \\ & = & & & (x_1 - x_2)(x_1+x_2)(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) \\ & \doteq & & & x_1^2 - x_2^2 + (x_1^2 - x_2^2)(\epsilon_1+\epsilon_2+\epsilon_3) \end{array} .$$

Für den relativen Fehler gilt

$$\frac{f_2(x_1,x_2) - \hat{f}_2(x_1,x_2)}{\hat{f}_2(x_1,x_2)} \quad \doteq \epsilon_1 + \epsilon_2 + \epsilon_3 \quad .$$

Hier findet also keine Fehlerverstärkung statt!

Dies liegt daran, dass die gefährlichen $\ominus, \oplus$-Operationen ($a \oplus b = a \ominus (-b)$) zuerst auf die Eingabe angewendet werden. □

Regel 3.6. Setze die potentiell gefährlichen Operationen $\oplus, \ominus$ möglichst früh ein. □

Nun berücksichtigen wir zusätzlich den Fehler in der Eingabe, d.h. $x_i = \mathrm{rd}(\hat{x}_i) = \hat{x}_i(1+\epsilon_{x_i})$

Beispiel 3.7 (Fortsetzung von Bsp 3.5 mit gerundeter Eingabe)**.** Für $f_2(x_1, x_2) = (x_1 \ominus x_2) \odot (x_1 \oplus x_2)$ aus (b) oben:

$$\begin{aligned} & f_2\left(\mathrm{rd}(\hat{x}_1), \mathrm{rd}(\hat{x}_2)\right) \\ & \quad \doteq \left[(\hat{x}_1(1+\epsilon_{x_1}))^2 - (\hat{x}_2(1+\epsilon_{x_2}))^2\right](1+\epsilon_1+\epsilon_2+\epsilon_3) \\ & \quad = \left[\hat{x}_1^2(1+2\epsilon_{x_1}+\epsilon_{x_1}^2) - \hat{x}_2^2(1+2\epsilon_{x_2}+\epsilon_{x_2}^2)\right](1+\epsilon_1+\epsilon_2+\epsilon_3) \\ & \quad \doteq \underbrace{\hat{x}_1^2 - \hat{x}_2^2}_{\hat{y}_2} + (\hat{x}_1^2 - \hat{x}_2^2)(\epsilon_1+\epsilon_2+\epsilon_3) + \hat{x}_1^2 2\epsilon_{x_1} - \hat{x}_2^2 2\epsilon_{x_2} \end{aligned}$$

daraus folgt der relative Fehler:

$$\frac{f_2(\mathrm{rd}(\hat{x}_1), \mathrm{rd}(\hat{x}_2)) - \hat{y}_2}{\hat{y}_2} \doteq \underbrace{\epsilon_1+\epsilon_2+\epsilon_3}_{\text{wie vorher}} + \underbrace{\frac{1}{1-(\frac{\hat{x}_2}{\hat{x}_1})^2}2\epsilon_{x_1} + \frac{1}{1-(\frac{\hat{x}_1}{\hat{x}_2})^2}2\epsilon_{x_2}}_{\text{Verstärkung der Eingabefehler}}$$

Interessanterweise erhält man für f_1 aus (a) oben das Ergebnis:

$$\frac{f_1(\mathrm{rd}(\hat{x}_1), \mathrm{rd}(\hat{x}_2)) - \hat{y}_1}{\hat{y}_1} \doteq \frac{1}{1 - (\frac{\hat{x}_2}{\hat{x}_1})^2}(2\epsilon_{x_1} + \epsilon_1 + \epsilon_3) + \frac{1}{1 - (\frac{\hat{x}_1}{\hat{x}_2})^2}(2\epsilon_{x_2} + \epsilon_2 + \epsilon_3)$$

Wir stellen fest: Unter Berücksichtigung von Eingabefehlern verhalten sich beide Algorithmen gleich (schlecht).

Dies liegt daran, dass die potentiell gefährlichen Operationen $\oplus, \ominus$ schon auf die fehlerbehafteten Operanden angewendet werden. □

3.3 Konditionsanalyse

Definition 3.8 (Konditionsanalyse)**.** Die Konditionsanalyse untersucht die numerische Aufgabe $\hat{y} = \hat{f}(\hat{x})$ auf Sensitivität bezüglich der Eingabedaten. Betrachtet wird also formal die Größe

$$\frac{\hat{f}_i(\hat{x}_1 + \Delta\hat{x}_1, \cdots, \hat{x}_m + \Delta\hat{x}_m) - \hat{f}_i(\hat{x}_1, \cdots, \hat{x}_m)}{\hat{f}_i(\hat{x}_1, \cdots, \hat{x}_m)} \qquad (\hat{f}_i \neq 0)$$

□

Achtung: Bei der Konditionsanalyse wird nur die numerische Aufgabe $\hat{f} : \mathbb{R}^m \to \mathbb{R}^n$ untersucht! Das numerische Berechnungsverfahren f welches $\hat{f}$ in Fließkommaarithmetik approximiert spielt keine Rolle! (Den Zusammenhang stellen wir unten her).

Um Schreibarbeit zu sparen lassen wir das $\hat{}$ bei allen Größen in diesem Abschnitt weg!

Wir benötigen einige Begriffe aus der Analysis.

Es sei also

$$f : \mathbb{R}^m \to \mathbb{R}^n$$

eine *zweimal stetig differenzierbare* Abbildung.

Nach dem Taylorschen Satz im $\mathbb{R}^m$ gilt dann mit $x, \Delta x \in \mathbb{R}^m$:

$$f_i(x + \Delta x) = f_i(x) + \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x)\Delta x_j + R_i^f(x; \Delta x) \qquad i = 1, \ldots n$$

wobei wir für das Restglied R_i^f annehmen, dass

$$R_i^f(x; \Delta x) = O(|\Delta x|^2)$$

wobei $|\Delta x| = \max_{j=1\ldots m} |\Delta x_j|$.

Definition 3.9 (Landausche[2] Symbole)**.** Man schreibt

$$g(t) = O(h(t)) \qquad (t \to 0)$$

[2]Edmund Georg Hermann Landau, 1877-1938, dt. Mathematiker.

falls für alle $t \in (0, t_0]$ (t_0 genügend klein) und einer Konstanten $c \geq 0$ gilt

$$|g(t)| \leq c|h(t)|.$$

Dies ist analog zur O-Notation bei der Komplexitätsanalyse von Algorithmen (nur geht dort $n \to \infty$). Entsprechend bedeutet

$$g(t) = o(h(t)) \qquad (t \to 0),$$

dass für alle $t \in (0, t_0]$ und einer Funktion $c(t)$, $c(t) \to 0$ für $t \to 0$, gilt

$$|g(t)| \leq c(t)|h(t)|.$$

Damit geht $g(t)$ „schneller als“ $h(t)$ gegen Null (falls h gegen 0 geht). □

Oben würde sogar $R_i^f(x; \Delta x) = o(|\Delta x|)$ genügen.

Aus der Taylorformel folgt

$$\Delta y_i := f_i(x + \Delta x) - f_i(x) \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \Delta x_j$$

und damit für den relativen Unterschied

$$\frac{\Delta y_i}{y_i} = \frac{f_i(x + \Delta x) - f_i(x)}{f_i(x)} \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \frac{\Delta x_j}{f_i(x)} = \sum_{j=1}^{m} \underbrace{\frac{\partial f_i}{\partial x_j} \frac{x_j}{f_i(x)}}_{=:k_{ij}(x)} \underbrace{\frac{\Delta x_j}{x_j}}_{\text{relativer Fehler in x}}$$

Definition 3.10 (Kondition)**.** Die Zahlen $k_{ij}(x)$ heißen Konditionszahlen. Die Aufgabe $y = f(x)$ heißt schlecht konditioniert wenn ein $|k_{ij}(x)| \gg 1$ ist, andernfalls gut konditioniert. Bei $|k_{ij}(x)| < 1$ liegt Fehlerdämpfung, bei $|k_{ij}(x)| > 1$ Fehlerverstärkung vor. □

Lemma 3.11 (Kondition der Grundoperationen)**.** Für $y = f(x_1, x_2) = x_1 + x_2$ ergibt sich

$$\begin{aligned} \frac{\Delta y}{y} &= 1 \cdot \frac{x_1}{x_1 + x_2} \cdot \frac{\Delta x_1}{x_1} + 1 \cdot \frac{x_2}{x_1 + x_2} \frac{\Delta x_2}{x_2} \\ &= \underbrace{\frac{1}{1 + \frac{x_2}{x_1}}}_{k_1} \frac{\Delta x_1}{x_1} + \underbrace{\frac{1}{1 + \frac{x_1}{x_2}}}_{k_2} \frac{\Delta x_2}{x_2} \end{aligned}$$

Die Addition ist schlecht konditioniert für $x_1 \approx -x_2$, die Subtraktion für $x_1 \approx x_2$.

Für $y = f(x_1, x_2) = x_1 x_2$ gilt

$$\frac{\Delta y}{y} = x_2 \cdot \frac{x_1}{x_1 x_2} \cdot \frac{\Delta x_1}{x_1} + x_1 \cdot \frac{x_2}{x_1 x_2} \cdot \frac{\Delta x_2}{x_2} = \underbrace{1}_{=k_1} \cdot \frac{\Delta x_1}{x_1} + \underbrace{1}_{=k_2} \cdot \frac{\Delta x_2}{x_2}$$

Die Multiplikation (und die Division) sind gut konditioniert. □

Machen wir noch ein

Beispiel 3.12. Bestimme die Kondition von $f(x_1, x_2) = x_1^2 - x_2^2$.

$$\begin{aligned}\frac{\Delta y}{y} &= 2x_1 \cdot \frac{x_1}{x_1^2 - x_2^2} \cdot \frac{\Delta x_1}{x_1} + (-2x_2) \cdot \frac{x_2}{x_1^2 - x_2^2} \cdot \frac{\Delta x_2}{x_2} \\ &= \underbrace{\frac{2}{1 - (\frac{x_2}{x_1})^2}}_{=k_1} \cdot \frac{\Delta x_1}{x_1} + \underbrace{\frac{2}{1 - (\frac{x_1}{x_2})^2}}_{=k_2} \cdot \frac{\Delta x_2}{x_2}\end{aligned}$$

Vergleich mit Beispiel 3.7 ergibt: k_1, k_2 sind genau die zusätzlichen Verstärkungsfaktoren bezüglich der Eingabe. □

Dieser Fehler lässt sich nicht vermeiden, er ist durch die *Aufgabe* und nicht durch den numerischen Algorithmus gegeben. Das motiviert die folgende Definition.

Definition 3.13 (Stabilität eines numerischen Verfahrens)**.** Ein Verfahren heißt numerisch stabil falls die im Lauf der Rechnung akkumuliert Rundungsfehler (Eingabe $\in \mathbb{F}$!) den durch die Konditionierung der numerischen Aufgabe unvermeidbaren Problemfehler nicht übersteigen. □

Kurz: Liefert die Rundungsfehleranalyse Verstärkungsfaktoren in der gleichen Größe wie die Konditionsanalyse ist alles in Ordnung.

Rundungsfehleranalyse und Konditionsanalyse ergänzen sich also gegenseitig.

Beispiel 3.14 (Anwendung auf 3.5 und 3.12)**.** Sowohl $(x_1 \odot x_1) \ominus (x_2 \odot x_2)$ als auch $(x_1 \ominus x_2) \odot (x_1 \oplus x_2)$ sind stabile Algorithmen zur Berechnung von $x_1^2 - x_2^2$, denn in beiden Fällen hat die Fehlerverstärkung die Form $\frac{1}{1-(\frac{x_1}{x_2})^2}$ bzw. $\frac{1}{1-(\frac{x_2}{x_1})^2}$. □

3.4 Rückwärtsfehleranalyse

Wir haben in diesem Kapitel die sog. *Vorwärtsanalyse* betrieben. Ausgehend von den gerundeten Eingaben $x = \mathrm{rd}(\hat{x})$ und dem mit Rundungsfehlern behafteten numerischen Verfahren f haben wir den Fehler im Ergebnis bestimt:

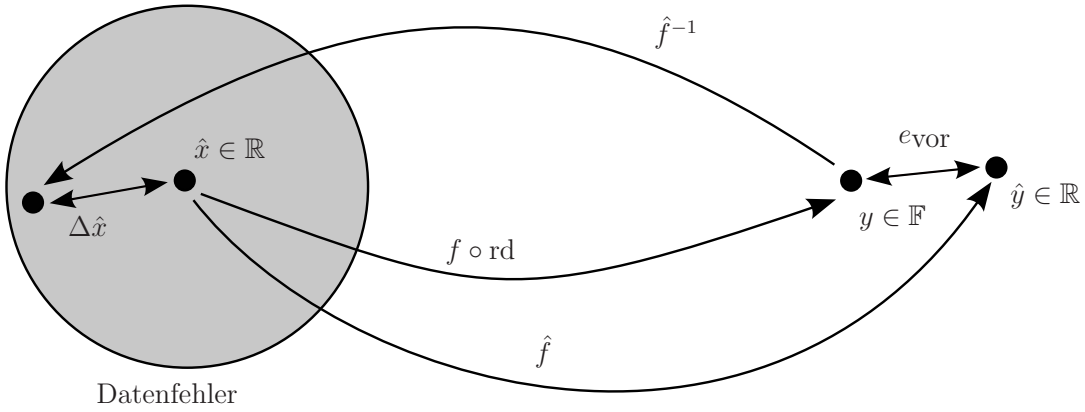
$$e_{\mathrm{vor}} = \hat{f}(\hat{x}) - f(\mathrm{rd}(\hat{x})).$$

Bei der Rückwärtsanalyse versucht man ein $\Delta\hat{x} \in \mathbb{R}^m$ zu finden so dass

$$\hat{f}(\hat{x} + \Delta\hat{x}) = f(\mathrm{rd}(\hat{x})). \tag{3.1}$$

Man stellt also „rückwärts“ die Frage: Welche Eingabe hätte denn mit der *exakten* Berechnung das Ergebnis des numerischen Verfahrens geliefert?

Graphisch wird die Sache klarer:



$\Delta\hat{x}$ erhält man durch Auflösen von (3.1):

$$\Delta\hat{x} = \hat{f}^{-1}(f(\text{rd}(\hat{x}))) - \hat{x} \quad .$$

Kann man zeigen, dass $\|\Delta\hat{x}\|$ in der Größenordnung des Datenfehlers liegt so ist das numerische Verfahren gutartig.

3.5 Zusammenfassung

- *Auslöschung* kann bei der Subtraktion zweier annähernd gleichgroßer *fehlerbehafteter* Zahlen entstehen. Bei Maschinenzahlen ist die Subtraktion bis auf Rundungsfehler genau.
- Diese Aussage erfordert exakt gerundete Maschinenoperationen.
- Bei der *Rundungsfehleranalyse* betrachtet man den Einfluss von Rundungsfehlern die während der numerischen Berechnung auftreten auf das Endergebnis. Fehler in der Eingabe bleiben unberücksichtigt.
- Bei der *Konditionsanalyse* betrachtet man den Einfluss von Fehlern in der Eingabe auf das Ergebnis der numerischen Aufgabe. Hier bleiben die Rundungsfehler der numerischen Berechnung unberücksichtigt.
- Ein numerisches Verfahren heißt *stabil*, wenn die Fehlerverstärkungsfaktoren aus der Rundungsfehleranalyse die aus der Konditionsanalyse nicht übersteigen.

4 Lagrange-Interpolation

4.1 Motivation und Aufgabenstellung

Funktionen, also Abbildungen $f : D \to W$, sind fundamentale Objekte der Mathematik.

Wir sind hier insbesondere an dem Fall $D \subseteq \mathbb{R}, \mathbb{C}$, also *überabzählbarer* Mengen interessiert.

Wie stellt man solche *kontinuierlichen* Funktionen im Rechner dar? Trick: Man *approximiert* f durch

$$f(x) \approx \sum_{i=0}^{n} a_i \varphi_i(x),$$

also mittels einem Satz von gegebenen *Basis*funktionen.

Es müssen nur die $n+1$ *Koeffizienten*, also (Fließkomma-) Zahlen, gespeichert werden.

Natürlich begeht man dabei einen Fehler, den *Approximationsfehler* (zusätzlich zum Rundungsfehler).

Die φ_i wählt man so, dass benötigte Operationen wie Auswertung, Differentiation oder Integration einfach sind.

Hier einige Anwendungen von kontinuierlichen Funktionen in der Informatik:

Kurvendarstellung Z. B. in Zeichenprogrammen, Fonts oder Datenformaten wie Postscript.

Computer Aided Design Darstellung von (dreidimensionalen) Körpern zur Anwendung in Fertigungstechnik oder Simulation.

Simulation Darstellung der Lösung von Differentialgleichungen, siehe z. B. das Pendel in der ersten Vorlesung.

Grafik, Visualisierung (Realistische, interaktive) Darstellung von komplexen Szenen auf dem Bildschirm.

Datenaufbereitung Gemessene Datenpunkte in funktionale Form bringen. Oft hat man viel mehr Datenpunkte als Koeffizienten.

Welche Funktionen mit endlich vielen Parametern nutzt man in der Praxis?

Hier eine kleine Auswahl:

(a) Polynome
$$p(x) = a_0 + a_1 x + \ldots + a_n x^n \quad .$$

(b) Rationale Funktionen
$$r(x) = \frac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_m x^m} \quad .$$

(c) Trigonometrische Polynome
$$t(x) = \frac{1}{2} a_0 + \sum_{k=1}^{n} \left(a_k \cos(kx) + b_k \sin(kx) \right) \quad .$$

(d) Exponentialsumme

$$e(x) = \sum_{k=1}^{n} a_k e^{b_k x} \quad .$$

Die Abbildungen 6 bis 8 zeigen eine Anwendung von Polynomen bei der Kurvenkompression in der Computergraphik.

Die Lage eines starren Körpers im Raum wird durch 6 Zahlen festgelegt (3 für die Position und 3 für die Orientierung), die sich mit der Zeit ändern können. Eine äquidistante Schrittweite erfordert einen hohen Speicheraufwand, um bei schnellen Positionsänderungen eine gute Genauigkeit erreichen zu können. Bei einer adaptiven Schrittweitenwahl werden möglichst wenig Zeitpunkte ausgewählt, aber so, dass ein vorgegebener Fehler nicht überschritten wird. Diese Anwendung haben Eric Schneider, Manuel Jerger und Benjamin Jillich im Rahmen eines Software-Praktikums im Sommersemester 2008 erarbeitet (Vielen Dank für die tollen Bilder!).

Sei nun $f : [a, b] \to \mathbb{R}$ eine gegebene Funktion. Diese soll mit einer Funktion $g(x, a_0, \ldots, a_n)$ mit n+1 Parametern $a_0, \ldots, a_n$ (z.B. g ein Polynom) dargestellt werden.

Definition 4.1 (Interpolation, Approximation)**.** Geschieht die Zuordnung durch Fixieren von Funktionswerten

$$g(x_i) = y_i := f(x_i) \qquad i = 0, \ldots, n$$

an den n+1 *paarweise verschiedenen Stützstellen* $x_i \in [a, b]$, spricht man von *Interpolation.*

Geschieht dies mittels

$$\max_{a \le x \le b} |f(x) - g(x)| \qquad \text{minimal für } g, \text{ oder}$$

$$\int_a^b |f(x) - g(x)|^2 \, dx \qquad \text{minimal für } g, \text{ oder andere Normen,}$$

so spricht man allgemeiner von Approximation.

Interpolation ist natürliche eine spezielle Approximation:

$$\max_{i=0,\ldots,n} |f(x_i) - g(x_i)| \qquad \text{minimal für } g.$$

Wir behandeln hier nur die Interpolation und (fast) nur in einer Raumdimension. □

Einen kleinen Ausflug in die Approximation wollen wir hier doch machen.

Die Taylorreihentwicklung[3] einer (genügend oft differenzierbaren Funktion) lautet für $x = x_0 + \Delta x$:

$$f(x) = f(x_0) + f'(x_0) \underbrace{(x - x_0)}_{\Delta x} + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x - x_0)^{n+1}.$$

Lässt man das *Restglied* fort, so erhält man die Approximation durch ein Polynom:

[3] Brook Taylor, 1685-1731, brit. Mathematiker.

Abbildung 6: Kurvenkompression in der Computergraphik: Die Szene.

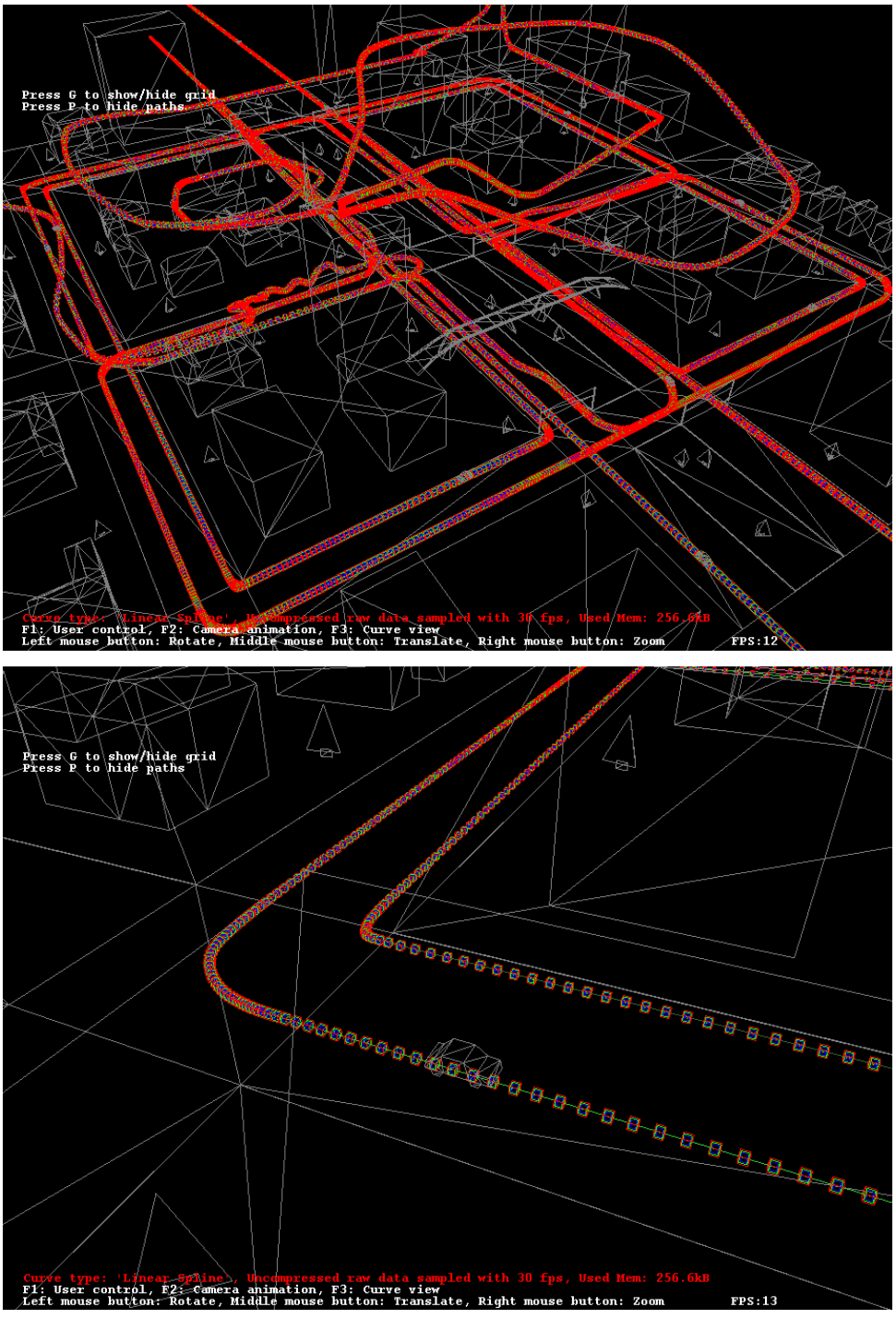


Abbildung 7: Kurvenkompression in der Computergraphik: Stützpunkte der unkomprimierten Kurve.

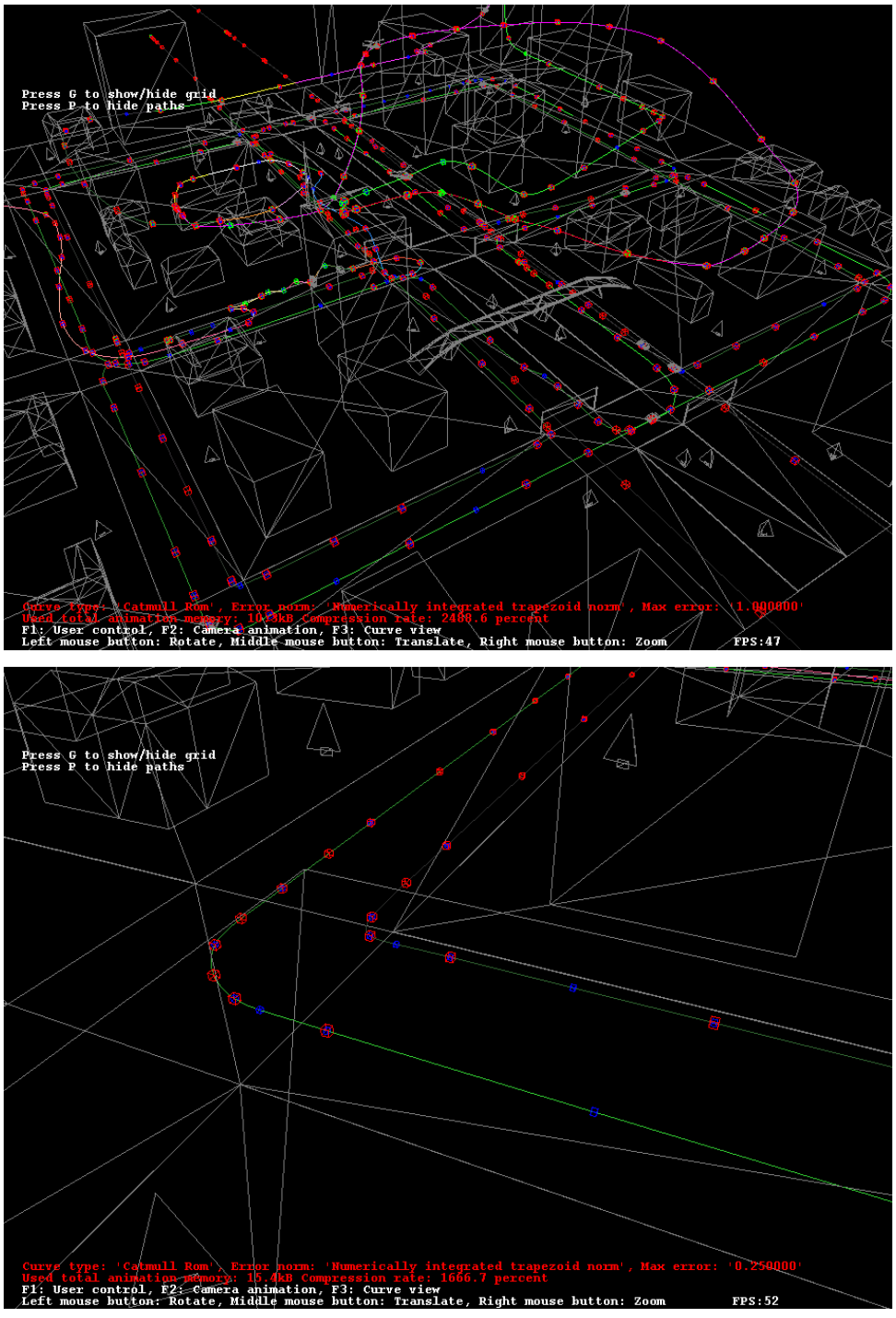


Abbildung 8: Kurvenkompression in der Computergraphik: Stützpunkte der komprimierten Kurve.

$$f(x) \approx p(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

Die Koeffizienten involvieren die Ableitungen von f am Punkt x_0.

Der Approximationsfehler entspricht gerade dem Restglied

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x - x_0)^{n+1} \qquad \text{für ein } \xi_x \in [a, b].$$

Falls $f^{(n+1)}(\xi) \leq M$, $\forall \xi \in [a, b]$ und alle n, kann man den Fehler für $n \to \infty$ beliebig klein machen auf $[a, b]$.

4.2 Polynome

Wenden wir uns nun der Interpolationsaufgabe mit Polynomen zu, d. h. wir suchen Koeffizienten $a_0, \ldots, a_n$ zu bestimmen, sodass

$$a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n = y_i := f(x_i), \qquad i = 0, \ldots, n,$$

für die paarweise verschiedenen Stützstellen x_i.

Schreibt man die Bedingungen für alle $i = 0, \ldots, n$ untereinander, erhält man ein lineares Gleichungssystem für die Koeffizienten a_i:

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad .$$

Diese Matrix heisst *Vandermondesche*[4] *Matrix*. Zu zeigen ist noch, dass diese Matrix regulär ist (für paarweise verschiedene x_i).

Bestimmung der Lösung des linearen Gleichungssystems erfordert $O(n^3)$ arithmetische Operationen. Unten werden wir geschicktere Arten zur Aufstellung des Interpolationspolynoms kennenlernen.

Zudem zeigt sich, dass die Vandermondesche Matrix „schwer“ zu lösen ist (in welchem Sinne, das kommt später).

Die Menge aller Polynome vom Grad kleiner gleich n über dem Körper $\mathbb{R}$ (geht auch über $\mathbb{C}$) lautet

$$P_n := \{p(x) = a_0 + a_1 x + \ldots + a_n x^n \,|\, a_i \in \mathbb{R},\ i \in 0, \ldots, n\}.$$

P_n ist ein $n + 1$-dimensionaler Vektorraum (über $\mathbb{R}$), d. h. man kann Polynome addieren und skalar multiplizieren.

P_n ist auch ein *Funktionenraum*, da die Elemente der Menge Funktionen sind.

[4] Alexandre-Théophile Vandermonde, 1735-1796, frz. Mathematiker.

Jedes Polynom $p(x) \in P_n$ kann durch einen Satz von $n+1$ linear unabhängigen Basispolynomen aus $\Phi_n = \{\varphi_0(x), \varphi_1(x), \ldots, \varphi_n(x)\}$ dargestellt werden:

$$p(x) = \sum_{i=0}^{n} \beta_i \varphi_i(x).$$

Die Wahl der Basispolynome ist beliebig (Voraussetzung: linear unabhängig). Oben haben wir die sog. *Monombasis* gewählt:

$$M_n = \left\{1, x, x^2, \ldots, x^n\right\} \qquad (x^k \text{ heißt } k\text{-tes Monom}).$$

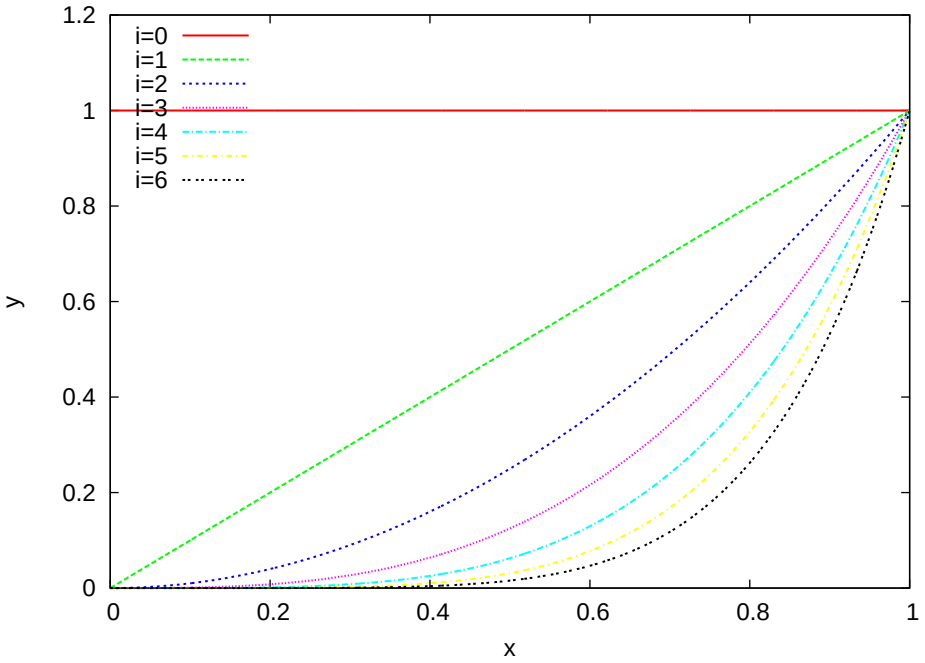


Abbildung 9: Die Monome bis zum Grad 6.

Die Abbildung 9 zeigt die Monome bis zum Grad 6.

Die Interpolationsaufgabe in beliebiger Basis lautet nun:

$$p(x_i) = \sum_{j=0}^{n} \beta_j \varphi_j(x_i) = y_i, \qquad i = 0, \ldots, n,$$

und liefert wieder ein lineares Gleichungssystem der Dimension $n+1$ nun für die Koeffizienten $\beta_0, \ldots, \beta_n$:

$$\begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \cdots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} .$$

Durch eine geschickte Wahl der Basispolynome kann man nun dafür sorgen, dass das lineare Gleichungssystem mit weniger Aufwand lösbar ist.

Geschickt wäre etwa ein dreieckförmiges oder gar diagonales System.

4.3 Lagrange-Interpolation

Definition 4.2 (Lagrange[5]-Polynome)**.** Man definiert die *Lagrangen Basispolynome* vom Grad n

$$L_i^{(n)}(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \qquad i = 0, \ldots, n \tag{4.1}$$

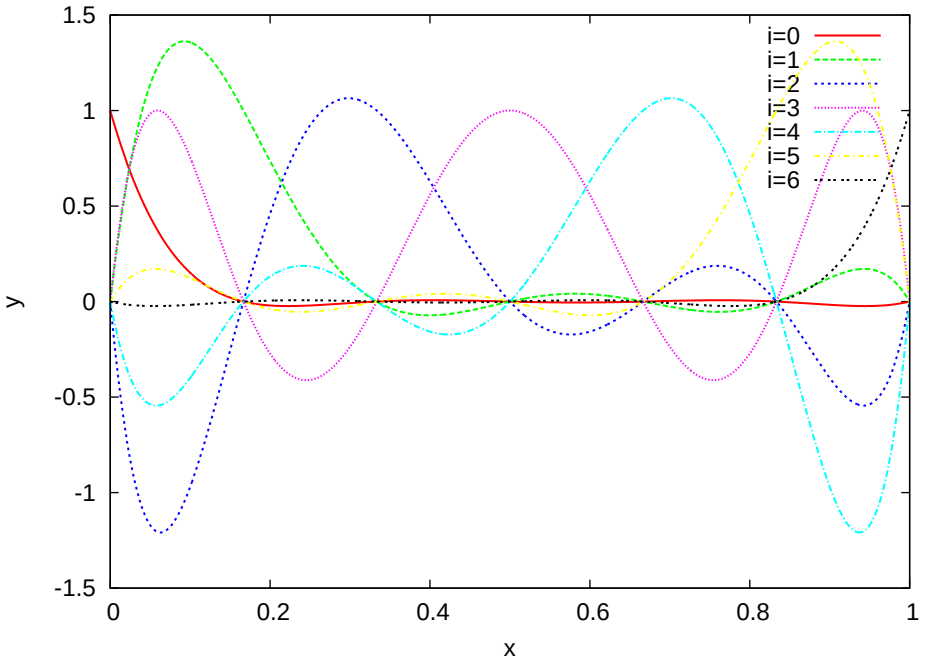


Abbildung 10: Die Lagrange-Polynome $L_i^{(6)}(x)$ vom Grad 6.

Abbildung 10 zeigt die Lagrange-Polynome vom Grad 6 bei äquidistanten Stützstellen auf $[0, 1]$.

Die Lagrange-Polynome haben die folgenden Eigenschaften:

(a) $L_i^{(n)}(x) \in P_n$, denn $\prod_{j=0, j\neq i}^{n}(x - x_j)$ ist ein Polynom vom Grad n.

(b) Es ist $L_i^{(n)}(x_k) = \delta_{ik} = \begin{cases} 1 & \text{für } i = k \\ 0 & \text{sonst} \end{cases}$, ($\delta_{ik}$ heißt Kronecker-Symbol[6]), denn für $k = i$

[5] Joseph Louis de Lagrange, 1736-1813, frz. Mathematiker.
[6] Leopold Kronecker, 1823-1891, dt. Mathematiker

gilt

$$L_i^{(n)}(x_k) = L_i^{(n)}(x_i) = \prod_{j=0, j\neq i}^{n} \frac{x_i - x_j}{x_i - x_j} = 1.$$

Für $k \neq i$ enthält das Produkt $\prod_{j=0,j\neq i}^{n}$ für $j = k$ den Faktor $\frac{x_k - x_k}{x_i - x_k} = 0$ und damit ist das ganze Produkt Null.

(c) Die $L_i^{(n)}$ bilden eine Basis von P_n.

Allgemein heißt ein Polynom $\varphi_i \neq 0$ linear abhängig von den Polynomen φ_k, $k \neq i$, falls es Koeffizienten β_k gibt, sodass $\varphi_i = \sum_{k\neq i} \beta_k \varphi_k$.

Für $\varphi_i = L_i^{(n)}$ kann dies aber nicht sein, denn es ist $L_i^{(n)}(x_i) = 1$ und $L_k^{(n)}(x_i) = 0$ für $k \neq i$. Die $L_i^{(n)}$ sind also linear unabhängig, es gibt $n+1$ Stück davon, sie bilden also eine Basis von P_n. □

Mit den Lagrange-Polynomen ist die Interpolationsaufgabe ganz simpel zu lösen. Man setzt

$$p(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x).$$

Wegen $L_i^{(n)}(x_k) = \delta_{ik}$ gilt dann $p(x_i) = y_i$.

Oder anders: Das lineare Gleichungssystem zur Interpolationsaufgabe in der Lagrange-Basis ist die Einheitsmatrix!

Beispiel 4.3. Zu interpolieren sei die folgende Wertetabelle mit 4 Einträgen:

x_i	y_i
0	1.0000
2	0.4546
7	0.0938
10	−0.0544

Abbildung 11 zeigt das zugehörige Interpolationspolynom sowie die skalierten Lagrange-Polynome $y_i L_i^{(3)}$. □

Satz 4.4 (Eindeutige Lösbarkeit der Polynominterpolation)**.** Zu der Tabelle (x_i, y_i), $i = 0, \ldots, n$, $x_i \neq x_j$ für $i \neq j$, gibt es genau ein $p \in P_n$, sodass $p(x_i) = y_i, \quad i = 0, \ldots, n$.

Beweis: Die Lagrange-Polynome bilden eine Basis von P_n. Daher gibt es genau eine Darstellung eines Polynomes zu dieser Basis □

Man kann die Eindeutigkeit der Polynominterpolation auch ohne Kenntnis einer Basis zeigen ([Sto05, S. 43]): Angenommen es gäbe zu $p \in P_n$ noch ein weiteres, von p verschiedenes $q \in P_n$, dann ist $r = p - q \in P_n$ und r hat die $n+1$ Nullstellen x_i. Nun hat ein Polynom vom Grad n aber höchstens n Nullstellen (Gaußscher Fundamentalsatz der Algebra) und somit muss $r \equiv 0$ sein. Das ist aber ein Widerspruch zu $p \neq q$.

Eine Folgerung hieraus ist, dass auch die Vandermondesche Matrix invertierbar ist (falls $x_i \neq x_j$). Das Interpolationspolynom ist immer das gleiche, es ist nur in einer anderen Basis dargestellt.

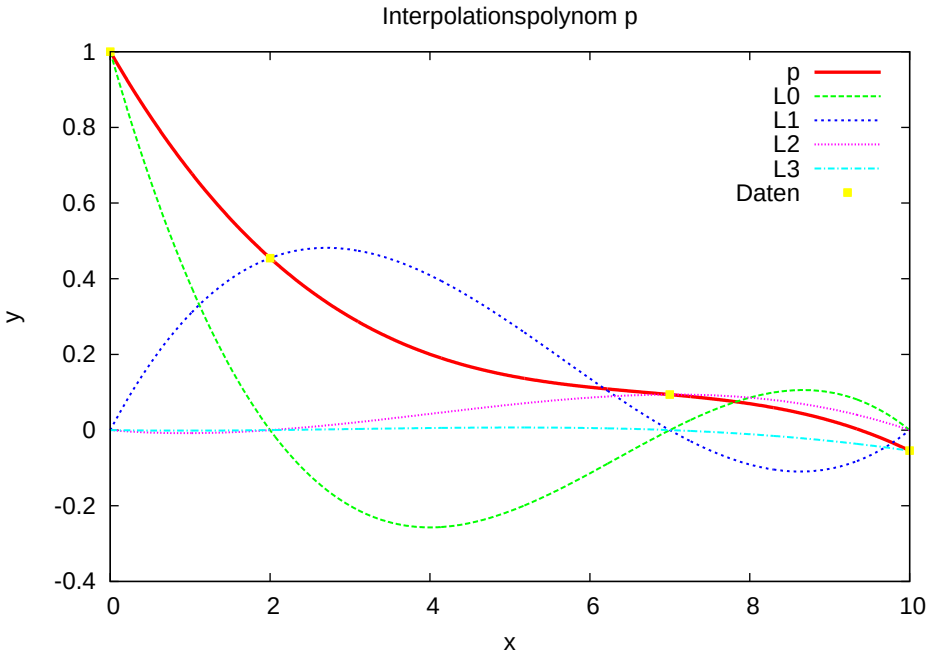


Abbildung 11: Interpolationspolynom zur Wertetabelle aus Beispiel 4.3.

4.4 Fehlerabschätzung

Der Fehler in den Interpolationspunkten $y_i - p(x_i)$ ist natürlich Null. Gilt $y_i = f(x_i)$ so liegt es nahe zu fragen, welchen Wert

$$e(x) = f(x) - p(x)$$

für $x \neq x_i$ annehmen kann?

In Beispiel 4.3 war $f(x) = \sin x/x$. Abbildung 12 zeigt den dabei gemachten Interpolationsfehler.

Es gilt folgender

Satz 4.5 (Interpolationsfehler)**.** Sei $f(x)$ n+1 mal stetig differenzierbar auf $[a, b]$, wobei $a = x_0 < x_1 < \ldots < x_n = b$. Dann gibt es zu jedem $x \in$ [a,b] ein $\xi_x \in (a, b)$, sodass

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

Beweis: Für $x \in \{x_0, \ldots, x_n\}$ liefert die Formel offensichtlich 0.

Wähle nun $x \in [a, b] \setminus \{x_0, \ldots, x_n\}$. Zu diesem x definiere die Funktion

$$F_x(t) = f(t) - p(t) - \underbrace{\frac{f(x) - p(x)}{l(x)}}_{\text{ist eine Zahl}} l(t) \qquad \text{mit } l(t) = \prod_{j=0}^{n} (t - x_j)$$

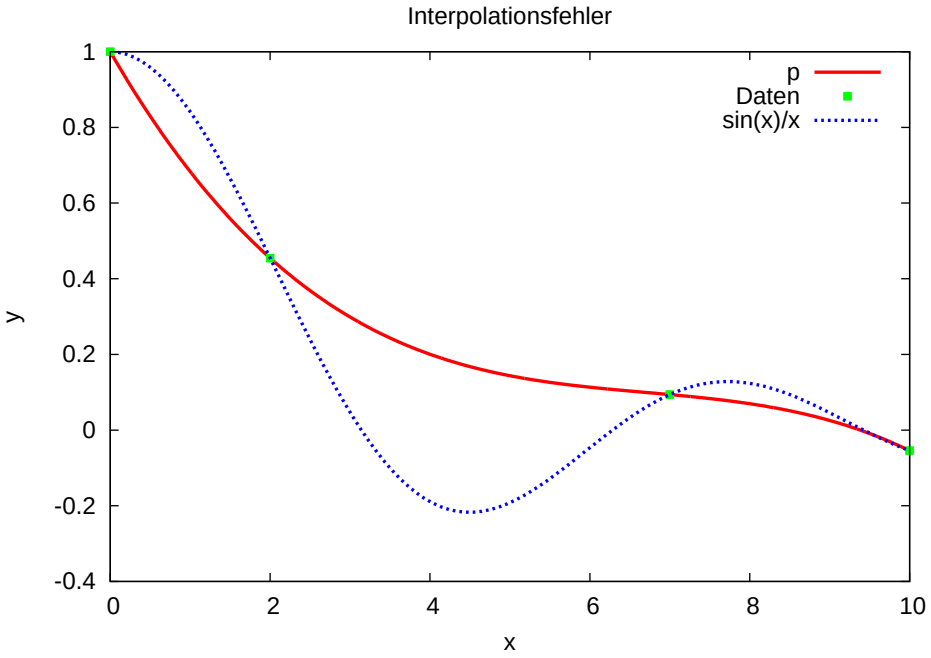


Abbildung 12: Illustration des Interpolationsfehlers.

$F_x(t)$ hat mindestens die n+2 Nullstellen $\{x_0, \ldots, x_n, x\}$, denn für x_i, $i = 0, \ldots, n$ ist

$$F_x(x_i) = \underbrace{f(x_i) - p(x_i)}_{=0} - \frac{f(x) - p(x)}{l(x)} \underbrace{l(x_i)}_{=0}$$

und für x gilt $F_x(x) = f(x) - p(x) - \frac{f(x)-p(x)}{l(x)} l(x) = 0$.

Der *Satz von Rolle*[7] sagt: $u(x)$ in $[a, b]$ stetig und in (a, b) differenzierbar sowie $u(a) = u(b) = 0$, dann gib es mindestens ein $x_0 \in (a, b)$ mit $u'(x) = 0$.

$\Rightarrow F_x^{(1)}(t)$ hat mindestens $n + 1$ Nullstellen, $F_x^{(2)}(t)$ hat mind. n Nullstellen,$\ldots, F_x^{(n+1)}(t)$ hat mind. eine Nullstelle. Diese Nullstelle sei ξ_x.

Für diese Nullstelle gilt dann

$$\begin{aligned} F_x^{(n+1)}(\xi_x) &= f^{(n+1)}(\xi_x) - \underbrace{p^{(n+1)}(\xi_x)}_{=0 \text{ da Grad } n} - \frac{f(x) - p(x)}{l(x)} \underbrace{l^{(n+1)}(\xi_x)}_{l(t)=t^{n+1}+\ldots} \\ &= f^{(n+1)}(\xi_x) - \frac{f(x) - p(x)}{l(x)} (n+1)! \\ &\stackrel{!}{=} 0 \quad . \end{aligned}$$

[7] Michel Rolle, 1652-1719, frz. Mathematiker.

Hier haben wir ausgenutzt, dass $l^{(n+1)} = \frac{d^{n+1}}{dt^{n+1}} t^{n+1} = (n+1)!$.

Schließlich liefert Auflösen nach $f(x) - p(x)$:

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

□

Man vergleiche das recht ähnliche Resultat bei der *Approximation* mit dem Taylorpolynom.

Wir nehmen an, die $n+1$-te Ableitungen der Funktion f sei beschränkt, weiter sei $x_{i+1} - x_i = h$ (äquidistant). Dann gilt

$$\begin{aligned} |f(x) - p(x)| &= \frac{\left|f^{(n+1)}(\xi_x)\right|}{(n+1)!} \underbrace{\prod_{j=0}^{n} |x - x_j|}_{\leq h \cdot h \cdot 2h \cdots nh} \\ &\leq \sup_{\xi \in (a,b)} \left|f^{(n+1)}(\xi)\right| \frac{1}{(n+1)!} h^{n+1} n! \\ &= \frac{M}{n+1} h^{n+1} \end{aligned}$$

Geht man jetzt bei *gleichem* n von $[a, b]$ zu $[a, (a+b)/2]$, so halbiert sich der Abstand, also $h' = h/2$ und der Fehler reduziert sich um $(1/2)^{n+1}$.

Lässt man das Intervall $[a, b]$ gleich und halbiert den Abstand ($n' = 2n$), so reduziert sich der Fehler sogar um mehr als $(1/2)^{2n+1}$.

Allerdings verdoppelt sich dann auch n und damit muss man die Beschränktheit der Ableitung $f^{(2n+1)}$ fordern.

Beispiel 4.6. Wir interpolieren die Funktionen $\sin(x)$ und $\sin(2x)$ im Intervall $[0, 2\pi]$ mit äquidistanten Stützstellen durch ein Polynom vom Grad n.

Mittels Kettenregel rechnet man nach

$$\frac{d^m}{dx^m} \sin(kx) = k^m \cdot (-1)^{(m/2)} \cdot \begin{cases} \sin(kx) & m \text{ gerade} \\ \cos(kx) & m \text{ ungerade} \end{cases}$$

(ganzzahlige Division in $m/2$!).

Somit gilt $\sup_{\xi \in [0,2\pi]} \left|\frac{d^m}{dx^m} \sin(kx)\right| \leq |k|^m$.

Berücksichtigt man $h = 2\pi/n$ und $k = 1$, bzw. $k = 2$ so erhalten wir mit der Abschätzung von oben

$$\begin{aligned} |\sin(x) - p(x)| &\leq \left(\frac{2\pi}{n}\right)^{n+1} \frac{1}{n+1}, \\ |\sin(2x) - p(x)| &\leq 2^{n+1} \left(\frac{2\pi}{n}\right)^{n+1} \frac{1}{n+1}. \end{aligned}$$

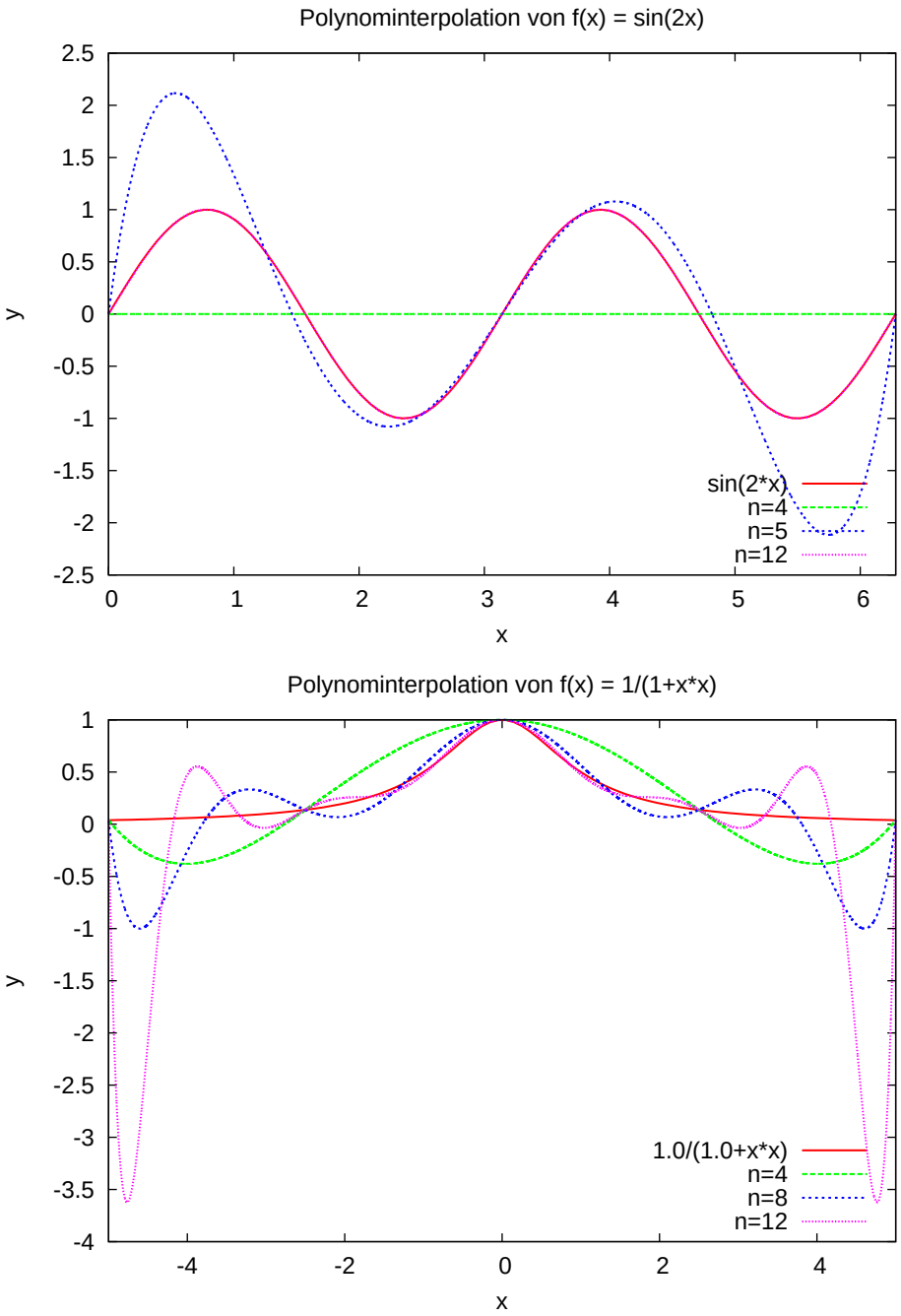


Abbildung 13: Interpolation der Funktionen $\sin(2x)$ (oben) und $\frac{1}{1+x^2}$ (unten) mit äquidistanten Stützstellen und verschiedenen Polynomgraden.

Numerisch erhält man die folgenden Werte:

n	$\max_{x\in[0,2\pi]} \lvert\sin(x) - p(x)\rvert$	$\max_{x\in[0,2\pi]} \lvert\sin(2x) - p(x)\rvert$
5	$2.67 \cdot 10^{-2}$	$1.29 \cdot 10^{0}$
6	$1.88 \cdot 10^{-2}$	$8.13 \cdot 10^{-1}$
7	$1.69 \cdot 10^{-3}$	$4.22 \cdot 10^{-1}$
8	$1.20 \cdot 10^{-3}$	$2.82 \cdot 10^{-1}$
9	$7.22 \cdot 10^{-5}$	$8.36 \cdot 10^{-2}$
10	$5.16 \cdot 10^{-5}$	$5.75 \cdot 10^{-2}$
11	$2.21 \cdot 10^{-6}$	$1.13 \cdot 10^{-2}$
12	$1.58 \cdot 10^{-6}$	$7.91 \cdot 10^{-3}$
13	$5.11 \cdot 10^{-8}$	$1.12 \cdot 10^{-3}$
14	$3.68 \cdot 10^{-8}$	$7.93 \cdot 10^{-4}$
15	$9.21 \cdot 10^{-10}$	$8.54 \cdot 10^{-5}$
16	$6.65 \cdot 10^{-10}$	$6.07 \cdot 10^{-5}$
17	$1.33 \cdot 10^{-11}$	$5.15 \cdot 10^{-6}$
18	$9.64 \cdot 10^{-12}$	$3.68 \cdot 10^{-6}$

Den Faktor 2^{n+1} kann man in der Tabelle klar erkennen. □

Im allgemeinen wachsen aber die n-ten Ableitungen zu schnell und der Fehler kann mit steigendem n sogar größer werden.

So erwähnt [Ran06] die Funktion $f(x) = \frac{1}{1+x^2}$, für die gilt

$$\left|f^{(n)}(x)\right| \approx 2^n n!\, O\left(\frac{|x|^n}{|1+x^2|^{n+1}}\right) \qquad n \to \infty.$$

Für $n \to \infty$ wächst die Ableitung für festes x immer stärker. Wollte man die Ableitung unter einer Schranke M hallten müsste man das Intervall immer kleiner machen. Deshalb ist für festes Intervall die Konvergenz nicht mehr gleichmäßig.

Die Beobachtungen in den Beispielen führen zu folgender

Regel 4.7 (Methoden hoher Ordnung)**.** Je höher der verwendete Polynomgrad in der Lagrange-Interpolation ist, desto mehr Ableitungen der zu interpolierenden Funktion müssen existieren und sie sollten nicht allzu groß sein. □

Da viele der im weiteren Verlauf der Vorlesung behandelten Verfahren auf Polynomen aufbauen, werden wir ähnlich formulierten Regeln noch öfters begegnen.

4.5 Kondition

Wir betrachten die Empfindlichkeit der Interpolationsaufgabe gegenüber den vorgegebenen Stützwerten y_i.

Nach Lagrange gilt:

$$p(x; y_0, \ldots, y_n) = \sum_{i=0}^{n} y_i L_i(x)$$

Damit gilt für ein Δy_i nach Einsetzen

$$\frac{p(x; y_0, \ldots, y_i + \Delta y_i, \ldots, y_n) - p(x; y_0, \ldots, y_i, \ldots, y_n)}{y_i} = L_i^{(n)}(x)\frac{\Delta y_i}{y_i}$$

da p linear in den y_i. Der Verstärkungsfaktor ist also gerade $L_i^{(n)}(x)$.

Für großes n können die Werte von $L_i^{(n)}$ sehr groß werden, insbesondere weit weg von x_i.

$\Rightarrow$ für n größer etwa 8 ist die Polynominterpolation sehr schlecht konditioniert.

Abbildung 14 illustriert das Wachsen der Lagrange-Polynome weit weg von der Stützstelle x an der $L_i^{(n)(x)} = 1$ gilt.

4.6 Horner Schema

Der Vollständigkeit halber sei noch die numerisch stabile *Auswertung* von Polynomen erwähnt.

Für $n = 3$ könnte man den Wert des Polynoms an der Stelle x folgendermaßen ausrechnen:

$$\begin{aligned} p(x) &= a_3x^3 + a_2x^2 + a_1x + a_0 \\ &= (a_3x + a_2)x^2 + a_1x + a_0 \\ &= ((a_3x + a_2)x + a_1)x + a_0 \quad . \end{aligned}$$

Allgemein erhalten wir die folgende Rekursion zur Bestimmung von $p(x)$

$$b_n = a_n; \qquad b_k = a_k + xb_{k+1} \qquad k = n-1, \ldots, 0; \qquad p(x) = b_0.$$

Dies nennt man das *Horner*[8] *Schema*.

4.7 Anwendung: Numerische Differentation

Interpolationspolynome kann man benutzen, um Ableitungen von tabellarisch gegebenen Funktionen $(x_i; y_i = f(x_i))$ zu berechnen.

Ebenso kann man damit die Ableitung von analytisch gegebenen Funktionen näherungsweise bestimmen.

Dazu betrachten wir die Lagrange-Interpolation näher:

$$P_n(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x); \quad L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)} = \underbrace{\left(\prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{(x_i - x_j)}\right)}_{\lambda_i \in \mathbb{R}} x^n + \underbrace{\ldots}_{x^{n-1}} \ldots$$

[8] William George Horner, 1786-1837, brit. Mathematiker.

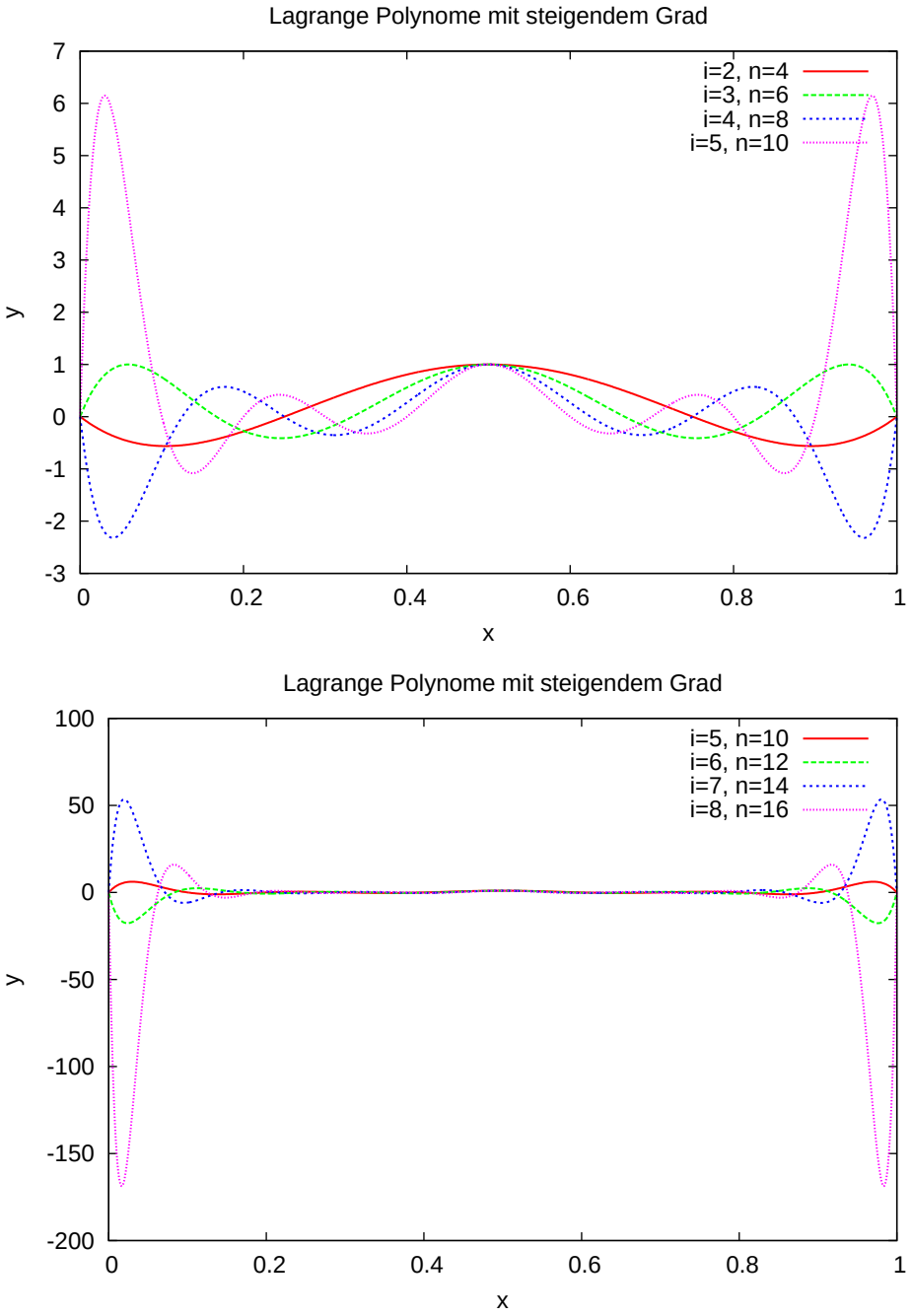


Abbildung 14: Die Lagrange-Polynome $L_{n/2}^{(n)}$ für $n = 4, 6, 8, 10, 12, 14, 16$.

Wenn wir nun das i-te Lagrange-Polynom n-mal nach x ableiten ergibt sich

$$\frac{d^n}{dx^n} L_i^{(n)}(x) = \lambda_i n!$$

und damit

$$\frac{d^n P_n(x)}{dx^n} = \sum_{i=0}^{n} y_i \, \lambda_i \, n! \approx f^{(n)}(x)$$

Wegen obiger Betrachtungen nehmen wir an, n ist nicht groß und die Stützstellen sind nahe zusammen.

Auskunft über den möglichen Fehler liefert folgender

Satz 4.8. Sei $f(x) \in C^n[a,b]$ mit $a = \min_i x_i, b = \max_i x_i$ Dann existiert ein $\xi \in (a,b)$ so dass

$$f^{(n)}(\xi) = \sum_{i=0}^{n} y_i \, \lambda_i \, n! \quad .$$

Beweis: Betrachte die Funktion $g(x) = f(x) - P_n(x)$. $g(x)$ hat mindestens die n+1 Nullstellen $\{x_0, x_1, \ldots, x_n\}$. Satz von Rolle liefert bei n-maliger Anwendung: g' hat n Nullstellen, g'' hat $n-1$ Nullstellen, $g^{(3)}$ hat $n-2$ Nullstellen, $\ldots g^{(n)}$ hat $n-(n-1) = 1$ Nullstelle. Diese liegt in (a,b) und wir nennen sie ξ. Für ξ gilt:

$$g^{(n)}(\xi) = f^{(n)}(\xi) - \sum_{i=0}^{n} y_i \, \lambda_i \, n! = 0 \quad .$$

Damit folgt die Behauptung. □

Im folgenden setzen wir äquidistante Stützstellen voraus, um einfachere Formeln für die Ableitungen zu erhalten. Dann gilt

$$\begin{aligned} \lambda_i &= \frac{1}{\underbrace{(x_i - x_0)\ldots(x_i - x_{i-1})}_{i \text{ Stück, positiv}} \underbrace{(x_i - x_{i+1})\ldots(x_i - x_n)}_{(n-i) \text{ Stück, negativ}}} \\ &= \frac{1}{h^n (-1)^{n-i} i!(n-i)!} = \frac{(-1)^{n-i}}{h^n n!} \binom{n}{i} \end{aligned}$$

und somit

$$f^{(n)}(x) \approx \underbrace{\frac{1}{h^n} \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} y_i}_{n\text{-ter Differenzenquotient}}$$

speziell gilt damit

$$f^{(1)}(x) = \frac{y_1 - y_0}{h}, f^{(2)}(x) \approx \frac{y_2 - 2y_1 + y_0}{h^2}, f^{(3)}(x) \approx \frac{y_3 - 3y_2 + 3y_1 - y_0}{h^3}$$

Bisher: n-te Ableitung aus Polynom vom Grad n. Man kann auch die m-te Ableitung aus einem Interpolationspolynom vom Grad $n > m$ ausrechnen. Der Wert hängt dann allerdings von der Auswertestelle ab (Warum oben nicht?).

Beispiel: Erste Ableitung ($m = 1$) aus Polynomgrad $n = 2$ also 3 Punkten. $x_i - x_{i-1} = h$ sei wieder äquidistant:

$$\begin{aligned} P_2(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= \frac{1}{2h^2}\big(y_0(x-x_1)(x-x_2) - 2y_1(x-x_0)(x-x_2) \\ &\quad + y_2(x-x_0)(x-x_1)\big) \quad . \end{aligned}$$

Ableiten liefert

$$P_2'(x) = \frac{1}{2h^2}\left(y_0(2x-x_1-x_2) - 2y_1(2x-x_0-x_2) + y_2(2x-x_0-x_1)\right) \quad .$$

Wertet man in der Mitte an x_1 aus, so erhält man

$$f'(x_1) \approx P_2'(x_1) = \frac{y_2 - y_0}{2h} \qquad \text{„zentraler Differenzenquotient“.}$$

Beispiel 4.9 (Zur numerischen Differentation)**.** Wir wollen die zweite Ableitung von $f(x) = \sinh(x)$ für $x = 0.6$ mit dem zweiten Differenzenquotient ermitteln:

$$\frac{d^2}{dx^2}\sinh(x) \approx \frac{\sinh(x+h) - 2\sinh(x) + \sinh(x-h)}{h^2}$$

zur Erinnerung:

$$\begin{aligned} \sinh(x) &= \frac{1}{2}(e^x - e^{-x}), \\ \frac{d}{dx}\sinh &= \cosh = \frac{1}{2}(e^x + e^{-x}), \\ \frac{d^2}{dx^2}\sinh(x) &= \sinh(x). \end{aligned}$$

Mit `double` Genauigkeit erhält man den Wert

$$\sinh(0.6) = 6,366535821482 \cdot 10^{-1}.$$

Dagegen liefert die numerische Differentiation die folgende Tabelle

h	Differenzenquotient		
$1 \cdot 10^{-1}$	6.3̲71	$\cdot 10^{-1}$	
$1 \cdot 10^{-2}$	6.366̲5888	$\cdot 10^{-1}$	
$1 \cdot 10^{-3}$	6.3665̲36352	$\cdot 10^{-1}$	
$1 \cdot 10^{-4}$	6.366535̲8540	$\cdot 10^{-1}$	Auslöschung
$1 \cdot 10^{-5}$	6.366̲5017	$\cdot 10^{-1}$	
$1 \cdot 10^{-6}$	6.3̲671	$\cdot 10^{-1}$	
$\vdots$			
$1 \cdot 10^{-10}$	1.1102	$\cdot 10^{4}$	!

Numerische Differentation ist sehr anfällig gegenüber Rundungsfehlern. Mögliche Abhilfe bietet die „Extrapolation“ (siehe Übung). □

4.8 Zusammenfassung

- Funktionen stellt man im Rechner als Linearkombination bekannter Basisfunktionen dar. Als Basisfunktionen benutzt man beispielsweise Polynome.
- Für die Bestimmung eines Polynoms, welches durch eine gegebene Menge von Datenpunkten geht, eignen sich besonders die Lagrangen Basispolynome, da sie eine direkte Angabe des Interpolationspolynoms ohne Lösen eines Gleichungssystems erlauben.
- Die Qualität der Interpolation einer gegebenen Funktion hängt von der Größe der höheren Ableitungen dieser Funktion ab. Wachsen diese sehr schnell, so sind Polynome hohen Grades (also viele Stützstellen) zu vermeiden.
- Mittels Polynomen kann man auch Formeln zur numerischen Differentiation einer Funktion herleiten. Hier ist insbesondere auf die Auswirkung von Rundungsfehlern zu achten.

5 Newton-Interpolation und Bernstein-Interpolation

5.1 Newton-Interpolation

Die Interpolation mit Lagrange-Polynomen hat einen entscheidenden Nachteil:

Verändert man die Anzahl der Stützstellen, so erhält man völlig neue Lagrange-Polynome. Lagrange-Polynome eignen sich also nicht zu einer inkrementellen Konstruktion des Interpolationspolynoms.

Die Lösung hat bereits Newton[9] gefunden, deswegen heißen die im folgenden eingeführten Basispolynome auch Newton-Polynome.

Definition 5.1 (Newton-Polynome)**.** Sei (x_i, y_i), $i = 0, \ldots, n$ eine Interpolationstabelle. Die Newton-Polynome sind gegeben durch

$$N_0(x) = 1 \quad \text{und} \quad N_i(x) = \prod_{j=0}^{i-1}(x - x_j) \quad i = 1, 2, \ldots, n.$$

□

Beispiel: $n = 2 : N_0(x) = 1, N_1(x) = (x - x_0), N_2(x) = (x - x_0)(x - x_1)$.

Die Newton-Polynome erlauben die rekursive Darstellung

$$N_0(x) = 1, \quad N_i(x) = (x - x_{i-1})N_{i-1}(x)$$

Damit gilt

$$N_i(x_k) = 0 \quad \text{für alle } i > k$$

denn es ist $N_{k+1}(x_k) = (x_k - x_k)N_k(x) = 0$ und der Term $(x - x_k)$ kommt in allen $N_i(x)$ mit $i \geq k + 1$, also $i > k$, vor.

Stellt man die Interpolationsaufgabe in der Newton-Basis

$$p(x) = \sum_{j=0}^{n} a_j N_j(x); \quad p(x_i) = y_i \quad i = 0, \ldots, n$$

so ergibt sich

$$\begin{aligned}
y_0 = p(x_0) &= a_0 \cdot 1 \quad (+0 \text{ da } N_i(x) = 0 \text{ für } i > 0) \\
&\Rightarrow a_0 = y_0 \\
y_1 = p(x_1) &= a_0 + a_1(x_1 - x_0) + 0 \\
&\Rightarrow a_1 = \frac{y_1 - a_0}{x_1 - x_0} \\
y_2 = p(x_2) &= a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_1)(x_2 - x_0) \\
&\Rightarrow a_2 = \frac{y_2 - a_1(x_2 - x_0) - a_0}{(x_2 - x_1)(x_2 - x_0)}
\end{aligned}$$

[9] Sir Isaac Newton, 1643-1727, engl. Physiker und Mathematiker.

usw.

Das LGS in der Newton-Basis hat also untere Dreiecksgestalt.

Hinzufügen eines Punktes $n+1$ ändert $a_0, \ldots, a_n$ nicht!

In der Praxis bestimmt man die a_i auf numerisch stabilere Weise:

Satz 5.2 (Dividierte Differenzen)**.** Das Interpolationspolynom in der Newton-Basis ist

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

wobei die Koeffizienten über die rekursive Darstellung definiert sind:

$$\forall i = 0, \ldots, n \qquad y[x_i] := y_i$$

$$\forall k = 1, \ldots, n-i \qquad y[x_i, \ldots, x_{i+k}] := \frac{\overbrace{y[x_{i+1}, \ldots, x_{i+k}]}^{\text{erstes weg}} - \overbrace{y[x_i, \ldots, x_{i+k-1}]}^{\text{letztes weg}}}{x_{i+k} - x_i}$$

Die Rekursion geht also über die Anzahl der Argumente.

Graphisch sieht das Rekursionsschema für den Fall $n = 3$ so aus:

$$\begin{array}{cccccccccc}
 & & \boxed{= a_0} & & \boxed{= a_1} & & \boxed{= a_2} & & \boxed{= a_3} & \\
y_0 & = & y[x_0] & \leftarrow & y[x_0, x_1] & \leftarrow & y[x_0, x_1, x_2] & \leftarrow & y[x_0, x_1, x_2, x_3] & \ldots \\
 & & & \swarrow & & \swarrow & & \swarrow & & \\
y_1 & = & y[x_1] & \leftarrow & y[x_1, x_2] & \leftarrow & y[x_1, x_2, x_3] & & \ldots & \\
 & & & \swarrow & & \swarrow & & & & \\
y_2 & = & y[x_2] & \leftarrow & y[x_2, x_3] & & \ldots & & & \\
 & & & \swarrow & & & & & & \\
y_3 & = & y[x_3] & & \ldots & & & & &
\end{array}$$

In der linken Spalte stehen die Interpolationswerte.

In der obersten Zeile stehen die zu bestimmenden Koeffizienten.

Praktisch berechnet man die Werte spaltenweise von links nach rechts.

Für den Fall $n = 4$ muss man das Tableau um eine „Diagonale“ erweitern. Die bisher berechneten Werte bleiben unangetastet.

Beweis der Rekursionsdarstellung: Nach [Ran06]. Sei $p_{i,i+k} \in P_k$, $0 \le i \le i+k \le n$ das Polynom, das die Punkte $(x_i, y_i), \ldots, (x_{i+k}, y_{i+k})$ interpoliert. Man zeigt

$$\begin{aligned} p_{i,i+k}(x) = y[x_i] + y[x_i, x_{i+1}](x - x_i) + \ldots \\ + y[x_i, \ldots, x_{i+k}](x - x_i) \ldots (x - x_{i+k-1}) \end{aligned} \tag{5.1}$$

Da $p = p_{0,n}$ beweist dies auch obige Aussage. Der Beweis erfolgt durch Induktion über den Grad des Polynoms k.

Für $k = 0$ gilt $y[x_i] = y_i$, und $p_{i,i}(x) = y[x_i] = y_i$ also alles klar.

Induktionsschritt $k-1 \to k$. Sei die Aussage also richtig für alle $p_{i,i+k-1}$ mit $0 \le i \le i+k-1 \le n$

(i) Konstruktionsgemäß gilt (vergleiche mit (5.1))

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

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

Ausmultiplizieren liefert $p_{i,i+k}(x) = p_{i,i+k-1} + ax^k + \dots$ und a ist der Koeffizient von x^k in $p_{i,i+k}(x)$ da $p_{i,i+k-1}$ nur ein Polynom vom Grad $k-1$ ist.

(ii) Aus der Induktionsannahme folgt mit der Überlegung aus (i)

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

wobei $\dots$ Polynome vom Grad kleiner $k-1$ sind.

(iii) Nun betrachte das folgende Polynom

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

Es gilt: q interpoliert die Punkte $(x_i,y_i)\dots(x_{i+k},y_{i+k})$!

Denn für $x = x_i$ oder $x = x_{i+k}$ ist einer der beiden Terme im Zähler 0 und die Induktionsvoraussetzung greift.

Für $x = x_j$, $j \ne i,k$ rechne $q(x_j) = \frac{(x_j-x_i)y_j-(x_j-x_{i+k})y_j}{x_{i+k}-x_i} = y_j$.

Andererseits ist $q(x) = p_{i,i+k-1} + (x-x_i)\frac{p_{i+1,i+k}(x)-p_{i,i+k-1}(x)}{x_{i+k}-x_i}$, also die Gestalt aus (i) !

Wir brauchen nur noch den führenden Koeffizienten aus dem zweiten Term zu bestimmen.

(iv) Dazu setze die Darstellung aus (ii) in die letzte Darstellung von q ein:

$$\begin{aligned} p_{i,i+k}(x) &= q(x) \\ &= p_{i,i+k-1} + \\ &\quad (x-x_i)\frac{y[x_{i+1},\dots,x_{i+k}]x^{k-1} + \dots - y[x_i,\dots,x_{i+k-1}]x^{k-1} - \dots}{x_{i+k}-x_i} \\ &= \underbrace{p_{i,i+k-1}}_{\text{Grad}<k} + x^k \underbrace{\left(\frac{y[x_{i+1},\dots,x_{i+k}] - y[x_i,\dots,x_{i+k-1}]}{x_{i+k}-x_i}\right)}_{\text{führender Term} =: a} + \underbrace{\dots}_{\text{Grad}<k} \end{aligned}$$

□

Beispiel 5.3. Es sei die Tabelle

x_i	0	1	2
y_i	0	1	4

zu interpolieren. Das Dividierte-Differenzen-Schema lautet

$y[x_0] = 0$	$y[x_0, x_1] = \frac{y[x_1]-y[x_0]}{x_1-x_0} = \frac{1-0}{1-0}$	$y[x_0, x_1, x_2] = \frac{y[x_1,x_2]-y[x_0,x_1]}{x_2-x_0} = \frac{3-1}{2-0}$
$y[x_1] = 1$	$y[x_1, x_2] = \frac{y[x_2]-y[x_1]}{x_2-x_1} = \frac{4-1}{2-1}$	
$y[x_2] = 4$		

Das Interpolationspolynom lautet dann

$$\begin{aligned} p(x) &= y[x_0] \cdot 1 + y[x_0, x_1](x - x_0) + y[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &= 0 \cdot 1 + 1 \cdot (x - x_0) + 1 \cdot (x - x_0)(x - x_1) \\ &= x + (x - 1)x = x + x^2 - x \\ &= x^2 \end{aligned}$$

□

5.2 Neville-Darstellung

Satz 5.4 (Neville[10]-Darstellung)**.** Die Polynome $p_{i,j}$, $0 \le i \le j \le n$ interpolieren die Punkte $(x_i, y_i), \ldots, (x_j, y_j)$ und sind rekursiv definiert:

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

Beweis: Das haben wir bereits im Beweis zu Satz 5.2 in (iii) bewiesen. □

Das Neville-Schema erlaubt die *Auswertung* des Interpolationspolynoms an einer Stelle ξ *ohne* die Koeffizienten des Interpolationspolynoms explizit zu berechnen.

Setze dazu einfach $x = \xi$ in obiger Rekursionsformel.

Dies bietet sich an, wenn das Polynom nur an einer oder wenigen Stellen ausgewertet werden soll (z. B. bei Extrapolationsverfahren).

Das Schema ist ganz ähnlich zu den Dividierten Differenzen, nur dass direkt die Polynomwerte in die Tabelle eingetragen werden:

x_0	$y_0 = p_{0,0}(x)$	$p_{0,1}(x)$	$\to$	$p_{0,2}(x)$	$\to$	$p_{0,3}(x)$	$\ldots$	$p_{0,n-1}(x)$	$\to$	$p_{0,n}(x)$
			$\nearrow$		$\nearrow$				$\nearrow$	
x_1	$y_1 = p_{1,1}(x)$	$p_{1,2}(x)$	$\to$	$p_{1,3}(x)$	$\to$	$p_{1,4}(x)$	$\ldots$	$p_{1,n}(x)$		
			$\nearrow$		$\nearrow$					
x_2	$y_2 = p_{2,2}(x)$	$p_{2,3}(x)$	$\to$	$p_{2,4}(x)$	$\to$	$p_{2,5}(x)$	$\ldots$			
$\vdots$	$\vdots$	$\vdots$								
$\vdots$	$\vdots$	$p_{n-1,n}(x)$								
x_n	$y_n = p_{n,n}(x)$									

[10] Eric Harold Neville, 1889-1961, engl. Mathematiker.

Beispiel 5.5. Es sei wieder die Tabelle

x_i	0	1	2
y_i	0	1	4

gegeben. Gesucht sei nur der Wert $p(\xi)$ für $\xi = 1/2$.

$\mathbf{x_2 = 2}$

$$p_{2,2}(\xi) = 4$$

$\mathbf{x_1 = 1}$

$$p_{1,1}(\xi) = 1$$
$$p_{1,2}(\xi) = p_{1,1}(\xi) + (x - x_1)\frac{p_{2,2}(\xi) - p_{1,1}(\xi)}{x_2 - x_1} = 1 + (\frac{1}{2} - 1)\frac{4-1}{2-1} = -\frac{1}{2}$$

$\mathbf{x_0 = 0}$

$$p_{0,0}(\xi) = 0$$
$$p_{0,1}(\xi) = p_{0,0}(\xi) + (x - x_0)\frac{p_{1,1}(\xi) - p_{0,0}(\xi)}{x_1 - x_0} = 0 + (\frac{1}{2} - 0)\frac{1-0}{1-0} = \frac{1}{2}$$
$$p_{0,2}(\xi) = p_{0,1}(\xi) + (x - x_0)\frac{p_{1,2}(\xi) - p_{0,1}(\xi)}{x_2 - x_0} = \frac{1}{2} + (\frac{1}{2} - 0)\frac{-\frac{1}{2} - \frac{1}{2}}{2-0} = \frac{1}{4}$$ □

5.3 Bernstein-Polynome

Definition 5.6 (Bernstein[11]-Polynome)**.** Die binomische Formel liefert

$$1 = ((1-t)+t)^n = \sum_{i=0}^{n} \underbrace{\binom{n}{i}(1-t)^{n-i}t^i}_{=:B_{in}(t)}$$

Die Polynome

$$B_{in}(t) = \binom{n}{i}(1-t)^{n-i}t^i \qquad i = 0, \ldots, n$$

heißen Bernstein-Polynome vom Grad n auf $[0,1]$. Mittels der Transformation $\varphi : [a,b] \to [0,1]$, $\varphi(u) = (u-a)/(b-a)$ kann man die Bernstein-Polynome auf $[a,b]$ erweitern:

$$\begin{aligned} B_{in}(u;a,b) &= B_{in}(\varphi(n)) = \tbinom{n}{i}(1 - \tfrac{u-a}{b-a})^{n-i}(\tfrac{u-a}{b-a})^i \\ &= \tbinom{n}{i}\tfrac{1}{(b-a)^n}(b-u)^{n-i}(u-a)^i \quad . \end{aligned}$$

Abbildung 15 zeigt die Bernstein-Polynome vom Grad 6 auf dem Intervall $[0,1]$.

Die Bernstein-Polynome haben einige schöne Eigenschaften, die wir zusammenfassen in

[11] Sergei Natanowitsch Bernstein, 1880-1968, russ. Mathematiker.

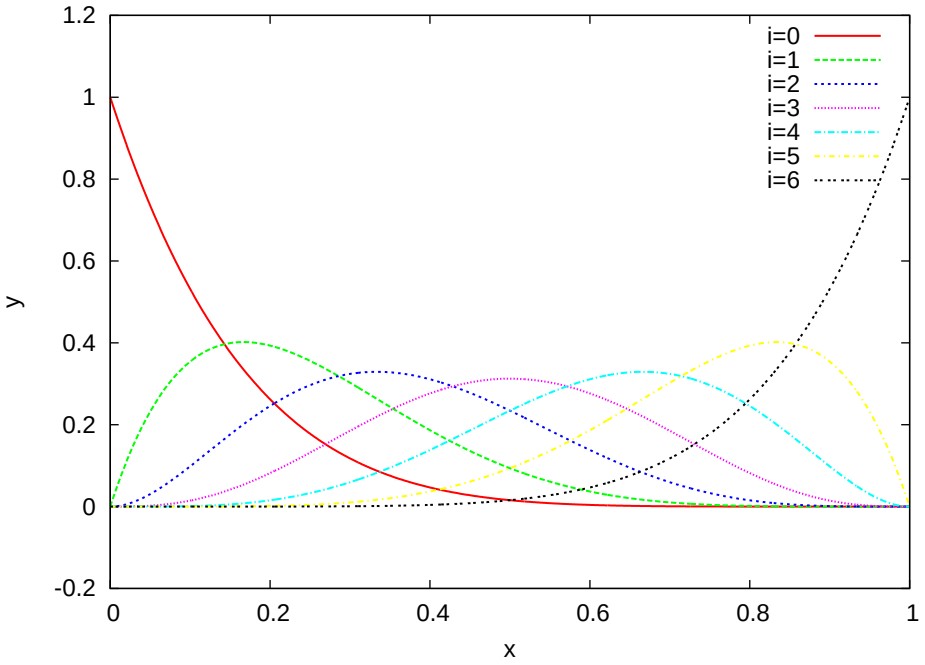


Abbildung 15: Die Bernstein-Polynome vom Grad 6.

Satz 5.7 (Eigenschaften der Bernstein-Polynome)**.** (i) $t = 0$ ist i-fache Nullstelle von B_{in}. Klar, da B_{in} den Term t^i enthält. Anders ausgedrückt: $\frac{d^j}{t^j} B_{in}(0) = 0$ für $j = 1, \ldots, i-1$.

(ii) $t = 1$ ist $n - i$-fache Nullstelle von B_{in}. Siehe (i).

(iii) Symmetrie: $B_{in}(t) = B_{n-i,n}(1-t)$. Folgt durch Einsetzen in die Definition.

(iv) Positivität: $0 \leq B_{in}(t) \leq 1$ für alle $t \in [0,1]$, $B_{in}(t) > 0$ für alle $t \in (0,1)$.

Für $t \in [0,1]$ ist $t \geq 0$ und $1 - t \geq 0$ daher auch $B_{in}(t) \geq 0$.

Wegen $\sum_{i=0}^{n} B_{in}(t) = 1$ ist $B_{in}(t) = 1 - \sum_{j=0, j\neq i}^{n} B_{jn}(t) \leq 1$.

$B_{in} > 0$ für $t \in (0,1)$ folgt aus (i),(ii), denn es gibt keine weiteren Nullstellen.

(v) B_{in} hat in $[0,1]$ genau ein Maximum in i/n. Die Ableitung des i-ten Bernstein-Polynoms vom Grad n

$$\begin{aligned} B'_{in}(t) &= \binom{n}{i} \left[-(n-i)(1-t)^{n-i-1} t^i + (1-t)^{n-i} t^{i-1} i \right] \\ &= \binom{n}{i} (1-t)^{n-i-1} t^{i-1} (i - nt) \end{aligned}$$

hat eine $i-1$-fache Nullstelle in 0, eine $n-i-1$-fache Nullstelle in 1 und eine einfache in $t = i/n$.

(vi) Die $\{B_{in}\}_{i=0}^{n}$ sind linear unabhängig und bilden eine Basis von P_n. Zu zeigen ist, dass aus $\sum_{i=0}^{n} b_i B_{in}(t) = 0$ für alle $t \in [0,1]$ zwingend folgt, dass $b_i = 0$.

Es ist

$$\frac{d^j}{t^j}\sum_{i=0}^{n} b_i B_{in}(t) = \sum_{i=0}^{n} b_i \frac{d^j}{t^j} B_{in}(t) = 0 \quad t \in [0,1].$$

Setze $j = 0$, $t = 0$. Es ist nur $B_{0n}(0) \neq 0$ also $b_0 = 0$.

Setze $j = 1$, $t = 0$. Es ist nur $B_{1n}(0) \neq 0$ also $b_1 = 0$. Usw.

(*vii*) Die Bernstein-Polynome können rekursiv über den Grad n dargestellt werden:

$$\begin{aligned} B_{0n}(t) &= (1-t)B_{0,n-1}(t) \\ B_{in}(t) &= tB_{i-1,n-1}(t) + (1-t)B_{i,n-1}(t) \\ B_{nn}(t) &= tB_{n-1,n-1}(t) \end{aligned}$$

Dies folgt aus der Rekursionsformel für Binomialkoeffizienten $\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}$ (das Pascalsche Dreieck).

(*viii*) Für die Ableitung gilt die Rekursionsformel $\quad (n \geq 1)$

$$B'_{in}(t) = \begin{cases} -nB_{0,n-1}(t) & i = 0 \\ n\left[B_{i-1,n-1}(t) - B_{i,n-1}(t)\right] & i = 1, \ldots, n-1 \\ nB_{n-1,n-1}(t) & i = n \end{cases}$$

Beachte: Das ist keine Rekursionsformel, die Ableitungen aus Ableitungen ausrechnet. Sondern die Ableitung wird durch Bernstein-Polynome niedrigeren Grades zusammengesetzt. □

Wie oben bewiesen bilden die B_{in} eine Basis von P_n. Jedes Polynom lässt sich also darstellen als

$$p(t) = \sum_{i=0}^{n} \beta_i B_{in}(t).$$

Diese Darstellung nennt man die Bézier[12]-Darstellung des Polynoms.

β_i heißt Bézier-Koeffizient, und die $(i/n, \beta_i)^T \in \mathbb{R}^2, \quad i = 0, \ldots, n$ heißen Bézier-Punkte.

Die Verbindung der Bézier-Punkte heißt Bézier-Polygon.

Beispiel 5.8. Für die Bézier-Punkte $(0,0), (1/3, -1), (2/3, -1), (1,1)$ lautet das Bézier-Polygon $t^3 + 3t^2 - 3t$.

Abbildung 16 zeigt das Bézier-Polynom und Bézier-Polygon.

□

Die Bézier-Punkte werden nicht exakt interpoliert, es handelt sich also um eine Approximation. Bernstein hat seine Polynome ursprünglich zum Beweis des Approximationssatzes von Weierstraß erfunden.

Man kann außerdem Folgendes zeigen:

- Das Bézier-Polynom liegt immer in der konvexen Hülle des Bézier-Polygons.

[12] Pierre Bézier, 1910-1999, frz. Ingenieur.

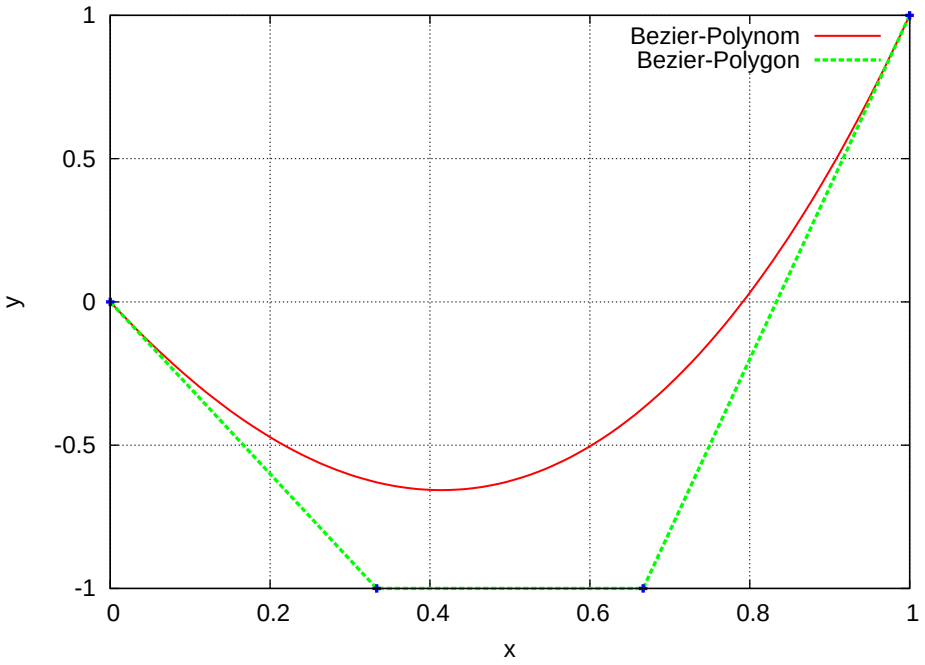


Abbildung 16: Das Bézier-Polynom und Bézier-Polygon zum Beispiel 5.8.

- Die Ableitung an den Endpunkten $0, 1$ stimmt mit den Steigungen des Bézier-Polygons überein (das kennt man aus Zeichenprogrammen). Es gilt unter Nutzung der Eigenschaften der Bernstein-Polynome:

$$\begin{aligned} p'(0) &= \sum_{i=0}^{n} \beta_i B'_{in}(0) = \beta_0 \left[-nB_{0,n-1}(0)\right] + \beta_1 \left[nB_{0,n-1}(0)\right] \\ &= n(\beta_1 - \beta_0) \end{aligned}$$

  Steigung des Bézier-Polygons ist:

$$\frac{\Delta y}{\Delta x} = \frac{(\beta_1 - \beta_0)}{i/n} = n(\beta_1 - \beta_0) = p'(0) \quad .$$

5.4 Algorithmus von de Casteljau

Der Erfolg der Bézier-Polynome basiert auf einem sehr effizienten Verfahren zur Auswertung der Polynome, bekannt als „Algorithmus von de Casteljau“.

Man nutzt die rekursive Darstellung der Bézier-Polynome:

$$\begin{aligned} p(t) &= \sum_{i=0}^{n} \beta_i^{(0)} B_{in}(t) \\ &= \beta_0^{(0)}(1-t)B_{0,n-1}(t) + \beta_1^{(0)}\Big(tB_{0,n-1}(t) + (1-t)B_{1,n-1}(t)\Big) \\ &\quad + \beta_2^{(0)}\Big(tB_{1,n-1}(t) + (1-t)B_{2,n-1}(t)\Big) + \ldots + \beta_n^{(0)}\Big(tB_{n-1,n-1}(t)\Big) \\ &= \sum_{i=0}^{n-1} \underbrace{\Big(\beta_i^{(0)}(1-t) + \beta_{i+1}^{(0)} t\Big)}_{=:\beta_i^{(1)}} B_{i,n-1}(t) \end{aligned}$$

Damit hat man nun ein Bézier-Polynome vom Grad $n-1$ mit neuen Koeffizienten auszuwerten, auf welches man denselben Trick rekursiv anwenden kann.

Es ergibt sich das Schema

$$\begin{aligned} \beta_i^{(0)}(t) &= \beta_i && i = 0, \ldots . n \\ \beta_i^{(k)}(t) &= \beta_i^{(k-1)}(1-t) + \beta_{i+1}^{(k-1)}(t) && i = 0, \ldots, n-k \end{aligned}$$

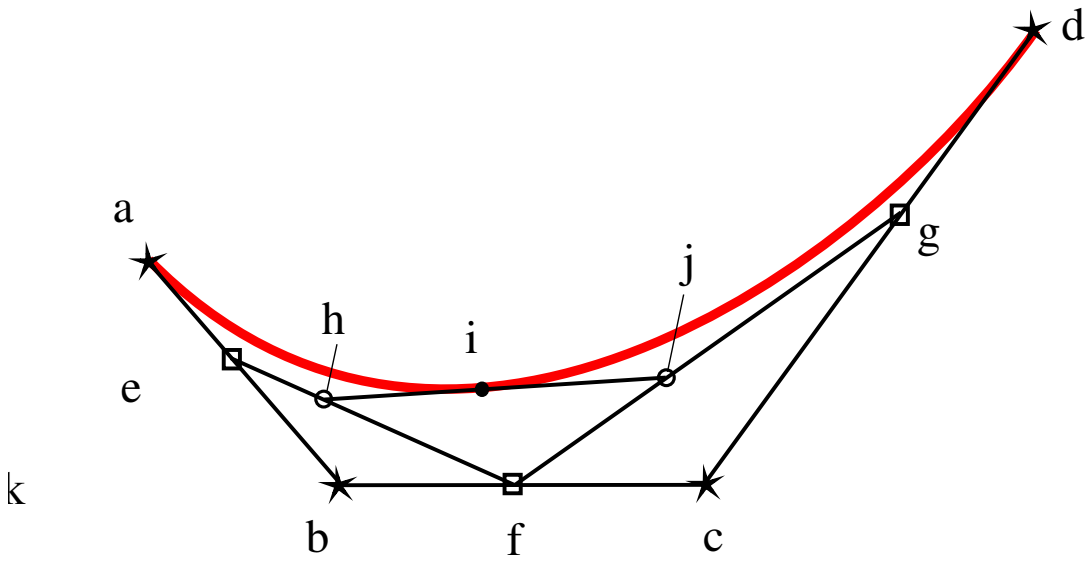
welches man wieder als Tableau ähnlich dem Neville-Schema sehen kann:

$$\begin{array}{cccccccccc} & & \boxed{k=0} & & \boxed{k=1} & & \boxed{k=2} & & & \boxed{k=n} \\ \beta_0 & = & \beta_0^{(0)} & \to & \beta_0^{(1)} & \to & \beta_0^{(2)} & \ldots & \to & \beta_0^{(n)} \\ & & & \nearrow & & \nearrow & & & \nearrow & \\ \beta_1 & = & \beta_1^{(0)} & \to & \beta_1^{(1)} & \to & \beta_1^{(2)} & & & \\ & & & \nearrow & & & & & & \\ \beta_2 & = & \beta_2^{(0)} & & & & & & & \\ \vdots & & & & & & & & & \\ & & & \nearrow & & & & & & \\ \beta_n & = & \beta_n^{(0)} & & & & & & & \end{array}$$

Schließlich gilt nach Konstruktion

$$p(t) = \beta_0^{(n)}(t).$$

Beispiel 5.9 (Zum Algorithmus von Casteljau)**.** Für die Auswertung an der Stelle $\xi = 1/2$ ergibt sich mit den Daten aus Beispiel 5.8.



5.5 Kurveninterpolation

Bis jetzt haben wir nur Funktionen $f : [a, b] \to \mathbb{R}$ betrachtet.

Was tut man, wenn man allgemein Kurven im $\mathbb{R}^m$, also Funktionen $f : [a, b] \to \mathbb{R}^m$ interpolieren (oder approximieren) möchte?

Gegeben seien $n+1$ Punkte $x^{(i)} \in \mathbb{R}^m$, $i = 0, \ldots, n$ (Superskript = Index, Subskript $x_k^{(i)}$ ist die k-te Komponente von $x^{(i)}$). Diese Punkte sollen durch eine Kurve $x : [a, b] \to \mathbb{R}^m$ verbunden werden.

Für die Kurve soll die Interpolationsbedingung

$$x(t_i) = x^{(i)} \quad \text{für } i = 0, \ldots, n \text{ und } a = t_0 < t_1 < \ldots < t_n = b$$

gelten.

Dies lässt sich auf m unabhängige Interpolationsaufgaben

$$x_k(t_i) = x_k^{(i)} \qquad i = 0, \ldots, n, \quad k = 1, \ldots, m$$

zurückführen, wofür man z.B. Lagrangeinterpolation nutzen kann.

Der Parameterbereich $[a, b]$ ist oft willkürlich, da nur die Punkte $x^{(i)}$ gegeben sind. Dann bietet sich die Bogenlänge an.

Die in vielen Zeichenprogrammen benutzten Bézierkurven erhält man mittels

$$x(t) = \sum_{i=0}^{n} x^{(i)} B_{in}(t),$$

wobei wir hier $a = 0$, $b = 1$ angenommen haben.

Hier werden die gegebenen Punkte aber (wie beim Bézier-Polynom) nicht exakt interpoliert, man spricht dann von Kontrollpunkten.

Der Casteljau-Algorithmus lässt sich völlig analog auch auf Kurven übertragen.

Schließlich haben erweiterte Verfahren über mehrdimensionalen Parameterbereichen vielfältige Anwendungen in der Computergeometrie bzw. im Computer Aided Design.

5.6 Zusammenfassung

- Mit der Newton-Interpolation erhält man ein inkrementelles Verfahren zur Polynominterpolation.
- Die Neville-Darstellung eignet sich insbesondere, wenn sehr wenige Auswertungen des Interpolationspolynoms gebraucht werden.
- Bernstein-Polynome und die daraus abgeleiteten Bézier-Polynome eignen sich sehr gut zur glatten Approximation gegebener Datenpunkte.
- Viele Zeichenprogramme verwenden die Bézier-Kurven zur Darstellung glatter Kurven, die man mittels Kontrollpunkten in ihrer Lage beeinflussen kann.

6 Stückweise Polynome

6.1 Einführung und Aufgabenstellung

Wir haben gesehen:

- Die Interpolation von Funktionen mit Polynomen vom Grad n erfordert die Existenz von Ableitungen bis zur Ordnung $n+1$.
- Selbst für simple Funktionen (wie z.B. $f(x) = \frac{1}{1+x^2}$) konvergiert die Polynominterpolation an *äquidistanten Stützstellen* nicht mehr gleichmäßig, da die hohen Ableitungen zu schnell wachsen.
- Die Polynominterpolation für großen Grad ist schlecht konditioniert in Bezug auf die Stützwerte (kleine Änderungen in den y_i können einen großen Effekt an anderer Stelle haben).

Was soll man also tun, wenn viele Stützstellen zu interpolieren sind oder die zu interpolierende Funktion nicht genügend oft differenzierbar ist?

Idee: Verwende Polynome abschnittsweise!

Sei $[a,b]$ ein Intervall mit einer gegebenen Unterteilung

$$a = x_0 < x_1 < \ldots < x_n = b.$$

$I_i = [x_{i-1}, x_i]$, $i = 1, \ldots, n$ heißt Teilintervall und wir setzen

$$h_i = x_i - x_{i-1}, \qquad h := \max_{i \in \{1,\ldots,n\}} h_i$$

für die Länge der Teilintervalle.

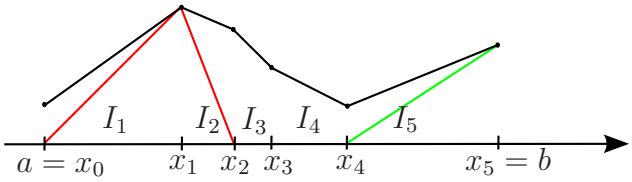
Zu $k, r \in \mathbb{N}_0$ definieren wir

$$S_h^{k,r}[a,b] = \{p \in C^r[a,b] \,|\, p|_{I_i} \in P_k\},$$

den Raum der global r mal stetig differenzierbaren Funktionen, die stückweise Polynome vom Grad k sind.

Beachte: Die Funktionen sind immer mindestens C^0.

Beispiel 6.1. $S_h^{1,0}$ heißt Raum der stückweise linearen und global stetigen Funktionen.



Eine Funktion $f \in S_h^{1,0}[a,b]$ ist eindeutig durch die Werte an den Knoten $x_0, \ldots, x_n$ definiert.

$S_h^{1,0}[a,b]$ ist ein $n+1$-dimensionaler Vektorraum. Als Basis wähle etwa $\varphi_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

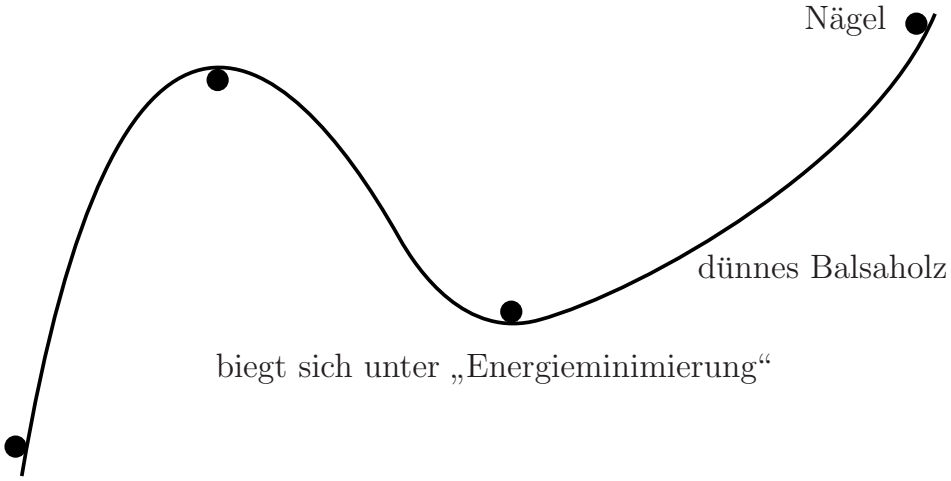
Diese Funktionen heißen „Hutfunktionen“.

$S_h^{1,0}$ spielt eine Rolle bei der numerischen Lösung partieller Differentialgleichungen. Hier interessieren uns Räume mit $r > 0$. □

6.2 Kubische Splines

In der Praxis wichtig ist $S_h^{3,2}[a,b]$, d. h. Polynomgrad 3 und global 2 mal stetig differenzierbar.

Geschichte: Zu Beginn des 20. Jahrhunderts konstruierte man glatte Kurven im Schiffs- und Flugzeugbau mit so genannten „Straklatten“ (engl. splines)



Welche Bedingungen legen eine Funktion $s(x) \in S_h^{3,2}[a,b]$ fest?

Zunächst ist $s(x)$ stückweise definiert, also gilt

$$p(x) = \begin{cases} p_i(x) & x \in [x_{i-1}, x_i), \quad i \in 1, \ldots, n, \\ p_n(x_n) & x = x_n \quad \text{(letzter Punkt)}, \end{cases}$$

mit $p_i(x) \in P_k$. An die p_i stellen wir die folgenden Bedingungen.

(*i*) *Interpolationsbedingungen (Stetigkeit)*:

$$i = 1, \ldots, n: \qquad \left.\begin{array}{rcl} p_i(x_{i-1}) & = & y_{i-1} \\ p_i(x_i) & = & y_i \end{array}\right\} \quad 2n \text{ Bedingungen.}$$

(*ii*) *Stetigkeitsbedingungen an die Ableitungen*:

$$\underbrace{i = 1, \ldots, n-1}_{\text{innere Knoten!}}: \qquad \left.\begin{array}{rcl} p_i'(x_i) & = & p_{i+1}'(x_i) \\ p_i''(x_i) & = & p_{i+1}''(x_i) \end{array}\right\} \; 2n-2 \text{ Bedingungen.}$$

(*iii*) *Randbedingungen* (für sog. natürliche Splines, andere sind möglich):

$$\left.\begin{array}{rcl} p_1''(x_0) & = & 0 \\ p_n''(x_n) & = & 0 \end{array}\right\} \; 2 \text{ Bedingungen.}$$

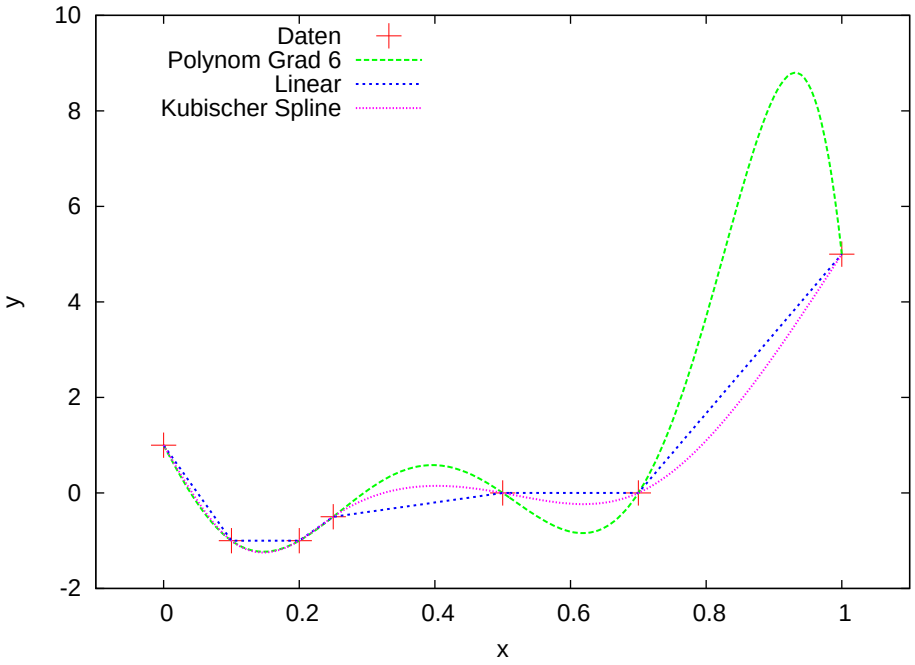


Abbildung 17: Vergleich der Interpolation mit Polynomen, stückweise linearen Funktionen und kubischen Splines.

Somit ergeben sich $4n$ Bedingungen für die $4n$ Freiheitsgrade (n Polynome $p_i(x)$ vom Grad $k = 3$).

Wie berechnet man die Koeffizienten der Polynome? Zuerst ein Beispiel.

Beispiel 6.2 (zu Kubischen Splines)**.** Es sei folgende Wertetabelle zu interpolieren:

x	0	0.1	0.2	0.25	0.5	0.7	1.0
y	1	-1	-1	-0.5	0	0	5

Wir verwenden ein Polynom vom Grad 6, stückweise lineare Funktionen sowie kubische Splines (mit natürlichen Randbedingungen).

Abbildung 17 zeigt die Interpolation mit Polynomen, stückweise linearen Funktionen und kubischen Splines anhand eines Beispiels. Deutlich zu erkennen ist der starke „Überschwinger“ bei der Polynominterpolation. □

Satz 6.3 (Berechnung kubischer Splines)**.** Wir schreiben die Teilpolynome der Splines in der Form

$$p_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3, \qquad i = 1, \ldots, n.$$

Die $a_2^{(i)}$ sind dann Lösung des linearen Gleichungssystems der Dimension $n-1$

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1}) a_2^{(i)} + h_{i+1} a_2^{(i+1)} = 3\left\{ \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right\} \quad i = 1, \ldots, n-1, \quad (6.1)$$

wobei $a_2^{(0)} = a_2^{(n)} = 0$ und $h_i = x_i - x_{i-1}$.

Die restlichen Koeffizienten können für $i = 1, \ldots, n$ aus den $a_2^{(i)}$ berechnet werden:

$$a_0^{(i)} = y_i \quad (6.2)$$

$$a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\left\{ 2a_2^{(i)} + a_2^{(i-1)} \right\}, \quad (6.3)$$

$$a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i}. \quad (6.4)$$

Dies wollen wir nun schrittweise herleiten. □

Beweis: Nach [Ran06, S. 45].

(*i*) Berechne zunächst die Ableitung von $p_i(x)$ (geht einfach, da $\frac{d}{dx}(x - x_i)^n = n(x - x_i)^{n-1}$) :

$$p_i'(x) = a_1^{(i)} + 2a_2^{(i)}(x - x_i) + 3a_3^{(i)}(x - x_i)^2$$
$$p_i''(x) = 2a_2^{(i)} + 6a_3^{(i)}(x - x_i)$$

(*ii*) Nun nutze die Interpolationsbedingungen: Einsetzen des Punktes x_i liefert

$$y_i = p_i(x_i) = a_0^{(i)} \rightarrow \boxed{a_0^{(i)} = y_i} \qquad i = 1, \ldots, n \quad (6.5)$$

was Aussage (6.2) des Satzes beweist.

Einsetzen des Punktes x_{i-1} liefert

$$y_{i-1} = p_i(x_{i-1}) = a_0^{(i)} - h_i a_1^{(i)} + h_i^2 a_2^{(i)} - h_i^3 a_3^{(i)} \qquad i = 1, \ldots, n$$

und damit wegen $a_0^{(i)} = y_i$

$$\boxed{y_{i-1} - y_i = -h_i a_1^{(i)} + h_i^2 a_2^{(i)} - h_i^3 a_3^{(i)}} \quad (6.6)$$

(*iii*) Einsetzen der Randbedingungen liefert

$$0 = p_1''(x_0) = 2a_2^{(1)} - 6h_1 a_3^{(1)} \rightarrow \boxed{a_2^{(1)} - 3h_1 a_3^{(1)} = 0} \quad (6.7)$$

$$0 = p_n''(x_n) = 2a_2^{(n)} \rightarrow \boxed{a_2^{(n)} = 0} \tag{6.8}$$

(iv) Stetigkeit der ersten Ableitung

$$p_i'(x_i) = p_{i+1}'(x_i) \qquad i = 1, \ldots, n-1 \text{ innere Punkte}$$

$$\boxed{a_1^{(i)} = a_1^{(i+1)} - 2h_{i+1}a_2^{(i+1)} + 3h_{i+1}^2 a_3^{(i+1)}} \tag{6.9}$$

(v) Stetigkeit der zweiten Ableitung

$$p_i''(x_i) = p_{i+1}''(x_i) \qquad i = 1, \ldots, n-1$$

$$2a_2^{(i)} = 2a_2^{(i+1)} - 6h_{i+1}a_3^{(i+1)} \rightarrow \boxed{a_2^{(i)} = a_2^{(i+1)} - 3h_{i+1}a_3^{(i+1)}} \tag{6.10}$$

(vi) Nun drücke $a_3^{(i)}$ durch $a_2^{(i)}$ aus. Dazu löse (6.10) aus (v) nach $a_3^{(i+1)}$ auf

$$a_3^{(i+1)} = \frac{a_2^{(i+1)} - a_2^{(i)}}{3h_{i+1}} \qquad i = 1, \ldots, n-1$$

(Beachte: (v) galt nur für die inneren Punkte).

Aus der Randbedingung (6.7) aus (iii) schließen wir

$$a_3^{(1)} = \frac{a_2^{(1)}}{3h_1}.$$

Indem wir formal $a_2^{(0)} = 0$ einführen (beachte: es gibt nur Koeffizienten für Superskript $1, \ldots, n$), können wir also schreiben

$$a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i} \qquad i = 1, \ldots, n.$$

Das beweist die Aussage (6.4) des Satzes.

(vii) Nun drücke die $a_1^{(i)}$ durch die $a_2^{(i)}$ aus. Dazu löse (6.6) aus (ii) nach $a_1^{(i)}$ auf.

$$a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} - h_i^2 a_3^{(i)} \qquad i = 1, \ldots, n$$

Nun drücke $a_3^{(i)}$ durch die $a_2^{(i)}$ aus

$$\begin{aligned} a_1^{(i)} &= \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} - h_i^2 \left(\frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i} \right) \qquad i = 1. \ldots, n \\ &= \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3} \left\{ 2a_2^{(i)} + a_2^{(i-1)} \right\} \qquad i = 1, \ldots, n. \end{aligned}$$

Das zeigt Aussage (6.3) des Satzes.

(*viii*) Nun setze die hergeleiteten Ausdrücke für $a_1^{(i)}$ und $a_3^{(i)}$ in die verbleibende Gleichung (6.9) aus (*iv*) ein

$$\frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\left\{2a_2^{(i)} + a_2^{(i-1)}\right\} = \frac{y_{i+1} - y_i}{h_{i+1}} + \frac{h_{i+1}}{3}\left\{2a_2^{(i+1)} + a_2^{(i)}\right\} - 2h_{i+1}a_2^{(i+1)} + 3h_{i+1}^2\left(\frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i}\right)$$

für $i = 1, \ldots n-1$.

Umordnen der a's nach links, und der y's nach rechts ergibt

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1})a_2^{(i)} + h_{i+1}a_2^{(i+1)} = 3\left\{\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right\} \qquad i = 1, \ldots n-1.$$

Dies ist die Aussage (6.1) des Satzes.

Dies sind $n-1$ Gleichungen für die $n-1$ Unbekannten $a_2^{(i)}$, $i = 1, \ldots n-1$ da wir formal $a_2^{(0)} = 0$ einführten und $a_2^{(n)} = 0$ nach (6.8) aus der Randbedingung. □

Das Gleichungssystem $Ax = b$ hat Tridiagonalgestalt und lautet:

$$A = \begin{bmatrix} 2(h_1+h_2) & h_2 & 0 & \ldots & 0 \\ h_2 & 2(h_2+h_3) & h_3 & & \\ 0 & h_3 & 2(h_3+h_4) & h_4 & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & & & h_{n-1} & 2(h_{n-1}+h_n) \end{bmatrix},$$

$$x = \begin{bmatrix} a_2^{(1)} \\ \vdots \\ a_2^{(n-1)} \end{bmatrix}, \qquad b = \begin{bmatrix} 3\left\{\frac{y_2-y_1}{h_2} - \frac{y_1-y_0}{h_1}\right\} \\ \vdots \\ 3\left\{\frac{y_n-y_{n-1}}{h_n} - \frac{y_{n-1}-y_{n-2}}{h_{n-1}}\right\} \end{bmatrix}.$$

Außerdem hat das lineare Gleichungssystem folgende Eigenschaften:

- $a_{ij} = a_{ji}$ (Symmetrie)
- $\sum_{j=1, i\neq j}^{n-1} |a_{ij}| < |a_{ii}|$ (strikte Diagonaldominanz)

Hieraus kann man mit Sätzen der linearen Algebra die eindeutige Lösbarkeit schließen.

Schließlich werden wir in einer späteren Vorlesung erfahren, dass dieses spezielle Gleichungssystem mit dem Aufwand $O(n)$ gelöst werden kann.

Auch für die kubischen Splines kann man wieder den Interpolationsfehler betrachten.

Satz 6.4 (Fehlerabschätzung für kubische Splines)**.** Sei $f \in C^4[a,b]$. Erfüllt der kubische Spline

$$s_n''(a) = f''(a) \quad \text{und} \quad s_n''(b) = f''(b)$$

(also in Erweiterung der natürlichen Randbedingungen oben) so gilt

$$\max_{a \le x \le b} |f(x) - s_n(x)| \le \frac{1}{2} h^4 \max_{a \le x \le b} |f^{(4)}(x)|. \tag{6.11}$$

Der Beweis kann hier in der Vorlesung nicht gegeben werden, wir verweisen auf [SW70]. □

Wir sehen: Schrittweite h und Differenzierbarkeitsordnung sind nun entkoppelt. Der (lokale) Polynomgrad geht in die Potenz von h ein.

Außerdem sind Splines sind wesentlich stabiler gegen Störungen in den Stützwerten y_i.

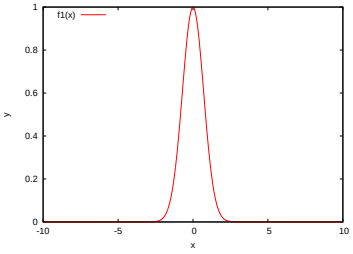
Beispiel 6.5 (Zur Konvergenzordnung stückweiser Polynome)**.** Wir betrachten die Interpolation der folgenden drei Funktionen

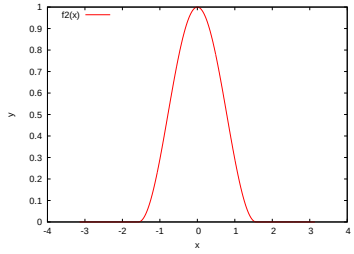
$$f_1(x) = \exp(-x^2) \qquad \text{in } [-10, 10], \tag{6.12}$$

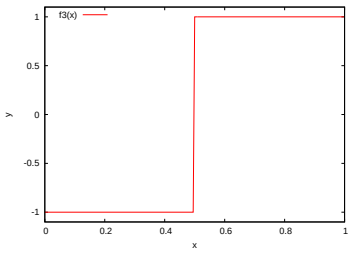
$$f_2(x) = \begin{cases} \cos^2(x) & |x| < \pi/2 \\ 0 & |x| \ge \pi/2 \end{cases} \qquad \text{in } [-\pi, \pi], \tag{6.13}$$

$$f_3(x) = \begin{cases} -1 & x < 1/2 \\ +1 & x \ge 1/2 \end{cases} \qquad \text{in } [0, 1], \tag{6.14}$$

mittels Polynomen, $S_h^{1,0}$ und $S_h^{3,2}$ (mit natürlichen Randbedingungen, alle Funktionen f_i erfüllen (näherungsweise) $f_i'' = 0$ an den Randpunkten).







Die Abbildung 18 zeigt die Interpolation der Funktion $f_1(x)$.

Die Abbildung 19 zeigt die Interpolation der Funktion $f_2(x)$.

Die Abbildung 20 zeigt die Interpolation der Funktion $f_3(x)$.

Wir lernen:

- Interpolation mit Polynomen steigenden Grades an äquidistanten Stützstellen schlägt in allen Fällen fehl, d. h. der Interpolationsfehler steigt mit dem Grad an.
- Kubische Splines konvergieren und liefern einen glatten Verlauf. Allerdings kommt es zu möglicherweise „unphysikalischen“ Unter- bzw. Überschwingern. Diese sind aber im Falle von $f_3(x)$ um die Sprungstelle lokalisiert.

- Stückweise lineare Funktionen haben diesen Defekt nicht.

Wir wollen nun den Interpolationsfehler noch experimentell bestimmen.

Fehler bei Interpolation der Funktion $f_1(x) = e^{-x^2}$:

n	$S_h^{1,0}$	$S_h^{3,2}$	P_n
4	$6.045_e - 01$	$7.420_e - 01$	$8.038_e - 01$
6	$4.447_e - 01$	$5.612_e - 01$	$9.999_e - 01$
8	$3.002_e - 01$	$3.918_e - 01$	$2.311_e + 00$
10	$1.774_e - 01$	$2.464_e - 01$	$5.949_e + 00$
16	$1.060_e - 01$	$2.753_e - 02$	
32	$6.946_e - 02$	$7.083_e - 03$	
64	$2.241_e - 02$	$3.316_e - 04$	
128	$5.974_e - 03$	$1.918_e - 05$	
256	$1.517_e - 03$	$1.173_e - 06$	
512	$3.809_e - 04$	$7.289_e - 08$	
1024	$9.533_e - 05$	$4.549_e - 09$	

Angegeben ist der maximale Fehler an einem Punkt. Polynome konvergieren nicht.

Stückweise linear konvergiert mit h^2 (d. h. $e_{2n}/e_n = (1/2)^2$), kubische Splines mit h^4 (d. h. $e_{2n}/e_n = (1/2)^4$).

In beiden Fällen gilt dies nur, wenn n genügend groß, man spricht von „asymptotischer“ Konvergenz.

Fehler bei Interpolation der Funktion $f_2(x) = \begin{cases} \cos^2(x) & x < \pi/2 \\ 0 & x \geq \pi/2 \end{cases}$:

n	$S_h^{1,0}$	$S_h^{3,2}$
4	$1.052e - 01$	$1.649e - 01$
8	$1.052e - 01$	$4.498e - 02$
16	$3.518e - 02$	$8.434e - 03$
32	$9.423e - 03$	$1.945e - 03$
64	$2.396e - 03$	$4.764e - 04$
128	$6.015e - 04$	$1.184e - 04$
256	$1.505e - 04$	$2.958e - 05$
512	$3.764e - 05$	$7.394e - 06$
1024	$9.412e - 06$	$1.848e - 06$

In diesem Fall konvergiert der maximale Fehler auch im Falle kubischer Splines nur mit h^2.

Dies liegt daran, dass $f_2''(x)$ unstetig am Punkt $x = \pi/2$ ist (springt von 2 auf 0).

Die dritte Ableitung existiert nicht mehr. □

Regel 6.6. Für die Interpolation mit stückweisen Polynomen merken wir uns:

Je höher der (abschnittsweise) Polynomgrad umso schneller konvergiert das Verfahren. Im allgemeinen erhält man $O(h^{k+1})$ Konvergenz für Polynome vom Grad k.

Dies gilt allerdings nur dann, wenn die zu interpolierende Funktion genügend of differenzierbar ist. Ist dies nicht der Fall so lohnt also auch die Verwendung von Polynomen hohen Grades nicht. □

Die oben eingeführten natürlichen kubischen Splines bilden den Funktionenraum

$$\tilde{S}_h^{3,2}[a,b] = \{p \in C^2[a,b] \,|\, p|_{I_i} \in P_3 \wedge p''(a) = p''(b) = 0\} \subset S_h^{3,2}[a,b].$$

Offensichtlich wird $s \in \tilde{S}_h^{3,2}[a,b]$ durch $n+1$ Werte an den Stützstellen x_i, $i = 0, \ldots, n$ eindeutig festgelegt, d. h. die Dimension von $\tilde{S}_h^{3,2}[a,b]$ ist $n+1$.

Oben haben wir eine Funktion $s \in \tilde{S}_h^{3,2}[a,b]$ mittels der Koeffizienten $a_k^{(i)}$ auf jedem Abschnitt $[x_{i-1}, x_i)$ festgelegt.

Eine andere Möglichkeit besteht darin, dass man eine Basis $\phi_0, \ldots, \phi_n$ von $\tilde{S}_h^{3,2}[a,b]$ wählt und dann die Koeffizienten β_i wie üblich mittels

$$\sum_{i=0}^{n} \beta_i \phi_i(x_i) = y_i$$

bestimmt.

In der Praxis wählt man die sogenannten „B-Splines“ als Basis. Diese haben folgenden Eigenschaften:

- Die ϕ_i sind an höchstens 5 aufeinanderfolgenden Stützstellen ungleich Null. Das Gleichungssystem hat 5-Diagonalgestalt.
- Die ϕ_i können rekursiv definiert werden.

Für weitere Einzelheiten verweisen wir auf [SK05].

6.3 Polynome in mehreren Raumdimensionen

Bisher: Nur eine Variable, z.B. Zeit, x-Position

Aber: Die Welt ist dreidimensional!

In der Anwendung treten oft Funktionen in mehreren Variablen auf. Polynome lassen sich entsprechend übertragen: z.B.

$$\begin{aligned} p(x_1, x_2) &= a_2 x_2 + a_1 x_1 + a_0 \\ q(x_1, x_2) &= a_3 x_1 x_2 + a_2 x_2 - a_1 x_1 + a_0 \\ r(x_1, x_2, x_3) &= a_5 x_1^2 + a_4 x_1 x_2 + a_3 x_3 + a_2 x_2 + a_1 x_1 + a_0 \end{aligned}$$

Wir wollen uns zunächst ansehen, wie man systematisch Polynome in mehr als einer Raumdimension definiert.

Dazu brauchen wir erst ein paar Bezeichnungen.

Definition 6.7 (Multiindex-Notation)**.** Es seien Vektoren

$$\underline{\alpha} = (\alpha_1, \ldots, \alpha_d)^T, \alpha_i \in \mathbb{N}_0, \qquad \underline{x} = (x_1, \ldots, x_d)^T, x_i \in \mathbb{R}$$

gegeben. Dann definieren wir

$$\underline{x}^{\underline{\alpha}} = \prod_{i=1}^{d} x_i^{\alpha_i}$$

Darüberhinaus setzen wir

$$\|\underline{\alpha}\|_1 = \sum_{i=1}^{d} \alpha_i, \qquad \|\underline{\alpha}\|_\infty = \max_{i=1\ldots d} \alpha_i.$$

□

Definition 6.8 (Mehrdimensionale Polynome)**.** Eine Funktion $\sum_{\underline{\alpha}\in A} a_{\underline{\alpha}} \underline{x}^{\underline{\alpha}}$ heißt Polynom in d Raumdimensionen.

Für die Menge A gibt es verschiedene Möglichkeiten.

Wir betrachten die beiden folgenden:

$$P_n^{(d)} = \left\{ v \,\middle|\, v = \sum_{\|\alpha\|_1 \le n} a_{\underline{\alpha}} \underline{x}^{\underline{\alpha}} \right\} \tag{6.15}$$

$$Q_n^{(d)} = \left\{ v \,\middle|\, v = \sum_{\|\alpha\|_\infty \le n} a_{\underline{\alpha}} \underline{x}^{\underline{\alpha}} \right\} \tag{6.16}$$

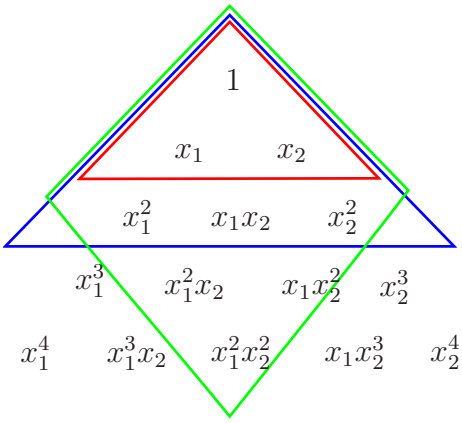
In zwei Raumdimensionen ($d = 2$) gilt dabei

$$\#P_n^{(2)} = \frac{(n+1)(n+2)}{2}, \qquad \#Q_n^{(2)} = (n+1)^2.$$

□

Wir verdeutlichen diese Konstruktion in 2 Raumdimensionen ($d = 2$).

Die Monome lassen sich folgendermaßen anordnen (wie im Pascal'schen Dreieck):



Es ist $P_1^{(2)}$ in rot, $P_2^{(2)}$ in blau und $Q_2^{(2)}$ in grün.

Lagrangeinterpolation lässt sich relativ leicht übertragen sofern man sich auf Q_n beschränkt.

Wir behandeln nur $d = 2$, die Erweiterung der Konstruktion auf größeres d gelingt aber leicht.

Die Koordinaten wollen wir mit (x, y) bezeichnen. Es seien

$$X = \{x_0, x_1, \ldots, x_n\}, \qquad Y = \{y_0, y_1, \ldots, y_n\}$$

die Unterteilungen für die x- respektive y-Richtung.

Für jede Unterteilung können wir die entsprechenden Lagrange-Basispolynome aufstellen:

$$L_i^{(x,n)}(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \qquad L_i^{(y,n)}(y) = \prod_{j=0, j\neq i}^{n} \frac{y - y_j}{y_i - y_j}.$$

Damit können wir dann zweidimensionale Lagrange-Polynome definieren mittels

$$L_{i,j}^{(n)}(x, y) = L_i^{(x,n)}(x) L_j^{(y,n)}(y) \quad . \tag{6.17}$$

Dies nennt man eine „Tensorproduktkonstruktion“.

Für diese Polynome gilt dann

$$L_{i,j}^{(n)}(x_r, y_s) = \begin{cases} 1 & r = i \wedge j = s \\ 0 & \text{sonst} \end{cases} \qquad \forall (r, s) \in \{0, \ldots, n\} \times \{0, \ldots, n\}.$$

Die $L_{i,j}^{(n)}$ bilden eine Basis von $Q_n^{(2)}$.

Die Konstruktion kann leicht auf unterschiedliche Zahl von Stützstellen in jede Richtung erweitert werden.

Beispiel 6.9. Wir betrachten $n = 2$, also $X = \{x_0, x_1, x_2\}$, $Y = \{y_0, y_1, y_2\}$. Für $i = 2, j = 1$ erhalten wir

$$\begin{aligned} L_{2,1}^{(2)}(x, y) &= L_2^{(x,2)}(x) L_1^{(y,2)}(y) \\ &= \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \frac{(y - y_0)(y - y_2)}{(y_1 - y_0)(y_1 - y_2)}. \end{aligned}$$

Die Abbildung 21 zeigt die Lagrange-Polynome $L_{2,1}^{(2)}$ und $L_{1,1}^{(2)}$. □

6.4 Zusammenfassung

- Zur Interpolation bei vielen Datenpunkten verwendet man stückweise polynomiale Funktionen.
- Dabei kann man mehr oder weniger viele Ableitungen der Interpolationsfunktion stetig halten.
- Wir haben Polynome in mehr als einer Variablen eingeführt, um Funktionen mit entsprechend vielen Variablen interpolieren zu können.
- Die Definition von Polynomen im $\mathbb{R}^d$ lässt einige Freiheit. Wir haben $P_n^{(d)}$ und $Q_n^{(d)}$ kennengelernt.
- Auch in mehreren Raumdimensionen lassen sich stückweise Polynome definieren, allerdings ist dies im allgemeinen wesentlich schwieriger, da die Abschnitte keine einfachen Intervalle mehr sind.

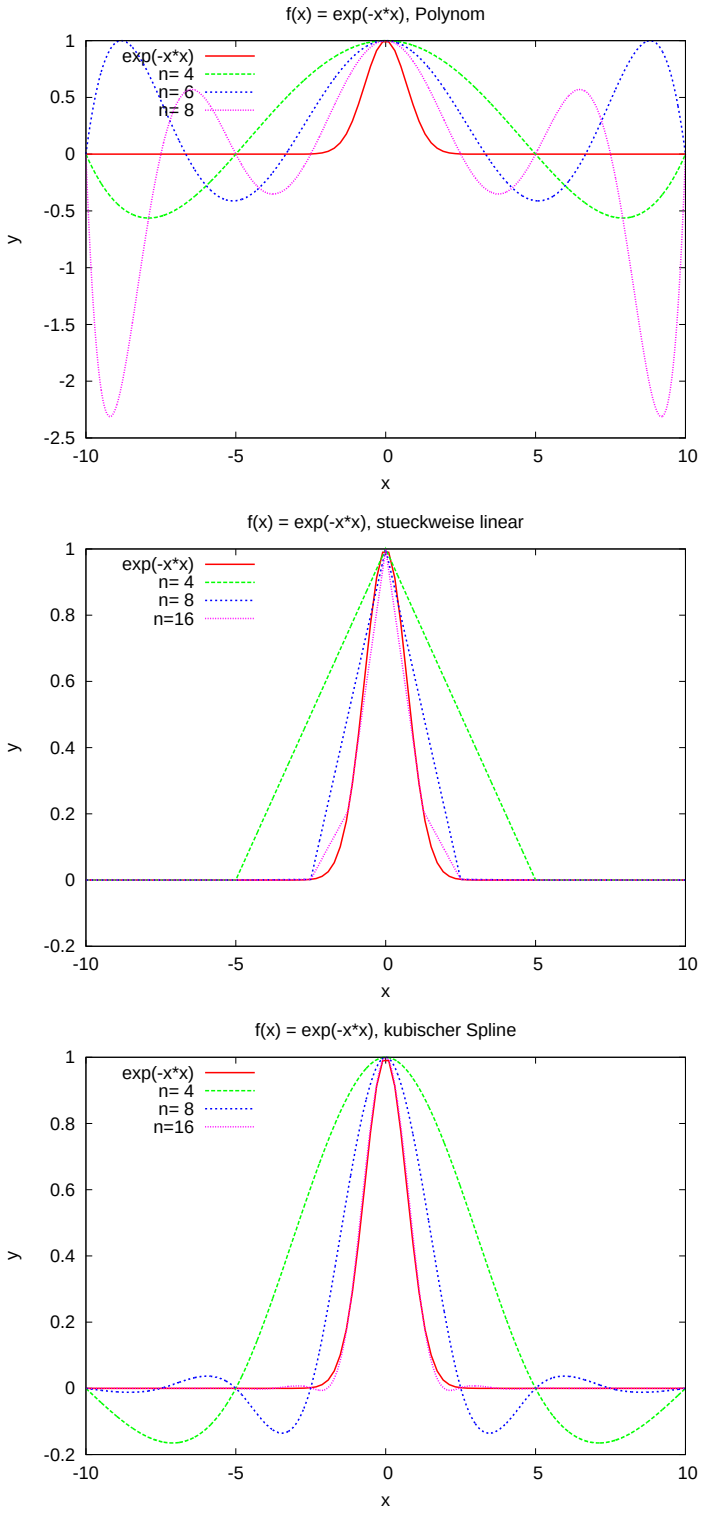


Abbildung 18: Interpolation der Funktion $f_1(x)$ mit Lagrange-Polynomen, stückweise linearen Funktionen und kubischen Splines.

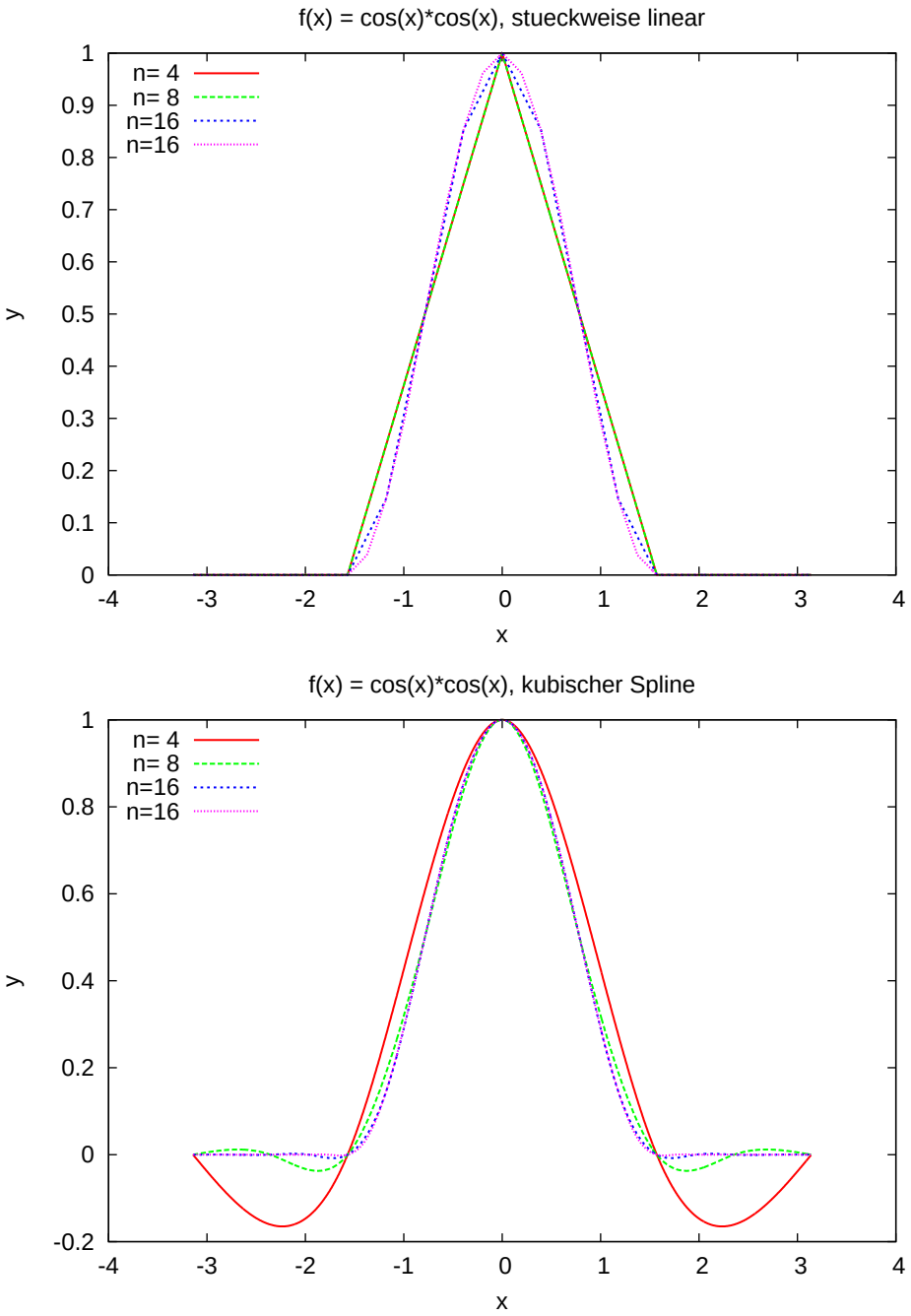


Abbildung 19: Interpolation der Funktion $f_2(x)$ mit Lagrange-Polynomen, stückweise linearen Funktionen und kubischen Splines.

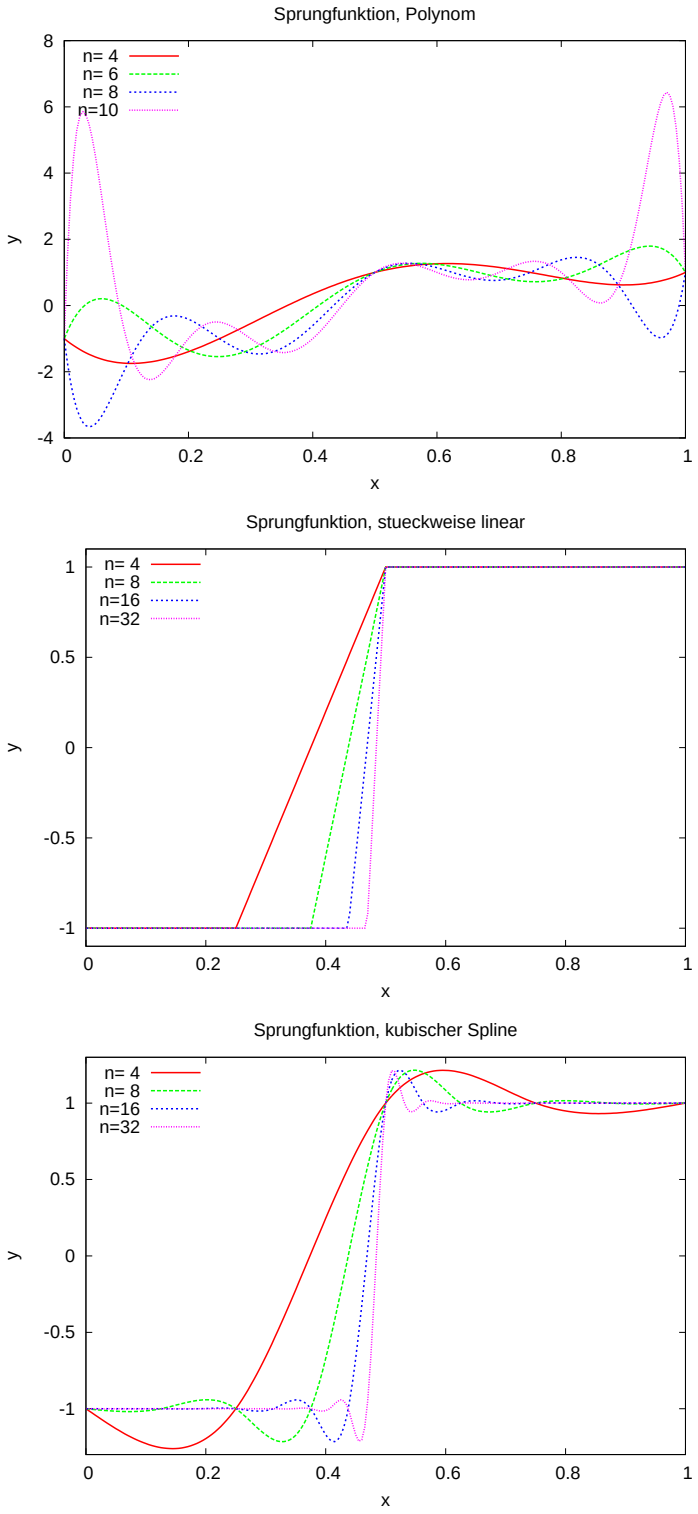


Abbildung 20: Interpolation der Funktion $f_3(x)$ mit Lagrange-Polynomen, stückweise linearen Funktionen und kubischen Splines.

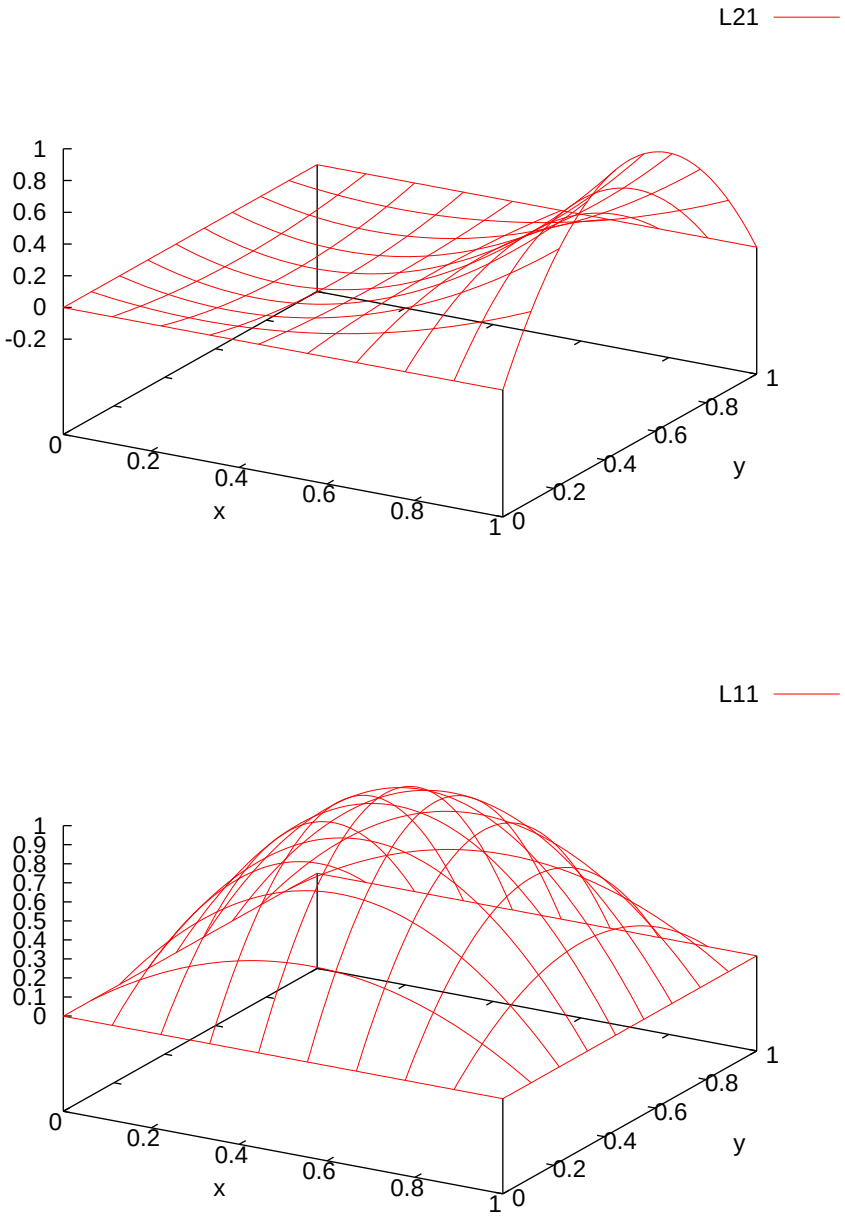


Abbildung 21: Die Lagrange-Polynome $L^{(2)}_{2,1}$ und $L^{(2)}_{1,1}$.

7 Trigonometrische Interpolation

7.1 Trigonometrische Polynome

In den Anwendungen treten oft „periodische“ Funktionen auf, d.h. für ein $\mathbb{R} \ni \omega > 0$ gilt

$$f(x+\omega) = f(x) \qquad x \in \mathbb{R}$$

Zur Darstellung solcher Funktionen im Rechner bietet sich die Interpolation mit trigonometrischen Summen

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \left\{ a_k \cos\left(\frac{kx2\pi}{\omega}\right) + b_k \sin\left(\frac{kx2\pi}{\omega}\right) \right\} \tag{7.1}$$

an, da jeder einzelne Summand bereits ω-periodisch ist ($\cos(k(x+\omega)2\pi/\omega) = \cos(kx2\pi/\omega + k2\pi)$.

(7.1) hat $2m+1$ freie Parameter, wir setzen deshalb ab sofort

$$\boxed{n := 2m}. \tag{7.2}$$

O.B.d.A. setzen wir ab sofort auch $\omega = 2\pi$, d.h. alle Funktionen sind 2π-periodisch. Als Stützstellen für die Interpolation verwenden wir

$$x_k = \frac{k}{n+1} 2\pi \qquad k = 0, \ldots, n. \tag{7.3}$$

Beachte: Wegen $f(x) = f(x+2\pi)$ ist $h = 1/n+1$.

Es zeigt sich, dass die Interpolationsaufgabe zunächst leichter im Bereich der komplexen Zahlen $\mathbb{C}$ zu lösen ist.

Dazu betrachtet man das komplexe trigonometrische Polynom

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx} \tag{7.4}$$

mit der imaginären Einheit $i = \sqrt{-1}$, komplexen Koeffizienten $c_k \in \mathbb{C}$ und der Eulerschen[13] Identität

$$e^{i\varphi} = \cos\varphi + i \sin\varphi.$$

Wir stellen zunächst einige Eigenschaften der komplexen Exponentialfunktion zusammen, die wir im folgenden benötigen.

Hilfssatz 7.1 (Komplexe Einheitswurzeln)**.** Setze $\mathbb{C} \ni w_k = e^{ix_k} = e^{i\frac{k2\pi}{n+1}}$ für alle $k \in \mathbb{Z}$. w_k nennt man k-te Einheitswurzel. Diese haben folgende Eigenschaften.

(a) $w_k^{n+1} - 1 = 0$ für alle $k \in \mathbb{Z}$.

Die w_k sind also Lösungen der Gleichung $w^{n+1} - 1 = 0$ in $\mathbb{C}$, denn

$$w_k^{n+1} = \left(e^{i\frac{k2\pi}{n+1}}\right)^{n+1} = e^{ik2\pi} = \cos k2\pi + i \sin k2\pi = 1.$$

[13] Leonhard Euler, 1707-1783, Schweitzer Mathematiker.

(b) $w_k^j = w_j^k$ für alle $k, j \in \mathbb{Z}$, denn

$$w_k^j = \left(e^{i\frac{k2\pi}{n+1}}\right)^j = e^{i\frac{kj2\pi}{n+1}} = \left(e^{i\frac{j2\pi}{n+1}}\right)^k = w_j^k.$$

(c) $w_k^{-j} = w_j^{-k}$ für alle $k, j \in \mathbb{Z}$. Zeigt man auch durch Einsetzen, ist aber nicht identisch (b).

(d) $w_k^j = w_k^{j \bmod (n+1)} = w_{k \bmod (n+1)}^j = w_{k \bmod (n+1)}^{j \bmod (n+1)}$.
Sei $j = r(n+1) + s$ mit $0 \le s < n+1$, dann rechnet man

$$w_k^j = e^{i\frac{k[r(n+1)+s]2\pi}{n+1}} = \underbrace{e^{ikr2\pi}}_{=1} e^{i\frac{ks2\pi}{n+1}} = w_k^{j \bmod (n+1)}.$$

Der Rest geht genauso.

(e) $\sum_{j=0}^{n} w_k^j = \begin{cases} n+1 & k = 0 \\ 0 & k \in \mathbb{Z} \setminus \{0\} \end{cases}$.
Ist $k = 0$ so ist $w_k = 1$ und $w_k^j = 1$ für alle j. Also ist $\sum_{j=0}^{n} w_k^j = n+1$.
Sei nun $k \neq 0$. Nach (a) ist jedes w_k Lösung von

$$w^{n+1} - 1 = (w-1)(w^n + w^{n-1} + \ldots + 1) = 0.$$

Für $k \neq 0$ ist $w_k \neq 1 \Rightarrow w_k - 1 \neq 0$. Also muss der zweite Faktor gleich 0 sein, mithin $\sum_{j=0}^{n} w_k^j = 0$.

Damit beweisen wir den folgenden

Satz 7.2 (Komplexe trigonometrische Interpolation)**.** Zu gegebenen Zahlen $y_0, \ldots, y_n \in \mathbb{C}$ gibt es genau eine Funktion der Gestalt

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}$$

die den Interpolationsbedingungen

$$t_n^*(x_j) = y_j \qquad j = 0, \ldots, n, \quad x_j = \frac{j}{n+1} 2\pi,$$

genügt. Die komplexen Koeffizienten c_k sind bestimmt durch

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j \underbrace{e^{-ijx_k}}_{=w_k^{-j}} \qquad k = 0, \ldots, n. \tag{7.5}$$

Beweis: Mit $w = e^{ix}$ gilt offensichtlich

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx} = \sum_{k=0}^{n} c_k \left(\underbrace{e^{ix}}_{w}\right)^k = \sum_{k=0}^{n} c_k w^k = p_n(w).$$

t_n^* ist also ein komplexes Polynom n-ten Grades in w. Die Interpolationsbedingungen lauten entsprechend

$$t_n^*(x_k) = p_n(x_k) = y_k \qquad k = 0, \ldots, n.$$

Die Polynominterpolation ist im Komplexen eindeutig, so auch t_n^*. Zur Berechnung der Koeffizienten. Für beliebiges k gilt

$$\sum_{j=0}^{n} y_j w_k^{-j} = \sum_{j=0}^{n} t_n^*(x_j) w_k^{-j} = \sum_{j=0}^{n} \left(\sum_{l=0}^{n} c_l \underbrace{e^{ilx_j}}_{=w_j^l} \right) w_k^{-j} = \sum_{l=0}^{n} c_l \left(\sum_{j=0}^{n} w_j^{l-k} \right).$$

Für die zweite Summe gilt $\sum_{j=0}^{n} w_j^{l-k} = \sum_{j=0}^{n} w_{l-k}^{j} = \begin{cases} n+1 & l = k \\ 0 & l \neq k \end{cases}$.

Damit bleibt aus der *äußeren* Summe nur ein Summand für $l = k$ übrig:

$$\sum_{j=0}^{n} y_j w_k^{-j} = c_k(n+1) \quad \Leftrightarrow \quad c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j w_k^{-j} = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k} \quad .$$

□

Damit haben wir insbesondere auch für *reelle* y_j die Interpolationsaufgabe gelöst.

Wegen

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx} = \sum_{k=0}^{n} c_k (\cos kx + i \sin kx)$$

ist das schon fast eine trigonometrische Summe.

Es stellt sich heraus, dass bei reellen Daten y_j die Koeffizienten c_k dergestalt sind, dass $t_n^*(x) \in \mathbb{R}$ für $x \in \mathbb{R}$.

7.2 Diskrete Fourier-Analyse

Aus dem komplexen trigonometrischen Interpolationspolynom kann auch die trigonometrische Summe mit ihren *reellen* Koeffizienten bestimmt werden.

Satz 7.3 (Diskrete Fourier[14]-Analyse)**.** Für $n \in \mathbb{N}_0$ gibt es zu gegebenen reellen Zahlen $y_0, \ldots, y_n$ genau ein trigonometrisches Polynom der Form

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \{a_k \cos(kx) + b_k \sin(kx)\} + \frac{\theta}{2} a_{m+1} \cos((m+1)x)$$

mit $t_n(x_j) = y_j \quad j = 0, \ldots, n$ sowie

$$\theta = 0,\ m = \frac{n}{2} \qquad n \text{ gerade} \qquad \rightarrow a_0, \ldots a_m, b_1 \ldots, b_m$$

$$\theta = 1,\ m = \frac{n-1}{2} \qquad n \text{ ungerade} \qquad \rightarrow a_0, \ldots a_{m+1}, b_1, \ldots, b_m$$

n gerade: $2 \cdot (n/2) + 1 = n + 1$, n ungerade: $2 \cdot (n-1)/2 + 2 = n + 1$.

[14] Jean-Baptiste de Fourier, 1768-1830, frz. Mathematiker und Physiker.

Die Koeffizienten werden bestimmt durch

$$a_k = \frac{2}{n+1}\sum_{j=0}^{n} y_j \cos(jx_k), \qquad b_k = \frac{2}{n+1}\sum_{j=0}^{n} y_j \sin(jx_k), \tag{7.6}$$

oder äquivalent dazu aus den Koeffizienten des komplexen Interpolationspolynoms:

$$a_0 = 2c_0, \tag{7.7a}$$
$$a_k = c_k + c_{n+1-k}, \qquad k = 1, \ldots, m, \tag{7.7b}$$
$$b_k = i(c_k - c_{n+1-k}), \qquad k = 1, \ldots, m, \tag{7.7c}$$
$$a_{m+1} = 2c_{m+1}, \qquad n = 2m+1 \quad (n \text{ ungerade}). \tag{7.7d}$$

Beweis: siehe [Ran06].

Wir zeigen hier nur, dass die Koeffizienten tatsächlich reell sind: Z. B. das a_k:

$$\begin{aligned} a_k &= c_k + c_{n+1-k} \\ &= \frac{1}{n+1}\sum_{j=0}^{n} y_j \left(e^{-ijx_k} + e^{-ijx_{(n+1)-k}}\right) = \frac{1}{n+1}\sum_{j=0}^{n} y_j \left(e^{-ijx_k} + e^{ijx_k}\right) \\ &= \frac{1}{n+1}\sum_{j=0}^{n} y_j [\cos(-jx_k) + i\sin(-jx_k) + \cos(jx_k) + i\sin(jx_k)] \\ &= \frac{2}{n+1}\sum_{j=0}^{n} y_j \cos(jx_k). \end{aligned}$$

□

7.3 Praktisches zur Diskreten Fourier Analyse

Abbildung 22 zeigt einige Beispiele für Spektren. Die Konstante im Zeitbereich hat einen Puls als Spektrum. Umgedreht hat ein Puls im Ortsbereich ein konstantes Spektrum. Schließlich wird noch das Spektrum eines Dreiecks- bzw. Rechtecksignals gezeigt.

Abbildung 23 zeigt die Interpolation von Dreick- bzw. Rechtecksignal bei Vorgabe von jeweils acht Datenpunkten.

Abbildung 24 illustriert die Verbesserung der Annäherung an die zu interpolierende Funktion bei steigendem Parameter n.

Abbildung 24 illustriert die Verbesserung der Annäherung bei unstetigen Funktionen, wenn an der Sprungstelle der Mittelwert vorgeschrieben wird. Wir verwenden einmal $n = 15$ (Sprungstelle ist Interpolationspunkt, Mittelwert wird vorgeschrieben) und $n = 16$ (Sprungstelle ist kein Interpolationspunkt).

7.4 Trigonometrische Approximation

Auch bei der trigonometrischen Interpolation kann man die Frage nach dem Interpolationsfehler zwischen den Stützpunkten stellen.

Wir betrachten dazu die folgende allgemeinere Aufgabe: Mit

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \{a_k \cos(kx) + b_k \sin(kx)\} \quad (2m+1 \text{ Parameter, } n = 2m)$$

und einer gegebenen Funktion $f(x)$ betrachte die Approximationsaufgabe

$$\|t_n(x) - f(x)\|_2 := \left[\int_{-\pi}^{\pi} (t_n(x) - f(x))^2 dx\right]^{1/2} \to \min \quad .$$

Es zeigt sich, dass

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx \qquad k = 0,\ldots,m \quad (m+1) \text{ Koeffizienten}$$
$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx \qquad k = 1,\ldots,m \quad m \text{ Koeffizienten}$$

das Problem löst.

Dabei ist

$$(f,g)_2 = \int_{-\pi}^{\pi} f(x)g(x)dx$$

das sogenannte L_2-Skalarprodukt auf dem Raum der quadratintegrierbaren Funktionen

$$L_2(a,b) = \{u : (a,b) \to \mathbb{R} \mid \int_{-\pi}^{\pi} u^2\, dx < \infty\}.$$

Die Funktionen $\cos(kx), \sin(kx)$ bilden ein „Orthogonalsystem“, d. h.

$$\int_{-\pi}^{\pi} \cos jx \cos kx\, dx = \begin{cases} 0 & j \neq k \\ 2\pi & j = k = 0 \\ \pi & j = k > 0 \end{cases}$$
$$\int_{-\pi}^{\pi} \sin jx \sin kx\, dx = \begin{cases} 0 & j \neq k, k, j > 0 \\ \pi & j = k > 0 \end{cases}$$
$$\int_{-\pi}^{\pi} \cos jx \sin kx\, dx = 0 \quad j \geq 0, k > 0 \quad .$$

Diese Konstruktion funktioniert für beliebig großes m.

Für $m = \infty$ spricht man auch von Fourier-Reihe.

Man kann weiter zeigen:

- Die Fourier-Reihe (unendliches m) konvergiert genau für die Funktionen aus $L_2(a,b)$.

- Für stückweise stetig differenzierbare Funktionen konvergiert die Fourier-Reihe gegen $f(x_0)$ falls $f(x)$ bei x_0 stetig, sonst gegen den Mittelwert aus links- und rechtsseitigem Grenzwert.
- Für endliches n löst $t_n(x)$ die anfangs gestellte Approximationsaufgabe.

Für Details sei auf [SK05] verwiesen.

Wir wollen nun zeigen welche Beziehung zwischen der Approximationsaufgabe und der Diskreten Fourier-Analyse besteht.

Für die endliche Fourier-Reihe sind die Koeffizienten mittels

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx \qquad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx$$

zu berechnen.

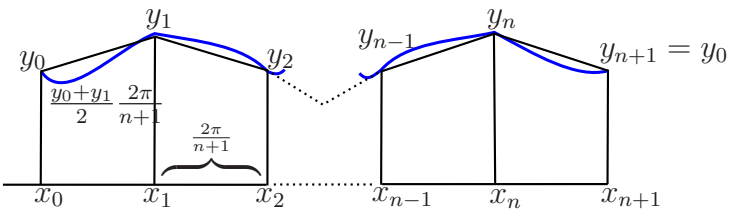
Berechnet man diese Integrale *näherungsweise* mit der *Trapezregel* (ausführlich nächste Stunde) so ergeben sich dieselben Koeffizienten wie bei der Diskreten Fourier-Analyse (also aus der trigonometrischen Interpolation):

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx \\ &\approx \frac{1}{\pi}\sum_{i=0}^{n}\left\{\frac{f(x_i)\cos(kx_i)+f(x_{i+1})\cos(kx_{i+1})}{2}\frac{2\pi}{n+1}\right\} \\ &= \frac{2\pi}{\pi(n+1)}\sum_{i=0}^{n} f(x_i)\cos(kx_i) \end{aligned}$$

Auch bei der Fourier-Reihe gibt es übrigens das Gibbsche Phänomen.

Bei der Trapezregel wird der Integrand stückweise linear angenähert und diese Funktion dann exakt integriert.

Als Stützwerte werden hier genau wieder die $x_i = -\pi + i\frac{2\pi}{n+1}$ gewählt.



7.5 Schnelle Fourier-Transformation

Wir gehen zurück zur komplexen trigonometrischen Interpolation $t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}$, welche an den Stellen $x_k = \frac{k2\pi}{n+1}$ den Wert y_k interpoliert.

Mit der Abkürzung $N = n+1$ lauten nach (7.5) die Koeffizenten:

$$c_k = \frac{1}{N}\sum_{j=0}^{N-1} y_j e^{-ijk2\pi/N} \qquad k = 0,\ldots,N-1 \quad \text{(Hintransformation)}. \tag{7.8}$$

Hat man die c_k berechnet so erhält man die y_j zurück mittels

$$y_j = \sum_{k=0}^{N-1} c_k e^{ikj2\pi/N} \qquad j = 0, \ldots, N-1 \quad \text{(Rücktransformation)}. \tag{7.9}$$

Diese beiden Gleichungen beschreiben eine bijektive Zuordnung der Daten y_j und der Koeffizienten c_k, die man auch als *diskrete Fourier-Transformation* (DFT) bezeichnet.

Fasse nun Daten und Koeffizienten in Vektoren aus $\mathbb{C}^N$ zusammen:

$$\underline{y} = (y_0, \ldots, y_{N-1})^T, \qquad \underline{c} = (c_0, \ldots, c_{N-1})^T \quad .$$

Hintransformation kann man als Matrix-Vektor-Produkt schreiben:

$$\underline{c} = N^{-1} W \underline{y}, \qquad \text{mit } (W)_{k,j} = e^{-ijk2\pi/N}.$$

Als Beispiel betrachte $N = 4$, dann hat man

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \frac{1}{N} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & w^{-1} & w^{-2} & w^{-3} \\ 1 & w^{-2} & 1 & w^{-2} \\ 1 & w^{-3} & w^{-2} & w^{-1} \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad .$$

mit der komplexen N-ten Einheitswurzel $w = e^{i2\pi/N}$

Schreibt man für die Rücktransformation $\underline{y} = U\underline{c}$ so gilt wegen $UN^{-1}W = I$ offensichtlich

$$W^{-1} = \frac{1}{N} U, \qquad (W^{-1})_{j,k} = \frac{1}{N} e^{ikj2\pi/N} = \frac{1}{N} w^{kj}.$$

Da die $N \times N$ Matrix W voll besetzt ist beträgt der Aufwand für Hin- und Rücktransformation je $O(N^2)$.

Der Berechnungsaufwand kann mittels der von Cooley[16] und Tukey[17] 1965 entwickelten schnellen Fourier-Transformation deutlich reduziert werden.

Wir betrachten c_k *ohne* den Vorfaktor $1/N$, also $\tilde{c}_k = \sum_{j=0}^{N-1} y_j e^{-ijk2\pi/N}$.

Falls N *gerade*, so gilt für die $\tilde{c}_k$:

$$\begin{aligned} \tilde{c}_k &= \underbrace{\sum_{j=0}^{N/2-1} y_{2j} e^{-i2jk2\pi/N}}_{\text{gerader Teil}} + \underbrace{\sum_{j=0}^{N/2-1} y_{2j+1} e^{-i(2j+1)k2\pi/N}}_{\text{ungerader Teil}} \\ &= \underbrace{\sum_{j=0}^{N/2-1} y_{2j} e^{-ijk2\pi/(N/2)}}_{=:\tilde{c}^g_k \text{ für } k=0,\ldots,N/2-1} + e^{-ik2\pi/N} \underbrace{\sum_{j=0}^{N/2} y_{2j+1} e^{-ijk2\pi/(N/2)}}_{=:\tilde{c}^u_k \text{ für } k=0,\ldots,N/2-1} \end{aligned}$$

[16] James Cooley, *1926, amerik. Mathermatiker.
[17] John Tukey, 1915-2000, amerik. Mathematiker.

Das gilt für alle $k \in 0, \ldots, N-1$ aber wegen der $N/2$-Periodizität von $e^{-ik2\pi/(N/2)}$ gilt

$$\tilde{c}^g_{k+N/2} = c^g_k, \qquad \tilde{c}^u_{k+N/2} = \tilde{c}^u_k, \qquad k = 0, \ldots, \frac{N}{2} - 1$$

Zusammen mit $e^{-i(k+N/2)2\pi/N} = e^{-ik2\pi/N} e^{-i\pi} = -e^{-ik2\pi/N}$ muss man also nur die Koeffizienten

$$\tilde{c}^g_k = \sum_{j=0}^{N/2-1} y_{2j} e^{-ijk2\pi/(N/2)}, \quad \tilde{c}^u_k = \sum_{j=0}^{N/2} y_{2j+1} e^{-ijk2\pi/(N/2)}, \quad 0 \le k < N/2$$

berechnen und setzt dann

$$\tilde{c}_k = \tilde{c}^g_k + e^{-ik2\pi/N} \tilde{c}^u_k, \qquad \tilde{c}_{k+N/2} = \tilde{c}^g_k - e^{-ik2\pi/N} \tilde{c}^u_k, \qquad 0 \le k < N/2. \tag{7.10}$$

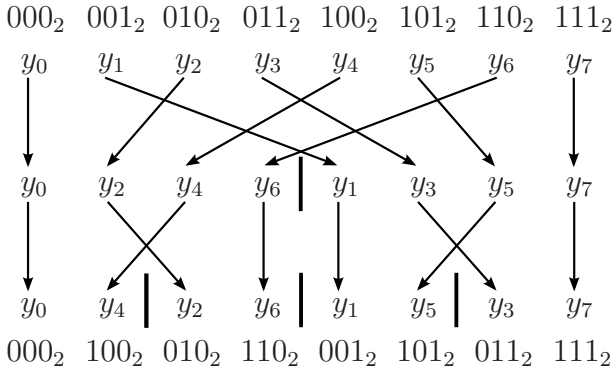
Somit wurde die DFT der Länge N auf 2 DFT der Länge $N/2$ zurückgeführt.

Für den Aufwand $A(N)$ in Anzahl Gleitkomma-Operationen gilt:

$$\begin{aligned} A(N) &= 2A\left(\frac{N}{2}\right) + cN = 2\left[2A\left(\frac{N}{4}\right) + c\frac{N}{2}\right] + cN \\ &= 4A\left(\frac{N}{4}\right) + cN + cN = \cdots \\ &= NA(0) + pN = O(N \log_2 N). \end{aligned}$$

Hierbei haben wir vorausgesetzt, dass N eine Zweierpotenz ist.

In der Abstiegsphase der Rekursion werden die Eingabedaten umsortiert bis $N = 2$ erreicht ist:

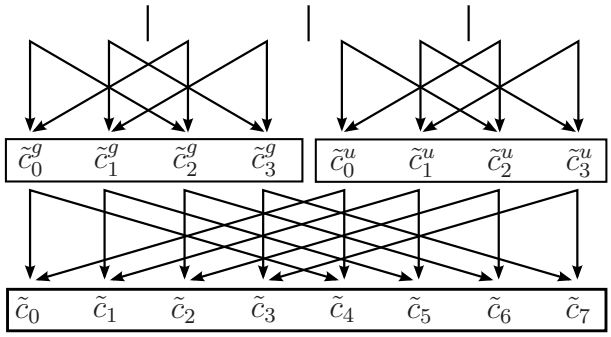


Die Umordnung der Indizes leistet die Operation „Bitreversal“:

$$(b_{x-1} \ldots b_0)_2 \to (b_0 \ldots b_{k-1})_2$$

Für $N = 2$ führt man die DFT direkt durch.

In der Aufstiegsphase erfolgt nur noch die Kombination der Koeffizienten mittels (7.10). Die Struktur dieser Berechnung ist:



Dieses Muster nennt man auch *perfect shuffle*.

7.6 Zusammenfassung

- Zur Interpolation von periodischen Funktionen haben wir in diesem Kapitel trigonometrische Summen kennengelernt.
- Mit dem Übergang zu komplexen Zahlen erkennt man, dass dies nichts anderes ist als eine komplexe Interpolationsaufgabe mit äquidistanten Stützstellen auf dem Einheitskreis in der komplexen Zahlenebene. Damit erhält man die Eindeutigkeit und eine Darstellung der Koeffizienten. Man spricht auch von (komplexer) Diskreter Fourier-Transformation.
- Aus den komplexen Koeffizienten erhält man auch die reellen Koeffizienten für die trigonometrische Summe (Diskrete Fourier-Analyse).
- Die Koeffizienten der Diskreten Fourier-Analyse erhält man auch durch näherungsweise Berechnung der Koeffizienten der Fourier-Reihe mittels der Trapezregel.
- Der Aufwand der DFT beträgt $O(N^2)$ bei N Datenpunkten. Mittels schneller Fourier-Transformation (FFT) kann man das auf $O(N \log_2 N)$ drücken.

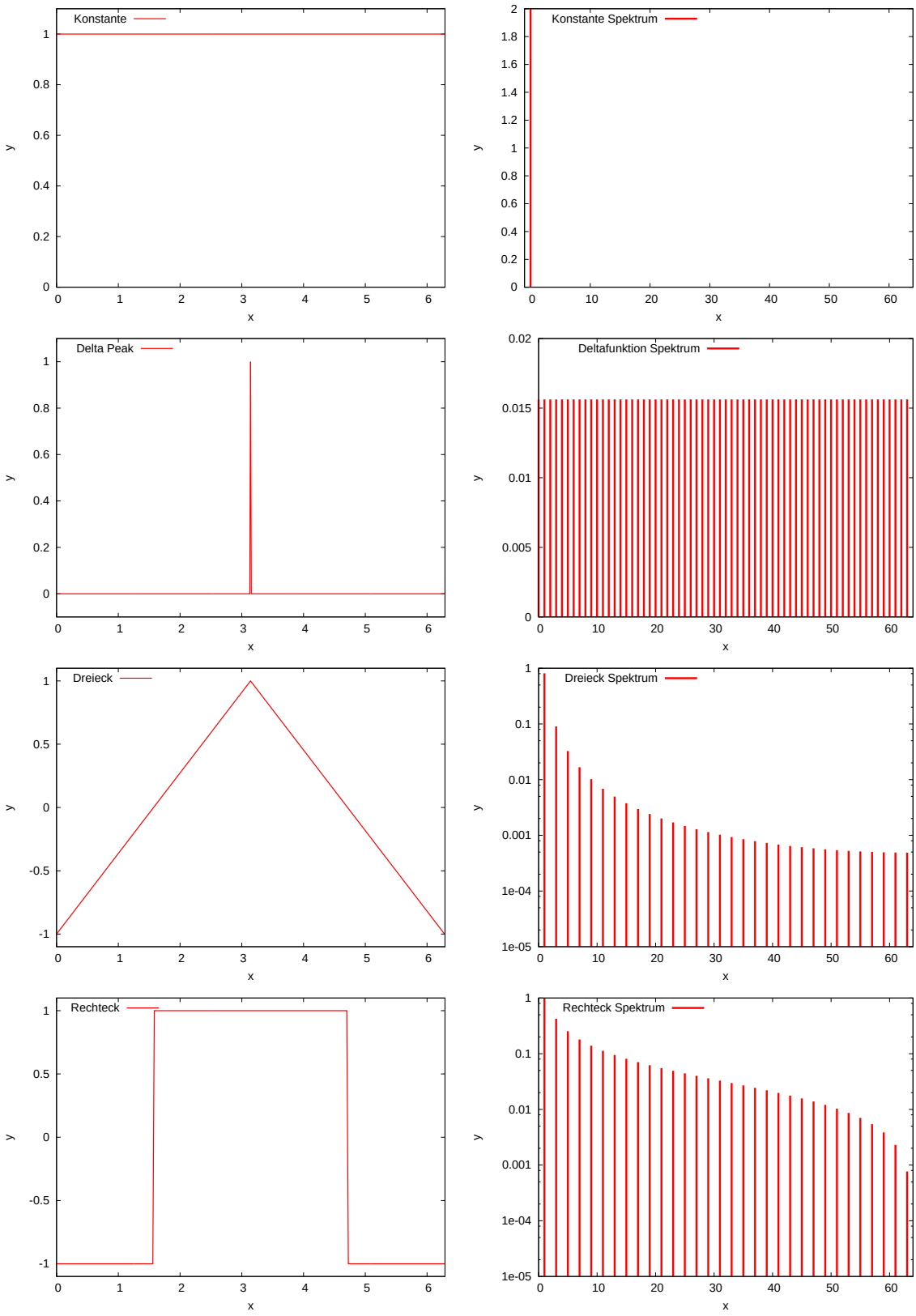


Abbildung 22: Spektren zu verschiedenen Funktionen.

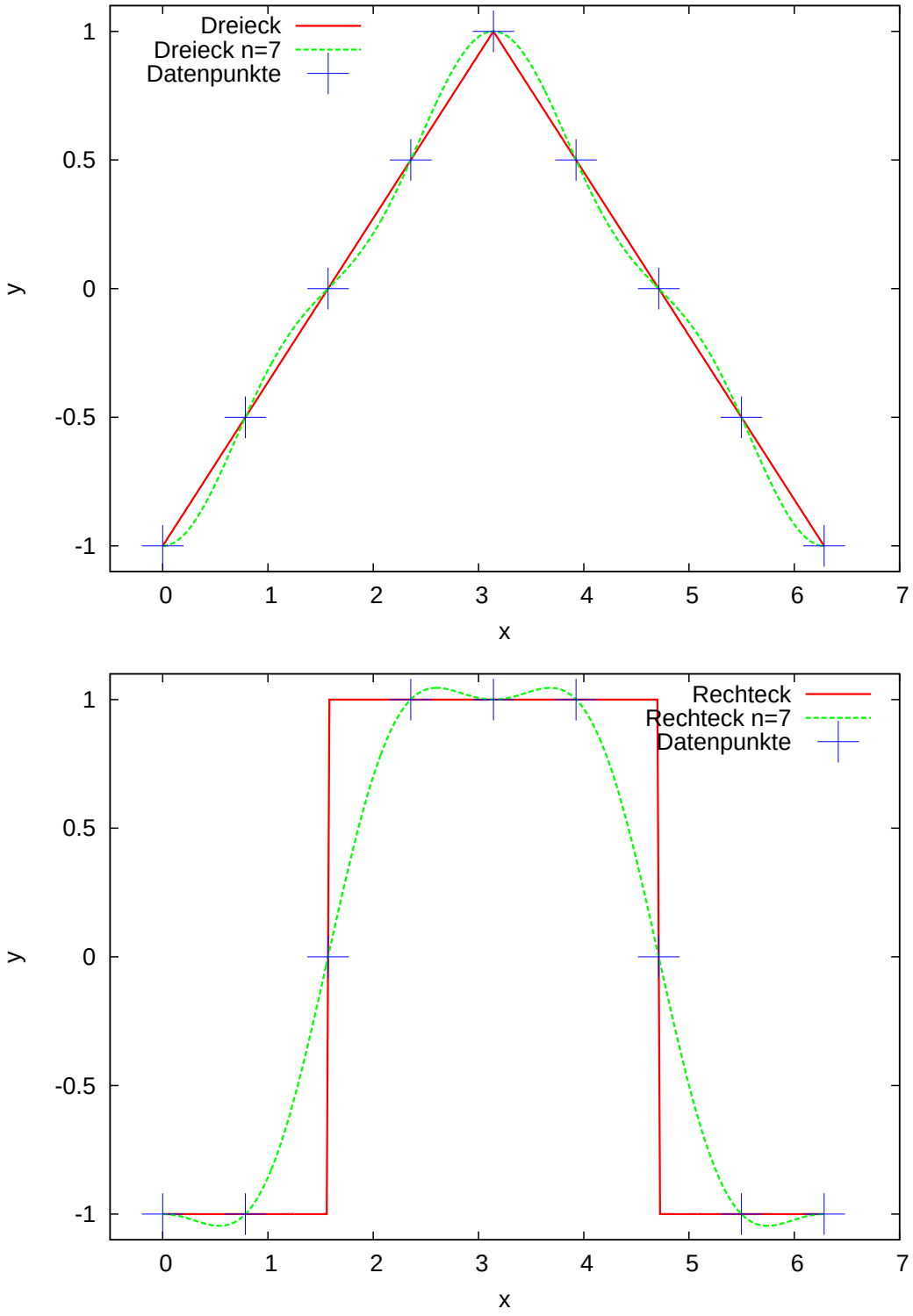


Abbildung 23: Interpolation verschiedener Funktionen.

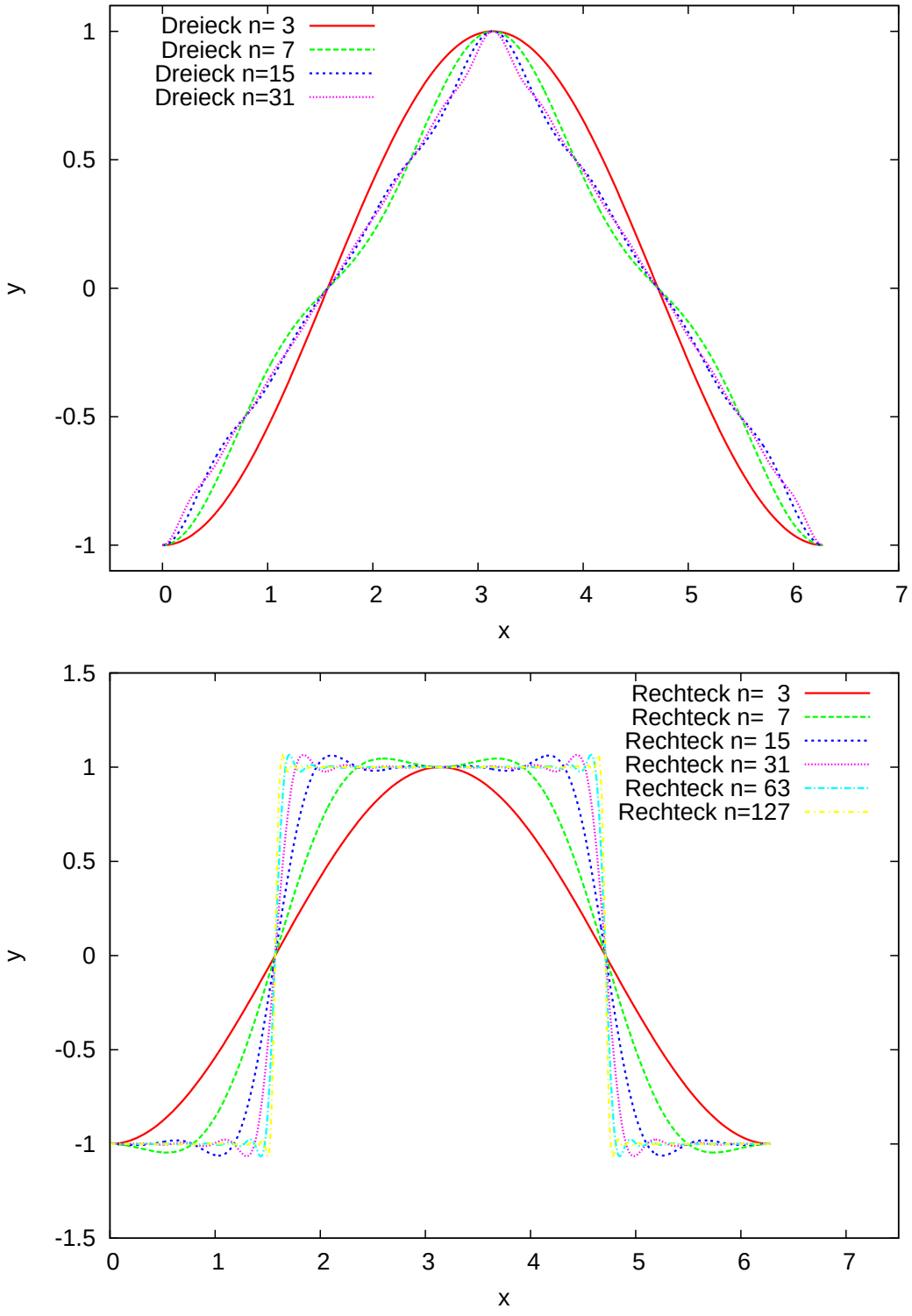


Abbildung 24: Approximation verschiedener Funktionen bei steigendem n.

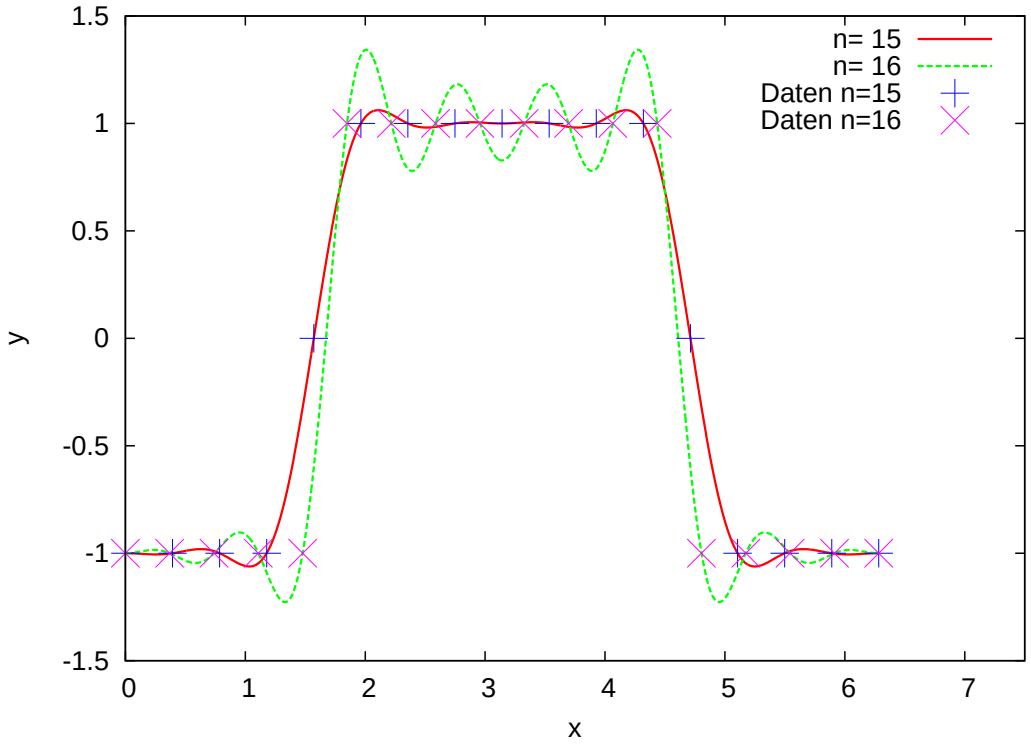


Abbildung 25: Interpolation einer unstetigen Funtion.

8 Quadraturen niedriger Ordnung

8.1 Die Integrationsaufgabe

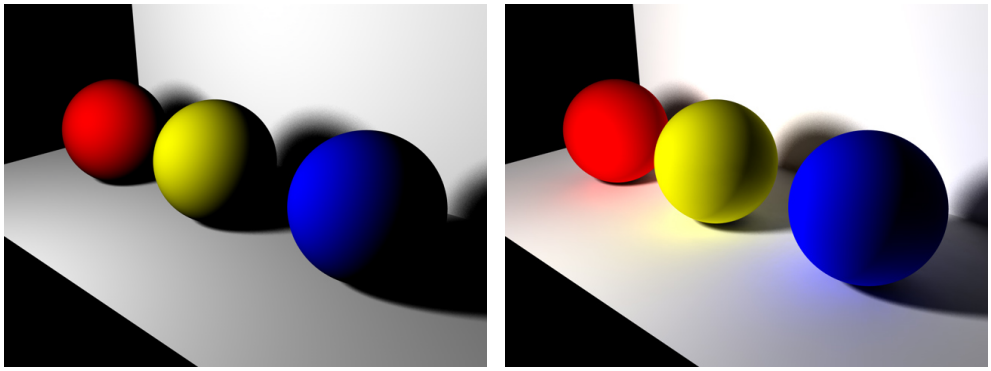
Klassische, seit dem Altertum wichtige Anwendungen sind die Berechnung von Flächeninhalten, Volumina oder Schwerpunkten. Der Schwerpunkt eines Körpers mit der (ortsabhängigen) Dichte $\rho(x)$ ist z. B. gegeben durch

$$\vec{s} = \frac{\int_\Omega \rho(\vec{x})\vec{x}d\vec{x}}{\int_\Omega \rho(\vec{x})d\vec{x}}.$$

In der Physik ist die Arbeit eine integrale Größe:

$$W = \int_\Gamma \vec{F}(\vec{s})\, d\vec{s}.$$

In der Stochastik werden wir Verteilungsfunktionen und Momente kennenlernen. Im Fall kontinuierlicher Zufallsgrößen sind dies Integrale.



(Quelle: Wikipedia)

Das Radiosity-Verfahren (rechtes Bild) ist ein Verfahren der Computergrafik zum Rendering von Szenen.

Im Vergleich zu Raytracing (linkes Bild, dem anderen großen Verfahren) ist es in der Lage, ideal diffuse Reflexion zu modellieren.

Beispiel: In einem durch ein Fenster beleuchteten Zimmer sind nicht nur die direkt vom Fenster aus sichtbaren Flächen erhellt.

Dieses Verhalten modelliert man mit der Radiosity-Gleichung:

$$B(x) = E(x) + \rho(x) \int_S B(x') \frac{1}{\pi r^2} \cos\phi_x \cos\phi_{x'} V(x, x')\, dA'.$$

$x \in S$ Punkt auf einer Oberfläche der Szene S.

$B(x)$ vom Punkt x abgestrahlte Energie (genauer: Leistung pro Fläche) in Form von Licht.

$E(x)$ Eigenstrahlung im Punkt x.

$\rho(x)$ Reflexionsfaktor in x.

r Abstand zwischen x und x'.

ϕ_x Winkel zwischen Normale und Verbindungslinie x, x'.

$V(x, x')$ 1 falls x von x' aus sichtbar und sonst 0.

Dies ist eine *Integralgleichung* für die unbekannte Funktion $B(x)$.

Diese kann durch Diskretisierung der Oberfläche S näherungsweise gelöst werden. Dies führt dann auf ein lineares Gleichungssystem.

Die *Variationsrechnung*, ein Zweig der Mathematik, beschäftigt sich mit „Funktionen von Funktionen“, oft eben Integrale über eine Funktion und/oder deren Ableitungen. Die Variationsrechnung hat unzählige praktische Anwendungen, etwa in der Mechanik.

Manche numerische Lösungsverfahren für partielle Differentialgleichungen benötigen auch die Berechnung von Integralen zur Aufstellung eines Gleichungssystems.

Schließlich ein letztes Beispiel: Die Koeffizienten der Fourierreihe berechnen sich über ein Integral.

Man sieht also, Integrale kommen in der technisch-wissenschaftlichen Praxis sehr häufig vor.

Viele Integrale lassen sich nicht geschlossen lösen, dann hilft nur noch eine numerische Berechnung.

Allerdings: Vor einer numerischen Berechnung sollten alle analytischen Möglichkeiten ausgeschöpft werden (Substitution, partielle Integration, Aufspaltung, insbesondere bei singulären Integranden).

Zu berechnen sei also ein Näherungswert des bestimmten Integrals

$$I(f) = \int_\Omega f(x)\,dx$$

für eine gegebene Funktion

$$f : \mathbb{R}^d \to \mathbb{R}$$

und ein Integrationsgebiet

$$\Omega \subset \mathbb{R}^d.$$

Man spricht auch von *numerischer Quadratur*.

Wir behandeln fast ausschließlich $d = 1$, sprechen den Fall $d > 1$ aber kurz an. Für große d (sog. hochdimensionale Integrale, z.B. $d > 5$) sind andere Methoden erforderlich!

Alle (von uns behandelten) Methoden haben folgende Form:

$$I(f) = \sum_{i=0}^{n} w_i f(x_i) + \text{ Fehler}, \qquad w_i \text{ heißt Gewicht und } x_i = \text{Stützstelle}.$$

Der Fehler hängt vom Verfahren und (höheren) Ableitungen von f ab.

Man möchte effiziente Verfahren: Möglichst kleiner Fehler bei möglichst wenig Funktionsauswertungen (Aufwand).

8.2 Newton-Cotes Formeln

Als erstes betrachten wir sog. Newton-Cotes[18] Formeln, die ein Spezialfall interpolatorischer Quadraturformeln sind.

Diese basieren auf folgender *Idee*: Interpoliere die $n+1$ Werte $(x_i, f(x_i))$, $i = 0, \ldots, n$, durch ein Polynom n-ten Grades und integriere dieses exakt.

Darstellung der Polynome mittels Lagrangeinterpolation liefert

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i^{(n)}(x), \qquad L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

also ist

$$I(f) \approx I^{(n)}(f) = \int_a^b p_n(x) dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_i^{(n)}(x)\, dx. \tag{8.1}$$

Definition 8.1 (Ordnung einer Quadraturformel)**.** Eine Quadraturformel $I^{(n)}(f)$ hat mindestens die Ordnung m, wenn sie Polynome vom Grad $m-1$ exakt integriert. □

Folgerung: Interpolatorische Quadraturformeln zu $n+1$ Stützstellen haben mindestens die Ordnung $n+1$. (Folgt aus Eindeutigkeit der Polynominterpolation, $n+1$ Stützstellen $\Rightarrow$ Grad $n \Rightarrow$ Ordnung $n+1$).

Diese Formulierung legt nahe, dass man mit $n+1$ Stützstellen auch eine Ordnung höher als $n+1$ erreichen kann, was tatsächlich der Fall ist.

Bemerkung 8.2. Es gilt

$$\sum_{i=0}^{n} w_i = b - a$$

da $f \equiv 1$ auf das Interpolationspolynom $p_n \equiv 1$ führt. □

Newton-Cotes-Formeln: Interpolatorische Quadratur zu *äquidistanten* Stützstellen. Man unterscheidet zwei Varianten:

Variante a) „Abgeschlossene“ Newton-Cotes-Formel (a, b Stützstellen)

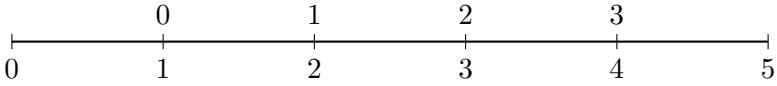
$$x_i = a + iH, \quad i = 0, \ldots, n, \quad H = \frac{b-a}{n}$$

Variante b) „Offene“ Newton-Cotes-Formel (a, b keine Stützstellen)

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

Stützpunkte der abgeschlossenen Formeln für $n = 5$ (unten) und der offenen Formeln für $n = 3$ (oben):

[18] Roger Cotes, 1682-1716, engl. Mathematiker



Beachte: Die offenen Formeln benutzen Werte des Interpolationspolynoms außerhalb des Intervalls der Stützstellen, das wird sich als ungünstig erweisen.

Berechnung der Gewichte (abgeschlossene Formeln): nach Konstruktion gilt

$$I^{(n)}(f) \overset{(8.1)}{=} \sum_{i=0}^{n} f(x_i) \int_a^b L_i^{(n)}(x)dx = (b-a) \sum_{i=0}^{n} \underbrace{\left(\frac{1}{b-a} \int_a^b L_i^{(n)}(x)dx \right)}_{=:w_i \text{ wird unabh. von } a,b} f(x_i).$$

Integration durch Substitution: $x = g(s) = a + sH \Rightarrow s = g^{-1}(x) = \frac{x-a}{H}$, $g'(x) = H$:

$$\begin{aligned} w_i &\overset{def}{=} \frac{1}{b-a} \int_a^b L_i^{(n)}(x)dx \overset{Subst.}{=} \frac{1}{b-a} \int_{g^{-1}(a)}^{g^{-1}(b)} L_i^{(n)}(a+sH) \cdot g'(s)ds \\ &= \frac{1}{b-a} \underbrace{\frac{b-a}{n}}_{H} \int_0^n \prod_{\substack{j=0 \\ i \neq j}}^{n} \frac{[a+sH-(a+jH)]}{[a+iH-(a+jH)]} ds \\ &= \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ i \neq j}}^{n} \frac{(s-j)}{(i-j)} \, ds. \end{aligned}$$

w_i ist unabhängig von a, b und kann für jedes n, i vorab berechnet werden.

Nach Auswertung erhält man folgende Formeln:

Beispiel 8.3 (Newton-Cotes Formeln)**.** *Abgeschlossene Formeln* $n = 1, 2, 3$, $H = \frac{b-a}{n}$

$$I^{(1)}(f) = \frac{b-a}{2} \{f(a) + f(b)\} \quad \text{(Trapez-, Sehnen-Trapezregel)},$$

$$I^{(2)}(f) = \frac{b-a}{6} \left\{ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right\} \text{(Simpson}^{19}\text{-, Keplersche}^{20}\text{ Fassregel)},$$

$$I^{(3)}(f) = \frac{b-a}{8} \{f(a) + 3f(a+H) + 3f(b-H) + f(b)\} \quad (\tfrac{3}{8}\text{-Regel}).$$

Offene Formeln $n = 0, 1, 2$, $H = \frac{b-a}{n+2}$

$$I^{(0)}(f) = (b-a) f\left(\frac{a+b}{2}\right) \quad \text{(Mittelpunkt-, Tangenten-Trapez-, Rechteckregel)},$$

$$I^{(1)}(f) = \frac{b-a}{2} \{f(a+H) + f(b-H)\},$$

$$I^{(2)}(f) = \frac{b-a}{3} \left\{ 2f(a+H) - f(\frac{a+b}{2}) + 2f(b-H) \right\}.$$

Weitere Beispiele in [Ran06]. □

Bemerkung 8.4 (Negative Gewichte)**.** Ab $n = 7$ für abgeschlossene und $n = 2$ für offene Newton-Cotes-Formeln treten negative Gewichte w_i auf. Dies ist aus folgenden Gründen ungünstig.

- Bei positivem Integranden f (Fläche, Volumen: $f \equiv 1$) und negativen Gewichten besteht wegen $\sum_{i=0}^{n} w_i = 1$ erhöhte Gefahr der Auslöschung.
- Konditionsbetrachtung: Sei $f(x_i)$ gestört um Δy_i mit $|\Delta y_i| \leq \varepsilon$ so gilt:

$$I^{(n)}(\tilde{f}) = \sum_{i=0}^{n} w_i \Big(f(x_i) + \Delta y_i \Big) = \underbrace{\sum_{i=0}^{n} w_i f(x_i)}_{I^{(n)}(f)} + \underbrace{\sum_{i=0}^{n} w_i \ \Delta y_i}_{\Delta I^{(n)}(f)}$$

  wobei wir die Differenz abschätzen können zu (Dreiecksungleichung):

$$\left| I^{(n)}(\tilde{f}) - I^{(n)}(f) \right| = \left| \Delta I^{(n)}(f) \right| = \left| \sum_{i=0}^{n} w_i \Delta y_i \right| \leq \varepsilon \sum_{i=0}^{n} |w_i|.$$

  Sind alle w_i positiv so gilt

$$\sum_{i=0}^{n} |w_i| = \sum_{i=0}^{n} w_i = b - a.$$

  Ansonsten kann der Verstärkungsfaktor größer werden. □

Welchen Fehler begeht man nun bei der numerischen Integration?

Satz 8.5 (Restglieder)**.** Es gelten folgende Restglieddarstellungen:

(*i*) Trapezregel:

$$I(f) - \frac{b-a}{2} \{ f(a) + f(b) \} = -\frac{(b-a)^3}{12} f''(\xi), \quad f \in C^2[a,b].$$

(*ii*) Simpson-Regel:

$$I(f) - \frac{b-a}{6} \left\{ f(a) + 4f\left(\frac{a+b}{2} \right) + f(b) \right\} = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \quad f \in C^4[a,b].$$

(*iii*) Mittelpunktregel:

$$I(f) - (b-a) f\left(\frac{a+b}{2} \right) = \frac{(b-a)^3}{24} f''(\xi), \quad f \in C^2[a,b].$$

Für gewisse Zwischenstellen $\xi \in [a,b]$.

Beweis: Der Fehler bei der Polynominterpolation (Grad n) war:

$$f(x) - p(x) = \frac{f^{(n+1)}(\eta(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_j) \qquad \text{Satz 4.5}$$

also

$$\int_a^b f(x) - p(x)dx = I(f) - I^{(n)}(f) = \frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\eta(x))\prod_{j=0}^{n}(x - x_j)dx.$$

(i) Also speziell für die Trapezregel:

$$I(f) - I^{(1)}(f) = \frac{1}{2}\int_a^b f''(\eta(x))\underbrace{(x-a)(x-b)}_{=:g(x)}dx.$$

Da f'' stetig und $g(x) \leq 0$, $\forall x \in [a,b]$, gilt der verallgemeinerte erste Mittelwertsatz der Integralrechnung

$$I(f) - I^{(1)}(f) = \frac{1}{2}f''(\xi)\int_a^b (x-a)(x-b)dx = -\frac{(b-a)^3}{12}f''(\xi) \qquad \xi \in [a,b].$$

(ii, iii) siehe [Ran06] (schwieriger wegen Vorzeichenwechseln von $g(x)$). □

Bemerkung 8.6.

- Mittelpunktregel hat den halben Fehler der Trapezregel bei nur einer f-Auswertung.
- Restglieder haben immer die typische Form

$$C(b-a)^{m+1}f^{(m)}(\xi)$$

□

8.3 Summierte Quadraturformeln

Erhöhen von n zur Genauigkeitssteigerung scheidet aus, da

- einige Gewichte w_i ab $n = 7$ (abgeschlossene Formeln, offene früher) negativ werden,
- die Lagrange-Interpolation zu äquidistanten Stützstellen nicht punktweise konvergiert.

Wie bei der Interpolation auch unterteilt man stattdessen das Integrationsintervall $[a, b]$ in N Teilintervalle

$$[x_i, x_{i+1}] \qquad x_i = a + ih, \quad i = 0, \dots N-1, \quad h = \frac{b-a}{N}$$

und wendet in jedem Teilintervall eine Quadraturformel fester Ordnung an:

$$I_h^{(n)} := \sum_{i=0}^{N-1} I^{(n)}_{[x_i, x_{i+1}]}(f).$$

Satz 8.7 (Restglied für summierte Quadraturen)**.** Gilt für die verwendete Quadraturformel die Fehlerdarstellung

$$I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f) = \alpha_n h^{m+2} f^{(m+1)}(\xi_i), \quad \xi_i \in [x_i, x_{i+1}],$$

für ein $m \geq n$, so gilt für die entsprechend summierte Formel

$$I(f) - I_h^{(n)}(f) = \alpha_n (b-a) h^{m+1} f^{(m+1)}(\xi). \quad \xi \in [a, b].$$

Die Trapezregel ($n = m = 1$) hat Ordnung 2 und konvergiert mit h^2, die Simpsonregel ($n = 2, m = 3$) hat Ordnung 4 und konvergiert mit h^4, man sieht also, dass der Ordnungsbegriff gerade so gewählt wurde, dass dieser Zusammenhang gilt.

Beweis: Zunächst sei der Zwischenwertsatz aus der Analysis wiederholt. Der lautet: $g(x)$ stetig auf $[\alpha, \beta]$, dann $\exists$ zu jedem $u \in [g(\alpha), g(\beta)]$ mindestens ein $\eta \in [\alpha, \beta]$ so dass $g(\eta) = u$. (Jeder Zwischenwert wird angenommen).

Seien nun N Werte $\xi_i \in [a, b]$, $i = 0, \dots, N-1$ mit $\xi_i \leq \xi_{i+1}$ gegeben. In jedem Intervall $[\xi_i, \xi_{i+1}]$ gilt der Zwischenwertsatz, g nimmt alle Werte zwischen $g_{\min} = \min_{i=0,\dots,N-1} g(\xi_i)$ und $g_{\max} = \max_{i=0,\dots,N-1} g(\xi_i)$ an.

Wegen $\frac{1}{N} \sum_{i=0}^{N-1} g(\xi_i) \in [g_{\min}, g_{\max}]$ gilt

$$\frac{1}{N} \sum_{i=0}^{N-1} g(\xi_i) = g(\xi) \quad \Leftrightarrow \quad \sum_{i=0}^{N-1} g(\xi_i) = N g(\xi) \qquad \xi \in [a, b].$$

Damit erhält man nun:

$$\begin{aligned} I(f) - I_h^{(n)}(f) &= \sum_{i=0}^{N-1} \alpha_n h^{m+2} f^{(m+1)}(\xi_i) = \alpha_n h^{m+2} \sum_{i=0}^{N-1} f^{(m+1)}(\xi_i) \\ &= \alpha_n h^{m+2} N f^{(m+1)}(\xi) = \alpha_n h^{m+2} \frac{b-a}{h} f^{(m+1)}(\xi) \\ &= \alpha_n (b-a) h^{(m+1)} f^{(m+1)}(\xi) \quad \text{mit } \xi \text{ abhängig von } N. \end{aligned}$$

□

Beispiel 8.8 (Einige summierte Quadraturformeln)**.**

(*i*) Summierte Trapezregel

$$\begin{aligned} I_h^{(1)}(f) &= \sum_{i=0}^{N-1} \underbrace{\frac{x_{i+1}-x_i}{2}}_{=h} \{f(x_i)+f(x_{i+1})\} \\ &= h\left\{\frac{f(a)}{2} + \sum_{i=1}^{N-1} f(x_i) + \frac{f(b)}{2}\right\} \\ I(f) - I_h^{(1)}(f) &= -\frac{(b-a)}{12}h^2 f''(\xi), \qquad \xi \in [a,b]. \end{aligned}$$

(*ii*) Summierte Simpson-Regel

$$\begin{aligned} I_h^{(2)}(f) &= \sum_{i=0}^{N-1} \frac{x_{i+1}-x_i}{6}\left\{f(x_i) + 4f\Big(\frac{x_i+x_{i+1}}{2}\Big) + f(x_{i+1})\right\} \\ &= h\left\{\frac{f(a)}{6} + \frac{1}{3}\sum_{i=1}^{N-1} f(x_i) + \frac{2}{3}\sum_{i=1}^{N-1} f\Big(\frac{x_i+x_{i+1}}{2}\Big) + \frac{f(b)}{6}\right\} \\ I(f) - I_h^{(2)}(f) &= -\frac{b-a}{2880}h^4 f^{(4)}(\xi), \qquad \xi \in [a,b] \end{aligned}$$

(*iii*) Summierte Mittelpunktregel

$$\begin{aligned} I_h^{(0)}(f) &= \sum_{i=0}^{N-1}(x_{i+1}-x_i) f\Big(\frac{x_i+x_{i+1}}{2}\Big) = h\sum_{i=0}^{N-1} f\Big(\frac{x_i+x_{i+1}}{2}\Big) \\ I(f) - I_h^{(0)}(f) &= \frac{b-a}{24}h^2 f''(\xi), \qquad \xi \in [a,b] \end{aligned}$$

□

Bemerkung 8.9. Die summierte Simpson-Regel lässt sich aus summierter Trapez- und Mittelpunktregel zusammensetzen:

$$I_h^{(2)}(f) = \frac{1}{3}I_h^{(1)}(f) + \frac{2}{3}I_h^{(0)}(f).$$

Die summierte Trapezregel zu dem nächst feineren Gitter erhält man aus summierter Trapez- und Mittelpunktregel des gröberen Gitters:

$$I_{\frac{h}{2}}^{(1)} = \frac{1}{2}I_h^{(1)}(f) + \frac{1}{2}I_h^{(0)}(f)$$

Diese Formeln erlauben eine ökonomischere Auswertung bei fortgesetztem Halbieren durch Wiederverwendung von Funktionswerten sowie eine Fehlerschätzung.

Beispiel 8.10 (Beispiele zu Quadraturformeln)**.** Wir betrachten folgende bestimmte Integrale:

(*i*) Eine einfache, unendlich oft differenzierbare Funktion:

$$\int_0^{\pi/2} \sin(x)\,dx = 1.$$

(*ii*) Eine glatte Funktion aber mit großen höheren Ableitungen:

$$\int_{-1}^{1} \frac{1}{10^{-5}+x^2}\,dx = 9.914588332462438 \cdot 10^2.$$

(*iii*) Eine nicht unendlich oft differenzierbare Funktion (Halbkreis):

$$\int_{-1}^{1} \sqrt{1-x^2}\,dx = \pi/2.$$

Summierte Trapezregel für (*i*).

I	Fehler	#Fktausw.
9.480594489685199e-01	5.1941e-02	3
9.871158009727754e-01	1.2884e-02	5
9.967851718861696e-01	3.2148e-03	9
9.991966804850723e-01	8.0332e-04	17
9.997991943200188e-01	2.0081e-04	33
9.999498000921015e-01	5.0200e-05	65
9.999874501175253e-01	1.2550e-05	129
9.999968625352869e-01	3.1375e-06	257
9.999992156341920e-01	7.8437e-07	513
...		
9.999999999995609e-01	4.3909e-13	524289
1.000000000003847e+00	3.8467e-12	1048577

Fehler viertelt sich jeweils, und das von Anfang an. Weniger als 10^{-13} wird mit `double` Genauigkeit nicht erreicht.

Summierte Simpsonregel für (*i*).

I	Fehler	#Fktausw.
1.000134584974194e+00	1.3458e-04	5
1.000008295523968e+00	8.2955e-06	9
1.000000516684706e+00	5.1668e-07	17
1.000000032265001e+00	3.2265e-08	33
1.000000002016129e+00	2.0161e-09	65
1.000000000126001e+00	1.2600e-10	129
1.000000000007874e+00	7.8739e-12	257
1.000000000000491e+00	4.9094e-13	513
1.000000000000030e+00	2.9976e-14	1025
1.000000000000006e+00	5.7732e-15	2049
1.000000000000002e+00	1.7764e-15	4097

Fehler reduziert sich jeweils um den Faktor $16 = (1/2)^4$, und das fast von Anfang an.

Summierte Trapezregel für (ii).

I	Fehler	#Fktausw.
1.000009999900001e+05	9.9010e+04	3
5.000449983500645e+04	4.9013e+04	5
2.501113751079469e+04	2.4020e+04	9
1.252430268327760e+04	1.1533e+04	17
6.300548144658167e+03	5.3091e+03	33
3.227572909110977e+03	2.2361e+03	65
1.765586982280199e+03	7.7413e+02	129
1.160976493727309e+03	1.6952e+02	257
1.003813438906513e+03	1.2355e+01	513
9.915347090712996e+02	7.5876e-02	1025
9.914588358257512e+02	2.5795e-06	2049
9.914588331667655e+02	7.9478e-08	4097
9.914588332263689e+02	1.9875e-08	8193
9.914588332412698e+02	4.9740e-09	16385

Fehlerverhalten am Anfang unklar, erst spät stellt sich h^2 ein.

Summierte Simpsonregel für (ii).

I	Fehler	#Fktausw.
3.333899978334190e+04	3.2348e+04	5
1.668001673605744e+04	1.5689e+04	9
8.362024407438566e+03	7.3706e+03	17
4.225963298451690e+03	3.2345e+03	33
2.203247830595247e+03	1.2118e+03	65
1.278258340003273e+03	2.8680e+02	129
9.594396642096787e+02	3.2019e+01	257
9.514257539662473e+02	4.0033e+01	513
9.874417991262286e+02	4.0170e+00	1025
9.914335447439017e+02	2.5289e-02	2049
9.914588322804369e+02	9.6581e-07	4097
9.914588332462367e+02	7.0486e-12	8193
9.914588332462367e+02	7.0486e-12	16385

Bis 4096 Auswertungen ist Simpson schlechter als Trapez. „Asymptotische Konvergenzrate“ stellt sich erst für genügend kleines h ein.

Summierte Trapezregel für (iii).

I	Fehler	#Fktausw.
1.000000000000000e+00	5.7080e-01	3
1.366025403784439e+00	2.0477e-01	5
1.497854534051220e+00	7.2942e-02	9
1.544909572178587e+00	2.5887e-02	17
1.561626518913870e+00	9.1698e-03	33
1.567551211438566e+00	3.2451e-03	65
1.569648456389842e+00	1.1479e-03	129
1.570390396198308e+00	4.0593e-04	257
1.570652791478614e+00	1.4354e-04	513
1.570745576359828e+00	5.0750e-05	1025
1.570778383269506e+00	1.7944e-05	2049
1.570789982705718e+00	6.3441e-06	4097
1.570794083803873e+00	2.2430e-06	8193
1.570795533774854e+00	7.9302e-07	16385

Die Konvergenzordnung h^2 wird nicht erreicht, sondern nur ein h^α mit $\alpha < 2$ (siehe unten).

Summierte Simpsonregel für (*iii*).

I	Fehler	#Fktausw.
1.488033871712585e+00	8.2762e-02	5
1.541797577473481e+00	2.8999e-02	9
1.560594584887709e+00	1.0202e-02	17
1.567198834492299e+00	3.5975e-03	33
1.569526108946797e+00	1.2702e-03	65
1.570347538040268e+00	4.4879e-04	129
1.570637709467796e+00	1.5862e-04	257
1.570740256572051e+00	5.6070e-05	513
1.570776504653564e+00	1.9822e-05	1025
1.570789318906069e+00	7.0079e-06	2049
1.570793849184461e+00	2.4776e-06	4097
1.570795450836595e+00	8.7596e-07	8193
1.570796017098507e+00	3.0970e-07	16385

Die Simpsonregel zeigt *dieselbe* Konvergenzordnung wie die Trapezregel!

Satz 8.7 liefert eine Konvergenzabschätzung der Form

$$|I(f) - I_h^{(n)}(f)| \leq Ch^{m+1}.$$

Für die summierte Trapezregel gilt $m = 1$, man spricht von h^2 Konvergenz, für die summierte Simpsonregel gilt $m = 3$, man hat h^4 Konvergenz.

Wir haben gesehen, dass die Konvergenzordnung kleiner als $m+1$ sein kann, wenn die Funktion nicht genügend oft differenzierbar ist. Experimentell können wir diese folgendermaßen bestimmen.

Mit dem Ansatz $e_h = |I(f) - I_h^{(n)}(f)| = Ch^\alpha$ gilt

$$\frac{e_{h/2}}{e_h} = \frac{C(h/2)^\alpha}{Ch^\alpha} = (1/2)^\alpha$$

und daraus erhalten wir

$$\alpha = \log\left(\frac{e_{h/2}}{e_h}\right) \Big/ \log\left(\frac{1}{2}\right).$$

Das so bestimmte α heißt *experimental order of convergence* (EOC).

Im letzen Beispiel oben erhalten wir $\alpha = 3/2$.

8.4 Fehlerkontrolle

Wir haben in Satz 8.7 gezeigt, wie der Fehler mit mehr Stützstellen abnimmt. Dies nennt man eine *a-priori Fehlerschranke*.

So erhalten wir etwa für die summierte Trapezregel:

$$|I(f) - I_h^{(1)}(f)| = |-\frac{b-a}{12} f''(\xi) h^2| \leq \underbrace{\frac{b-a}{12} \max_{\xi\in[a,b]} |f''(\xi)|}_{=:C} h^2.$$

C ist allerdings im allgemeinen schwer zu bestimmen.

In der Praxis würde man aber gerne wissen, bei wievielen Stützstellen (bei welchem h) der Fehler kleiner als eine vorgegebene *Toleranz* ist.

Dazu wollen wir eine Methode zur *a-posteriori Fehlerschätzung* vorstellen.

Idee: Die Simpson-Summe konvergiert schneller als die Trapezsumme (f genügend glatt), hat also für genügend kleines h einen kleineren Fehler.

Wir wollen den Fehler in der Trapezsumme zum Gitter $h/2$ abschätzen. Dazu „schieben“ wir die Auswertung der Simpsonsumme dazwischen:

$$|I(f) - I_{\frac{h}{2}}^{(1)}(f)| = |I(f) - I_h^{(2)}(f) + I_h^{(2)}(f) - I_{\frac{h}{2}}^{(1)}(f)|.$$

Nun nutze die Dreiecksungleichung:

$$|I(f) - I_{\frac{h}{2}}^{(1)}(f)| \leq |I(f) - I_h^{(2)}(f)| + |I_h^{(2)}(f) - I_{\frac{h}{2}}^{(1)}(f)|.$$

Nun *nimmt man an*, dass die Simpsonsumme genauer ist als die Trapezsumme: $|I(f) - I_h^{(2)}(f)| \leq \omega |I(f) - I_{\frac{h}{2}}^{(1)}(f)|$ mit $0 < \omega < 1$:

$$|I(f) - I_{\frac{h}{2}}^{(1)}(f)| \leq \omega |I(f) - I_{\frac{h}{2}}^{(1)}(f)| + |I_h^{(2)}(f) - I_{\frac{h}{2}}^{(1)}(f)|.$$

Auflösen nach dem Fehler in der Trapezsumme liefert:

$$|I(f) - I_{\frac{h}{2}}^{(1)}(f)| \leq \frac{1}{1-\omega} |I_h^{(2)}(f) - I_{\frac{h}{2}}^{(1)}(f)|.$$

Besonders ökonomisch lässt sich die Fehlerkontrolle zusammen mit Bemerkung 8.9 umsetzen (deshalb haben wir oben die Trapezregel zu $h/2$ und die Simpsonsumme zu h verwendet):

$h = b - a$; $N = 1$; $I1 = h(f(a) + f(b))/2$;
while $(h > \varepsilon)$ **do**
  $I0 = 0$;
  for $(i = 0, i < N, i = i + 1)$ **do**
    $I0 = I0 + hf(a + (i + 0.5)h)$; {Mittelpunktsumme}
  end for
  $I2 = \frac{1}{3}I1 + \frac{2}{3}I0$; {Simpson-Summe zu h}
  $I1 = \frac{1}{2}I1 + \frac{1}{2}I0$; {Trapez-Summe zu $h/2$}
  $h = \frac{1}{2}h$; $N = 2N$;
  if $(\frac{1}{1-\omega}(I2 - I1) \leq TOL)$ **then**
    return $I1$; {Liefere Trapez-Summe zu h}
  end if
end while

Die Fehlerkontrolle lässt sich so ohne zusätzlichen Aufwand erledigen.

8.5 Zusammenfassung

- In diesem Abschnitt haben wir die Newton-Cotes Formeln zur numerischen Quadratur kennengelernt. Diese basieren auf der Polynominterpolation zu äquidistanten Stützwerten und exakter Integration des Interpolationspolynoms.
- Man unterscheidet abgeschlossene und offene Formeln.
- Die Schwierigkeiten der Polynominterpolation bei hohem Grad übertragen sich auf die Quadratur. Das äußert sich z. B. in negativen Gewichten.
- Deshalb verwendet man nicht zu hohen Grad und zusätzlich eine Aufspaltung des Integrals in Teilintervalle.
- Schließlich haben wir ein einfaches Verfahren zur Fehlerkontrolle kennengelernt.

9 Quadraturen höherer Ordnung

Newton-Cotes Formeln zur numerischen Quadratur eignen sich nicht zur Integration mit hoher Ordnung, da spätestens ab $n = 7$ negative Gewichte auftreten.

Eine weitere Frage ist, ob die Konvergenzordnung der Newton-Cotes Formeln schon optimal ist für die gegebene Anzahl von Stützstellen (Funktionsauswertungen). Antwort: Nein!

Gibt es Verfahren, die höhere Ordnung ohne die Nachteile der Newton-Cotes Formeln erreichen?

Wir werden in diesem Abschnitt zwei Ansätze vorstellen, um hohe Ordnung zu erreichen.

9.1 Romberg-Integration

Satz 9.1 (Euler-MacLaurinsche[21] Summenformel)**.** Sei $f \in C^{2m+2}[a,b]$ und einmal integrierbar, dann gilt

$$\begin{aligned} \underbrace{I_h^{(1)}(f)}_{Trapezsumme!} &= \int_a^b f(t)dt + \sum_{k=1}^{m} h^{2k} \frac{B_{2k}}{(2k)!} \Big(f^{(2k-1)}(b) - f^{(2k-1)}(a) \Big) \\ &\quad + \underbrace{h^{2m+2} \frac{B_{2m+2}}{(2m+2)!}(b-a) f^{(2m+2)}(\zeta)}_{Restglied} \qquad \zeta \in [a,b] \end{aligned} \tag{9.1}$$

mit $B_0 = 1$, $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, ..., den Bernoulli-Zahlen[22]. Diese sind die konstanten Glieder $B_{2k} = B_{2k}(0)$ der Bernoulli-Polynome

$$\begin{aligned} B_1(x) &= x - \frac{1}{2}, & B_k'(x) &= kB_{k-1}(x), & k &> 0, \\ B_{2k+1}(0) &= B_{2k+1}(1) = 0, & & & k &> 0. \end{aligned}$$

Beweis: hier nur ein kleiner Hinweis zum Beweis, Rest siehe [Sto05, Kap. 3.3].

Setze $B_1(x) = x - \frac{1}{2} \Rightarrow B_1'(x) \equiv 1$, also haben wir

$$\int_0^1 \underbrace{B_1'(t)}_{\equiv 1} g(t)dt \underset{\substack{\text{partielle} \\ \text{Integration}}}{=} \underbrace{\Big[B_1(t)g(t) \Big]_0^1}_{g(1)/2 + g(0)/2} - \underbrace{\int_0^1 B_1(t)g'(t)dt}_{\text{rekursive Anwendung}}$$

Setze $B_2'(x) = 2B_1(x) \Leftrightarrow B_1(x) = \frac{1}{2}B_2'(x)$

$$\int_0^1 B_1(t)g'(t)dt = \frac{1}{2}\int_0^1 B_2'(t)g'(t)dt = \frac{1}{2}\Big[B_2(t)g'(t) \Big]_0^1 - \frac{1}{2}\int_0^1 B_2(t)g''(t)dt$$

[21] Colin Maclaurin, 1698-1746, schottischer Mathematiker

[22] Jacob Bernoulli, 1655-1705, schweitzer Mathematiker.

Setze $B_3'(x) = 3B_2(x)$ und $B_3(0) = B_3(1) = 0$ (dies legt auch die Konstante in $B_2(t)$ fest)

$$\int_0^1 B_2(t)g''(t)dt = \frac{1}{3}\int_0^1 B_3'(t)g''(t)dt = \underbrace{\frac{1}{3}\Big[B_3(t)g''(t)\Big]_0^1}_{0 \text{ wg. Normierungsbed.}} - \frac{1}{3}\int_0^1 B_3(t)g'''(t)dt$$

und so fort.

Einsetzen der Gleichungen ineinander liefert:

$$\int_0^1 g(t)dt = \underbrace{\frac{g(0)}{2} + \frac{g(1)}{2}}_{\to I_1^{(1)}} - \underbrace{\frac{1}{2}\Big(B_2(1)g'(1) - B_2(0)g'(0)\Big)}_{\sum_{k=1}^{1}\ldots, \text{ wobei } h=1} - \underbrace{\frac{1}{3}\int_0^1 B_3(t)g'''(t)dt}_{\text{Restglied}}$$

Die ersten drei Terme haben also schon die Gestalt aus (9.1) für den Fall $a = 0, b = 1, h = 1$.

Es werden nur die geraden Bernoullipolynome B_{2k} benötigt, da die ungeraden aufgrund der Normierungsbedingung immer wegfallen.

Der Rest ergibt sich mittels

- weiterer partieller Integration,
- Anwendung auf Teilintervalle $[x_k, x_{k+1}], x_k = kh$,
- Transformation auf allgemeines Intervall $[a, b]$
- und sorgfältige Anwendung des Mittelwertsatzes auf das Restglied.

□

Satz 9.1 bedeutet, dass für die Trapezsumme folgende Darstellung gilt:

$$I_h^{(1)}(f) = \underbrace{\tau_0}_{\int_a^b f(x)dx} + \underbrace{\tau_1 h^2 + \tau_2 h^4 \ldots + \tau_m h^{2m}}_{\substack{\text{unabhängig von } h! \\ \text{nur abh. von } f,a,b}} + \alpha_{m+1}(h)h^{2m+2} \tag{9.2}$$

wobei die Konstanten τ_i *nicht* von der Schrittweite h abhängen.

Diese sog. *asymptotische Entwicklung* stellt ein Polynom in $y = h^2$ dar.

Bei der „Extrapolation“ kombiniert man $I_h^{(1)}$ für verschiedene h so, dass möglichst viele Terme außer τ_0 Null werden.

Dazu das

Beispiel 9.2. Wir setzen eine Linearkombination zweier Schrittweiten h_1, h_2 an:

$$a_1 I_{h_1}^{(1)}(f) + a_2 I_{h_2}^{(1)}(f) = (a_1 + a_2)\tau_0 + (a_1 h_1^2 + a_2 h_2^2)\tau_1 + O(h^4).$$

Nun wähle a_1, a_2 so, dass

$$a_1 + a_2 = 1 \qquad \text{und} \qquad a_1 h_1^2 + a_2 h_2^2 = 0.$$

Dabei sind h_1^2 und h_2^2 bekannte Zahlen.

Dieses lineare Gleichungssystem hat die Lösung

$$a_1 = \frac{1}{1-(h_1/h_2)^2}, \qquad a_2 = \frac{1}{1-(h_2/h_1)^2}.$$

Für $h_1 = h$ und $h_2 = h/2$ gilt speziell

$$\frac{4}{3} I^{(1)}_{h/2}(f) - \frac{1}{3} I^{(1)}_h(f) = I(f) + O(h^4).$$

□

Anstatt die Bedingungen an die Koeffizienten in Form eines linearen Gleichungssystemes aufzustellen, kann man auch so vorgehen.

Man betrachte die Interpolationsaufgabe

$$p(h_i^2) = I^{(1)}_{h_i} \qquad i = 0, \ldots, n,$$

d. h. dem Quadrat der Schrittweite wird der Wert der entsprechenden Trapezsumme zugeordnet.

Dann ist das Auswerten dieses Interpolationspolynoms an der Stelle 0 äquivalent zur Elimination der Terme $\tau_1 h^2, \ldots, \tau_n h^2$ in der asymptotischen Entwicklung.

Da man außerhalb des Bereiches der Stützstellen auswertet, spricht man von Extrapolation.

Speziell in ihrer Anwendung auf die Berechnung von Integralen mittels Trapezsummen unterschiedlicher Schrittweiten heißt dieses Verfahren Romberg-Integration[23].

Bei Anwendung auf allgemeine Diskretisierungsverfahren spricht man auch von Richardson-Extrapolation[24].

Zur praktischen Durchführung eignet sich besonders das Neville-Schema aus Satz 5.4, welches angepasst lautet:

$$\begin{aligned} i = 0, \ldots, n: &\qquad p_{i,i}(h_i^2) = I^{(1)}_{h_i}(f) \quad (p_{i,i} \text{ sind Konstanten}) \\ k = 0, \ldots, n-i: &\qquad p_{i,i+k}(0) = p_{i,i+k-1}(0) - h_i^2 \frac{p_{i+1,i+k}(0) - p_{i,i+k-1}(0)}{h_{i+k}^2 - h_i^2} \end{aligned}$$

$p_{i,i+k}$ interpoliert Stützstellen h_i^2 bis h_{i+k}^2

Bei Interpolation von $n+1$ Werten

$$(h_i^2, I^{(1)}_{h_i}(f)), \qquad i = 0, \ldots, n$$

gilt dann

$$p(0) = \tau_0 + O(h^{2n+2})$$

Aber: Anwendung dieser Methode erfordert dann eben auch $f \in C^{2n+2}[a,b]$.

[23] Werner Romberg, 1909-2003, deutscher Mathematiker.
[24] Lewis Fry Richardson, 1881-1953, brit. Mathematiker.

9.2 Gauss-Integration

Negative Gewichte bei Newton-Cotes treten wegen äquidistanter Stützstellen auf (Lagrange-Polynome oszillieren am Rand stark).

Frage: Kann man die Situation durch Wahl *nicht* äquidistanter Stützstellen verbessern?

Idee: Bestimme Gewichte w_i *und* Stützstellen x_i so, dass Polynome möglichst hohen Grades exakt integriert werden.

Beispiel 9.3. Finde x_1, x_2, w_1, w_2 so, dass $p_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ in $I = [-1, 1]$ exakt integriert wird:

$$\int_{-1}^{1} p_3(x)dx = \sum_{i=1}^{2} w_i p_3(x_i)$$

$$\Leftrightarrow \quad a_0 \underbrace{\int_{-1}^{1} 1dx}_{2} + a_1 \underbrace{\int_{-1}^{1} xdx}_{0} + a_2 \underbrace{\int_{-1}^{1} x^2 dx}_{2/3} + a_3 \underbrace{\int_{-1}^{1} x^3 dx}_{0} =$$

$$w_1(a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3) + w_2(a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^3)$$

Koeffizientenvergleich (für die a_i !) ergibt vier nichtlineare Gleichungen:

$$\left.\begin{array}{rcl} 2a_0 & = & a_0(w_1 + w_2) \\ 0a_1 & = & a_1(w_1 x_1 + w_2 x_2) \\ \frac{2}{3}a_2 & = & a_2(w_1 x_1^2 + w_2 x_2^2) \\ 0a_3 & = & a_3(w_1 x_1^3 + w_2 x_2^3) \end{array}\right\} \Rightarrow w_1 = w_2 = 1 \qquad x_1 = \frac{-1}{\sqrt{3}}, x_2 = \frac{1}{\sqrt{3}}.$$

Damit hat man also mit zwei Stützstellen eine Quadraturformel mit Ordnung 4 erhalten.

Zum Vergleich: Die Trapezregel erreicht mit zwei Stützstellen nur Ordnung 2. Newton-Cotes Formeln haben allgemein bei $n + 1$ Stützstellen nur die Ordnung $n + 1$ (das ist das Minimum). □

Im folgenden beschränken wir uns auf Quadraturformeln für das Einheitsintervall $[-1, 1]$. Integrale über $[a, b]$ berechnet man per Transformation.

Man kann folgende Aussage zu nichtäquidistanten interpolatorischen Quadraturformeln zeigen.

Satz 9.4. Die maximale Ordnung einer Quadraturformeln mit $n + 1$ Stützstellen ist $2n + 2$ (d. h. Polynome vom Grad $2n + 1$ werden exakt integriert).

Beweis: Angenommen die Ordnung wäre $2n + 3$, d. h. ein Polynom mit Grad $2n + 2$, also insbesondere

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

würde exakt integriert werden. Dann gilt

- $q(x)$ hat Grad $2(n + 1) = 2n + 2$.
- $q(x) \geq 0$, $\forall x$, und $q(x) \not\equiv 0 \Rightarrow \int_{-1}^{1} q(x)dx > 0$.

- andererseits gilt $q(x_i) = 0$ an den Stützstellen und damit $\sum w_i q(x_i) = 0$ also Widerspruch!

□

Die Bestimmung der Gewichte und Stützstellen im allgemeinen Fall zeigt der

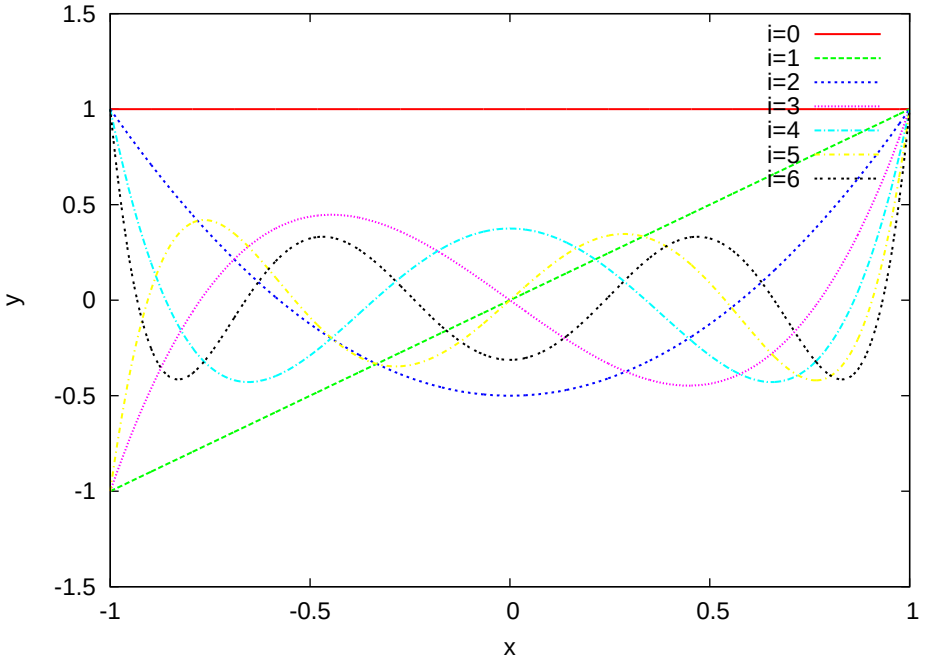
Satz 9.5 (Gauß-Quadratur)**.** Es gibt genau eine interpolatorische Quadraturformel zu $n+1$ paarweise verschiedenen Stützstellen in $[-1,1]$ mit der Ordnung $2n+2$ (d. h. der maximal möglichen Ordnung). Ihre Stützstellen sind die Nullstellen $\lambda_0, \dots \lambda_n \in (-1,1)$ des (n+1)-ten Legendrepolynoms[25] $L_{n+1} \in P_{n+1}$. Die Gewichte erhält man mittels

$$w_i = \int_{-1}^{1} \prod_{\substack{i=0 \\ j \neq i}}^{n} \Big(\frac{x - \lambda_j}{\lambda_i - \lambda_j} \Big)^2 dx > 0 \qquad i = 0, \dots, n$$

Beweis: siehe [Ran06, Satz 3.4]. □

Die Legendrepolynome lauten

$$L_0(x) = 1, \qquad L_1(x) = x, \qquad L_{n+1}(x) = \frac{2n+1}{n+1} \cdot x \cdot L_n(x) - \frac{n}{n+1} L_{n-1}(x)$$
$$(n+1) L_{n+1}(x) = (2n+1) \cdot x \cdot L_n(x) - n L_{n-1}(x)$$



Die Legendre-Polynome bilden ein Orthogonalsystem in folgendem Sinne:

$$\int_{-1}^{1} L_n(x) L_m(x)\, dx = 0 \qquad n \neq m.$$

[25] Adrien-Marie Legendre, 1752-1833, frz. Mathematiker.

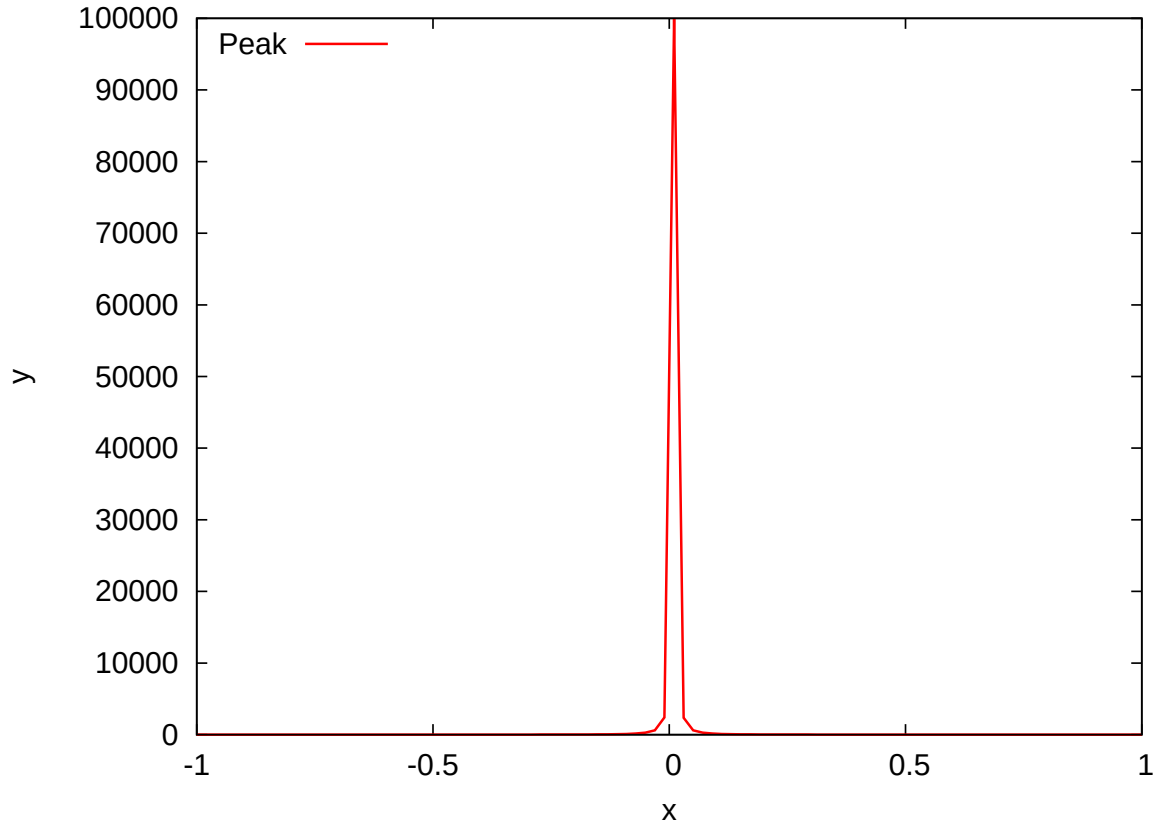
Beispiel 9.6. Für $n = 1$, $n = 2$ ergibt sich: $h = \frac{b-a}{2}$, $c = \frac{b+a}{2}$ und

$$\begin{aligned} I^{(1)}(f) &= \tfrac{b-a}{2}\left\{f(c-\sqrt{1/3}h) + f(c+\sqrt{1/3}h)\right\} && \text{Ordnung 4} \\ I^{(2)}(f) &= \tfrac{b-a}{18}\left\{5f(c-\sqrt{3/5}h) + 8f(c) + 5f(c+\sqrt{3/5}h)\right\} && \text{Ordnung 6} \end{aligned}$$

Hier wurde schon auf das allgemeine Intervall $[a, b]$ transformiert. □

9.3 Adaptive Quadratur

Quadratur mit konstanter Schrittweite ist bei manchen Integranden ineffizient. Betrachte z. B. $f(x) = \frac{1}{10^{-5}+x^2}$.



In so einem Fall möchte man die Schrittweite „adaptiv“, d. h. angepasst an den speziellen Integranden wählen.

Dazu bedient man sich eines lokalen „Fehlerschätzers“ (oder Indikators), der angibt, an welcher Stelle die Schrittweite weiter verkleinert werden muss.

Es bietet sich an, dies noch mit einer Fehlerkontrolle zu kombinieren. Grob ergibt sich das folgende Vorgehen:

(1) Wähle eine Unterteilung $G_0 = \{x_i^{(0)} \mid i = 0, \ldots, N_0\}$. Berechne das Integral I_0 bezüglich der Unterteilung G_0. Setze $k = 0$.

(2) Berechne eine Schätzung für den Fehler E_k in I_k. Falls $E_k < TOL$ sind wir fertig.

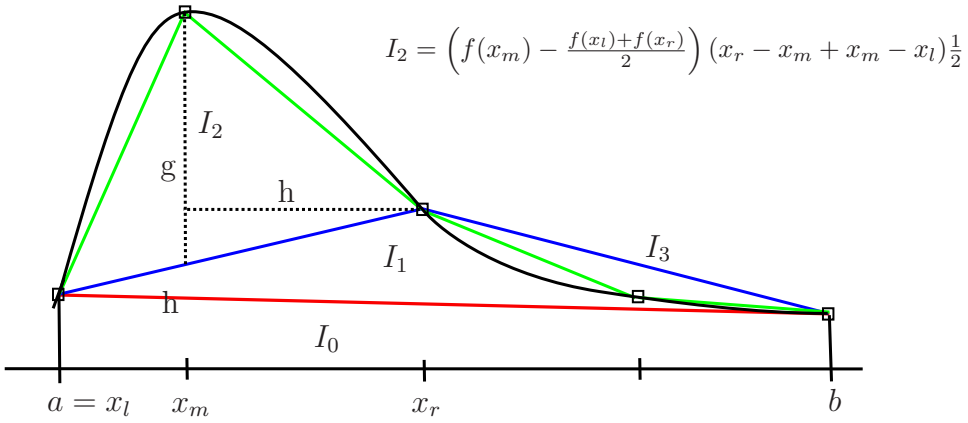
(3) Verfeinere die Unterteilung G_k zu G_{k+1} durch Hinzufügen von Punkten angepasst an den Integranden und setze $k = k + 1$.

(4) Berechne I_k zu G_k und gehe nach (2).

Im folgenden beschränken wir uns auf einen einfachen Algorithmus ohne globale Fehlerkontrolle.

Peter Bastian

Wir betrachten das Prinzip von Archimedes[26]:



Wir können das Integral (die Fläche) „hierarchisch“ zerlegen in die Anteile

$$I = I_0 + I_1 + \ldots$$

wobei

I_0 das Trapez $(a,0),(b,0),(b,f(b)),(a,f(a))$.

I_1 das Dreieck $(a,f(a)),(b,f(b)),((a+b)/2,f((a+b)/2))$.

I_2 das Dreieck $(a,f(a)),((a+b)/2,f((a+b)/2)),(a+1/4(b-a),f(a+1/4(b-a)))$.

usw.

Die hierarchischen Zuwächse berechnen sich wie folgt. Seien die Punkte (x_l, f_l) und (x_r, f_r) gegeben, so berechnet sich die Fläche des Dreiecks mit $(x_m, f(x_m))$, $x_m = (x_l + x_r)/2$ mittels

$$I_\Delta = \left(f(x_m) - \frac{f(x_l) + f(x_r)}{2}\right)\frac{x_r - x_l}{2}$$

[26] Archimedes von Syrakus, 287 v. Chr.-212 v. Chr., griechischer Mathematiker, Physiker und Ingenieur.

(Fläche eines Dreiecks : $A = g \cdot h/2$).

Der hierarchische Zuwachs I_Δ dient gleichzeitig als lokaler Fehlerindikator. Ist er klein genug, so ist die Funktion dort gut angenähert und das Teilintervall muss *nicht* weiter verfeinert werden.

Es bietet sich die Formulierung als rekursive Funktion an:

archi $(x_l,\, x_r,\, f_l,\, f_r,\, l)$:
$x_m = (x_e + x_r)/2,\; f_m = f(x_m)$;
$s = (f_m - (f_l + f_r)/2) \cdot (x_r - x_l)/2$;
if $(s \geq \mathrm{TOL} \vee l < l_{\min})$ **then**
  return s+archi$(x_l, x_m, f_l, f_m, l+1)$+archi$(x_m, x_r, f_m, f_r, l+1)$;
else
  return s;
end if

Das Integral berechnet sich dann via

$$I = (b-a)(f(a) + f(b))/2 + \mathrm{archi}(a, b, f(a), f(b), 0).$$

Beispiel 9.7 (Beispiel zur numerischen Quadratur)**.** Wir betrachten dieselben Funktionen wie in Beispiel 8.10:

(i) Eine einfache, unendlich oft differenzierbare Funktion:

$$\int_0^{\pi/2} \sin(x)\, dx = 1.$$

(ii) Eine glatte Funktion aber mit großen höheren Ableitungen:

$$\int_{-1}^{1} \frac{1}{10^{-5} + x^2}\, dx = 9.914588332462438 \cdot 10^2.$$

(iii) Eine nicht unendlich oft differenzierbare Funktion (Halbkreis):

$$\int_{-1}^{1} \sqrt{1 - x^2}\, dx = \pi/2.$$

Verschiedene Quadraturen für (i) aus Beispiel 9.7.

Methode	I	Fehler	#Fktausw.
Gauss4	9.999101667698898e-01	8.9833e-05	4
	9.999944679581383e-01	5.5320e-06	8
	9.999996555171785e-01	3.4448e-07	16
	9.999999784895880e-01	2.1510e-08	32
Gauss6	1.000000118910998e+00	1.1891e-07	6
	1.000000001828737e+00	1.8287e-09	12
	1.000000000028461e+00	2.8461e-11	24
	1.000000000000444e+00	4.4409e-13	48
Archi	9.480594489685199e-01	5.1941e-02	3
	9.871158009727754e-01	1.2884e-02	5
	9.967851718861697e-01	3.2148e-03	9
	9.990131153231707e-01	9.8688e-04	15
	9.997876171856270e-01	2.1238e-04	31

Das Verfahren hoher Ordnung zahlt sich aus.

Verschiedene Quadraturen (*ii*) aus Beispiel 9.7.

Methode	I	Fehler	#Fktausw.
Trapez	1.000009999900001e+05	9.9010e+04	3
	3.227572909110977e+03	2.2361e+03	65
	1.765586982280199e+03	7.7413e+02	129
	1.160976493727309e+03	1.6952e+02	257
	1.003813438906513e+03	1.2355e+01	513
	9.915347090712996e+02	7.5876e-02	1025
	9.914588358257512e+02	2.5795e-06	2049
Archi	1.767335925226728e+03	7.7588e+02	25
	1.004348965298925e+03	1.2890e+01	37
	9.946212584262852e+02	3.1624e+00	81
	9.922788393957054e+02	8.2001e-01	173
	9.916266302474447e+02	1.6780e-01	361
	9.914967844457766e+02	3.7951e-02	769
	9.914672523966888e+02	8.4192e-03	1625
	9.914606991793892e+02	1.8659e-03	3465
	9.914592092819358e+02	3.7604e-04	7629

Fehlerreduktion mit Archi ist von Anfang an quadratisch, allerdings „überholt“ die Trapezsumme dann kräftig.

Verschiedene Quadraturen für (*iii*) aus Beispiel 9.7.

Methode	I	Fehler	#Fktausw.
Gauss4	1.592226038754547e+00	2.1430e-02	4
	1.570801362699711e+00	5.0359e-06	1024
	1.570798107100650e+00	1.7803e-06	2048
	1.570796956200537e+00	6.2941e-07	4096
	1.570796549318533e+00	2.2252e-07	8192
Gauss6	1.578036347519909e+00	7.2400e-03	6
	1.570801237513435e+00	4.9107e-06	768
	1.570798062869299e+00	1.7361e-06	1536
	1.570796940567470e+00	6.1377e-07	3072
	1.570796543792309e+00	2.1700e-07	6144
Archi	1.366025403784439e+00	2.0477e-01	5
	1.570774639679624e+00	2.1687e-05	365
	1.570791453003758e+00	4.8738e-06	765
	1.570795219591928e+00	1.1072e-06	1605
	1.570796082320714e+00	2.4447e-07	3433

Hohe Ordnung lohnt sich nicht wegen mangelnder Differenzierbarkeit.

9.4 Mehrdimensionale Quadratur

In der Praxis sind oft Integrale in mehr als einer Raumdimension zu berechnen. Wie macht man das?

Betrachten wir zunächst das Quadrat $[-1,1]\times[-1,1]$. Am simpelsten ist die Produktintegration zu realisieren:

$$\begin{aligned}\int_{-1}^{1}\int_{-1}^{1} f(x,y)\,dxdy &\approx \sum_{i=1}^{n} w_i \int_{-1}^{1} f(x,y_i)\,dx \\ &\approx \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{n} w_j f(x_j,y_i)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j f(x_j,y_i)\end{aligned}$$

Dies lässt sich einfach auf d Raumdimensionen verallgemeinern.

Zur Integration über komplex berandete Gebiete nutzt man den Transformationssatz für Integrale:

$$\int_\Omega f(x,y)dxdy = \int_{-1}^{1}\int_{-1}^{1} f\Big(\varphi(\xi,\eta),\psi(\xi,\eta)\Big)\left|\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right| d\xi d\eta$$

wobei die Transformation

$$\begin{pmatrix}\varphi(\xi,\eta)\\ \psi(\xi,\eta)\end{pmatrix} : [-1,1]\times[-1,1]\to\Omega$$

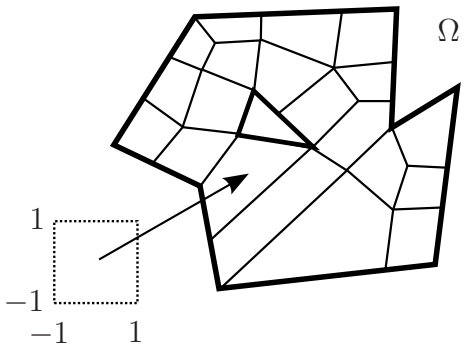
das Gebiet $[-1,1] \times [-1,1]$ auf Ω abbildet.

Weiter ist

$$\left|\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right| = det \begin{pmatrix} \frac{\partial\varphi}{\partial\xi}(\xi,\eta) & \frac{\partial\psi}{\partial\xi}(\xi,\eta) \\ \frac{\partial\varphi}{\partial\eta}(\xi,\eta) & \frac{\partial\psi}{\partial\eta}(\xi,\eta) \end{pmatrix} \neq 0$$

die Determinante der (transponierten) Jacobimatrix[27] der Transformation.

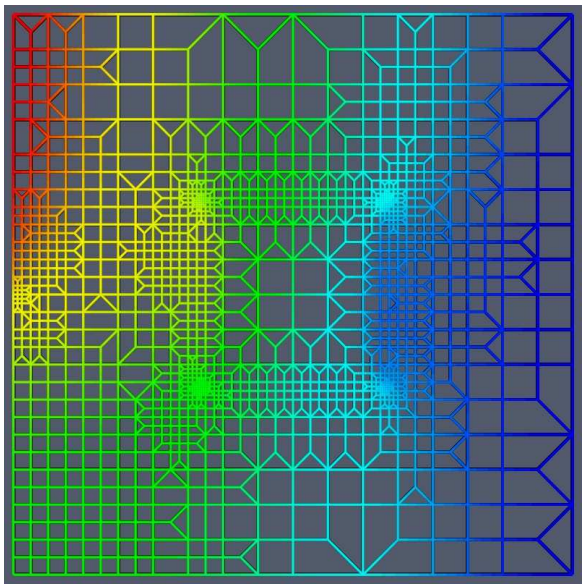
Bei komplizierteren Gebieten wendet man das stückweise an:

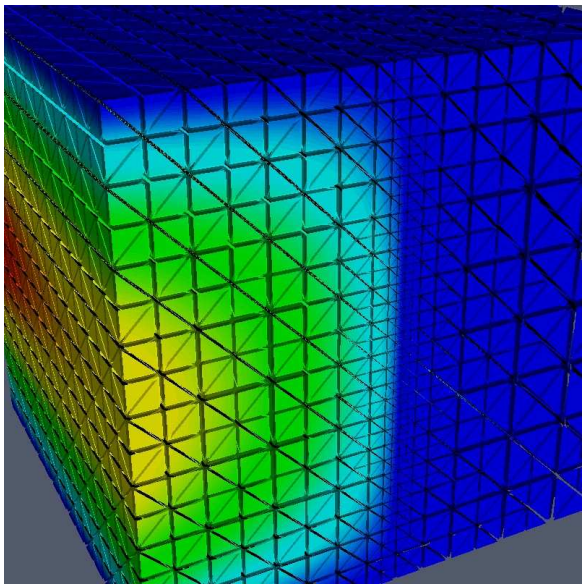


Auch direkte Integrationsformeln für Dreiecke (Simplizes) sind möglich.

Die Zerlegung eines Gebietes in Teilgebiete einfacher geometrischer Gestalt (Dreicke, Vierecke, Tetraeder, Hexaeder, ...) nennt man Triangulierung oder Gittergenerierung. Dies ist insbesondere in drei Raumdimensionen eine schwierige Aufgabe.

Auch in mehr als einer Raumdimensionen kann man hierarchisch adaptiv verfeinern:





[27] Carl Gustav Jacob Jacobi, 1804-1851, dt. Mathematiker

Links: Adaptives Dreiecks- und Viereckgitter. Rechts: Adaptives Tetraedergitter mit Bisektionsverfeinerung.

9.5 Zusammenfassung

- In diesem Abschnitt haben wir zwei Verfahren zur numerischen Quadratur mit hoher Ordnung vorgestellt: Die auf Extrapolation beruhende Romberg-Integration und die auf nichtäquidistanter interpolatorischer Quadratur beruhende Gauss-Integration.
- Beide Verfahren sind bei genügender Differenzierbarkeit des Integranden in der Lage beliebig hohe Ordnung zu erreichen.
- Mit dem Prinzip von Archimedes haben wir das wichtige Gebiet der adaptiven Verfahren illustriert.
- Schließlich haben wir noch kurz vorgestellt, wie man Integrale über mehrdimensionale Integrationsbereiche berechnet.

10 Lineare Gleichungssysteme und Gauß-Elimination

10.1 Motivation

Einige Anwendungen linearer Gleichungssysteme haben wir schon kennengelernt:

- Bestimmen der Koeffizienten in einer Basisdarstellung.
- Interpolation mit kubischen Splines.
- Lineare Integralgleichungen, wie sie etwa bei der Radiosity-Methode auftreten.

Weitere sind:

- Implizite Lösungsverfahren für lineare gewöhnliche Differentialgleichungssysteme.
- Manche Lösungsverfahren für lineare partielle Differentialgleichungen führen auf teils extrem große lineare Gleichungssysteme. Dies liegt daran, dass der *Diskretisierungsfehler* direkt mit der Größe des linearen Gleichungssystemes gekoppelt ist.
- Das numerische Lösen nichtlinearer algebraischer Gleichungssysteme erfordert das mehrfache Lösen von linearen Gleichungssystemen.

Man sieht, lineare Gleichungssysteme sind wirklich das Arbeitspferd der numerischen Simulation.

Auch die Leistung der größten Supercomputer der Welt wird mit der Anzahl Fließkommaoperationen pro Sekunde gemessen, die bei der Lösung von linearen Gleichungssystemen erreicht wird (Linpack Benchmark).

Das zur Zeit schnellste System (Liste vom Juni 2007) ist die IBM BlueGene/L am Lawrence Livermore Lab, USA, mit 131072 Prozessoren. Diese Maschine erreicht $2.8 \cdot 10^{14}$ Fließkommaoperationen pro Sekunde (0.28 PF) bei der Lösung eines Gleichungssystemes mit der Dimension $2 \cdot 10^6$. Siehe `http://www.top500.org/`.

Sehr oft treten lineare Gleichungssysteme als Teilprobleme in einer größeren Aufgabe auf. Wir wollen nun zwei Anwendungen betrachten, die direkt auf lineare Gleichungssysteme führen.

Ausgleichsrechnung Ein Polynom $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ vom Grad n wird durch $n+1$ paarweise verschiedene Datenpunkte (x_i, y_i), $i = 0, \ldots, n$ eindeutig bestimmt.

Oft sind die Datenpunkte fehlerbehaftet und man versucht, durch Messung von $m+1 > n+1$ Datenpunkten (zu mindestens $n+1$ verschiedenen x_i) den Messfehler „auszumitteln“.

Eine Möglichkeit dies zu tun ist es, die $n+1$ Koeffizienten $a_0, \ldots, a_n$ derart zu bestimmen, dass die Funktion

$$g(a_0, \ldots, a_n) = \sum_{i=0}^{m} w_i [p(x_i) - y_i]^2 \quad \rightarrow \quad \min$$

minimiert wird. Dies nennt man „Methode der kleinsten Quadrate“ oder „Ausgleichsrechnung“. Diese hat der 24-jährige Gauß[28] benutzt, um die Bahnelemente des Asteroiden Ceres zu bestimmen.

[28] Carl Friedrich Gauß, 1777-1855, dt. Mathematiker.

Die Zahlen w_i sind Gewichte, mit denen man das Vertrauen in die Messungen quantifizieren kann.

Notwendig für ein Minimum von g ist die Bedingung

$$\frac{\partial g}{\partial a_j} = 0, \qquad j = 0, \ldots, n.$$

Nach Einsetzen erhalten wir:

$$\begin{aligned}\frac{\partial}{\partial a_j}\left[\sum_{i=0}^{m} w_i \left(\sum_{k=0}^{n} a_k x_i^k - y_i\right)^2\right] &= \sum_{i=0}^{m} w_i 2 \left(\sum_{k=0}^{n} a_k x_i^k - y_i\right) x_i^j \\ = \sum_{i=0}^{m}\sum_{k=0}^{n} 2 w_i a_k x_i^{k+j} - \sum_{i=0}^{m} 2 w_i x_i^j y_i &\stackrel{!}{=} 0 \qquad j = 0, \ldots, n.\end{aligned}$$

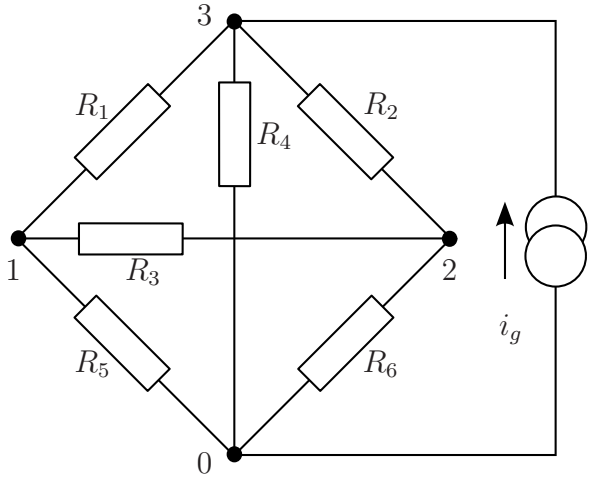
Dies führt auf das lineare Gleichungssystem

$$\sum_{k=0}^{n} a_k \underbrace{\left(\sum_{i=0}^{m} w_i x_i^{k+j}\right)}_{n_{j,k}} = \underbrace{\sum_{i=0}^{m} w_i x_i^j y_i}_{b_j} \qquad j = 0, \ldots, n.$$

Die Gleichungen $Na = b$ heißen auch „Normalengleichungen“.

Wegen $N^T = N$ ist N symmetrisch. Es gilt auch $\xi^T N \xi > 0$, $\forall \xi \neq 0$, also N positiv definit. Daraus folgt die eindeutige Lösbarkeit von $Na = b$.

Netzwerkanalyse Eine praktische Aufgabenstellung, die direkt auf lineare Gleichungssysteme führt, ist die elektrische Netzwerkanalyse.



Zu bestimmen seien alle Zweigströme und Spannungen in dem oben angegebenen Netzwerk.

Für dieses und noch sehr viel allgemeinere Netzwerke wurden in der Elektrotechnik diverse Analyseverfahren entwickelt.

Das *Knotenpotentialverfahren* führt auf das lineare Gleichungssystem

$$\begin{bmatrix} \frac{1}{R_1}+\frac{1}{R_3}+\frac{1}{R_5} & -\frac{1}{R_3} & -\frac{1}{R_1} \\ -\frac{1}{R_3} & \frac{1}{R_2}+\frac{1}{R_3}+\frac{1}{R_6} & -\frac{1}{R_2} \\ -\frac{1}{R_1} & -\frac{1}{R_2} & \frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_4} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ i_g \end{bmatrix}$$

für die Knotenpotentiale u_1, u_2, u_3. Implizit gilt $u_0 = 0$.

Die Zweigspannungen ergeben sich dann als

$$u_{10} = u_1 - u_0, \quad u_{12} = u_1 - u_2, \quad \text{usw.}$$

Der Fluss von Wasser in Rohrleitungssystemen lässt sich ganz ähnlich analysieren.

Netzwerke bestehend aus Widerständen, Kondensatoren und Spulen lassen sich bei harmonischer Anregung im eingeschwungenen Zustand mit komplexwertigen linearen Gleichungssystemen beschreiben.

Dies setzt sog. ideale Netzwerkelemente voraus.

10.2 Aufgabenstellung

Wir wollen nun also das lineare Gleichungssystem

$$Ax = b$$

lösen mit

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbf{R}^{m\times n}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbf{R}^n.$$

$Ax = b$ heißt

- unterbestimmt falls $m < n$,
- quadratisch falls $m = n$
- und überbestimmt falls $m > n$.

$Ax = b$ ist genau dann lösbar (für beliebige m, n), wenn

$$\text{Rang}(A) = \text{Rang}([A, b])$$

(Vorsicht: hier steht nicht *eindeutig*).

Im quadratischen Fall (den wir hier nur betrachten wollen) sind folgende Aussagen äquivalent:

(i) $Ax = b$ ist für jedes b eindeutig lösbar, d. h. „regulär“,

(ii) $\text{Rang}(A) = n$,

(iii) $\det(A) \neq 0$,

(iv) alle Eigenwerte von A sind ungleich Null.

Hinsichtlich A unterscheidet man:

vollbesetzte Matrizen Anzahl Nichtnullelemente von A ist $O(n^2)$. Dies bedeutet insbesondere:

- Nehme auf Nullen keine besondere Rücksicht.
- Datenstruktur zur Speicherung der Matrix ist ein zweidimensionales Feld.

dünnbesetzte Matrizen Anzahl Nichtnullelemente von A ist $O(n)$ oder höchsten $O(n \log n)$. Das hat zur Konsequenz:

- Es lohnt auf die Nullen Rücksicht zu nehmen und sie *nicht* zu speichern.
- Dies erfordert spezielle Datenstrukturen. Hier gibt es verschiedene Möglichkeiten, je nach Struktur (Bandmatrix, Blockmatrix, beliebig).

Bei vielen in der Praxis auftretenden dünnbesetzten Matrizen gilt, dass die Anzahl der Nichtnullelemente pro Zeile konstant ist (unabhängig von n).

Hinsichtlich der Lösungsverfahren unterscheidet man:

direkte Verfahren Liefern in exakter Arithmetik nach vorab bekannter Zahl von Rechenoperationen für jedes invertierbare A und b die Lösung x.

iterative Verfahren Diese konstruieren ausgehend von einem beliebigen Startwert x^0 eine Folge

$$x^0, x^1, \ldots, x^k, \ldots \qquad \text{mit } \|x - x^k\| \to 0 \text{ für } k \to \infty.$$

Diese Verfahren sind vor allem geeignet für dünnbesetzte Matrizen.

Oft ist bei diesen Verfahren der Aufwand pro Schritt $O(n)$ und die wesentliche Frage ist wieviele Schritte notwendig sind.

10.3 Kondition der Lösung linearer Gleichungssysteme

Bevor wir uns der Lösung von $Ax = b$ zuwenden, untersuchen wir die Kondition dieser Aufgabe.

Wir interessieren uns also für die Auswirkung von Änderungen in A bzw. b auf das Ergebnis x.

Um das quantifizieren zu können, benötigen wir den Begriff der Norm auf Vektoren und Matrizen.

Definition 10.1 (Norm eines Vektors)**.** Eine Abbildung $\|.\| : \mathbf{R}^n \to \mathbf{R}_+$ heißt *Norm*, wenn sie folgende Eigenschaften erfüllt:

(i)	$\|x\| > 0 \quad x \in \mathbf{R}^n \backslash \{0\}$	(Definitheit)
(ii)	$\|\alpha x\| = \|\alpha\| \|x\|, \quad x \in \mathbf{R}^n, \alpha \in \mathbf{R}$	(positive Homogenität)
(iii)	$\|x + y\| \leq \|x\| + \|y\|, \quad x, y \in \mathbf{R}^n$	(Dreiecksungleichung)

□

Die drei wichtigsten Vektornormen sind:

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

$$\|x\|_2 = \Big(\sum_{i=0}^{n} x_i^2\Big)^{\frac{1}{2}} \qquad (\text{Euklidische Norm, } l_2 \text{ -Norm})$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \qquad (\text{Maximum-Norm, } l_\infty \text{ -Norm})$$

Auch $\mathbf{R}^{n\times n}$ stellt einen Vektorraum (der Dimension n^2) dar und Normen können entsprechend Definition 10.1 vereinbart werden. Man benötigt aber in der Praxis noch zusätzliche Eigenschaften.

Definition 10.2. Eine Norm $\|.\| : \mathbf{R}^{n\times n} \to \mathbf{R}_+$ heißt *verträglich* mit der Vektornorm $\|.\| : \mathbf{R}^n \to \mathbf{R}_+$, wenn gilt

$$\|Ax\| \le \|A\|\|x\| \qquad \forall x \in \mathbf{R}^n.$$

□

Definition 10.3. Eine Norm $\|.\| : \mathbf{R}^{n\times n} \to \mathbf{R}_+$ heißt *Matrizennorm* (oder submultiplikativ), wenn gilt

$$\|AB\| \le \|A\|\|B\| \qquad A, B \in \mathbf{R}^{n\times n}.$$

□

Die *Frobeniusnorm*

$$\|A\|_{Fr} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$$

ist eine mit der euklidischen Norm verträgliche Matrizennorm.

Zu einer beliebigen Vektornorm $\|.\|$ erhält man immer eine verträgliche Matrizennorm mittels:

Definition 10.4.

$$\|A\| := \sup_{x\in\mathbf{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbf{R}^n, \|x\|=1} \|Ax\|$$

heißt natürliche (oder zugeordnete) Matrixnorm zu $\|.\|$. □

Wichtige Matrizennormen sind:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \qquad \text{Spaltensummennorm, zug. zu } \|.\|_1$$

$$\|A\|_2 = \max\{|\lambda|^{\frac{1}{2}} \mid \lambda \text{ Eigenwert von } A^TA\} \qquad \text{Spektralnorm, zug. zu } \|.\|_2$$

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \qquad \text{Zeilensummennorm, zug. zu } \|.\|_\infty$$

Damit können wir die folgende Aussage zur Kondition von $Ax = b$ machen.

Satz 10.5 (Störungssatz)**.** Die Matrix $A \in \mathbf{R}^{n\times n}$ sei regulär und es gelte für die Störungsmatrix

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}.$$

Ist nun $Ax = b$, so gilt für die Lösung des gestörten Systems $(A+\delta A)(x+\delta x) = (b+\delta b)$ folgende Abschätzung:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\left\{\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right\} \quad .$$

Dabei ist $\|.\|$ eine beliebige Vektornorm mit verträglicher Matrizennorm und

$$\kappa(A) = \|A^{-1}\|\|A\|$$

die sogenannte *Konditionszahl* von A.

Beweis: siehe [Ran06, Satz 4.1]. □

Regel 10.6. Der Fehler in der Eingabe sei $\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}$, $\frac{\|\delta b\|}{\|b\|} \approx 10^{-k}$ und die Kondition der Matrix A sei $\kappa(A) \approx 10^s$ wobei $0 \leq 10^s \cdot 10^{-k} \ll 1$ gelten soll. Dann gilt

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{10^s}{1-10^s\cdot 10^{-k}} \cdot 2 \cdot 10^{-k} \approx 10^{s-k}$$

Man verliert also s Stellen Genauigkeit! (Vorher war der Fehler in der k-ten Nachkommastelle, dann ist er in der Stelle $k-s$).

Man kann auch zeigen, dass die Abschätzung im wesentlichen scharf ist [Ran06, S. 111]. □

Beispiel 10.7 (Kondition und Determinante)**.** (a) Wir betrachten die folgende 2×2 Matrix:

$$A = \begin{bmatrix} -1 & 1 \\ 1+\varepsilon & -1 \end{bmatrix}, \qquad A^{-1} = \frac{1}{\varepsilon}\begin{bmatrix} -1 & -1 \\ -(1+\varepsilon) & -1 \end{bmatrix} \quad .$$

Für die Kondition gilt also

$$\|A\|_\infty = \max\{2, 2+\varepsilon\}, \quad \|A^{-1}\|_\infty = \frac{1}{\varepsilon}\max\{2, 2+\varepsilon\}, \quad \kappa(A) = \frac{(2+\varepsilon)^2}{\varepsilon}.$$

Hier ist $\det(A) = \varepsilon$ die Determinante von A. Im allgemeinen ist die Kleinheit der Determinante aber kein gutes Maß für die Kondition wie folgendes Beispiel zeigt.

(b) Betrachte

$$B = \begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{bmatrix}, \qquad B^{-1} = \begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{bmatrix} \quad .$$

und damit $\kappa(B) = 1$ *obwohl* $\det(B) = 10^{-20}$. □

10.4 Gauß-Elimination

Wir beschränken uns hier auf quadratische Systeme ($m = n$). Besonders leicht lösbar sind Gleichungsysteme mit einer *oberen Dreiecksmatrix*:

$$a_{ij} = 0 \quad \forall i > j \quad \text{und} \quad a_{ii} \neq 0 \quad \forall i = 1, \ldots, n.$$

also

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

Sog. *Rückwärtseinsetzen* führt zu dem Verfahren

$$x_n = \frac{b_n}{a_{nn}}; \quad i = n-1, \ldots, 1: \quad x_i = \frac{\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right)}{a_{ii}}$$

Um so ein Gleichungssystem zu lösen, benötigt man

$$F_{\text{Back}}(n) = \sum_{i=0}^{n-1}(2i+1) = n(n-1) + n = n^2$$

arithmetische Operationen.

Die Gauß-Elimination formt das gegebene System $Ax = b$, welches *eindeutig lösbar* sein soll, schrittweise so um, dass eine obere Dreiecksmatrix entsteht.

Hierzu benutzt man die elementaren Umformungen:

(i) Vertauschen zweier Gleichungen

(ii) Addition des Vielfachen einer Gleichung zu einer anderen.

Keine dieser Umformungen ändert die Lösung des linearen Gleichungssystems.

Zur kompakteren Notation ordnet man die Matrix A und die rechte Seite b in einer einzigen $n \times (n+1)$ Matrix an:

$$\left[A^{(0)}, b^{(0)}\right] = \left[A, b\right].$$

Oben kommt ein Superskript (0) dran, da es sich hier um die Ausgangssituation handelt.

Wir beschreiben nun die Schritte des Verfahrens.

Bestimme $r \in \{1, \ldots, n\}$ so, dass $a_{r1}^{(0)} \neq 0$. Vertausche die Zeilen r und 1. Das Ergebnis bekommt nach dieser Operation eine Schlange drüber:

$$\left[\tilde{A}^{(0)}, \tilde{b}^{(0)}\right] = \left[\begin{array}{cccc|c} \boxed{\tilde{a}_{11}^{(0)} \neq 0} & \tilde{a}_{12}^{(0)} & \cdots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ \vdots & & & & \\ \tilde{a}_{n1}^{(0)} & \cdots & \cdots & \tilde{a}_{nn}^{(0)} & \tilde{b}_n^{(0)} \end{array}\right]$$

Das neue Element $\tilde{a}_{11}$ nach dem Vertauschen heißt *Pivotelement*. Dies ist immer möglich, denn sonst wäre A nicht regulär.

Für alle $i \in \{2, \ldots, n\}$ *subtrahiere* nun das $\frac{\tilde{a}_{i1}^{(0)}}{\tilde{a}_{11}^{(0)}} =: q_{i1}$-fache der ersten Zeile von der i-ten Zeile. In Formeln:

$$q_{i1} = \frac{\tilde{a}_{i1}^{(0)}}{\tilde{a}_{11}^{(0)}}; \quad j = 1, \ldots, n: \quad a_{ij}^{(1)} = \tilde{a}_{ij}^{(0)} - q_{i1} \cdot \tilde{a}_{1j}^{(0)}, \quad b_i^{(1)} = \tilde{b}_i^{(0)} - q_{i1} \cdot \tilde{b}_i^{(0)}$$

Die so entstandene Matrix trägt den Superskript (1) (ohne Schlange).

Wegen

$$a_{i1}^{(1)} = \tilde{a}_{i1}^{(0)} - \frac{\tilde{a}_{i1}^{(0)}}{\tilde{a}_{11}^{(0)}} \tilde{a}_{11}^{(0)} = 0$$

gilt:

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

Wir haben also die erste Spalte unterhalb des Pivotelementes zu Null eliminiert.

Nun verfahre ebenso mit der Teilmatrix, die durch Streichen der ersten Zeile bzw. Spalte entsteht.

D. h. sorge für $\tilde{a}_{22}^{(1)} \neq 0$ und eliminiere $\tilde{a}_{32}^{(1)} \ldots \tilde{a}_{n2}^{(1)}$.

Nach k solchen Schritten ergibt sich schließlich folgende Situation:

$$\left[A^{(k)}, b^{(k)}\right] = \left[\begin{array}{ccccccc|c} \tilde{a}_{11}^{(0)} & & & & & & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & \tilde{a}_{22}^{(1)} & & & & & \tilde{a}_{2n}^{(1)} & \tilde{b}_2^{(1)} \\ \vdots & & \ddots & & & & & \\ 0 & \ldots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & & b_k^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k)} & & a_{nn}^{(k)} & & b_n^{(k)} \end{array}\right].$$

Als Algorithmus schreibt sich das Ganze so:

for $(k = 1; k < n; k = k + 1)$ **do**
  Finde $r \in \{$k,...,n$\}$ so dass $a_{rk} \neq 0$
  und vertausche Zeilen k und r {sorge dafür, dass $a_{kk} \neq 0$}
  for $(i = k + 1; i \leq n; i = i + 1)$ **do**
    $q_{ik} = a_{ik}/a_{kk}$;
    for $(j = k + 1; j \leq n; j = j + 1)$ **do**
      $a_{ij} = a_{ij} - q_{ik} \cdot a_{kj}$;
    end for

$b_i = b_i - q_{ik} b_k$;
end for
end for

Bemerkung 10.8. Elemente von $A^{(k)}$ werden jeweils mit denen von $A^{(k+1)}$ überschrieben. Das ursprüngliche A und b stehen somit *nicht* mehr zur Verfügung.

Der gegebene Algorithmus ist nicht numerisch stabil gegenüber Rundungsfehlern. Dazu nächstes Mal mehr. □

Für den Aufwand erhält man:

$$\begin{aligned} F_{\text{Gauß}}(n) &= \sum_{k=1}^{n-1} \left\{ \underbrace{n-k}_{\text{Multiplikatoren } q_{ik}} + (n-k)[2+2(n-k)] \right\} \\ &= 2\sum_{k=1}^{n-1} (n-k)^2 + O(n^2) \\ &= \frac{2}{3} n^3 + O(n^2) \quad . \end{aligned}$$

Der oben angegebene naive Algorithmus nutzt den Cache in heutigen Prozessoren für große n nicht gut aus.

Es gibt jedoch cache-optimale Implementierungen, die die Tatsache, dass $O(n^3)$ Operationen auf $O(n^2)$ Daten (Speicher für A, b) ausgeführt werden, ausnutzen können.

Die gesamte Prozedur zur Lösung von $Ax = b$ besteht somit aus:

(i) Bringe A auf obere Dreiecksgestalt.

(ii) Löse Dreieckssystem durch Rückwärtseinsetzen.

Beispiel 10.9. Wir geben ein Beispiel zur Gauß-Elimination. Hier sind keine Zeilenvertauschungen notwendig. Das Pivotelement ist jeweils durch einen Kasten gekennzeichnet.

$$\left[\begin{array}{cccc|c} \boxed{2} & 4 & 6 & 8 & 40 \\ 16 & 33 & 50 & 67 & 330 \\ 4 & 15 & 31 & 44 & 167 \\ 10 & 29 & 63 & 97 & 350 \end{array}\right] \rightarrow \left[\begin{array}{cccc|c} 2 & 4 & 6 & 8 & 40 \\ 0 & \boxed{1} & 2 & 3 & 10 \\ 0 & 7 & 19 & 28 & 87 \\ 0 & 9 & 33 & 57 & 150 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cccc|c} 2 & 4 & 6 & 8 & 40 \\ 0 & 1 & 2 & 3 & 10 \\ 0 & 0 & \boxed{5} & 7 & 17 \\ 0 & 0 & 15 & 30 & 60 \end{array}\right] \rightarrow \left[\begin{array}{cccc|c} 2 & 4 & 6 & 8 & 40 \\ 0 & 1 & 2 & 3 & 10 \\ 0 & 0 & 5 & 7 & 17 \\ 0 & 0 & 0 & 9 & 9 \end{array}\right]$$

Schließlich liefert Rückwärtseinsetzen:

$$x_4 = 9/9 = \boxed{1}, \qquad x_3 = (17 - 7 \cdot 1)/5 = \boxed{2},$$
$$x_2 = (10 - 2 \cdot 2 - 3 \cdot 1)/1 = \boxed{3}, \qquad x_1 = (40 - 4 \cdot 3 - 6 \cdot 2 - 8 \cdot 1)/2 = \boxed{4}.$$

□

10.5 Zusammenfassung

- Lineare Gleichungssysteme $Ax = b$ lösen ist *die* Standardaufgabe im Wissenschaftlichen Rechnen.
- Die Konditionierung der Aufgabe hängt von der Konditionszahl $\kappa(A) = \|A\|\|A^{-1}\|$ der Matrix A ab.
- Als ein erstes Lösungsverfahren, welches insbesondere für dichtbesetzte Matrizen geeignet ist, haben wir das Gaußsche Eliminationsverfahren kennengelernt.
- Der Aufwand für diese Methode beträgt $O(n^3)$ bei $A \in \mathbf{R}^{n\times n}$.

11 Pivotisierung und LR-Zerlegung

11.1 Pivotisierung

Für eine *reguläre* Matrix A führt die Gauß-Elimination in *exakter* Arithmetik immer auf eine obere Dreiecksmatrix.

Dabei wird in der Restspalte immer ein nichtverschwindendes Pivotelement gefunden.

Die Sache ändert sich, wenn wir das Gaußsche Verfahren in Fließkommaarithmetik durchführen.

Hier kann selbst ein einzelner Rundungsfehler fatale Auswirkungen haben. Dazu betrachten wir ein Beispiel.

Beispiel 11.1 (aus [GO96])**.** Wir betrachten das 2×2 System

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{11.1}$$

In exakter Arithmetik führt das Gaußsche Verfahren nach Elimination von a_{21} auf

$$\begin{bmatrix} -10^{-5} & 1 \\ 0 & 1 + 2 \cdot 10^5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \cdot 10^5 \end{bmatrix}$$

mit der Lösung

$$x_1 = -0.4999975, \qquad x_2 = 0.999995 \quad .$$

Nun führen wir das Verfahren in $\mathbb{F}(10, 4, 1)$ durch. Beim Multiplikator

$$q_{21} = (0.2 \cdot 10^1) \oslash (-0.1 \cdot 10^{-4}) = -0.2 \cdot 10^6$$

ergibt sich kein Rundungsfehler.

Für das neue a_{22} ergibt sich

$$\begin{aligned} a_{22}^{(1)} &= 0.1 \cdot 10^1 \ominus (-0.2 \cdot 10^6) \odot (0.1 \cdot 10^1) \\ &= 0.1 \cdot 10^1 \oplus 0.2 \cdot 10^6 = \boxed{0.2 \cdot 10^6}. \end{aligned}$$

Hier wurde auf vier Stellen gerundet.

Damit ergibt sich (ohne Fehler)

$$b_2^{(1)} = -(-0.2 \cdot 10^6) \odot (0.1 \cdot 10^1) = 0.2 \cdot 10^6$$

und

$$\begin{aligned} x_2 &= b_2^{(1)} \oslash a_{22}^{(1)} = 0.2 \cdot 10^6 \oslash 0.2 \cdot 10^6 = \boxed{1}, \\ x_1 &= (0.1 \cdot 10^1 \ominus 0.1 \cdot 10^1 \odot 1) \oslash (-0.1 \cdot 10^{-4}) = \boxed{0} \quad . \end{aligned}$$

Es ist also *keine* Stelle im Ergebnis korrekt obwohl nur an einer *einzigen* Stelle (in der Berechnung von $a_{22}^{(1)}$) ein Rundungsfehler eingeführt wurde.

Darüberhinaus überprüfe man, dass für die Kondition von A gilt:

$$\kappa(A) = 3 \quad .$$

Demnach ist das System gut konditioniert! Der Algorithmus, so wie er ist, ist numerisch nicht stabil.

Das Problem ist offensichtlich der große Multiplikator q_{21} der aus dem sehr kleinen a_{11} resultiert und der dafür sorgt, dass das ursprüngliche a_{22} in $a_{22}^{(1)}$ vollkommen ignoriert wird.

Im Prinzip haben wir in Fließkommaarithmetik das System

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

exakt gelöst, was eine völlig andere Lösung hat als das ursprüngliche (11.1) (Rückwärtsanalyse).

Der große Multiplikator kann ganz einfach vermieden werden indem man eine Zeilenvertauschung durchführt, d. h. wir lösen

$$\begin{bmatrix} 2 & 1 \\ -10^{-5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{11.2}$$

Nun erhält man

$$\begin{aligned}
q_{21} &= -0.1 \cdot 10^{-4} \oslash 0.2 \cdot 10^{1} = -0.5 \cdot 10^{-5}, \\
a_{22}^{(1)} &= 0.1 \cdot 10^{1} \ominus (-0.5 \cdot 10^{-5}) \odot 0.1 \cdot 10^{1} = 0.1 \cdot 10^{1} \oplus 0.5 \cdot 10^{-5} = 0.1 \cdot 10^{1}, \\
b_2^{(1)} &= 0.1 \cdot 10^{1} \ominus 0.5 \cdot 10^{-5} \odot 0 = 0.1 \cdot 10^{1}, \\
x_2 &= 0.1 \cdot 10^{1} \oslash 0.1 \cdot 10^{1} = \boxed{1}, \\
x_1 &= (0 \ominus 0.1 \cdot 10^{1} \odot 0.1 \cdot 10^{1}) \oslash 0.2 \cdot 10^{1} = \boxed{-0.5},
\end{aligned}$$

was in $\mathbb{F}(10, 4, 1)$ völlig in Ordnung ist. □

Dies legt den folgenden Algorithmus nahe. Als Pivotelement wählen wir immer das betragsgrößte Element der Spalte.

for $(k = 1; k < n; k = k + 1)$ **do**
  Finde $r \in$ {k,...,n} so dass $|a_{rk}|$ maximal ist
  und vertausche Zeilen k und r
  if $(a_{kk} = 0)$ **then**
    STOP, Matrix ist singulär;
  end if
  for $(i = k + 1; i \leq n; i = i + 1)$ **do**
    $q_{ik} = a_{ik}/a_{kk}$;
    for $(j = k + 1; j \leq n; j = j + 1)$ **do**
      $a_{ij} = a_{ij} - q_{ik} \cdot a_{kj}$;
      $b_i = b_i - q_{ik} b_k$;

end for
end for
end for

Man nennt dieses Vorgehen *Spaltenpivotisierung.*

Im Prinzip kann man auch das betragsgrößte Element aus der kompletten Restmatrix $\{a_{ij}|i,j \geq k\}$ bestimmen und als Pivotelement verwendet. Dann spricht man von *Totalpivotisierung.*

Dies erfordert zusätzlich noch Spaltenvertauschung (Umnummerieren von Unbekannten) und ist deshalb etwas aufwendiger zu realisieren.

Die Spaltenpivotisierung alleine ist allerdings immer noch nicht ausreichend, wie das folgende Beispiel zeigt.

Beispiel 11.2 (ebenfalls aus [GO96])**.** Wir betrachten das 2×2 System

$$\begin{bmatrix} 10 & -10^6 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -10^6 \\ 0 \end{bmatrix}.$$

welches aus (11.1) durch Multiplikation der ersten Zeile mit -10^6 entsteht.

Die Spaltenpivotisierung erfordert keine Vertauschung. Allerdings entsteht für $a_{22}^{(1)} = 1 + 2 \cdot 10^5$ genau dasselbe Problem wie oben! □

Die Spaltenpivotisierung ist effektiver, wenn man das Gleichungssystem *vor* der Elimination so skaliert, dass die betragsmäßigen Zeilensummen der Elemente in etwa gleich sind.

Dies erreicht man durch Multiplikation mit einer Diagonalmatrix von links:

$$Ax = b \quad \rightarrow \quad DAx = Db \qquad \text{mit } d_{ii} = \Big(\sum_{j=1}^{n} |a_{ij}|\Big)^{-1}.$$

Es gibt auch Gleichungssysteme, die immer ohne Pivotisierung eliminiert werden können. Dies sind:

- Reguläre, diagonaldominante Matrizen:

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \leq |a_{ii}| \qquad \forall i = 1, \ldots, n$$

  (gilt $<$ statt $\leq$ ist die Matrix automatisch regulär).

- Positiv definite Matrizen:

$$\langle x, Ax \rangle > 0 \qquad \forall x \in \mathbb{R}^n \setminus \{0\}, \qquad \langle x, y \rangle = \sum_{i=1}^{n} x_i y_i.$$

11.2 LR-Zerlegung

Wir betrachten die Gauß-Elimination ohne Zeilenvertauschung nun in Matrixform.

Die Elimination des Elementes a_{ik} lässt sich als Matrixmultiplikation von links schreiben. Sei

$$q_{ik} = \frac{a_{ik}}{a_{kk}} \text{ und } Q_{ik} \text{ die Matrix } (Q_{ik})_{\alpha,\beta} = \begin{cases} q_{ik} & \alpha = i \wedge \beta = k \\ 0 & \text{sonst} \end{cases}$$

dann beschreibt

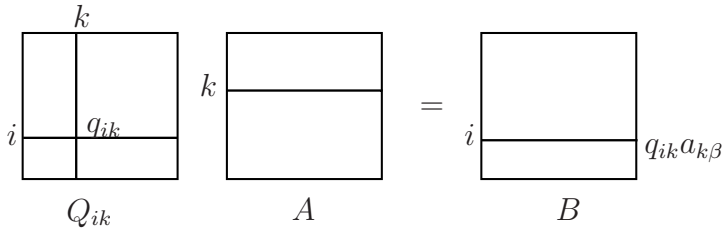
$$(I - Q_{ik})A$$

die Subtraktion des q_{ik}-fachen der k-ten Zeile von A von der i-ten Zeile von A.

Das sieht man wie folgt:

$$(I - Q_{ik})A = A - \underbrace{Q_{ik}A}_{B}$$

$$b_{\alpha\beta} = \begin{cases} 0 & \alpha \neq i \\ q_{ik}a_{k\beta} & \alpha = i \leftarrow \text{ i-te Zeile} \end{cases}$$



Damit gilt für die komplette Gauß-Elimination auf obere Dreiecksgestalt:

$$\underbrace{(I - Q_{n,n-1})}_{\text{letztes zu eliminierendes Element}} \cdots \underbrace{(I - Q_{32})}_{a_{32}} \cdots \underbrace{(I - Q_{31})}_{a_{31}} \underbrace{(I - Q_{21})}_{\text{Elim. von } a_{21}} A = \underbrace{R}_{\substack{\text{rechte obere Dreiecksmatrix.} \\ \text{Ergebnis der Gauß-Elim.}}} . \quad (11.3)$$

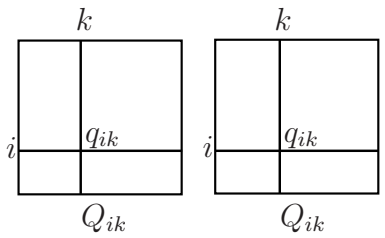
Die Matrix $(I - Q_{ik})$ hat eine einfache Inverse, denn

$$(I - Q_{ik})(I + Q_{ik}) = I + Q_{ik} - Q_{ik} - \underbrace{Q_{ik}Q_{ik}}_{=0 \text{ da } i>k} = I$$

und (es genügt $(Q_{ik}Q_{ik})_{ik}$ zu betrachten)

$$\sum_{\alpha=1}^{n} (Q_{ik})_{i\alpha}(Q_{ik})_{\alpha k} \neq 0 \Leftrightarrow \alpha = k \wedge \alpha = i$$

was aber wegen $i > k$ unmöglich ist.



Dies zeigt $(I - Q_{ik})^{-1} = I + Q_{ik}$

Damit können wir (11.3) nach A auflösen, indem man von links mit den ganzen Inversen multipliziert:

$$(I + Q_{2,1}) \cdots \underbrace{(I + Q_{n,n-1})(I - Q_{n,n-1})}_{=I} \cdots (I - Q_{2,1})A = \underbrace{(I + Q_{2,1}) \cdots (I + Q_{n,n-2})(I + Q_{n,n-1})}_{=L} R. \quad (11.4)$$

Es ergibt sich die sog. *LR-Zerlegung*

$$A = LR$$

mit einer unteren Dreiecksmatrix L (dies ist noch zu zeigen) und einer oberen Dreiecksmatrix R.

Die Matrix L hat folgende Gestalt:

$$L = (I + Q_{2,1}) \cdots (I + Q_{n,n-2})(I + Q_{n,n-1}) = I + Q_{2,1} + \ldots Q_{n,n-2} + Q_{n,n-1} \quad (11.5)$$

Damit ist L eine untere Dreiecksmatrix und $(L)_{\alpha,\alpha} = 1$, denn die $Q_{i,k}$ sind strikte untere Dreiecksmatrizen.

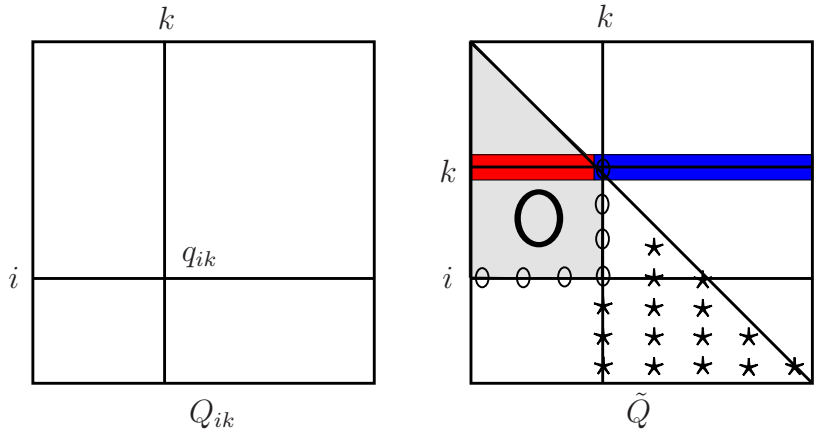
Wir zeigen dies durch Induktion über die umgekehrte Reihenfolge der Elimination.

(i) $I + Q_{n,n-1}$ aus dem letzten Schritt hat die geforderte Gestalt.

(ii) Betrachte den Schritt

$$(I + Q_{ik})(I + \underbrace{Q_{i+1,k} + \ldots + Q_{n,n-1}}_{\tilde{Q}}) = I + Q_{i,k} + \tilde{Q} + Q_{i,k}\tilde{Q}.$$

Wir zeigen nun, dass $Q_{i,k}\tilde{Q} = 0$.



Es gilt $\left(Q_{i,k}\tilde{Q}\right)_{\alpha,\beta} = \begin{cases} 0 & \alpha \neq i, \text{ da nur } q_{i,k} \neq 0 \\ q_{i,k}\tilde{q}_{k,\beta} & \alpha = i,\ \beta \text{ beliebig} \end{cases}$. Also betrachte $\tilde{q}_{k,\beta}$:

$\underline{\beta \geq k}$: $\tilde{q}_{k,\beta} = 0$, da $\tilde{Q}$ strikte untere Dreiecksmatrix (blauer Teil in der Abbildung).

$\underline{\beta < k}$: im Schritt (i,k), $k < i$, sind in $\tilde{Q}$ die Elemente zu den Indizes $\{(i',k') \,|\, k' < i' \wedge i' \leq i \wedge k' \leq k\}$ sicher noch Null (grauer Bereich in der Abbildung). Damit ist aber auch $\tilde{q}_{k,\beta} = 0$ (roter Bereich im Bild).

Die algorithmische Formulierung der LR-Zerlegung (ohne Pivotisierung) lautet:

for $(k = 1; k < n; k = k + 1)$ **do**
  for $(i = k + 1; i \leq n; i = i + 1)$ **do**
    $a_{ik} = a_{ik}/a_{kk}$; {Überschreibe a_{ik} mit $q_{ik} = l_{ik}$}
    for $(j = k + 1; j \leq n; j = j + 1)$ **do**
      $a_{ij} = a_{ij} - a_{ik}a_{kj}$;
    end for
  end for
end for

Am Ende gilt

$$r_{\alpha\beta} = a_{\alpha\beta} \text{ für } \beta \geq \alpha \text{ (oberes Dreieck)},$$
$$l_{\alpha\beta} = a_{\alpha\beta} \text{ für } \beta < \alpha \text{ (striktes unteres Dreieck)}.$$

$l_{\alpha\alpha} = 1$ speichert man nicht explizit ab.

Beispiel 11.3. Hier das Beispiel der LR-Zerlegung einer 4×4 Matrix, die wir schon aus Beispiel 10.9 kennen.

Der Unterschied ist, dass die Multiplikatoren im unteren Dreieck gespeichert werden und dass

die rechte Seite b wegfällt.

$$\begin{bmatrix} \boxed{2} & 4 & 6 & 8 \\ 16 & 33 & 50 & 67 \\ 4 & 15 & 31 & 44 \\ 10 & 29 & 63 & 97 \end{bmatrix} \qquad \rightarrow \begin{bmatrix} 2 & 4 & 6 & 8 \\ 8 & \boxed{1} & 2 & 3 \\ 2 & 7 & 19 & 28 \\ 5 & 9 & 33 & 57 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 2 & 4 & 6 & 8 \\ 8 & 1 & 2 & 3 \\ 2 & 7 & \boxed{5} & 7 \\ 5 & 9 & 15 & 30 \end{bmatrix} \qquad \rightarrow \begin{bmatrix} 2 & 4 & 6 & 8 \\ 8 & 1 & 2 & 3 \\ 2 & 7 & 5 & 7 \\ 5 & 9 & 3 & 9 \end{bmatrix}$$

Man überprüfe, dass $LR = A$ gilt. □

Bemerkung 11.4. Die LR-Zerlegung lässt sich auch mit Pivotisierung durchführen. Man erhält dann eine Zerlegung

$$PA = LR$$

wobei P eine Permutationsmatrix ist, die die Zeilenvertauschungen beschreibt.

Ausserdem sind unterschiedliche Eliminationsreihenfolgen möglich. Bisher wurde die „spaltenorientierte“ Variante betrachtet. □

Bemerkung 11.5. Ist A eine symmetrische und positiv definite Matrix, d. h. $A = A^T$ und $x^T Ax > 0 \; \forall x \neq 0$, dann kann A zerlegt werden in

$$A = LDL^T$$

wobei D die Diagonale von R aus der LR-Zerlegung ist.

Dies nennt man die Cholesky-Zerlegung. Der Aufwand zur Berechnung ist halb so groß wie bei der LR-Zerlegung.

Warum? Es gilt

$$A = LR = LD \underbrace{D^{-1}R}_{=L^T} \qquad \text{aus Symmetriegründen.}$$

Man kann auch alternativ schreiben

$$A = \tilde{L}\tilde{L}^T \qquad \text{mit } \tilde{L} = LD^{1/2},\ \left(D^{1/2}\right)_{i,i} = \sqrt{d_{i,i}}.$$

□

Zur Lösung von $Ax = b$ setzt man

$$Ax = L \underbrace{Rx}_{=:y} = b,$$

und muss dann zwei Dreieckssysteme lösen:

(i) $Ly = b$, gefolgt von

(ii) $Rx = y$.

Die Auflösung der beiden Dreieckssystem als Algorithmus:

for $(i = 1; i \leq n; i = i + 1)$ **do**
  for $(j = 1; j < i; j = j + 1)$ **do**
    $b_i = b_i - l_{i,j} y_j$;
    $y_i = b_i$; {da $l_{ii} = 1$}
  end for
end for
for $(i = n; i \geq 1; i = i - 1)$ **do**
  for $(j = i + 1; i \leq n; j = j + 1)$ **do**
    $y - i = y_i - r_{i,j} x_{j,i}$;
    $x_i = y_i / r_{i,i}$;
  end for
end for

Der Aufwand beträgt

- $\frac{2}{3}n^3$ für die LR-Zerlegung und
- $2n^2$ für das Auflösen der beiden Dreieckssysteme.

Dies lohnt sich vor allem, wenn man dasselbe Gleichungssystem zu mehreren rechten Seiten lösen muss.

11.3 Berechnung der Inversen

Es sei $e_j = (0, \ldots, \underbrace{1}_{j\text{-te Komp.}}, \ldots, 0)^T$ der j-te Einheitsvektor.

Für eine beliebige Matrix B gilt dann $Be_j = j$-te *Spalte* von B.

Folglich ergibt das *Lösen* von

$$Aa_j = e_j \qquad \Leftrightarrow \qquad a_j = A^{-1} e_j$$

die j-te Spalte von A^{-1}.

Dies ergibt folgenden Algorithmus zur Bestimmung von A^{-1}:

1. Berechne die LR-Zerlegung von A.
2. Löse $Aa_j = e_j$ für $j = 1, \ldots, n$.
3. Setze A^{-1} spaltenweise aus den a_j zusammen.

Der Aufwand beträgt somit

$$\frac{2}{3}n^3 + n \cdot 2n^2 = \frac{8}{3}n^3$$

Berechnung der Inversen erfordert somit etwa den *vierfachen* Aufwand des Lösens eines linearen Gleichungssystemes.

Es ist also im allgemeinen keine gute Idee, erst A^{-1} auszurechnen, um dann $Ax = b$ mittels $x = A^{-1}b$ zu berechnen.

11.4 Rangbestimmung

Führt das Eliminationsverfahren (oder die LR-Zerlegung zum Schluss auf $a_{n,n} = r_{n,n} \neq 0$, so war die Ausgangsmatrix A regulär und es gilt

$$Rang(A) = n.$$

Wir nehmen hier exakte Arithmetik an (keine Rundungsfehler).

Kann dagegen im Schritt k (also bei der Bestimmung von $a_{k,k}$) mittels *Totalpivotisierung* (wichtig! Spaltenpivotisierung nicht ausreichend!) kein $a_{k,k} \neq 0$ bestimmt werden, so ist

$$\text{Rang}(A) = k - 1 \quad .$$

Bemerkung: Dieser Algorithmus ist sehr empfindlich gegen Rundungsfehler und es gibt bessere für A mit großer Kondition $\kappa(A)$.

11.5 Tridiagonalsysteme

Es gibt viele Algorithmen für lineare Gleichungssysteme mit spezieller Gestalt. Wir betrachten hier Gleichungssysteme mit Tridiagonalgestalt, wie sie bei den kubischen Splines auftraten.

Sei also eine Matrix mit Tridiagonalgestalt gegeben:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & & & & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & & & \\ & \vdots & \vdots & \vdots & & \\ & & & & & a_{n-1,n} \\ 0 & & & & a_{n,n-1} & a_{n,n} \end{bmatrix}.$$

Diese sind ein Spezialfall von Bandmatrizen:

$$a_{i,j} = 0 \qquad \text{für } j < i - m_l \text{ und } j > i + m_r$$

wobei $m = m_l + m_r + 1$ Bandbreite heißt.

Ist die Gauß-Elimination für eine Tridiagonalmatrix *ohne* Pivotisierung durchführbar (z.B. bei Diagonaldominanz), so ergibt sich der folgende einfache Algorithmus (einfach die Nullstruktur beachten):

for $(i = 1; i < n; i = i + 1)$ **do**
  $q = q_{i+1,i}/q_{i,i}$;
  $a_{i+1,i+1} = a_{i+1,i+1} - q \cdot a_{i,i+1}$;
  $b_{i+1} = b_{i+1} - q \cdot b_i$;
end for
$x_n = b_n/a_{n,n}$;
for $(i = n - 1; i \geq 1; i = i - 1)$ **do**
  $x_i = (b_i - a_{i,i+1} \cdot x_{i+1})/a_{i,i}$;

end for

Dieses Verfahren ist auch als „Thomas-Algorithmus“ bekannt.

Der Aufwand beträgt

$$(n-1)\cdot 5 + 1 + (n-1)\cdot 3 = 8(n-1) + 1 = O(n)$$

arithmetische Operationen.

11.6 Zusammenfassung

- Mittels Beispielen wurde motiviert, dass die (Teil-) Pivotisierung notwendig ist, um die Auswirkungen von Rundungsfehlern in der Gauß-Elimination zu vermeiden.
- Trotzdem können sich bei schlecht konditionierten Systemen Rundungsfehler akkumulieren.
- Die LR-Zerlegung wurde hergeleitet. Diese hat denselben Aufwand wie die Gauß-Elimination und wird bei mehreren rechten Seiten vorteilhaft.
- Schließlich haben wir noch kurz speziellere Probleme wie Inversenbildung, Rangbestimmung und Tridiagonalsysteme behandelt.

12 Iterative Lösung linearer Gleichungssysteme

12.1 Dünnbesetzte Matrizen

Wir sind wieder interessiert an der Lösung von

$$Ax = b, \qquad A \in \mathbf{R}^{n\times n}, \quad x, b \in \mathbf{R}^n$$

Definition 12.1 (Dünnbesetzte Matrix)**.** Eine $n \times n$ Matrix A heißt dünn besetzt, wenn sie $O(n)$ statt n^2 Einträgen hat. □

Hierbei denkt man an eine parametrisierte Schar von Matrizen (z.B. Bandmatrizen), sonst macht das $O(n)$ keinen Sinn.

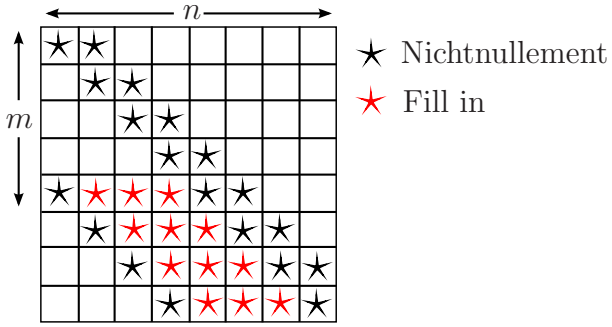
Typisch ist etwa eine konstante Zahl von Einträgen pro Zeile unabhängig von n. Dies tritt etwa bei der Diskretisierung von partiellen Differentialgleichungen auf.

Der Rechenaufwand wird reduziert, wenn man nur mit den Nichtnullelementen rechnet.

Beispiel: Gauß-Elimination für Tridiagonalmatrix → Aufwand $O(n)$.

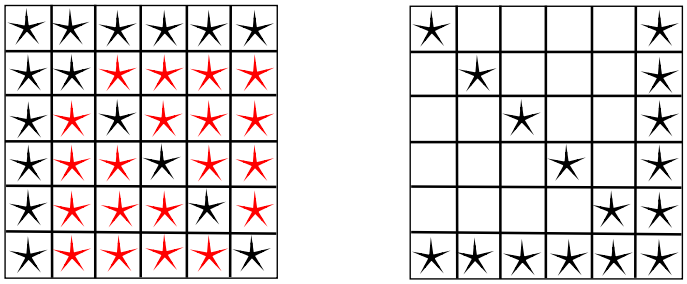
Aber: im Allgemeinen ist das leider nicht so einfach. Es entsteht ein sog. *Fill in*.

Betrachte A mit maximal 3 Elementen pro Zeile:



Im L-Faktor entstehen $O(m \cdot n)$ zusätzliche Einträge.

Extrem ist folgendes Beispiel:



Links entsteht aus der dünnbesetzten Matrix mit cirka $3n$ Einträgen eine *vollbesetzte* Matrix.

Rechts ist sind erste und letzte Zeile sowie erste und letzte Spalte vertauscht worden. Nun entsteht überhaupt kein Fill in!

Wir lernen: Menge des Fill in hängt von der Anordnung ab. Somit kann man auch nach einer optimalen Anordung fragen (führt auf interessantes diskretes Optimierungsproblem).

Eine grundsätzlich andere Idee sind *iterative* Lösungsverfahren.

Ausgehend von einem Startwert $x^{(0)} \in \mathbf{R}^n$ konstruiert man eine Folge

$$x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots$$

mit der Eigenschaft

$$\lim_{k\to\infty} x^{(k)} = x.$$

Vorsicht: k ist der Iterationsindex, keine Potenz! (deswegen Klammern!)

Typischerweise ist der Aufwand zur Berechnung von $x^{(k)}$ nur $O(n)$, entscheidend ist nun die *Anzahl von Iterationen*, die man benötigt, bis die Norm des Fehlers

$$\|x - x^{(k)}\|$$

klein genug ist.

Weiter ist wichtig: Um die Nullstruktur effektiv auszunutzen, benötigt man spezielle Datenstrukturen.

12.2 Relaxationsverfahren

Eine simple Idee zur Konstruktion von Iterationsverfahren ist die folgende.

Beachte die i-te Gleichung in $Ax = b$:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i$$

und löse nach x_i auf:

$$x_i = \frac{1}{a_{ii}} \Big(b_i - \sum_{j \neq i} a_{ij} x_j \Big).$$

Voraussetzung ist $a_{ii} \neq 0 \quad \forall i = 1 \ldots n$. Das geht also nur für bestimmte Matrizen.

Nun bearbeite alle Zeilen *der Reihe nach*:

gegeben $x^{(k)}$
for $(i = 1;\ i \leq n;\ i = i + 1)$ **do**
  $x_i^{(k+1)} = \frac{1}{a_{ii}} \Big(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \Big)$
end for

liefert $x^{(k+1)}$.

Dieses Verfahren heißt Einzelschritt oder Gauß-Seidel[29] Verfahren und gehört zu den Relaxationsverfahren.

Der Aufwand zur Berechnung von $x^{(k+1)}$ aus $x^{(k)}$ ist proportional zur Anzahl der Nichtnullelemente, also $O(n)$.

Es stellen sich die Fragen:

- Unter welchen Bedingungen gilt $\lim_{k\to\infty} x^{(k)} = x$?
- Wie viele Iterationen benötigt man, um

  $$\|x - x^{(k)}\| \leq \varepsilon$$

  zu erreichen für gegebenes ε?
- Wie stellt man (effizient) fest, dass $\|x - x^{(k)}\| \leq \varepsilon$ erreicht ist (x ist unbekannt!) ?

Bevor wir diese Fragen untersuchen, wollen wir noch weitere Relaxationsverfahren angeben.

Jacobi- oder Gesamtschrittverfahren

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\Big)$$

gedämpftes Jacobi-Verfahren Für $\omega > 0$

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\Big)$$

SOR (successive overrelaxation) Verfahren Für $\omega \in (0,2)$

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\Big)$$

gedämpftes Richardson Verfahren Für $\omega > 0$

$$x_i^{(k+1)} = (1-\omega a_{ii})x_i^{(k)} + \omega\Big(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\Big)$$

Dies ist nicht unmittelbar einsichtig, sondern wird weiter unten klar.

[29] Philipp Ludwig von Seidel, 1821-1896, dt. Mathematiker.

12.3 Matrixschreibweise der Relaxationsverfahren

Wir wollen nun die Relaxationsverfahren kompakter schreiben.

Dazu zerlege

$$A = L + D + U$$

mit

$$l_{ij} = \begin{cases} a_{ij} & i > j \\ 0 & sonst \end{cases}, \quad d_{ij} = \begin{cases} a_{ij} & i = j \\ 0 & sonst \end{cases}, \quad u_{ij} = \begin{cases} a_{ij} & i < j \\ 0 & sonst \end{cases}$$

also unteres Dreieck, Diagonale und oberes Dreieck.

Für das gedämpfte Jacobi-Verfahren erhalten wir

$$\begin{aligned} x_i^{(k+1)} &= (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\Big) \qquad \forall i = 1,\dots,n \\ \Leftrightarrow x^{(k+1)} &= (1-\omega)x^{(k)} + \omega D^{-1}\left(b - (L+U)x^{(k)}\right) \\ &= x^{(k)} - \omega D^{-1}Dx^{(k)} + \omega D^{-1}\left(b - (L+U)x^{(k)}\right) \\ &= x^{(k)} - \omega D^{-1}\left(b - Ax^{(k)}\right) \end{aligned}$$

Auch das Gauß-Seidel Verfahren lässt sich auf ähnliche Form bringen:

$$\begin{aligned} x_i^{(k+1)} &= \frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\Big) \qquad \forall i = 1,\dots,n \\ \Leftrightarrow \sum_{j\leq i} a_{ij}x_j^{(k+1)} &= b_i - \sum_{j>i} a_{ij}x_j^{(k)} \qquad \forall i = 1,\dots,n \\ \Leftrightarrow (L+D)x^{(k+1)} &= b - Ux^{(k)} \\ \Leftrightarrow x^{(k+1)} &= (L+D)^{-1}(b - Ux^{(k)}) \\ &= x^{(k)} - (L+D)^{-1}(L+D)x^{(k)} + (L+D)^{-1}(b - Ux^{(k)}) \\ &= x^{(k)} + (L+D)^{-1}(b - Ax^{(k)}). \end{aligned}$$

Die folgende Herleitung zeigt, dass diese Formulierung kein „Zufall“ ist.

Sei

$$e^{(k)} := x - x^{(k)}$$

der *Fehler* in der k-ten Iteration. Wir erhalten aufgrund der Linearität:

$$Ae^{(k)} = A(x - x^{(k)}) = Ax - Ax^{(k)} = b - Ax^{(k)} =: d^{(k)} \tag{12.1}$$

Die Größe $d^{(k)} = b - Ax^{(k)}$ heißt *Defekt* und ist leicht berechenbar.

Aus gegebenem $x^{(k)}$ ließe sich x mittels

$$x = x^{(k)} + e^{(k)} = x^{(k)} + A^{-1}(b - Ax^{(k)})$$

ausberechnen.

Allerdings ist Lösen von $Ae = d$ nicht leichter als $Ax = b$.

Die Idee ist nun A in der Fehlergleichung (12.1) durch eine Matrix M zu ersetzen, sodass

- $M \approx A$, aber
- M leichter invertierbar.

Somit erhält man das Iterationsverfahren

$$x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)}). \tag{12.2}$$

Die Größe $v = M^{-1}(b - Ax^{(k)})$, also die Lösung des Systems

$$Mv = d^{(k)}$$

heißt *Korrektur*.

Alle bisherigen Verfahren lassen sich so schreiben:

$M = \omega^{-1}D$	: gedämpftes Jacobi-Verfahren
$M = L + D$	: Gauß-Seidel
$M = \omega^{-1}I$	: Richardson-Iteration
$M = L + \omega^{-1}D$	: SOR Verfahren

12.4 Konvergenzanalyse

Wir wollen nun überlegen, unter welchen Umständen ein Relaxationsverfahren konvergiert.

Für das allgemeine Iterationsverfahren ergibt sich

$$\begin{aligned}
& x^{(k+1)} = x^{(k)} + M^{-1}\Big(b - Ax^{(k)}\Big) \\
\Leftrightarrow \quad & x - x^{(k+1)} = x - x^{(k)} - M^{-1}\Big(b - Ax^{(k)}\Big) \\
& e^{(k+1)} = e^{(k)} - M^{-1}\Big(Ax - Ax^{(k)}\Big) \\
& \qquad = e^{(k)} - M^{-1}A\Big(x - x^{(k)}\Big) \\
& \qquad = \underbrace{\Big(I - M^{-1}A\Big)}_{=:S} e^{(k)} \quad .
\end{aligned}$$

Es ergibt sich die *Fehlerfortpflanzungsgleichung*

$$e^{(k+1)} = Se^{(k)}$$

mit der *Iterationsmatrix* $S = I - M^{-1}A$.

Rekursives Einsetzen ergibt:

$$e^{(k)} = Se^{(k-1)} = S^2e^{(k-2)} = \ldots = S^ke^{(0)}.$$

Gilt $\lim_{k\to\infty} S^k = 0$ (Nullmatrix), so konvergiert das Verfahren *unabhängig vom Startwert* $x^{(0)}$.

Wegen des linearen Zusammenhanges $e^{(k+1)} = Se^{(k)}$ heißen diese Verfahren auch *lineare Iterationsverfahren.*

Eine allgemeine Auskunft über die Konvergenz gibt der nun folgende Satz.

Satz 12.2. Ein Iterationsverfahren der Form $x^{(k+1)} = x^{(k)} + M^{-1}\left(b - Ax^{(k)}\right)$ konvergiert unabhängig vom Startwert *genau dann wenn* $\varrho(S) < 1$ mit

$$\varrho(S) = \max\{|\lambda| \,|\, \lambda \text{ ist Eigenwert von } S\}$$

dem *Spektralradius* einer Matrix.

Teilbeweis: Ist S diagonalisierbar (n linear unabhängige Eigenvektoren), also $S = TDT^{-1}$, so gilt

$$S^k = TDT^{-1}TDT^{-1}\ldots TDT^{-1} = TD^kT^{-1} \qquad \text{mit}$$

$$D^k = \begin{pmatrix} \lambda_1^k & & & \\ & \lambda_2^k & & \\ & & \ddots & \\ & & & \lambda_n^k \end{pmatrix} \to 0 \quad \Leftrightarrow \quad |\lambda_i| < 1 \quad \forall i = 1 \ldots n.$$

Für den allgemeinen Fall sei auf [Ran06, Satz 6.1] verwiesen. □

Konkret erfordert die Anwendung dieses Satzes also Aussagen über die Eigenwerte von $S = I - M^{-1}A$. Dies ist im allgemeinen nicht einfach.

Relativ leicht ist die Richardson-Iteration für symmetrisch positiv definite Matrizen zu analysieren.

Satz 12.3. Sei A symmetrisch und positiv definit, dann konvergiert die gedämpfte Richardson-Iteration für genügend kleines $\omega > 0$.

Beweis: Aus A s.p.d. folgt alle Eigenwerte sind reell und positiv, also gilt für das *Spektrum* von A

$$\sigma(A) = \{\lambda_{\min}(A) = \lambda_1, \lambda_2, \ldots, \lambda_n = \lambda_{\max}(A)\}$$

mit

$$0 < \lambda_i \leq \lambda_{i+1} \qquad \forall i = 1, \ldots, n-1.$$

Das gedämpfte Richardson-Verfahren, $M = \omega^{-1}I$, hat die Iterationsmatrix

$$S_\omega = I - M^{-1}A = I - \omega A$$

und diese hat das Spektrum

$$\sigma(S_\omega) = \{\mu_i \,|\, \mu_i = 1 - \omega\lambda_i \text{ mit } \lambda_i \in \sigma(A)\}.$$

Wählt man jetzt speziell $\omega = \frac{1}{\lambda_{\max}(A)}$ so ergibt sich

$$0 = 1 - \frac{\lambda_{\max}}{\lambda_{\max}} \leq \mu_i \leq 1 - \frac{\lambda_{\min}}{\lambda_{\max}} = 1 - \frac{1}{\kappa_2(A)} \qquad \kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \quad .$$
„Spektralkondition“

Wegen $\kappa_2(A) \geq 1$ gilt $\varrho(S) < 1$. □

Bemerkung 12.4. Für typische Anwendungen, etwa bei der numerischen Lösung partieller Differentialgleichungen, werden die Matrizen *sehr* groß und die spektrale Kondition steigt mit n an.

So gilt etwa bei Lösung der Laplacegleichung in $\Omega \subset \mathbf{R}^d$ mit „Finiten Differenzen“

$$\kappa_2(A) = O\left(n^{2/d}\right).$$

Damit konvergiert das Verfahren umso schlechter je größer das Problem ist. □

Bemerkung 12.5. Zur praktischen Anwendung der Richardson-Iteration benötigt man eine Schätzung für $\lambda_{\max}(A)$. Man kann zeigen:

$$\lambda_{\max}(A) \leq \max_{i=1,\ldots,n} \Big(a_{ii} + \sum_{j \neq i} |a_{ij}| \Big).$$

(Satz von Gerschgorin). □

12.5 Diagonaldominante Matrizen

Wir geben nun ein weiteres Konvergenzresultat für das Jacobi-, bzw. Gauß-Seidel Verfahren an.

Dieses Resultat zeigt auch, dass die Symmetrie keine notwendige Vorrausetzung ist.

Satz 12.6. Erfüllt die Matrix A die *starke Zeilensummenbedingung*

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}| \qquad \forall i = 1, \ldots, n$$

so konvergieren sowohl das Jacobi als auch das Gauß-Seidel Verfahren.

Beweis: Aus $e^{(k+1)} = Se^{(k)}$ folgt durch Bilden der Norm

$$\|e^{(k+1)}\|_\infty = \|Se^{(k)}\|_\infty \leq \|S\|_\infty \|e^{(k)}\|_\infty$$

für jede verträgliche Matrixnorm. Wir verwenden hier die Maximumnorm mit Zeilensummennorm als zugeordneter Matrixnorm, siehe Definition 10.4.

Wir zeigen nun $\|S\|_\infty < 1$, woraus unmittelbar die Konvergenz folgt.

Jacobi-Verfahren: Es gilt

$$S = I - D^{-1}A = I - D^{-1}(L + D + U) = -D^{-1}(L + U)$$

und damit

$$\|S\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |s_{ij}| = \max_{i=1,\dots,n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}| < 1.$$

Gauß-Seidel: Für S gilt die Darstellung

$$\begin{aligned} & S = I - (L+D)^{-1}A \\ \Leftrightarrow \quad & (L+D)S = (L+D) - A = L + D - (L+D+U) = -U \\ \Leftrightarrow \quad & DS = -U - LS \\ \Leftrightarrow \quad & S = -D^{-1}(U + LS) \end{aligned}$$

Komponentenweise heißt das für Sx:

$$(Sx)_i = \sum_{j=1}^{n} s_{ij}x_j = -\frac{1}{a_{ii}}\Big(\sum_{j>i} a_{ij}x_j + \sum_{j<i} a_{ij}(Sx)_j\Big) \qquad \text{Rekursion für } (Sx)_i.$$

Betrag bilden und Dreiecksungleichung ergibt:

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|}\Big(\sum_{j>i} |a_{ij}||x_j| + \sum_{j<i} |a_{ij}||(Sx)_j|\Big).$$

Per Induktion zeigen wir nun :

$$|(Sx)_i| < \|x\|_\infty \qquad \forall i = 1, \dots, n.$$

Sei $i = 1$. Dann haben wir

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|}\Big(\sum_{j>i} |a_{ij}||x_j|\Big) \leq \|x\|_\infty \frac{1}{|a_{ii}|} \sum_{j>i} |a_{ij}| < \|x\|_\infty.$$

Bis $i-1$ sei die Annahme bewiesen. Für die i-te Zeile gilt dann

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|}\Big(\sum_{j>i} |a_{ij}| \underbrace{|x_j|}_{\leq \|x\|_\infty} + \sum_{j<i} |a_{ij}| \underbrace{|(Sx)_j|}_{\leq \|x\|_\infty}\Big) \leq \|x\|_\infty \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}| < \|x\|_\infty.$$

Damit haben wir $\|Sx\|_\infty < \|x\|_\infty$ gezeigt.

Nun setzen wir das in die Definition der Matrixnorm ein:

$$\|S\|_\infty = \sup_{x\neq 0} \frac{\|Sx\|_\infty}{\|x\|_\infty} < 1.$$

Das Resultat kann unter gewissen zusätzlichen Voraussetzungen (Irreduzibilität) auf den Fall $\sum_{j\neq i} |a_{ij}| \leq |a_{ii}|$ verallgemeinert werden (schwach diagonaldominante Matrizen). □

Wir merken uns:

- Iterationsverfahren konvergieren nur für bestimmte Klassen von Matrizen.
- Für ingenieurrelevante Probleme sind oft keine Konvergenzaussagen möglich.

12.6 Praktische Realisierung

Wir brauchen noch ein Kriterium, wann die Iteration abgebrochen werden kann.

Aus $Ae^{(k)} = d^{(k)} \Leftrightarrow e^{(k)} = A^{-1}d^{(k)}$ erhalten wir

$$\|e^{(k)}\| \leq \|A^{-1}\|\|d^{(k)}\| \qquad \text{(für jede verträgliche Matrixnorm)}$$

Es liegt nahe, den Defekt $d^{(k)} = b - Ax^{(k)}$ als Abbruchkriterium heranzuziehen.

Wegen $\|b\| = \|Ax\| \leq \|A\|\|x\| \Leftrightarrow \|x\| \geq \frac{\|b\|}{\|A\|}$ erhalten wir außerdem für den interessanteren relativen Fehler:

$$\frac{\|e^{(k)}\|}{\|x\|} \leq \frac{\|A^{-1}\|\|d^{(k)}\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|d^{(k)}\|}{\|b\|} = \kappa(A)\frac{\|d^{(k)}\|}{\|b\|}.$$

Bei großer Konditionszahl kann deshalb der Fehler trotz kleinem $\|d^{(k)}\|$ groß sein. Eine Schätzung für $\kappa(A)$ ist außerdem schwer erhältlich.

In der Praxis verwendet man häufig eine relative Abbruchbedingung der Form

$$\|d^{(k)}\| < \varepsilon\|d^{(0)}\|$$

wobei $d^{(0)}$ der Defekt zum Startwert $x^{(0)}$ ist.

Mit einem geeigneten ε erhalten wir dann folgenden Algorithmus:

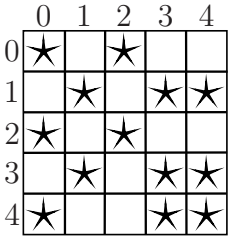
Gegeben seien x, b;
Berechne $d = b - Ax$; (Anfangsdefekt)
Setze $d0 = \|d\|$;
while ($\|d\| \geq \varepsilon \cdot d0$) **do**
  Löse $Mv = d$;
  Setze $x = x + v$;
  Setze $d = d - Av$;
end while

Diese Version vermeidet Rundungsfehler in der Berechnung des Defektes.

12.7 Datenstrukturen für dünnbesetzte Matrizen

Wie nutzt man die Nullstruktur der Matrix A nun effektiv aus?

A kann *nicht* mehr als zweidimensionales Feld gespeichert werden. Eine sehr beliebte Datenstruktur ist „compressed row storage“ (CRS).



Sei $A \in \mathbf{R}^{n\times n}$ und $m = \#$ Nichtnullelemente von A.
Feld $a[m]$ enthält *zeilenweise* alle Nichtnullelemente

$$a[m] = \boxed{a_{00}\mid a_{02}\mid a_{11}\mid a_{13}\mid a_{14}\mid a_{20}\mid a_{22}\mid \ldots \mid a_{44}}$$

Feld $j[m]$ enthält zeilenweise jeweils die zugehörigen Spaltenindizes

$$j[m] = \boxed{0\mid 2\mid 1\mid 3\mid 4\mid 0\mid 2\mid \ldots \mid 4}$$

Feld $r[n+1]$ enthält die Startindizes für die jede Zeile

$$r[n+1] = \boxed{0\mid 2\mid 5\mid 7\mid 10\mid 13}$$

Damit programmiert man die Matrix-Vektor-Multiplikation $y = Ax$ als:

for $(i = 0;\ i < n;\ i = i + 1)$ **do**
  $y[i] = 0;$
  for $(k = r[i];\ k < r[i+1];\ k = k + 1)$ **do**
    $y[i] = y[i] + a[k] \cdot x[j[k]];$
  end for
end for

12.8 Abstiegsverfahren

Wir kommen nun zu einer weiteren Klasse von Iterationsverfahren zur Lösung von linearen Gleichungssystemen, den sogenannten Abstiegsverfahren.

Diese formulieren die Lösung des linearen Gleichungssystems

$$Ax = b$$

als Minimierungsaufgabe um:

Satz 12.7. Sei A eine symmetrisch positiv definite, $n \times n$ Matrix, dann nimmt das Funktional $F : \mathbf{R}^n \to \mathbf{R}$

$$F(x) = \frac{1}{2}x^T Ax - b^T x$$

sein eindeutiges Minimum in $x^* = A^{-1}b$ an.

Beweis: Für ein beliebiges x setze $x = x^* + v$. Dann gilt:

$$\begin{aligned}
F(x) &= \frac{1}{2}(x^* + v)^T A(x^* + v) - b^T(x^* + v) \\
&= \frac{1}{2}\left[(x^*)^T Ax^* + \underbrace{(x^*)^T Av + v^T Ax^*}_{2v^T Ax^*} + v^T Av\right] - b^T x^* - b^T v \\
&= \underbrace{\frac{1}{2}(x^*)^T Ax^* - b^T x^*}_{} + v^T \underbrace{(Ax^* - b)}_{=0} + \frac{1}{2}v^T Av \\
&= F(x^*) + \frac{1}{2}v^T Av \quad .
\end{aligned}$$

Da A s.p.d. ist $v^T Av > 0$ für alle $v \neq 0$ und es gilt $F(x) > F(x^*)$ für alle $x \neq x^*$, also ist x^* ein Minimum von F.

Eindeutigkeit: Sei x' weiteres Minimum, dann gilt für $x' = x^* + (x' - x^*)$

$$F(x') = F(x^*) + \frac{1}{2}(x' - x^*)^T A(x' - x^*) > F(x^*)$$

und somit Widerspruch zur Annahme, dass x' ein Minimum ist. □

Diese Charakterisierung nutzt man nun folgendermaßen aus.

Sei $p^{(k)} \in \mathbf{R}^n$, $p^{(k)} \neq 0$ ein beliebiger Vektor, eine sog. „Suchrichtung“, dann minimiere F entlang der Geraden

$$x^{(k)} + \alpha p^{(k)},$$

das heißt

$$\text{finde } \alpha^{(k)}, \text{ sodass } F(x^{(k)} + \alpha p^{(k)}) \text{ mininimal wird.}$$

Diese eindimensionale Minimierungsaufgabe kann man einfach lösen:

$$F(x^{(k)} + \underbrace{\alpha p^{(k)}}_{=v}) = F(x^{(k)}) + \alpha (p^{(k)})^T (\underbrace{Ax^{(k)} - b}_{-d^{(k)}}) + \frac{\alpha^2}{2}(p^{(k)})^T A p^{(k)}$$

(folgt aus dem Beweis oben) und damit

$$\frac{d}{d\alpha} F(x^{(k)} + \alpha p^{(k)}) = (p^{(k)})^T (Ax^{(k)} - b) + \alpha (p^{(k)})^T A p^{(k)} \stackrel{!}{=} 0$$

$$\Leftrightarrow \qquad \alpha^{(k)} = \frac{(p^{(k)})^T (b - Ax^{(k)})}{(p^{(k)})^T A p^{(k)}}$$

Wie wählt man nun die Suchrichtung $p^{(k)}$ im konkreten Fall?

Man erinnere sich: Der Gradient $\nabla F(x_0)$ einer Funktion $F : \mathbf{R}^n \to \mathbf{R}$ im Punkt x_0 ist ein Vektor, der senkrecht auf der Niveaulinie $\{x \in \mathbf{R}^n \mid F(x) = F(x_0)\}$ steht und in Richtung des größten Anstiegs von F zeigt.

Methode des steilsten Abstiegs: Wähle die negative Gradientenrichtung, also

$$p^{(k)} = -\nabla F(x^{(k)}) = - \begin{pmatrix} \frac{\partial F}{\partial x_1}(x^{(k)}) \\ \vdots \\ \frac{\partial F}{\partial x_n}(x^{(k)}) \end{pmatrix}.$$

Man rechnet für das Funktional F nach:

$$-\nabla F(x^{(k)}) = b - Ax^{(k)} \quad \text{der Defekt!}$$

In algorithmischer Form lautet das Gradientenverfahren wie folgt:

Gegeben $x^{(0)}$;
Berechne $d^{(0)} = b - Ax^{(0)}$;
for $(k = 0, 1, \ldots)$ **do**
  $q = Ad^{(k)}$;
  $\alpha^{(k)} = \frac{(d^{(k)})^T d^{(k)}}{(d^{(k)})^T q}$;
  $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$;
  $d^{(k+1)} = d^{(k)} - \alpha^{(k)} q$;
end for

Der Aufwand pro Iteration ist im wesentlichen ein Matrix-Vektor-Produkt.

Für die Konvergenz des Verfahrens kann man zeigen:

$$\|x - x^{(k+1)}\|_A \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \|x - x^{(k)}\|_A$$

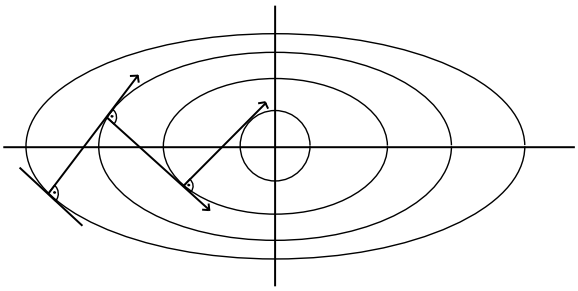
(wobei $\|x\|_A = \sqrt{x^T Ax}$ die Energienorm ist).

Das Gradientenverfahren konvergiert nicht besser als das Gauß-Seidel-Verfahren. Die Konvergenzprobleme macht das folgende Beispiel anschaulich.

Beispiel 12.8 (Zum Gradientenverfahren)**.** Betrachte

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow F(x) = x_1^2 + \frac{1}{2} x_2^2$$

mit Minimum in $(0, 0)^T$. Die Höhenlinien von F sind Ellipsen.



Es ist $p^{(k+1)} \perp p^{(k)}$ aber $p^{(k+2)}$ beinahe parallel zu $p^{(k)} \rightarrow$. Dieser Effekt wird umso stärker je exzentrischer die Ellipsen sind, d.h. je unterschiedlicher die Eigenwerte von A. □

Es gibt Verbesserungen des Gradientenverfahrens, z. B. das Verfahren der konjugierten Gradienten, die diesen Effekt vermeiden.

12.9 Zusammenfassung

- Dünnbesetzte Matrizen sind solche, die nur $O(n)$ Nichtnullelemente haben, wenn $n \times n$ die Dimension der Matrix ist.
- Direkte Lösungsverfahren wie die LR-Zerlegung führen bei dünnbesetzten Matrizen oft zu einem Fill in der Matrix und damit zu unvertretbar hohem Aufwand.

- Iterative Verfahren eignen sich für solche Matrizen besser, da der Aufwand pro Schritt typischerweise nur $O(n)$ ist. Allerdings ist die Konvergenz nur für gewisse Klassen von Matrizen gewährleistet.
- Wir haben zwei Klassen von Iterationsverfahren kennengelernt, die Relaxationsverfahren und die Abstiegsverfahren.

13 Lösung nichtlinearer Gleichungssysteme

13.1 Aufgabenstellung

Auch wenn lineare Modell so bequem und einfach zu lösen sind: Die Welt ist nichtlinear!

Seit $f : I = [a, b] \to \mathbf{R}$ eine stetige Funktion.

Wir interessieren uns für die Lösung der Aufgabe

$$\text{Finde } x \in [a, b] \quad : \quad f(x) = 0,$$

wir suchen also eine „Nullstelle“ einer Funktion.

Ein Beispiel für eine solche Aufgabe hatten wir schon: Die Stützstellen bei der Gauss-Quadratur sind die Nullstellen der Legendrepolynome.

In der Praxis tritt diese Aufgabe häufig in höherdimensionalen Räumen auf, also

$$\text{Finde } x_1, \dots x_n \quad : \quad f_i(x_1, \dots x_n) = 0 \qquad \forall i = 1, \dots, n.$$

Wenn man die Komponenten f_i zu einer vektorwertigen Funktion $\underline{f} : \mathbf{R}^n \to \mathbf{R}$ zusammenfasst, schreibt sich das kurz als

$$\underline{f}(\underline{x}) = 0.$$

Zunächst beschränken wir uns aber auf $n = 1$.

13.2 Intervallschachtelung (Bisektion)

Als erste Methode zur Lösung nichtlinearer Gleichungen betrachten wir die Bisektion. Diese ist sehr ähnlich zur binären Suche.

Idee: Angenommen es existiert ein Teilintervall $I_0 = [a_0, b_0]$, sodass $f(a_0), f(b_0)$ verschiedenes Vorzeichen haben, also $f(a_0) \cdot f(b_0) < 0$. So hat wegen dem Zwischenwertsatz (für stetige Funktionen) f mindestens eine Nullstelle in $[a_0, b_0]$.

Dies führt zu folgendem Algorithmus:

Gegeben: $I_0 = [a_0, b_0]$ mit $f(a_0) \cdot f(b_0) < 0$ und Toleranz ε;
for $(t = 0, 1, \dots)$ **do**
  $x_t = \frac{1}{2}(a_t + b_t)$; {Mittelpunkt des Intervalls}
  if $(f(x_t) = 0)$ **then**
    break; {fertig!}
  end if
  if $(f(a_t)f(x_t) < 0)$ **then**
    $a_{t+1} = a_t$; $b_{t+1} = x_t$; {Nullstelle in $[a_t, x_t]$}
  else
    $a_{t+1} = x_t$; $b_{t+1} = b_t$; $\{f(x_t)f(b_t) < 0$ da $VZ(x_t) = VZ(a_t)!\}$
  end if

```
    if (b_t − a_t < ε) then
      break; {Fehler ist akzeptabel}
    end if
  end for
```

Nun zur Analyse des Verfahrens.

In jedem Schritt gilt

$$a_t \leq a_{t+1} < b_{t+1} \leq b_t$$

und

$$|b_{t+1} - a_{t+1}| = \frac{1}{2}|b_t - a_t| = \left(\frac{1}{2}\right)^{t+1} |b_0 - a_0|.$$

Wir halten folgende Eigenschaften fest:

- Die Konvergenzrate ist $\frac{1}{2}$ pro Schritt.
- Die Bisektion ist numerisch sehr stabil (unanfällig gegen Rundungsfehler) und insbesondere bei monotonen Funktionen die Methode der Wahl.
- Ein Nachteil ist, dass die Methode nur für reelle Funktionen (also etwa nicht für komplexwertige) anwendbar ist.

13.3 Fixpunktiteration

Wir geben nun ein weiteres Verfahren an, welches die Nullstellensuche in eine Fixpunktsuche umformuliert.

Zu gegebenem $f : I \to \mathbf{R}$ betrachte die Hilfsfunktion

$$g(x) = x + \sigma f(x) \qquad \text{mit } 0 \neq \sigma \in \mathbf{R}.$$

Offensichtlich gilt

$$\begin{aligned} g(x) = x \quad &\Leftrightarrow \quad x + \sigma f(x) = x \\ &\Leftrightarrow \quad \sigma f(x) = 0 \\ &\Leftrightarrow \quad f(x) = 0\ . \end{aligned}$$

Die Suche nach Nullstellen von f ist also äquivalent zur Suche nach Fixpunkten

$$g(x) = x$$

von g.

Diese Suche nach Fixpunkten untersucht der folgende Satz.

Satz 13.1 (Banachscher[30] Fixpunktsatz)**.** Es sei $I \subset \mathbf{R}$ ein nichtleeres, abgeschlossenes Intervall und $g : I \to I$ eine „Lipschitz[31]-stetige“ Abbildung

$$|g(x) - g(y)| \leq q|x - y| \qquad x, y \in I$$

mit $q < 1$ (Kontraktion). Dann konvergiert die durch

$$x^{(t+1)} = g(x^{(t)})$$

generierte Folge für beliebige Startwerte gegen den eindeutigen Fixpunkt $z \in I$.

Für den Fehler gilt:

$$|x^{(t)} - z| \leq \frac{q}{1-q}|x^{(t)} - x^{(t-1)}| \leq \frac{q^t}{1-q}|x^{(1)} - x^{(0)}|.$$

Beweis: Da $g : I \to I$ ist $x^{(t)} = g(x^{(t-1)}) = g(g(x^{(t-2)})) = \ldots g^t(x^{(0)})$ wohldefiniert.

Weiter gilt:

$$|x^{(t+1)} - x^{(t)}| = |g(x^{(t)}) - g(x^{(t-1)})| \leq q|x^{(t)} - x^{(t+1)}| \leq \ldots \leq q^t|x^{(1)} - x^{(0)}|$$

Wir zeigen nun, dass die $x^{(t)}$ eine Cauchy-Folge bilden. Seien $\varepsilon > 0$ und $m \geq 1$ gegeben

$$\begin{aligned}
|x^{(t+m)} - x^{(t)}| &\leq |x^{(t+m)} - x^{(t+m-1)} + x^{(t+m-1)} - x^{(t+m-2)} + \ldots + x^{(t+1)} - x^{(t)}| \\
&\leq |x^{(t+m)} - x^{(t+m-1)}| + |x^{(t+m-1)} - x^{(t+m-2)}| + \ldots + |x^{(t+1)} - x^{(t)}| \\
&\leq q^{t+m-1}|x^{(1)} - x^{(0)}| + q^{t+m-2}|x^{(1)} - x^{(0)}| + \ldots + q^t|x^{(1)} - x^{(0)}| \\
&\leq (q^{t+m-1} + q^{t+m-2} + \ldots + q^t)|x^{(1)} - x^{(0)}| \\
&\leq q^t \frac{1-q^m}{1-q}|x^{(1)} - x^{(0)}| \leq \varepsilon \qquad \text{für } t \geq t(\varepsilon) \text{ groß genug.}
\end{aligned}$$

Wegen dem Vollständigkeitsaxiom konvergiert jede Cauchy-Folge gegen einen Grenzwert $z \in \mathbf{R}$.

Wegen $g : I \to I$ und I abgeschlossen gilt $z \in I$.

Fehlerabschätzung:

$$\begin{aligned}
|x^{(t+m)} - x^{(t)}| &\leq |x^{(t+m)} - x^{(t+m-1)}| + \ldots + |x^{(t+1)} + x^{(t)}| \quad \text{(wie oben)} \\
&\leq q^m|x^{(t)} - x^{(t-1)}| + \ldots + q|x^{(t)} - x^{(t-1)}| \\
&\leq (q^m + \ldots + q)|x^{(t)} - x^{(t-1)}| \\
&\leq \frac{q}{1-q}|x^{(t)} - x^{(t-1)}|
\end{aligned}$$

Für $m \to \infty$ konvergiert $x^{(t+m)}$ gegen z, die rechte Seite ist unabhängig von m, also folgt

$$|z - x^{(t)}| \leq \frac{q}{1-q}|x^{(t)} - x^{(t-1)}| \leq \frac{q^t}{1-q}|x^{(1)} - x^{(0)}|.$$

[30] Stefan Banach, 1892-1945, polnischer Mathematiker.
[31] Rudolf O. S. Lipschitz, 1832-1903, dt. Mathematiker.

Eindeutigkeit: Sei $z' \neq z$ ein weiterer Fixpunkt so gilt

$$|z - z'| = |g(z) - g(z')| \leq q|z - z'| \quad \Leftrightarrow \quad 1 \leq q \quad (z - z' \neq 0).$$

Dies ist ein Widerspruch zu $q < 1$ (Lipschitz). Also ist $z = z'$. □

Bemerkung 13.2. Ein hinreichende Bedingung für die Lipschitz-Stetigkeit von g ist $|g'(x)| \leq q$ für alle $x \in I$.

Aus dem Mittelwertsatz der Differentialrechnung folgern wir:

$$\frac{g(x) - g(y)}{x - y} = g'(\xi) \Leftrightarrow g(x) - g(y) = g'(\xi)(x - y)$$
$$\Rightarrow |g(x) - g(y)| = |g'(\xi)||x - y|$$

und somit die Lipschitzstetigkeit falls $g'(x) \leq q$ für alle $x \in I$.

Ist ausserdem $q < 1$ so hat man die Kontraktionseigenschaft. □

Bemerkung 13.3. $|g'(x)| \leq q$ ist nur eine hinreichende Bedingung für die Lipschitz-Stetigkeit.

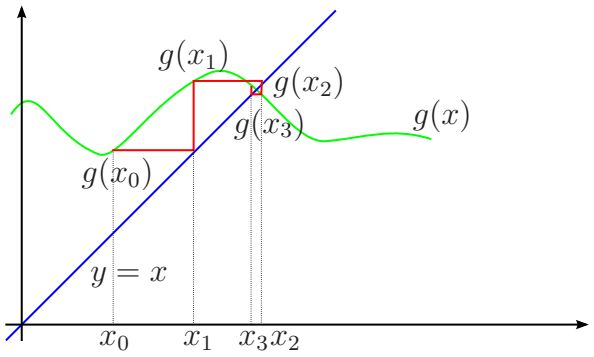
So haben wir etwa für die Funktion $|x|$:

$$| \, |x| - |y| \, | \leq |x - y|$$

also Lipschitz-Stetigkeit mit Konstante 1.

Es ist gerade die Stärke des Banachschen Fixpunktsatzes, dass die Differenzierbarkeit der Iterationsfunktion g *nicht* erforderlich ist. □

Geometrische Interpretation der Fixpunktiteration



Bemerkung 13.4. Der Banachsche Fixpunktsatz kann auf Funktionen $g : G \to \mathbf{R}^n$, $G \subseteq \mathbf{R}^n$, erweitert werden. Entsprechend ist wieder

$$\|g(x) - g(y)\| \leq q\|x - y\|, \qquad x, y \in G$$

mit einem $q < 1$ zu fordern.

Oben haben wir die iterative Lösung von $Ax = b$ untersucht. Dies entspricht einer Nullstellensuche $f(x) = b - Ax = 0$.

Die linearen Iterationsverfahren lauteten

$$x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)}) = \underbrace{(I - M^{-1}A)}_{S} x^{(k)} + \underbrace{M^{-1}b}_{c} = g(x^{(k)}).$$

Untersuchen wir die Lipschitz-Stetigkeit von g:

$$\|g(x) - g(y)\| = \|Sx - Sy\| = \|S(x - y)\| \leq \|S\|\|x - y\|.$$

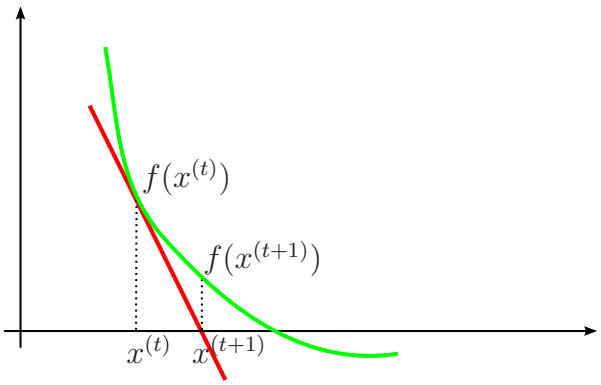
Für $\|S\| < 1$ erhalten wir Konvergenz unabhängig vom Startwert. □

13.4 Newton-Verfahren

Wir kehren zurück zur Nullstellensuche $f(x) = 0$.

Für das Newton-Verfahren wollen wir mit der geometrischen Idee beginnen.

Am aktuellen Punkt $x^{(t)}$ ersetze die Funktion f durch ihre Tangente und berechne deren Nullstelle. Das ist $x^{(t+1)}$.



Formal lautet die Gleichung für die Tangente im Punkt $x^{(t)}$

$$T(x) = f'(x^{(t)})(x - x^{(t)}) + f(x^{(t)}).$$

Die Nullstelle der Tangente erhalten wir mittels

$$\begin{aligned} T(x^{(t)}) = 0 \quad &\Leftrightarrow \quad f'(x^{(t)})(x^{(t+1)} - x^{(t)}) + f(x^{(t)}) = 0 \\ &\Leftrightarrow \quad x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}. \end{aligned}$$

Voraussetzung ist natürlich, dass $f'(x^{(t)}) \neq 0$, d. h. es liegt insbesondere eine *einfache* Nullstelle im Punkt x vor.

Die Konvergenzeigenschaften des Newton-Verfahrens beschreibt der folgende Satz.

Satz 13.5. Die Funktion $f \in C^2[a,b]$ habe in (a,b) eine Nullstelle z und es sei

$$m := \min_{a\le x\le b} |f'(x)| > 0, \qquad M := \max_{a\le x\le b} |f''(x)|.$$

Sei $\varrho > 0$ so gewählt, dass

$$q = \frac{M}{2m}\varrho < 1, \quad K_\varrho(z) = \{x \in \mathbf{R} \,|\, |x-z| \le \varrho\} \subset [a,b].$$

Dann sind für jeden Startwert $x^{(0)} \in K_\varrho(z)$ die Newton-Iterierten $x^{(t)} \in K_\varrho(z)$ definiert und es gelten die Abschätzungen

$$|x^{(t)} - z| \le \frac{2m}{M} q^{(2^t)} \quad \text{bzw.} \quad |x^{(t)} - z| \le \frac{M}{2m}|x^{(t)} - x^{(t-1)}|^2.$$

Beweis: Siehe [Ran06, Satz 5.1]. □

Bemerkung 13.6.

- Das Newton-Verfahren konvergiert „quadratisch“:

  $$|x^{(t)} - z| \le C|x^{(t)} - x^{(t-1)}|^2, \qquad |x^{(t)} - z| \le q^{2^t}.$$

  Bisektion und Fixpunktiteration konvergieren nur „linear“:

  $$|x^{(t)} - z| \le C|x^{(t)} - x^{(t-1)}|, \qquad |x^{(t)} - z| \le Cq^t.$$

  Z.B. $C = 1, q = 0.1$:

t	linear	quadratisch
1	10^{-1}	10^{-1}
2	10^{-2}	10^{-2}
3	10^{-3}	10^{-4}
4	10^{-4}	10^{-8}

  Bei linearer Konvergenz ist die Zahl der gültigen Ziffern proportional zu t, bei quadratischer Konvergenz verdoppelt sich die Zahl der gültigen Ziffern asymptotisch in jedem Schritt!

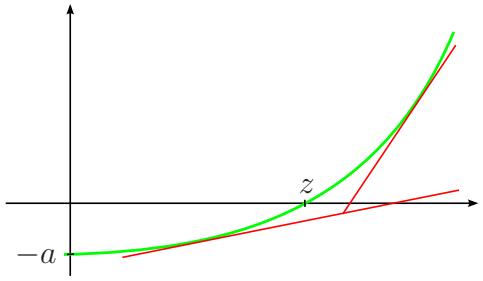
- Der Nachteil des Newton-Verfahrens ist die *lokale* Konvergenz, d. h. der Startwert muss hinreichend nahe an der Lösung liegen. □

Beispiel 13.7. Wir betrachten die Wurzelberechnung, d. h. die Lösung von

$$f(x) = x^n - a = 0 \qquad \text{für } a > 0.$$

Das Newton-Verfahren lautet

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} = x^{(t)} - \frac{(x^{(t)})^n - a}{n(x^{(t)})^{n-1}}.$$



Konvergiert für jedes $x^{(0)} > 0$ gegen die positive Wurzel, denn falls $x^{(0)} < z$ gilt $x^{(1)} > z$ und für $x^{(t)} > z$ fällt die Folge *monoton*.

Für den Fall $n = 2$ (d.h. $x^2 - a = 0$) ergibt sich quadratische Konvergenz falls

$$|x^{(t)} - \sqrt{a}| < 2\sqrt{a}.$$

□

Wir behandeln nun noch einige Varianten des Newton-Verfahrens.

Das sog. *gedämpfte Newton-Verfahren* lautet

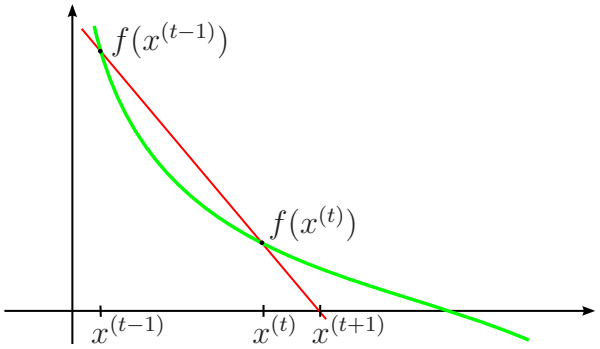
$$x^{(t+1)} = x^{(t)} - \lambda^{(t)} \frac{f(x^{(t)})}{f'(x^{(t)})}$$

mit $\lambda^{(t)} \in (0, 1]$.

Man addiert also nicht die volle Korrektur sondern „dämpft“ diese mit dem Faktor $\lambda^{(t)}$.

Bei geeigneter Wahl von $\lambda^{(t)}$ kann man den „Konvergenzbereich“ des Newton-Verfahrens vergrößern.

In der sog. *Sekantenmethode* vermeidet man die Berechnung von Ableitungen durch Verwendung zweier aufeinanderfolgender Punkte:



Man ersetzt also die Tangente durch die Sekante:

$$T(x) = \frac{f(x^{(t)}) - f(x^{(t-1)})}{x^{(t)} - x^{(t-1)}}(x - x^{(t)}) + f(x^{(t)})$$

$$\Rightarrow \qquad x^{(t+1)} = x^{(t)} - f(x^{(t)}) \frac{x^{(t)} - x^{(t-1)}}{f(x^{(t)}) - f(x^{(t-1)})}.$$

Für die Konvergenz der Sekantenmethode kann man zeigen:

$$|x^{(t)} - z| \leq \frac{2m}{M} q^{\gamma_t} \qquad \gamma_t\text{: Fibonacci-Zahlen } \gamma_0 = \gamma_1 = 1,\ \gamma_{t+1} = \gamma_t + \gamma_{t-1}.$$

Dies entspricht einer Konvergenzordnung

$$|x^{(t)} - z| \leq C|x^{(t)} - x^{(t-1)}|^s, \quad \text{mit } s = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618 \text{ „Goldener Schnitt“}.$$

Ein Problem der Sekantenmethode ist die Empfindlichkeit gegenüber Auslöschung.

Weitere Alternative: Berechne $f'(x^{(t)})$ im Newton-Verfahren durch numerische Differentiation:

- Aufwendiger als Sekantenmethode, aber quadratische Konvergenz.
- Auch empfindlich gegen Auslöschung.

Falls $f'(x)$ nicht exakt berechnet wird, spricht man oft von *Quasi*-Newton-Verfahren.

13.5 Newton-Verfahren im $\mathbf{R}^n$

Wir wenden uns nun der Lösung von

$$\begin{aligned} & f_i(x_1, \ldots, x_n) = 0 \qquad i = 1, \ldots, n \\ \Leftrightarrow \qquad & \underline{f}(\underline{x}) = 0 \text{ mit } \underline{x} = (x_1, \ldots, x_n)^T \text{ und } \underline{f} : \mathbf{R}^n \to \mathbf{R}^n \end{aligned}$$

zu.

Taylorreihe im $\mathbf{R}^n$ liefert Verallgemeinerung der Tangente:

$$\underline{f}(\underline{x} + \underline{\Delta x}) = \underline{f}(\underline{x}) + \underline{J}(\underline{x})\underline{\Delta x} + \quad \text{Restglied.}$$

Hierbei ist $\underline{J}(\underline{x})$ die *Jacobimatrix* an der Stelle $\underline{x}$:

$$(\underline{J}(\underline{x}))_{i,j} = \frac{\partial f_i}{\partial x_j}(\underline{x}) \quad \in \mathbf{R}^{n \times n}.$$

Nullstelle der „Tangente“ liefert:

$$\begin{aligned} & \underline{f}(\underline{x}^{(t)}) + \underline{J}(\underline{x}^{(t)})(\underline{x}^{(t+1)} - \underline{x}^{(t)}) \stackrel{!}{=} 0 \\ & \qquad \Leftrightarrow \quad x^{(t+1)} = \underline{x}^{(t)} - (\underline{J}(\underline{x}))^{-1} \underline{f}(\underline{x}^{(t)}). \end{aligned}$$

Jeder Schritt des Newton-Verfahrens erfordert das Lösen eines linearen Gleichungssystems

$$\underline{J}(\underline{x}^{(t+1)})v = \underline{f}(\underline{x}^{(t)}).$$

Hierfür setzt man wieder direkte oder iterative Verfahren ein.

Bei den inexakten oder Quasi-Newton-Verfahren wird dieses lineare Gleichungssystem

- nur näherungsweise gelöst, oder
- die Jacobi-Matrix nicht in jedem Schritt neu aufgestellt.

13.6 Zusammenfassung

- Nichtlineare algebraische Gleichungen können nur iterativ gelöst werden. Daher ist bei allen vorgestellten Methoden die Konvergenz gegen eine Lösung nur unter einschränkenden Voraussetzungen sichergestellt.
- Für monontone Funktionen bietet sich die Bisektion an.
- Die Fixpunktiteration erfordert, dass die Verfahrensfunktion eine Kontraktion darstellt. Dafür konvergiert sie unabhängig vom Startwert.
- Das Newtonverfahren erfordert Differenzierbarkeit der nichtlinearen Funktion und konvergiert nur wenn der Startwert genügend nahe an der Lösung liegt. Dafür ist es wegen der quadratischen Konvergenz sehr schnell.
- Sowohl Fixpunktiteration als auch das Newton-Verfahren können auf Systeme erweitert werden.

14 Einführung in Gewöhnliche Differentialgleichungen

14.1 Motivation

Angenommen wir wollen das Wachstum einer Population von Bakterien, Füchsen, ..., in Abhängigkeit der Zeit ermitteln.

Sei $y(t) : [a, b] \to \mathbf{R}$ die Anzahl der Individuen der Population zur Zeit t. Dabei machen wir zwei Annahmen:

- $[a, b]$ ist das Zeitintervall, in dem uns $y(t)$ interessiert.
- Die Zahl der Individuen ist eine kontinuierliche Größe.
- Wir vernachlässigen die räumliche Verteilung, indem wir uns auf einen kleinen Raumbereich beschränken (etwa eine Petrischale).

Sei nun Δt ein kleines Zeitintervall, dann machen wir die folgende *Modellannahme*:

> Die Zunahme der Zahl der Individuen in Δt is proportional zu Δt und der Zahl der Individuen zur Zeit t.

In Formeln übersetzt heisst das

$$\underbrace{y(t+\Delta t)}_{\#\text{ Individuen am Ende des Intervalls}} = \underbrace{y(t)}_{\#\text{ Individuen am Anfang des Intervalls}} + \underbrace{\lambda \Delta t y(t)}_{\text{Zuwachs}}.$$

λ heisst Wachstumsrate.

Stellen wir die Gleichung etwas um so erhalten wir

$$\frac{y(t+\Delta t) - y(t)}{\Delta t} = \lambda y(t)$$

und für den Limes Δt gegen Null ergibt sich schließlich

$$\frac{dy(t)}{dt} = y'(t) = \lambda y(t). \tag{14.1}$$

So eine Gleichung nennt man eine *Differentialgleichung*, weil die unbekannte Funktion durch eine Bedingung an die Ableitung festgelegt wird.

Eine Funktion $y(t) : [a, b] \to \mathbf{R}$ heißt *Lösung* der Differentialgleichung, falls sie die Gleichung *für alle* $t \in [a, b]$ erfüllt.

Wir wollen uns nun die Lösungsmenge dieser Differentialgleichung überlegen.

Eine Lösung errät man leicht. Wegen

$$\frac{d}{dt} e^{\lambda t} = \lambda e^{\lambda t}$$

ist $y(t) = e^{\lambda t}$ offensichtlich eine Lösung.

Diese Lösung kann man auch mit einer beliebigen Konstanten multiplizieren und erhält weitere Lösungen

$$\frac{d}{dt}\underbrace{Ce^{\lambda t}}_{y(t)} = \lambda\underbrace{Ce^{\lambda t}}_{y(t)}.$$

Sind dies nun alle Lösungen? Dazu überlegt man folgendermaßen. Sei $y(t)$ eine beliebige Lösung von (14.1). Dann gilt:

$$\frac{d}{dt}\left(y(t)e^{-\lambda t}\right) = y'(t)e^{-\lambda t} - y(t)\lambda e^{-\lambda t} = \underbrace{\left(y'(t) - \lambda y(t)\right)}_{=0 \text{ da } y \text{ Lösung}} e^{-\lambda t} = 0.$$

Wegen $\frac{d}{dt}\left(y(t)e^{-\lambda t}\right) = 0$ muss $y(t)e^{-\lambda t} = C$ sein für alle t. Somit haben *alle* Lösungen von (14.1) die Gestalt

$$y(t) = Ce^{\lambda t}.$$

Um die Lösung unserer Beispielgleichung eindeutig festzulegen, müssen wir eine zusätzliche Bedingung stellen, die die Konstante C in der allgemeinen Lösung festlegt.

Eine natürliche Bedingung ist die Zahl der Individuen zu Beginn des Zeitintervalles $[a, b]$, also

$$y(a) = Y.$$

Dies nennt man einen *Anfangswert*.

Hieraus erhält man leicht

$$y(a) = Ce^{\lambda a} = Y \qquad \Leftrightarrow \qquad C = Ye^{-\lambda a}.$$

Damit hat das sogenannte *Anfangswertproblem*

$$y'(t) = \lambda y(t), \qquad y(a) = Y$$

die Lösung

$$y(t) = Ye^{\lambda(t-a)}.$$

Das Wachstum einer Population erfordert Ressourcen, etwa Energie, z. B. in Form von Nahrung.

Exponentielles Wachstum erfordert demnach unbegrenzte Verfügbarkeit von Ressourcen, was in der Realität nicht beliebig lange möglich ist.

Ein realisitischeres Modell für Wachstum nimmt an, dass es eine obere Grenze für die Größe der Population gibt.

Legen wir diese Größe willkürlich auf 1 fest (entsprechend 100%), so erhält man die Differentialgleichung:

$$y'(t) = \lambda(1 - y(t))y(t)$$

Die neue Wachstumsrate $\lambda(1 - y(t))$ wird Null, wenn $y(t) = 1$ erreicht.

Dies nennt man das *logistische Wachstumsmodell*.

Die Lösungsmenge dieser Gleichung ist viel schwieriger zu ermitteln. Wen es interessiert: [TT07, S. 176].

Für viele Differentialgleichungen lässt sich die Lösungsmenge überhaupt nicht in geschlossener Form angeben und man ist auf eine *numerische Lösung* angewiesen.

14.2 Problemstellung

Wir wollen in der Vorlesung Anfangswertaufgaben (AWA) der folgenden Form behandeln:

Finde $y(x) \in C^1[a, b]$, sodass

$$\begin{aligned} y'(x) &= f(x, y(x)), \qquad x \in [a, b], \\ y(a) &= Y \qquad \text{(Anfangswert)} . \end{aligned} \tag{14.2}$$

Diese Differentialgleichung ist

- *gewöhnlich*, da y nur eine Funktion in *einer* Variablen ist, d.h. der Definitionsbereich ist eindimensional.
- *skalar*, da der Wertebereich eindimensional ist.
- *explizit*, da die Gleichung in der Form $y'(x) = \ldots$ ist.
- *erster Ordnung*, da als höchste Ableitung nur eine erste Ableitung von y vorkommt.

Wir wollen kurz darauf eingehen, welche Verallgemeinerungen hiervon es gibt.

Systeme gewöhnlicher Differentialgleichungen Gesucht sind $m > 1$ Funktionen $y_i \in C^1[a, b]$, sodass

$$\begin{aligned} y_i'(x) &= f_i(x, y_1(x), \ldots, y_m(x)), \qquad i = 1, \ldots, n, x \in [a, b], \\ y_i(a) &= Y_i \qquad i = 1, \ldots, n, \text{ (Anfangswerte)} . \end{aligned} \tag{14.3}$$

Mittels

$$\begin{aligned} y &: \mathbf{R} \to \mathbf{R}^n, & y(x) &= (y_1(x), \ldots, y_m(x))^T, \\ f &: \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n, & f(x, y) &= (f_1(x, y), \ldots, f_m(x, y))^T, \end{aligned}$$

kann man das in vektorieller Form schreiben als

$$\begin{aligned} y'(x) &= f(x, y(x)), \qquad x \in [a, b], \\ y(a) &= Y \qquad \text{(Anfangswert).} \end{aligned} \tag{14.4}$$

Für jede Komponente ist eine Anfangsbedingung erforderlich.

Gewöhnliche Differentialgleichungen höherer Ordnung Hier sucht man eine $n > 1$ mal stetig differenzierbare Funktion $y(x) \in C^n[a, b]$, sodass

$$\begin{aligned} \frac{d^n y}{dx^n}(x) &= f(x, y(x), \frac{dy}{dx}(x), \ldots, \frac{d^{n-1}y}{dx^{n-1}}(x)) \qquad x \in [a, b] \\ y(a) = Y_0, \frac{dy}{dx}(a) &= Y_1, \ldots, \frac{d^{n-1}y}{dx^{n-1}}(a) = Y_{n-1} \quad \text{(Anfangswerte)}. \end{aligned} \tag{14.5}$$

Hier sind n Anfangsbedingungen erforderlich.

Eine skalare, gewöhnliche Differentialgleichung n-ter Ordnung kann immer auf ein System erster Ordnung mit n Komponenten reduziert werden.

Dazu führt man die Ableitungen bis zur Ordnung $n - 1$ als zusätzliche Unbekannte ein:

$$w_0(x) = y(x), \quad w_1(x) = \frac{dy}{dx}, \quad \ldots, \quad w_{n-1}(x) = \frac{d^{n-1}y}{dx^{(n-1)}}.$$

Damit erhält man dann das System

$$\begin{array}{llll} w_0'(x) = w_1(x), & \ldots & w_{n-2}'(x) = w_{n-1}(x), & w_{n-1}'(x) = f(x, w_0(x), \ldots, w_{n-1}(x)), \\ w_0(a) = Y_0, & \ldots & w_{n-2}(a) = Y_{n-2}, & w_{n-1}'(a) = Y_{n-1}. \end{array}$$

Aufgrund dieses Tricks werden wir gewöhnliche Differentialgleichungen höherer Ordnung nicht weiter behandeln und annehmen, dass man sie entsprechend auf ein System reduziert.

Numerisch muss das nicht unbedingt geschickt sein, wie wir in der Einführungsvorlesung bei dem Pendel gesehen haben.

Man kann auch zeigen, dass sich Systeme von Differentialgleichungen mit m Komponenten n-ter Ordnung auf ein System erster Ordnung mit mn Komponenten reduzieren lassen, [SK05].

Randwertprobleme Bei gewöhnlichen Differentialgleichungen höherer Ordnung muss man nicht alle zusätzlichen Bedingungen am Anfang des Intervalls stellen.

Wir betrachten eine Gleichung zweiter Ordnung der folgenden Form. Finde $y \in C^2[a, b]$, sodass

$$\begin{aligned} y''(x) &= f(x, y(x), y'(x)) \qquad x \in [a, b], \\ y(a) = Y_a, &\quad y(b) = Y_b. \end{aligned}$$

Hier ist also y am Anfang und Ende des Intervalls vorgegeben. Man spricht dann von einem *Randwertproblem*.

Solche Aufgaben wollen wir hier nicht behandeln.

Differential-algebraische Systeme Hier hat man zusätzlich zur gewöhnlichen Differentialgleichung noch eine algebraische Nebenbedingung:

$$\begin{aligned} y' &= f(x, y(x), z(x)) \qquad x \in [a, b], \\ 0 &= g(x, y(x), z(x)) \qquad x \in [a, b], \\ y(a) &= 0. \end{aligned}$$

Hierbei sind $y(x), z(x)$ und entsprechend $f(x, y, z), g(x, y, z)$ vektorwertige Funktionen.

Implizite Form der Differentialgleichung Hat die gewöhnliche Differentialgleichung die Form

$$\begin{aligned} F(x, y(x), y'(x)) &= 0, \qquad x \in [a, b], \\ y(a) &= Y, \end{aligned}$$

spricht man von einer gewöhnlichen Differentialgleichung erster Ordnung in impliziter Form.

Hierbei seien y und entsprechend F als vektorwertig angenommen.

Oft sind Systeme in der impliziten Form differentiell-algebraisch, nämlich dann, wenn die Jacobimatrix von F bezüglich des dritten Arguments y' singulär wird.

Partielle Differentialgleichungen Gesucht ist eine Funktion in *mehr als einer* Variablen und es sind Bedingungen an die partiellen Ableitungen gegeben.

Finde $y \in C^2([a, b] \times [c, d])$, sodass

$$\begin{aligned} \frac{\partial^2 y}{\partial x_1^2}(x_1, x_2) + \frac{\partial^2 y}{\partial x_2^2}(x_1, x_2) &= f(x_1, x_2) && \forall (x_1, x_2) \in (a, b) \times (c, d) \\ y(x_1, x_2) &= g(x_1, x_2) && (x_1, x_2) \in [a, b] \times [c, d] \\ & && \quad \wedge (x_1 \in \{a, b\} \vee x_2 \in \{c, d\}) \end{aligned}$$

Das ist die sogenannte "Poisson-Gleichung" in zwei Raumdimensionen. Partielle Differentialgleichungen behandeln wir in dieser Vorlesung nicht!

Delay-Gleichungen Eine sogenannte *Delay-Gleichung* hat die Form

$$\begin{aligned} y'(x) &= f(x, y(x), y(x - \tau)) \qquad x \in [a, b] \\ y(x) &= g(x) \qquad x \in [a - \tau, b] \end{aligned}$$

$y'(x)$ hängt also nicht nur von $y(x)$ sondern auch von $y(x - \tau)$ ab.

Behandeln wir auch nicht!

14.3 Weitere Beispiele für gewöhnliche Differentialgleichungen

Beispiel 14.1 (Einfache Reaktion)**.** Zwei Stoffe A und B reagieren zu einem Stoff C, also $A + B \to C$.

Es seien

$c_A(t)$ Konzentration (Stoffmenge, z.B. Mol pro Volumen) von A zur Zeit t.

$c_B(t)$ Konzentration von B zur Zeit t.

$c_C(t)$ Konzentration von C zur Zeit t.

Für die Änderung der Konzentration von Stoff A in einem Zeitintervall Δt nehmen wir an, dass diese proportional zu Δt und dem Produkt $c_A(t)c_B(t)$ ist (da die Atome/Moleküle beider Stoffe zusammenkommen müssen):

$$\begin{aligned} & c_A(t+\Delta t) = c_A(t) - k\Delta t c_A(t)c_B(t) \\ \Leftrightarrow \qquad & c_A'(t) = -k c_A(t)c_B(t) \end{aligned}$$

mit einem $k > 0$ (Reaktionsrate). Für die anderen beiden Komponenten erhält man analog $c_B'(t) = -kc_A(t)c_B(t)$, $c_C'(t) = kc_A(t)c_B(t)$. □

Beispiel 14.2 (Komplexe Reaktion)**.** Eine Gleichgewichtsreaktion der Form

$$aA + bB \underset{k_2}{\overset{k_1}{\rightleftharpoons}} cC + dD$$

wird modelliert durch das System

$$\begin{aligned} c_A'(t) &= R(t) \\ c_B'(t) &= R(t) \\ c_C'(t) &= -R(t) \\ c_D'(t) &= -R(t) \end{aligned} \qquad \text{mit} \qquad \begin{aligned} R(t) &= -k_1(c_A(t))^a(c_B(t))^b + k_2(c_C(t))^c(c_D(t))^d \\ c_i(t_0) &= C_i; \quad i \in \{A,B,C,D\} \end{aligned}$$

$c_i(t)$ ist die Konzentration von Stoff i. Im chemischen Gleichgewicht gilt

$$c_i'(t) = 0 \quad \Leftrightarrow \quad R(t) = 0 \quad \Leftrightarrow \quad \frac{c_A^a c_B^b}{c_C^c c_D^d} = \frac{k_2}{k_1} = K_{\text{eq}}.$$

Dies ist das *Massenwirkungsgesetz.* □

Beispiel 14.3 (N-Körper Problem, Astronomie)**.** Betrachte die Bewegung von N Körpern mit den Massen m_i unter ihrem eigenen Schwerefeld.

Unbekannt: Positionen $x_i(t) \in \mathbf{R}^3$ und Geschwindigkeiten $v_i(t) \in \mathbf{R}^3$.

Position und Geschwindigkeit hängen zusammen über

$$\frac{dx_i(t)}{dt} = v_i(t); \qquad x_i(t_0) = x_{i,0}; \qquad i = 1, \ldots, N$$

Gravitationskraft und zweites Newtonsches Gesetz gibt:

$$\sum_{\substack{1\le j\le N\\ j\neq i}} \frac{\gamma m_j m_i}{\|x_j - x_i\|^2}\frac{(x_j - x_i)}{\|x_j - x_i\|} = \vec{F}_i(t) = m_i a_i(t) = m_i \frac{dv_i(t)}{dt}, \qquad i = 1, \ldots, N$$

$$\Rightarrow \frac{dv_i(t)}{dt} = \sum_{\substack{1\le j\le N\\ j\neq i}} \frac{\gamma m_j (x_j - x_i)}{\|x_j - x_i\|^3}; \qquad v_i(t_0) = v_{i,0}; \qquad i = 1, \ldots, N$$

Also $6N$ gekoppelte, nichtlineare Differentialgleichungen. □

Wie man sieht, spielen gewöhnliche Differentialgleichungen in vielen verschiedenen Gebieten eine Rolle. Und wir haben nur eine winzige Auswahl gegeben.

Bevor wir uns an die numerische Lösung wagen, ist zu klären, wann wir überhaupt erwarten können, dass eine Anfangswertaufgabe eine Lösung besitzt und ob diese eindeutig ist.

14.4 Zur Theorie gewöhnlicher Differentialgleichungen

Satz 14.4 (Existenzsatz von Peano)**.** Sei $f : \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n$ stetig (in allen Argumenten) auf dem Streifen

$$D = \{(x, y) \in \mathbf{R} \times \mathbf{R}^n \big| |x - a| \le \alpha; \|y - Y\| \le \beta\}$$

dann hat das Problem

$$y'(x) = f(x, y(x)), \qquad y(a) = Y$$

eine Lösung auf dem Intervall $I = [a - T, a + T]$ für ein gewisses $T \le \alpha$, welches von α, β und f abhängt.

Beweis: Siehe [Ran]. □

Satz 14.5 (Stabilitätssatz)**.** Die Funktion $f : [a, b] \times \mathbf{R}^n \to \mathbf{R}^n$ sei Lipschitzstetig in der Variablen y, d.h. es existiere eine Konstante $L > 0$, sodass

$$\|f(x, y) - f(x, z)\| \le L\|y - z\| \qquad \forall x \in [a, b], y, z \in \mathbf{R}^n. \tag{14.6}$$

Weiter seien $y(x), z(x)$ Lösungen des Anfangswertproblems zu den zwei Startwerten Y bzw. Z dann gilt

$$\|y(x) - z(x)\| \le \|Y - Z\| e^{L|x-a|}.$$

Beweis: Siehe [Ran]. □

Eine Folgerung aus dem Stabilitätssatz ist die Eindeutigkeit der Lösung, falls sie existiert.

Allerdings *können* die Lösungen für kleine Änderungen in den Startwerten exponentiell schnell auseinander laufen.

14.5 Zusammenfassung

- Gewöhnliche Differentialgleichungen beschreiben eine Vielzahl von Vorgängen in den Natur- und Ingenieurwissenschaften.
- Wir betrachten hier vor allem skalare Anfangswertaufgaben erster Ordnung.
- Stetigkeit der Funktion f sichert lokale Existenz und Lipschitz-Stetigkeit sichert Eindeutigkeit der Lösung.

15 Einige einfache Verfahren

15.1 Expliziter Euler

Wir betrachten die AWA

$$y'(x) = f(x, y(x)) \quad \text{in } [a, b], \qquad y(a) = Y.$$

Aus der Taylorentwicklung

$$y(x + h) = y(x) + hy'(x) + \frac{1}{2}h^2 y''(x + \xi h), \qquad \text{für ein } \xi \in [0, 1],$$

erhalten wir für die erste Ableitung:

$$y'(x) = \underbrace{\frac{y(x + h) - y(x)}{h}}_{\text{Differenzenquotient}} + O(h). \tag{15.1}$$

Der Differenzenquotient, die sog. *Vorwärtsdifferenz*, liefert somit eine Approximation erster Ordnung an die Ableitung.

Idee: Ersetze $y'(x)$ in der AWA durch den Differenzenquotienten und vernachlässige den Fehlerterm.

Dazu wählen wir eine Unterteilung des Intervalles $[a, b]$:

$$a = x^{(0)} < x^{(1)} < \ldots < x^{(N_h - 1)} < x^{(N_h)} = b$$

und setzen

$$h^{(j)} = x^{(j)} - x^{(j-1)}, \qquad h := \max_{i \in \{1, \ldots, N_h\}} h^{(j)}.$$

Äquidistante Gitterpunkte: $h = (b - a)/N_h$ und $x^{(j)} = a + jh$.

Einsetzen des Differenzenquotienten in die AWA liefert

$$\frac{y(x^{(j+1)}) - y(x^{(j)})}{h^{(j+1)}} + O(h) = f(x^{(j)}, y(x^{(j)})) \quad \Rightarrow \quad \frac{y_h^{(j+1)} - y_h^{(j)}}{h^{(j+1)}} = f(x^{(j)}, y_h^{(j)}).$$

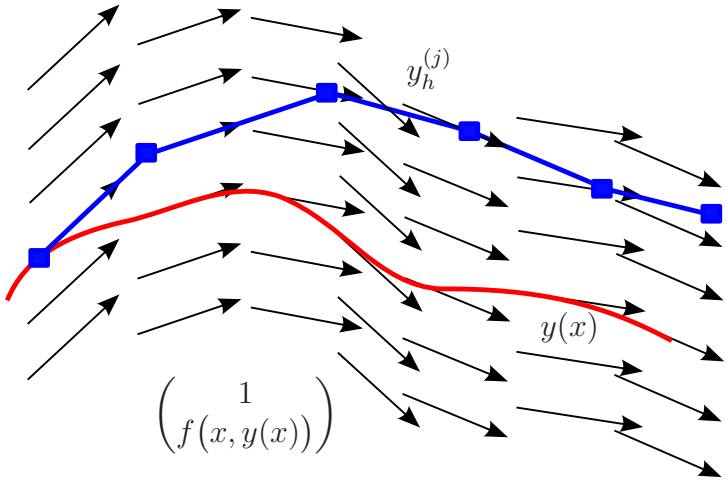
Umstellen und Hinzufügen der Anfangsbedingung liefert eine Rekursionsformel für die unbekannten Werte $y_h^{(j)}$:

$$\begin{aligned} y_h^{(j+1)} &= y_h^{(j)} + h^{(j+1)} f(x^{(j)}, y_h^{(j)}), \qquad i = 0, \ldots, N_h - 1, \\ y_h^{(0)} &= Y. \end{aligned} \tag{15.2}$$

Hier haben wir y durch die sogenannte *Gitterfunktion* y_h ersetzt, da wegen Weglassen des Fehlerterms nur $y_h^{(j)} \approx y(x^{(j)})$ gilt.

Dies ist ein sogenanntes *explizites Verfahren*, da der unbekannte Funktionswert $y_h^{(j+1)}$ alleine auf der linken Seite steht.

Das Verfahren erlaubt eine einfache Interpretation im skalaren Fall:



Man spricht daher auch von „Eulerschem Polygonzugverfahren“.

15.2 Impliziter Euler

Hier verwenden wir die Taylorentwicklung

$$y(x-h) = y(x) - hy'(x) + \frac{1}{2}h^2 y''(x-\xi h), \qquad \text{für ein } \xi \in [0,1],$$

und erhalten für die erste Ableitung:

$$y'(x) = \frac{y(x) - y(x-h)}{h} + O(h).$$

Dies bezeichnet man als *Rückwärtsdifferenz*.

Einsetzen des Differenzenquotienten in die AWA liefert

$$\frac{y(x^{(j+1)}) - y(x^{(j)})}{h^{(j+1)}} + O(h) = f(x^{(j+1)}, y(x^{(j+1)})) \quad \Rightarrow \quad \frac{y_h^{(j+1)} - y_h^{(j)}}{h^{(j+1)}} = f(x^{(j+1)}, y_h^{(j+1)}).$$

Somit ergibt sich für die Werte an den Gitterpunkten $x^{(j)}$:

$$\begin{aligned} y_h^{(j+1)} - h^{(j+1)} f(x^{(j+1)}, y_h^{(j+1)}) &= y_h^{(j)} \\ y_h^{(0)} &= Y; \end{aligned} \tag{15.3}$$

Dieses Verfahren nennt man *implizit*, da nicht sofort nach $y_h^{(j+1)}$ aufgelöst werden kann.

Implizite Verfahren erfordern im allgemeinen die Lösung eines nichtlinearen algebraischen Gleichungssystems:

$$F(u) = u - h^{(j+1)} f(x^{(j+1)}, u) - y_h^{(j)} = 0.$$

Eine Möglichkeit zur Lösung ist die Fixpunktiteration

$$u^{(k+1)} = g(u^{(k)}); \qquad g(u) = u - F(u),$$

also

$$g(u) = u - \left[u - h^{(j+1)} f(x^{(j+1)}, u) - y_h^{(j)}\right] = h^{(j+1)} f(x^{(j+1)}, u) + y_h^{(j)}.$$

Für die Lipschitz-Stetigkeit von g rechnet man

$$\begin{aligned} \|g(u) - g(u')\| &= \|h^{(j+1)} f(x^{(j+1)}, u) + y_h^{(j)} - h^{(j+1)} f(x^{(j+1)}, u') - y_h^{(j)}\| \\ &= h^{(j+1)} \|f(x^{(j+1)}, u) - f(x^{(j+1)}, u')\| \\ &\leq h^{(j+1)} L \|u - u'\|. \end{aligned}$$

Für genügend kleines h lässt sich das nichtlineare System immer per Fixpunktiteration lösen.

Allerdings ist diese Variante im allgemeinen nicht effizient und man verwendet eher das Newton-Verfahren.

15.3 Trapezregel

Aus $y'(x) = f(x, y(x))$ folgt durch Integration über ein Teilintervall

$$\begin{aligned} & \int_{x^{(j)}}^{x^{(j+1)}} y'(x) dx = \int_{x^{(j)}}^{x^{(j+1)}} f(x, y(x))\, dx \\ \Leftrightarrow \qquad & y\left(x^{(j+1)}\right) - y\left(x^{(j)}\right) = \int_{x^{(j)}}^{x^{(j+1)}} f(x, y(x))\, dx. \end{aligned}$$

Nun ersetze das Integral rechts durch Auswertung mittels Trapezregel:

$$y(x^{(j+1)}) - y(x^{(j)}) = \frac{h^{(j+1)}}{2} \left\{ f(x^{(j)}, y(x^{(j)})) + f(x^{(j+1)}, y(x^{(j+1)})) \right\} + O(h^3) \tag{15.4}$$

Im Vergleich zum expliziten bzw. impliziten Euler ist dieses Verfahren eine Ordnung genauer.

Weglassen des Restglieds ergibt das implizite Rekursionsschema für y_h:

$$\begin{gathered} y_h^{(j+1)} - \frac{h^{(j+1)}}{2} f(x^{(j+1)}, y_h^{(j+1)}) = y_h^{(j)} + \frac{h^{(j+1)}}{2} f(x^{(j)}, y_h^{(j)}) \\ y_h^{(0)} = Y. \end{gathered}$$

15.4 Mittelpunktregel

Die bisher behandelten Verfahren sind alle so genannte *Einschrittverfahren*, da aus $y_h^{(j)}$ das $y_h^{(j+1)}$ berechnet wird.

Nun zeigen wir ein erstes Beispiel für ein Mehrschrittverfahren.

Durch Integration des AWP über *zwei* Teilintervalle erhalten wir:

$$\int_{x^{(j)}}^{x^{(j+2)}} y'(x)dx = \int_{x^{(j)}}^{x^{(j+2)}} f(x, y(x))\, dx.$$

Die Mittelpunktregel liefert bei $h^{(j+1)} = h^{(j+2)}$ (!)

$$y(x^{(j+2)}) - y(x^{(j)}) = 2h^{(j+1)} f(x^{(j+1)}, y(x^{(j+1)})) + O(h^3). \tag{15.5}$$

Dies führt dann zu der folgenden Rekursionsformel

$$\begin{aligned} y_h^{(j+2)} &= y_h^{(j)} + 2h^{(j+1)} f(x^{(j+1)}, y_h^{(j+1)}) \\ y_h^{(0)} &= Y; \\ y_h^{(1)} &= Y + h^{(1)} f(a, y_0); \qquad \text{expliziter Euler für } y_h^{(1)}; \end{aligned}$$

Es handelt sich hier um ein explizites Zweischrittverfahren mit besserer Fehlerordnung als das explizite Euler-Verfahren.

15.5 Anwendung auf ein Modellproblem

Beispiel 15.1. Wir lösen nun die AWA

$$y'(x) = \lambda y(x) \quad \text{in } [a, b], \qquad y(a) = Y,$$

$\mathbf{R} \ni \lambda < 0$ mit der exakten Lösung

$$y(x) = Y e^{\lambda(x-a)}.$$

Für die verschiedenen oben behandelten Verfahren ergibt sich unter Einsetzen von $f(x, y) = \lambda y$ für eine äquidistante Schrittweite:

Expliziter Euler

$$y_h^{(j+1)} = y_h^{(j)} + h\lambda y_h^{(j)} = (1 + h\lambda) y_h^{(j)}.$$

Impliziter Euler Da f linear ist kann man auflösen:

$$\begin{aligned} y_h^{(j+1)} - h\lambda y_h^{(j+1)} &= y_h^{(j)} \\ \Leftrightarrow \qquad y_h^{(j+1)} &= \left(\frac{1}{1 - h\lambda}\right) y_h^{(j)} \end{aligned}$$

Trapezregel

$$\begin{aligned} y_h^{(j+1)} - \frac{h}{2}\lambda y_h^{(j+1)} &= y_h^{(j)} + \frac{h}{2}\lambda y_h^{(j)} \\ \Leftrightarrow \qquad y_h^{(j+1)} &= \left(\frac{1+\frac{h}{2}\lambda}{1-\frac{h}{2}\lambda}\right) y_h^{(j)} \end{aligned}$$

Wegen $y(x+h) = Ye^{\lambda(x+h-a)} = Ye^{\lambda(x-a)}e^{\lambda h} = e^{\lambda h}y(x)$ sind die Faktoren

$$(1+h\lambda), \qquad \frac{1}{1-h\lambda}, \qquad \frac{1+\frac{h}{2}\lambda}{1-\frac{h}{2}\lambda}$$

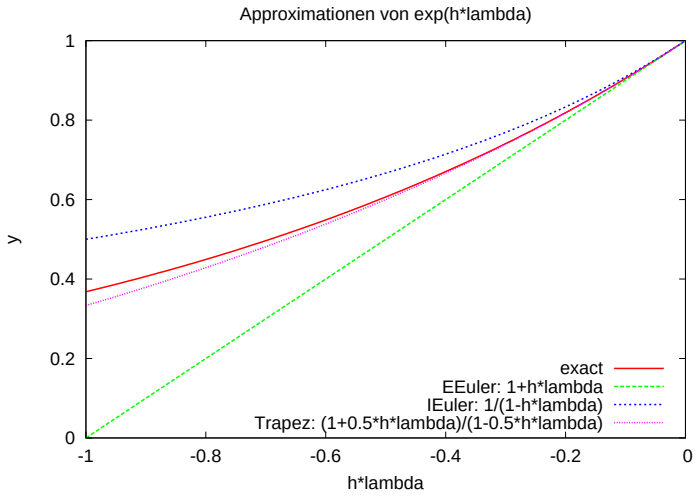
alles verschiedene Approximationen von $e^{\lambda h}$.

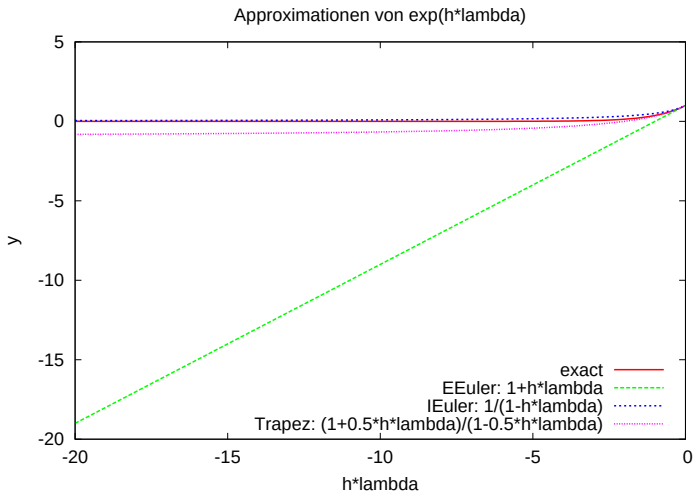
Mittelpunktregel

$$\begin{aligned} y_h^{(j+2)} &= y_h^{(j)} + 2h\lambda y_h^{(j+1)} \\ y_h^{(0)} &= y_0; \\ y_h^{(1)} &= y_0 + h\lambda y_0; \qquad \text{(expliziter Euler)} \end{aligned}$$

Wir betrachten nun

- Approximationen von $e^{h\lambda}$.
- Näherungslösung und Fehler beim Modellproblem für expliziten Euler und Trapezregel.
- Fehlerordnung bei den verschiedenen Verfahren.





Der explizite Euler besitzt eine deutlich schlechtere Approximation an $e^{h\lambda}$.

Hier der Fehler bei Anwendung des expliziten Euler auf das Modellproblem mit $\lambda = -1$ und Schrittweite $h = 0.1$:

```
0.000000000000000e+00 1.000000000000000e+00
0.000000000000000e+00 1.000000000000000e-01
9.000000000000000e-01 4.837418035959495e-03
2.000000000000000e-01 8.100000000000001e-01
8.730753077981768e-03 3.000000000000000e-01
7.290000000000001e-01 1.181822068171778e-02
4.000000000000000e-01 6.561000000000001e-01
1.422004603563920e-02 5.000000000000000e-01
5.904900000000001e-01 1.604065971263335e-02
6.000000000000000e-01 5.314410000000001e-01
1.737063609402634e-02 7.000000000000000e-01
4.782969000000000e-01 1.828840379140950e-02
7.999999999999999e-01 4.304672100000000e-01
1.886175411722157e-02 8.999999999999999e-01
3.874204890000000e-01 1.914917074059913e-02
```

```
9.999999999999999e-01 3.486784401000000e-01
1.920100107144235e-02 1.100000000000000e+00
3.138105960900001e-01 1.906048760807955e-02
1.200000000000000e+00 2.824295364810001e-01
1.876467543120208e-02 1.300000000000000e+00
2.541865828329000e-01 1.834521020111257e-02
1.400000000000000e+00 2.287679245496100e-01
1.782903939199643e-02 1.500000000000000e+00
2.058911320946490e-01 1.723902805378077e-02
1.600000000000000e+00 1.853020188851841e-01
1.659449910947125e-02 1.700000000000000e+00
1.667718169966657e-01 1.591170705606890e-02
1.800000000000000e+00 1.500946352969991e-01
1.520425292458732e-02 1.900000000000001e+00
1.350851717672992e-01 1.448344745533578e-02
2.000000000000000e+00 1.215766545905693e-01
1.375862864604334e-02
```

Fehler nimmt mit x ab.

Hier der Fehler bei Anwendung der Trapezregel auf das Modellproblem mit $\lambda = -1$ und Schrittweite $h = 0.1$:

```
0.000000000000000e+00 1.000000000000000e+00
0.000000000000000e+00 1.000000000000000e-01
9.047619047619048e-01 7.551327405475039e-05
2.000000000000000e-01 8.185941043083900e-01
1.366487695918517e-04 3.000000000000000e-01
7.406327610409242e-01 1.854596407936393e-04
4.000000000000000e-01 6.700963076084553e-01
2.237384271840392e-04 5.000000000000000e-01
6.062776116457452e-01 2.530480668881951e-04
6.000000000000000e-01 5.485368867271028e-01
2.747493669236212e-04 7.000000000000000e-01
4.962952784673787e-01 2.900253240308293e-04
7.999999999999999e-01 4.490290614704855e-01
2.999026467361277e-04 8.999999999999999e-01
4.062643889494869e-01 3.052707911123109e-04
9.999999999999999e-01 3.675725423828691e-01
3.068987885733176e-04 1.100000000000000e+00
3.325656335845006e-01 3.054501135790200e-04
1.200000000000000e+00 3.008927161002624e-01
3.014958119397226e-04 1.300000000000000e+00
2.722362669478565e-01 2.955260861561282e-04
1.400000000000000e+00 2.463090034290130e-01
2.879605125934437e-04 1.500000000000000e+00
2.228510031024403e-01 2.791570459894643e-04
1.600000000000000e+00 2.016270980450651e-01
2.694199495902883e-04 1.700000000000000e+00
1.824245172788684e-01 2.590067738661994e-04
1.800000000000000e+00 1.650507537284999e-01
2.481344930864993e-04 1.900000000000001e+00
1.493316343257857e-01 2.369848968493127e-04
2.000000000000000e+00 1.351095739138061e-01
2.257093228065499e-04
```

Fehler kleiner als bei explizitem Euler.

Methode	u(4.0)	Fehler
EEuler	1.478088294143459e-02	3.534755947299559e-03
	1.795055327504517e-02	3.650856136897633e-04
	1.827901982748948e-02	3.661906125073527e-05
IEuler	2.209492815217999e-02	3.779289263445847e-03
	1.868316662016864e-02	3.675277314337082e-04
	1.835228237083395e-02	3.664348209373933e-05
Trapez	1.825459696317023e-02	6.104192556392191e-05
	1.831502836845542e-02	6.105202795189668e-07
	1.831563278352137e-02	6.105218844365545e-09
Midpoint	-9.224910943198056e-02	1.105647483207147e-01
	1.697950575278587e-02	1.336133135949061e-03
	1.830203341571142e-02	1.360547302880002e-05

- Expliziter/Impliziter Euler konvergieren mit Ordnung h.
- Trapez- und Mittelpunktregel konvergieren mit h^2, Mittelpunktregel hat eine schlechtere Konstante.

15.6 Lineare Mehrschrittverfahren

Wir betrachten nun die allgemeine Konstruktion von Mehrschrittverfahren als eine wichtige Verfahrensklasse.

Auch hier spielen die Lagrange-Polynome wieder eine Rolle.

Der Einfachheit halber beschränken wir uns auf *äquidistante* Gitter,

$$x^{(j)} = a + jh, \qquad h = (b-a)/N.$$

Eine Erweiterung auf nichtäquidistante Gitter ist jedoch möglich.

Weiter behandeln wir hier nur den skalaren Fall. Die Erweiterung auf Systeme erfolgt durch komponentenweise Anwendung.

Es gibt zwei Konstruktionsmethoden für lineare Mehrschrittverfahren:

- mittels Integration oder
- mittels Differentiation.

Für ein $\sigma \in \mathbf{N}$ haben wir

$$\begin{aligned} \int_{x^{(n-\sigma)}}^{x^{(n)}} y'(x)\,dx &= \int_{x^{(n-\sigma)}}^{x^{(n)}} f(x, y(x))\,dx, \\ \Leftrightarrow y\left(x^{(n)}\right) &= y\left(x^{(n-\sigma)}\right) + \int_{x^{(n-\sigma)}}^{x^{(n)}} f(x, y(x))\,dx \end{aligned} \tag{15.6}$$

Idee: Lege nun ein Polynom vom Grad $m \geq 0$ durch die Werte von f an den Stellen $x^{(k-m)}, \ldots, x^{(k)}$:

$$p_{m,k} = \sum_{\mu=0}^{m} f\left(x^{(k-\mu)}, y\left(x^{(k-\mu)}\right)\right) L_\mu^{(m,k)}(x). \tag{15.7}$$

Beachte:

- Das Polynom benutzt die Stützwerte $k-m, \ldots, k$, wobei k eventuell verschieden von n ist, und auch m verschieden von σ sein kann!
- $L_\mu^{(m,k)}(x^{(k-\nu)}) = \delta_{\mu\nu}$, $\nu = 0, \ldots, m$ sind Lagrangepolynome.

Für den Interpolationsfehler des Polynoms erhalten wir:

$$f(x, y(x)) - p_{m,k}(x) = \frac{L(t)}{(m+1)!} f^{(m+1)}(\xi_x, y(\xi_x)) = \frac{L(t)}{(m+1)!} y^{(m+2)}(\xi_x)$$

und

$$L(t) = \prod_{i=0}^{m} (x - x^{(k-i)}).$$

Einsetzen in (15.6) liefert

$$y\left(x^{(n)}\right) = y\left(x^{(n-\sigma)}\right) + \sum_{\mu=0}^{m} f\left(x^{(k-\mu)}, y\left(x^{(k-\mu)}\right)\right) \int_{x^{(n-\sigma)}}^{x^{(n)}} L_\mu^{(m,k)}(x)\, dx + E, \tag{15.8}$$

wobei

$$E = \frac{y^{(m+2)}(\xi_x)}{(m+1)!} \int_{x^{(n-\sigma)}}^{x^{(n)}} \prod_{i=0}^{m} (x - x^{(k-i)})\, dx = O(h^{m+2}).$$

Das Integral in der letzten Beziehung rechnet man wieder zweckmäßig mittels der Transformation $g(s) = a + (n - \sigma + s)h$ aus.

Durch Weglassen des Fehlerterms erhält man wieder eine Rekursionsgleichung für die Gitterfunktion y_h:

$$\begin{aligned} y_h^{(n)} &= y_h^{(n-\sigma)} + h \sum_{\mu=0}^{m} \underbrace{f\left(x^{(k-\mu)}, y_h^{(k-\mu)}\right)}_{=f_h^{(k-\mu)}} \underbrace{h^{-1} \int_{x^{(n-\sigma)}}^{x^{(n)}} L_\mu^{(m,k)}(x)\, dx}_{=\beta^{(\mu)}} \\ &= y_h^{(n-\sigma)} + h \sum_{\mu=0}^{m} \beta^{(\mu)} f_h^{(k-\mu)}. \end{aligned} \tag{15.9}$$

Dabei hat man σ, k und m als freie Parameter in der Methode.

Je nach Wahl von σ und k erhält man unterschiedliche Klassen von Verfahren, durch Wahl von m erhält man Verfahren unterschiedlicher Ordnung innerhalb der jeweiligen Klasse.

Adams[32]-Bashforth[33]-Formeln Für $\sigma = 1$ und $k = n - 1$ erhält man die Adams-Bashforth-Formeln, die alle *explizit* sind. Für $m = 0, \ldots, 3$ ergibt sich:

$$
\begin{aligned}
m = 0: &\qquad y_h^{(n)} = y_h^{(n-1)} + h f_h^{(n-1)} \qquad \text{(expliziter Euler)} \\
m = 1: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{2}\left(3 f_h^{(n-1)} - f_h^{(n-2)}\right) \\
m = 2: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{12}\left(23 f_h^{(n-1)} - 16 f_h^{(n-2)} + 5 f_h^{(n-3)}\right) \\
m = 3: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{24}\left(55 f_h^{(n-1)} - 59 f_h^{(n-2)} + 37 f_h^{(n-3)} - 9 f_h^{(n-4)}\right).
\end{aligned}
$$

Startwertbestimmung. Offensichtlich benötigt man bei der Anwendung der Methode m zusätzliche Startwerte. Diese muss man sich mit einem Einschrittverfahren verschaffen. Hier ist auf die entsprechende Ordnung zu achten.

Adams-Moulton[34]-Formeln Für $\sigma = 1$ und $k = n$ erhält man die Adams-Moulton-Formeln, die alle *implizit* sind. Für $m = 0, \ldots, 3$ ergibt sich:

$$
\begin{aligned}
m = 0: &\qquad y_h^{(n)} = y_h^{(n-1)} + h f_h^{(n)} \qquad \text{(impliziter Euler)} \\
m = 1: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{2}\left(f_h^{(n)} + f_h^{(n-1)}\right) \qquad \text{(Trapezregel)} \\
m = 2: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{12}\left(5 f_h^{(n)} + 8 f_h^{(n-1)} - f_h^{(n-2)}\right) \\
m = 3: &\qquad y_h^{(n)} = y_h^{(n-1)} + \frac{h}{24}\left(9 f_h^{(n)} + 19 f_h^{(n-1)} - 5 f_h^{(n-2)} + f_h^{(n-3)}\right).
\end{aligned}
$$

Nyström-Formeln Für $\sigma = 2$ und $k = n - 1$ erhält man die Nyström-Formeln, die alle *explizit* sind.

Für $m = 0$ ergibt sich:

$$
m = 0: \qquad y_h^{(n)} = y_h^{(n-2)} + 2h f_h^{(n-1)} \qquad \text{(Mittelpunktsregel)}.
$$

Milne-Simpson-Formeln Für $\sigma = 2$ und $k = n$ erhält man die Milne-Simpson-Formeln, die alle *implizit* sind.

Für $m = 2$ ergibt sich:

$$
m = 2: \qquad y_h^{(n)} = y_h^{(n-2)} + \frac{h}{3}\left(f_h^{(n)} + 4 f_h^{(n-1)} + f_h^{(n-2)}\right) \qquad \text{(Simpson-Regel)}.
$$

[32] John Couch Adams, 1819-1892, brit. Mathematiker und Astronom.
[33] Francis Bashforth, 1819-1912.
[34] Forest Ray Moulton, 1872-1952, amerik. Astronom.

Ein weiterer Zugang zu Mehrschrittformeln ergibt sich über die Differentiation.

Hierzu legt man ein Polynom m-ten Grades durch $m+1$ Werte von y:

$$p_{m,k} = \sum_{\mu=0}^{m} y(x^{(k-\mu)}) L_\mu^{(m,k)}.$$

Dieses lässt sich einfach differenzieren und man erhält bei Auswertung an der Stelle $x^{(n)}$:

$$\underbrace{\sum_{\mu=0}^{m} y(x^{(k-\mu)}) L_\mu^{(m,k)\prime}(x^{(n)})}_{p'_{m,k}(x^{(n)})} = f(x^{(n)}, y(x^{(n)})) + O(h^{m+1}). \tag{15.10}$$

(Dies folgt durch Differenzieren der Fehlerdarstellung des Polynoms).

Rückwärtsdifferenzenformeln Für die Wahl $k = n$ erhält man die sogenannten *Rückwärtsdifferenzenformeln* (engl.: *backward difference formulas*):

$$\begin{aligned}
m = 1: &\quad y_h^{(n)} - y_h^{(n-1)} = h f_h^{(n)} \qquad \text{(impliziter Euler)} \\
m = 2: &\quad y_h^{(n)} - \frac{4}{3} y_h^{(n-1)} + \frac{1}{3} y_h^{(n-2)} = \frac{2}{3} h f_h^{(n)} \\
m = 3: &\quad y_h^{(n)} - \frac{18}{11} y_h^{(n-1)} + \frac{9}{11} y_h^{(n-2)} - \frac{2}{11} y_h^{(n-3)} = \frac{6}{11} h f_h^{(n)} \\
m = 4: &\quad y_h^{(n)} - \frac{48}{25} y_h^{(n-1)} + \frac{36}{25} y_h^{(n-2)} - \frac{16}{25} y_h^{(n-3)} + \frac{3}{25} y_h^{(n-4)} = \frac{12}{25} h f_h^{(n)}.
\end{aligned}$$

15.7 Zusammenfassung

- In diesem Abschnitt haben wir einige der einfachsten numerischen Lösungsverfahren für gewöhnliche Differentialgleichungen hergeleitet.
- Explizites und implizites Eulerverfahren basieren auf der Taylorreihenentwicklung und Trapez- sowie Mittelpunktsregel auf entsprechenden Auswertungen des Integrals.
- Dann haben wir verschiedene lineare Mehrschrittverfahren hergeleitet.
- Explizite Verfahren liefern direkt eine Approximation der Funktion aus Funktionswerten zu früheren Werten. Implizite Verfahren erfordern die Lösung einer nichtlinearen algebraischen Gleichung.
- Alle hier behandelten Verfahren lassen sich entsprechend auf Systeme von Differentialgleichungen erweitern.

16 Konvergenz, Stabilität und dynamische Systeme

16.1 Konvergenz von Einschrittverfahren

Alle Einschrittverfahren lassen sich in die folgende Form bringen:

$$y_h^{(j+1)} = y_h^{(j)} + h^{(j+1)} \Phi_h \left(x^{(j)}, y_h^{(j)}, x^{(j+1)}, y_h^{(j+1)} \right) \tag{16.1}$$

bringen. Beispiele:

$$\begin{array}{ll} \text{expliziter Euler} & \Phi_h \left(x^{(j)}, y_h^{(j)}, x^{(j+1)}, y_h^{(j+1)} \right) = f(x^{(j)}, y_h^{(j)}) \\ \text{impliziter Euler} & \Phi_h \left(x^{(j)}, y_h^{(j)}, x^{(j+1)}, y_h^{(j+1)} \right) = f(x^{(j+1)}, y_h^{(j+1)}) \end{array}$$

Φ_h heißt *Verfahrensfunktion*.

Eine wichtige Rolle bei der Analyse der Verfahren spielt der

Definition 16.1 (Lokaler Diskretisierungsfehler)**.** Die Größe

$$\tau_h^{(j)} := \frac{y(x^{(j+1)}) - y(x^{(j)})}{h^{(j+1)}} - \Phi_h(x^{(j)}, y(x^{(j)}), x^{(j+1)}, y(x^{(j+1)}))$$

heißt lokaler Diskretisierungsfehler. □

τ_h^j entsteht durch Einsetzen der *exakten Lösung* y an den Gitterpunkten in die Verfahrensfunktion:

$$\underbrace{y(x^{(j)})}_{\text{exakte Lsg. in } x^{(j)}} + h^{(j+1)} \underbrace{\Phi_h(x^{(j)}, y(x^{(j)}), x^{(j+1)}, y(x^{(j+1)}))}_{\Phi_h \text{ mit exakten Werten ausgewertet}} = \underbrace{u^*}_{\text{nach einem Schritt}}.$$

Nun vergleiche u^* mit dem exakten Wert $y(x^{(j+1)})$ und teile durch h:

$$\frac{y(x^{(j+1)}) - u^*}{h^{(j+1)}} = \frac{y(x^{(j+1)}) - y(x^{(j)}) - h^{(j+1)} \Phi_h \left(x^{(j)}, y(x^{(j)}), x^{(j+1)}, y(x^{(j+1)}) \right)}{h^{(j+1)}} = \tau_h^{(j)}.$$

Beispiel 16.2. Für den expliziten Euler $\Phi_h(x, y, x', y') = f(x, y)$ erhalten wir:

$$\begin{aligned} \tau_h^{(j)} &= \frac{y(x^{(j+1)}) - y(x^{(j)})}{h^{(j+1)}} - \Phi_h(x^{(j)}, y^{(j)}, x^{(j+1)}, y^{(j+1)}) \\ &= \frac{y(x^{(j+1)}) - y(x^{(j)})}{h^{(j+1)}} - f(x^{(j)}, y^{(j)}) \\ &= \underbrace{y'(x^{(j)}) + O(h)}_{\text{Taylor, (15.1)}} - f(x^{(j)}, y^{(j)}) \\ &= O(h) \end{aligned}$$

Für die Trapezregel $\Phi_h(x,y,x',y') = \frac{1}{2}\left(f(x,y)+f(x',y')\right)$ erhalten wir

$$\begin{aligned}\tau_h^{(j)} &= \frac{y(x^{(j+1)})-y(x^{(j)})}{h^{(j+1)}} - \frac{1}{2}\left(f(x^{(j)},y^{(j)})+f(x^{(j+1)},y^{(j+1)})\right)\\ &= \frac{1}{h^{(j+1)}}\underbrace{\left[\int_{x^{(j)}}^{x^{(j+1)}} f(x,y(x))dx - \frac{h^{(j+1)}}{2}\left(f(x^{(j)},y^{(j)})+f(x^{(j+1)},y^{(j+1)})\right)\right]}_{\text{Trapezregel } O(h^3),\ (15.4)}\\ &= O(h^2)\end{aligned}$$

□

Dies führt zu den beiden folgenden Definitionen

Definition 16.3 (Konsistenz)**.** Man nennt ein Verfahren *konsistent*, falls für den lokalen Diskretisierungsfehler gilt

$$\max_j \|\tau_h^{(j)}\| \le \gamma(h) \qquad \text{mit } \lim_{h\to 0}\gamma(h) = 0.$$

Hierbei ist $\|.\|$ im Fall von Systemen eine beliebige Vektornorm, sonst der Betrag. □

Definition 16.4 (Konsistenzordnung)**.** Ein Verfahren heißt *konsistent von der Ordnung p*, falls

$$\max_j \|\tau_h^{(j)}\| \le Kh^p \qquad \text{für } h \le h_0$$

und K unabhängig von j und h. □

Definition 16.5 (Konvergenz)**.** Schließlich heißt ein Verfahren *konvergent*, falls gilt

$$\max_j \|y_h^{(i)} - y(x^{(i)})\| \le \gamma(h) \text{ mit } \lim_{h\to 0}\gamma(h) = 0.$$

Dies ist nicht unbedingt das selbe γ wie oben. □

Definition 16.6 (Konvergenzordnung)**.** Ein Verfahren hat die *Konvergenzordnung p*, falls

$$\max_j \|y_h^{(i)} - y(x^{(i)})\| \le Kh^p \text{ für } h \le h_0.$$

Dies ist nicht unbedingt das selbe K wie oben. □

Man bezeichnet die Konsistenz auch als *lokale* Konvergenz und die eben definierte Konvergenz als *globale* Konvergenz.

Mit diesen Definitionen formuliert man den folgenden Satz.

Satz 16.7 (Konvergenzsatz)**.** Das Einschrittverfahren sei konsistent von der Ordnung p und die Funktion Φ_h erfülle die Lipschitzbedingung

$$\|\Phi_h(x,y,x',y') - \Phi_h(x,\tilde{y},x',\tilde{y}')\| \le L\max(\|y-\tilde{y}\|, \|y'-\tilde{y}'\|). \tag{16.2}$$

Dann gilt für einen *festen* Punkt $\bar{x} \in [a,b]$ und $h = \frac{\bar{x}-a}{N}$ (also $\bar{x} = a+Nh$) die Abschätzung

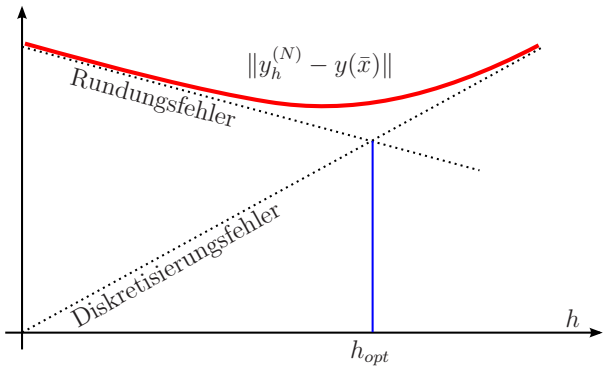
$$\|y_h^{(N)} - y(\bar{x})\| \le ch^p \frac{e^{L|\bar{x}-a|}-1}{L}. \tag{16.3}$$

Wichtig ist hierbei, dass $\bar{x} \in [a,b]$ fest gewählt ist und $h \to 0$ geht. □

Die Lipschitzbedingung der Verfahrensfunktion folgt üblicherweise direkt aus der Lipschitzbedingung von f, die man ohnehin für die Eindeutigkeit braucht.

Dieses Resultat bedeutet: Bei Einschrittverfahren ist die Konvergenzordnung gleich der Konsistenzordnung.

Bemerkung 16.8. Bei Fließkomma-Arithmetik fester Stellenzahl gibt es eine optimale Schrittweite.



□

16.2 Runge-Kutta-Verfahren

Wie konstruiert man nun Einschrittverfahren hoher Ordnung?

Allgemeine *Runge*[35]*-Kutta*[36]*-Verfahren* haben die Form

$$y_h^{(j+1)} = y_h^{(j)} + h^{(j+1)} \sum_{l=1}^{m} \gamma_l k_l(x^{(j)}, y_h^{(j)}) \text{ mit}$$

$$k_l(x^{(j)}, y_h^{(j)}) = f\left(x^{(j)} + \alpha_l h^{(j+1)}, y_h^{(j)} + h^{(j+1)} \sum_{r=1}^{m} \beta_{lr} k_r(x^{(j)}, y_h^{(j)})\right).$$

m ist die *Stufenzahl*, γ_l, α_l, β_{lr} parametrisieren die Verfahren.

Man unterscheidet folgende Klassen:

explizit $\beta_{lr} = 0$ für $r \geq l$.

diagonal implizit $\beta_{lr} = 0$ für $r > l$. Erfordert das m-malige Lösen eines nichtlinearen Gleichungssystems wie bei implizitem Euler.

implizit Dies erfordert das Lösen eines nichtlinearen Systems der m-fachen Größe wie bei implizitem Euler.

[35] Carl Runge, 1856-1927, dt. Mathematiker
[36] Martin Wilhelm Kutta, 1867-1944, dt. Mathematiker

Beispiel 16.9 (Einige Runge-Kutta-Verfahren)**. Verfahren von Heun.** Dieses explizite zweistufige Verfahren der Ordnung 2 lautet

$$\begin{aligned} y_h^{(j+1)} &= y_h^{(j)} + h^{(j+1)} \left(\frac{1}{2}k_1 + \frac{1}{2}k_2 \right) \\ k_1 &= f(x^{(j)}, y_h^{(j)}) \\ k_2 &= f(x^{(j+1)}, y_h^{(j)} + h^{(j+1)}k_1). \end{aligned}$$

$\frac{1}{2}k_1 + \frac{1}{2}k_2$ stellt ein „verbesserte Steigung“ dar.

Impliziter Euler. Kann als einstufiges implizites Runge-Kutta-Verfahren der Ordnung 1 geschrieben werden:

$$\begin{aligned} y_h^{(j+1)} &= y_h^{(j)} + h^{(j+1)} f\left(x^{(j+1)}, y_h^{(j+1)}\right) = y_h^{(j)} + h^{(j+1)}k_1 \\ \text{mit} \quad k_1 &= f\left(x^{(j)} + h^{(j+1)}, y_h^{(j)} + h^{(j+1)}k_1\right). \end{aligned}$$

Wie versteht man das? Impliziter Euler löst das System

$$y_h^{(j+1)} - h^{(j+1)} f\left(x^{(j+1)}, y_h^{(j+1)}\right) = y_h^{(j)}$$

nach $y_h^{(j+1)}$ auf. Setzen wir formal $y_h^{(j+1)} = y_h^{(j)} + h^{(j+1)}k_1$ in diese Beziehung ein, so ist dies äquivalent zu

$$k_1 = f\left(x^{(j+1)}, y_h^{(j)} + h^{(j+1)}k_1\right)$$

wie oben behauptet.

Verfahren von Alexander. Zweistufiges diagonal implizites Runge-Kutta-Verfahren der Ordnung 2:

$$\begin{aligned} y_h^{(j+1)} &= y_h^{(j)} + h^{(j+1)} \left[(1-\alpha)k_1 + \alpha k_2\right] \\ k_1 &= f\left(x^{(j)} + \alpha h^{(j+1)}, y_h^{(j)} + \alpha h^{(j+1)}k_1\right) \\ k_2 &= f\left(x^{(j)} + h^{(j+1)}, y_h^{(j)} + h^{(j+1)}[(1-\alpha)k_1 + \alpha k_2]\right) \end{aligned}$$

mit $\alpha = 1 - \sqrt{2}/2$.

Klassisches Runge-Kutta-Verfahren. Vierstufiges explizites Verfahren der Ordnung 4:

$$\begin{aligned} y_h^{(j+1)} &= y_h^{(j)} + \frac{h^{(j+1)}}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(x^{(j)}, y_h^{(j)}) \\ k_2 &= f(x^{(j)} + \frac{h^{(j+1)}}{2}, y_h^{(j)} + \frac{h^{(j+1)}}{2}k_1) \\ k_3 &= f(x^{(j)} + \frac{h^{(j+1)}}{2}, y_h^{(j)} + \frac{h^{(j+1)}}{2}k_2) \\ k_4 &= f(x^{(j)} + h^{(j+1)}, y_h^{(j)} + h^{(j+1)}k_3). \end{aligned}$$

□

16.3 Verfahrensstabilität

Beispiel 16.10. Wir wenden das explizite bzw. implizite Euler Verfahren auf das berühmte Modellproblem

$$y'(x) = \lambda y(x), \qquad \mathbf{R} \ni \lambda < 0$$

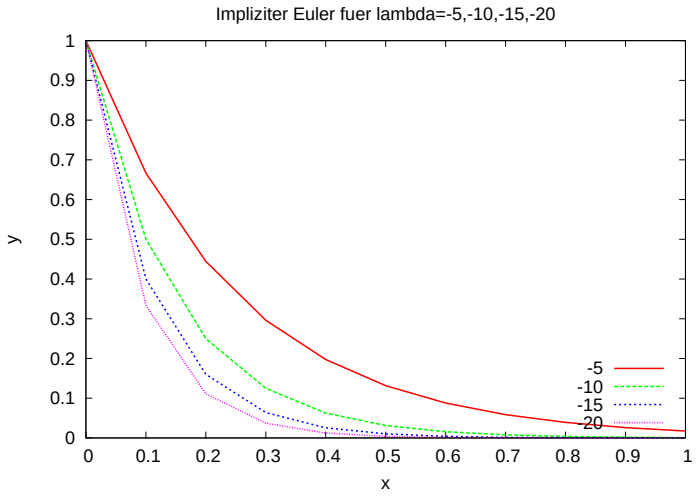
an.

Für $\lambda < 0$ sind die Lösungen unterschiedlich schnell abklingende e-Funktionen.

Die numerischen Verfahren lauten wie bekannt:

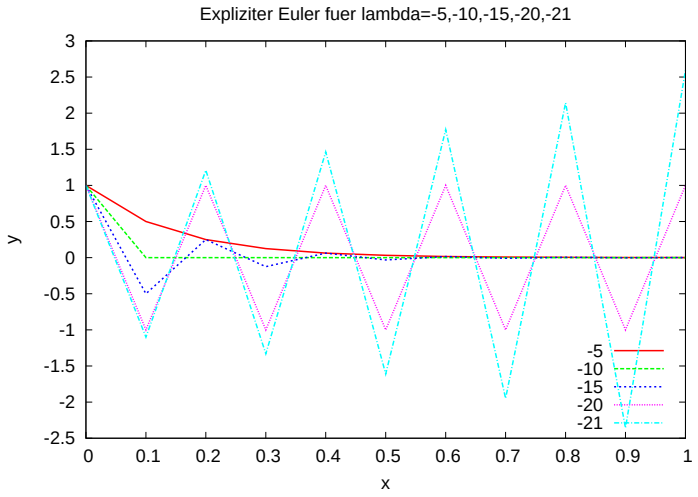
$$y_h^{(j+1)} = (1 + h\lambda) y_h^{(j)}, \qquad y_h^{(j+1)} = \left(\frac{1}{1 - h\lambda}\right) y_h^{(j)}.$$

Zunächst der *implizite* Euler für $h\lambda = -0.5, -1, -1.5, -2$:



Das Verfahren zeigt qualitativ das richtige Verhalten.

Nun der *explizite* Euler für $h\lambda = -0.5, -1, -1.5, -2, -2.1$:



Das Verfahren zeigt für große $|h\lambda|$ qualitativ das falsche Verhalten. Woran liegt das? □

Die Rekursionsgleichung für den expliziten Euler lautet

$$y_h^{(j+1)} = (1 + h\lambda) y_h^{(j)}.$$

Im Fall $\lambda < 0, y(0) > 0$ sind die Lösungen $y(x)$ monoton fallend und positiv.

Die numerische Lösung ist für $h\lambda < 0$ *nicht wachsend*, nur dann wenn

$$|1 + h\lambda| < 1 \quad \Leftrightarrow \quad 1 + h\lambda > -1 \quad \Leftrightarrow \quad h < -\frac{2}{\lambda}.$$

Die Schrittweit h muss also genügend klein sein bei gegebenem λ.

Für den impliziten Euler dagegen gilt folgendes:

$$y_h^{(j+1)} = \frac{1}{1 - h\lambda} y_h^{(j)}$$

und

$$\left| \frac{1}{1 - h\lambda} \right| < 1 \quad \Leftrightarrow \quad |1 - h\lambda| > 1 \to 1 - h\lambda > 1 \text{ gilt } \forall h > 0, \lambda < 0.$$

Im allgemeinen kann $\lambda \in \mathbf{C}$ sein. Dies führt zu der folgenden Definition.

Definition 16.11 (A-Stabilität)**.** Ein numerisches Verfahren heißt absolut stabil, kurz A-stabil, wenn es angewandt auf das AWP

$$y' = \lambda y, \quad y(0) = Y, \quad \lambda \in \mathbf{C}, \Re(\lambda) \leq 0$$

für jede Schrittweite h eine nicht wachsende Folge

$$|y_h^{(j)}| \leq |Y|, \quad j \geq 0$$

liefert. □

Bemerkung 16.12. Über die bisher behandelten Verfahren lässt sich folgendes sagen:

- Kein explizites Runge-Kutta-Verfahren ist A-stabil.
- Kein explizites lineares Mehrschrittverfahren ist A-stabil
- Impliziter Euler, Trapezregel, Alexander und BDF-2 sind A-stabil.
- Es gibt kein implizites, A-stabiles lineares Mehrschrittverfahren mit Konsistenzordnung größer 2.
- Es gibt A-stabile implizite Runge-Kutta-Verfahren beliebig hoher Ordnung.

Es gibt außerdem noch eine ganze Reihe weiterer Stabilitätsdefinitionen.

□

16.4 Steife Systeme

Beispiel 16.13 (aus [Sim])**.** Das lineare *System* gewöhnlicher Differentialgleichungen

$$\begin{pmatrix} y_1'(x) \\ y_2'(x) \end{pmatrix} = \begin{pmatrix} 998 & 1998 \\ -999 & -1999 \end{pmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}, \qquad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

hat die Lösung

$$y_1(x) = 2e^{-x} - e^{-1000x}, \qquad y_2(x) = -1e^{-x} + e^{-1000x}.$$

Die Lösung besteht aus zwei Anteilen: e^{-x} langsam abklingend, e^{-1000x} schnell abklingend.

Löst man mit explizitem Euler, so muss gelten:

$$h < \min\left(-\frac{2}{-1}, -\frac{2}{-1000}\right) = \frac{1}{500}.$$

Andererseits ist für $x = \frac{2}{100}$ bereits

$$e^{-1000\frac{2}{100}} = e^{-20} \approx 2.06 \cdot 10^{-9}, \quad e^{-\frac{2}{100}} \approx 0.98$$

d.h. man „sieht“ in der Lösung nach kurzer Zeit nur den langsamen Anteil und möchte mit großem h rechnen. □

Systeme mit sehr unterschiedlich schnellen Lösungsanteilen heißen *steif*. Es existieren allerdings verschiedene Definitionen, der Begriff ist schwer zu fassen.

Zur numerischen Lösung steifer Systeme benötigt man möglichst stabile Verfahren, etwa A-stabile Verfahren.

Steife Systeme werden deshalb besser mit impliziten Verfahren gelöst.

16.5 Inhärente Instabilität

Satz 14.5 sagte: Seien $y(x), z(x)$ Lösungen eines AWPs zu *zwei verschiedenen* Startwerten Y, Z so gilt

$$\|y(x) - z(x)\| \leq \|Y - Z\| e^{L|x-a|}.$$

Wir zeigen per Beispiel, dass diese Abschätzung scharf ist.

Sei $F \in C^1[a, b]$ beliebig. Das AWP

$$y'(x) = \lambda\{y(x) - F(x)\} + F'(x), \qquad y(a) = Y$$

hat die Lösung

$$y(x) = (Y - F(a))\, e^{\lambda(x-a)} + F(x)$$

Beweis: Einsetzen liefert

$$\begin{aligned} y'(x) &= (Y - F(a))\,\lambda e^{\lambda(x-a)} + F'(x) \\ &= \lambda\{\underbrace{(Y - F(a))e^{\lambda(x-a)} + F(x)}_{=y(x)} - F(x)\} + F'(x) \\ \text{sowie } y(a) &= (Y - F(a)) \cdot 1 + F(a) = Y. \end{aligned}$$

Für den speziellen Anfangswert $Y = F(a)$ gilt

$$y(x) = F(x).$$

Für $Y = F(a) + \varepsilon$ hingegen gilt

$$y_\varepsilon(x) = \varepsilon e^{\lambda(x-a)} + F(x).$$

Für $\lambda > 0$ laufen die Lösungen exponentiell auseinander, das Problem ist schlecht konditioniert.

Unabhängig vom verwendeten Verfahren kann man das Problem nur für relativ kleine Intervalle $[a, b]$ befriedigend genau lösen.

Man sagt, das Problem sei *inhärent instabil*.

16.6 Dynamische Systeme

Zum Schluss wollen wir noch einen kleinen Ausflug in die Welt der *dynamischen Systeme* machen.

Darunter versteht man das Studium des *qualitativen Verhaltens* nichtlinearer Systeme gewöhnlicher Differentialgleichungen.

Betrachten wir etwa das *autonome* System

$$y'(x) = f(y(x)), \qquad y(a) = Y, \qquad y(x) \in \mathbf{R}^n.$$

(autonom: f hängt nicht von x ab).

Zunächst bestimmt man sog. *kritische Punkte* oder *Knoten* in denen gilt

$$f(y_s) = 0.$$

In y_s ist die Lösung stationär, d.h. sie ändert sich nicht.

Nun studiert man das Verhalten der Lösung in der Umgebung der stationären Punkte durch lineare Stabilitätsanalyse:

Stabiler Punkt Alle Lösungen in der Nähe des Punktes laufen in diesen hinein. Beispiel: Abnehmende Population.

Instabiler Punkt Alle Lösung in der Nähe des Punktes laufen von diesem weg. Beispiel: Auf der Spitze stehendes Pendel mit starrer Stange.

Sattelpunkt Es gibt in der Umgebung des Punktes Lösungen, die sowohl hin als auch weg laufen.

Zentrumsknoten Lösungen laufen periodisch um den stationären Punkt. Beispiel: Pendel ohne Reibung.

Stabiler Strudelpunkt Lösungen laufen oszillatorisch auf den stationären Punkt zu. Beispiel: Pendel mit Dämpfung.

Instabiler Strudelpunkt Lösungen laufen oszillatorisch vom stationären Punkt weg.

All diese Phänomene kann man bereits bei linearen Systemen $y(x)' = Ay(x)$ beobachten.

Bei *nichtlinearen* Systemen gibt es noch weitere Möglichkeiten, etwa den *Grenzzyklus*: Periodische Bewegung, in die Lösungen hineinlaufen (stabiler Grenzzyklus) oder von dem Lösungen weglaufen (instabiler Grenzzyklus).

Kritische Punkte und Grenzzyklen heißen auch *Attraktoren* des dynamischen Systems.

Besondere Attraktoren stellen die *strange attractors* dar, die bei *chaotischen* Systemen auftreten können. Hier nähern sich die Lösungen einer Punktmenge, sind aber nicht periodisch.

Wir werden nun einige Beispiele für das komplexe Verhalten nichtlinearer Systeme geben.

Beispiel 16.14. Wir betrachten das gravitative N-Körper-Problem.

Für $i = 1, \dots, N$ suchen wir $x_i : [a, b] \to \mathbf{R}^3$ und $v_i : [a, b] \to \mathbf{R}^3$:

$$\begin{aligned} \frac{dx_i(t)}{dt} &= v_i(t), & x_i(a) &= X_i, \\ \frac{dv_i(t)}{dt} &= \sum_{\substack{1 \le j \le N \\ j \ne i}} \frac{\gamma m_j (x_j - x_i)}{\|x_j - x_i\|^3}, & v_i(a) &= V_i. \end{aligned}$$

Das System hat keine stationären Punkte, da die Kraft nicht Null werden kann.

Die Bilder zu folgenden Aussagen findet man in Abbildung 16.6.

Im Fall $N = 2$ gibt es im wesentlichen zwei Typen von Lösungen. Die beiden Körper sind gravitativ gebunden und umkreisen sich.

Oder sie entfernen sich voneinander bei genügend hoher Geschwindigkeit.

Auch im Fall $N = 3$ gibt es periodische Lösungen. Hier ist eine davon: Jeder Körper läuft auf einer Ellipsenbahn.

Hier ist noch eine: Die Körper (gleicher Masse) laufen auf einer 8-förmigen Bahn. Diese Lösung ist stabil und wurde erst 1993 gefunden!

Meistens sind die Lösung aber instabil und ein Körper wird aus dem System geschleudert.

Beim eingeschränkten 3-Körper-Problem ist ein Körper sehr viel leichter als als die anderen beiden (die einander umkreisen).

Beim 3-Körper-Problem können chaotische Lösungen auftreten. Ist diese Lösung periodisch oder chaotisch?

□

Beispiel 16.15 (Lorenz[37]-System)**.** Das klassische Beispiel für ein chaotisches dynamisches System ist das Lorenz-System. Es lautet:

$$\begin{aligned} y_1'(x) &= -10y_1(x) + 10y_2(x) \\ y_2'(x) &= 28y_1(x) - y_2(x) - y_1(x)y_3(x) \\ y_3'(x) &= y_1(x)y_2(x) - \frac{8}{3}y_3(x). \end{aligned}$$

Als Startwert verwenden wir $Y = (1, 2, 3)^T$.

Die Bilder zu folgenden Aussagen findet man in Abbildung 16.6.

Die Lösung umkreist zwei Punkte im $\mathbf{R}^3$, wobei z. B. die Zeiten zu denen die Lösung von der Umkreisung des einen auf die Umkreisung des anderen Punktes sehr sensitiv von den Anfangsbedingungen abhängt.

Hier sieht man die Lösung zu zwei leicht verschiedenen Anfangsbedingungen übereinandergezeichnet. Nach relativ kurzer Zeit haben sich die Lösungen weit auseinanderentwickelt. Das System ist inhärent instabil.

□

Bei chaotischen Systemen führen winzige Unterschiede in den Anfangsbedingung nach einer gewissen Zeit zu sehr unterschiedlichen Lösungen (sog. *Schmetterlingseffekt*).

Sie lassen sich deshalb nur für eine gewisse Zeit vorhersagen. So ist etwa das Wetter ein chaotisches System (Lorenz hat sein System in den 60er Jahren durch starke Vereinfachung der Navier-Stokes-Gleichungen, der Grundgleichung der Strömungsmechanik, entwickelt).

Diese Systeme erscheinen zufällig im Sinne von nicht vorhersagbar.

Wichtig ist jedoch, dass diese Systeme vollkommen *deterministisch* sind. Startet man mit gleichen Anfangswerten, erhält man jedes Mal dieselbe Lösung!

Eine Möglichkeit, trotzdem zu gewissen Aussagen über solche System zu gelangen, stellt die Stochastik dar, zu der wir nun kommen werden.

Im Zusammenhang mit dem Wetter gibt etwa der „Hundertjährige Kalender“ Aussagen über das mittlere Verhalten des Systems.

16.7 Zusammenfassung

- Der lokale Diskretisierungsfehler gibt Auskunft über das Konvergenzverhalten eines Einschrittverfahrens (sofern man die Lipschitzstetigkeit der Verfahrensfunktion voraussetzt).

[37] Edward Norton Lorenz, geb. 1917, amerk. Mathematiker und Meteorologe.

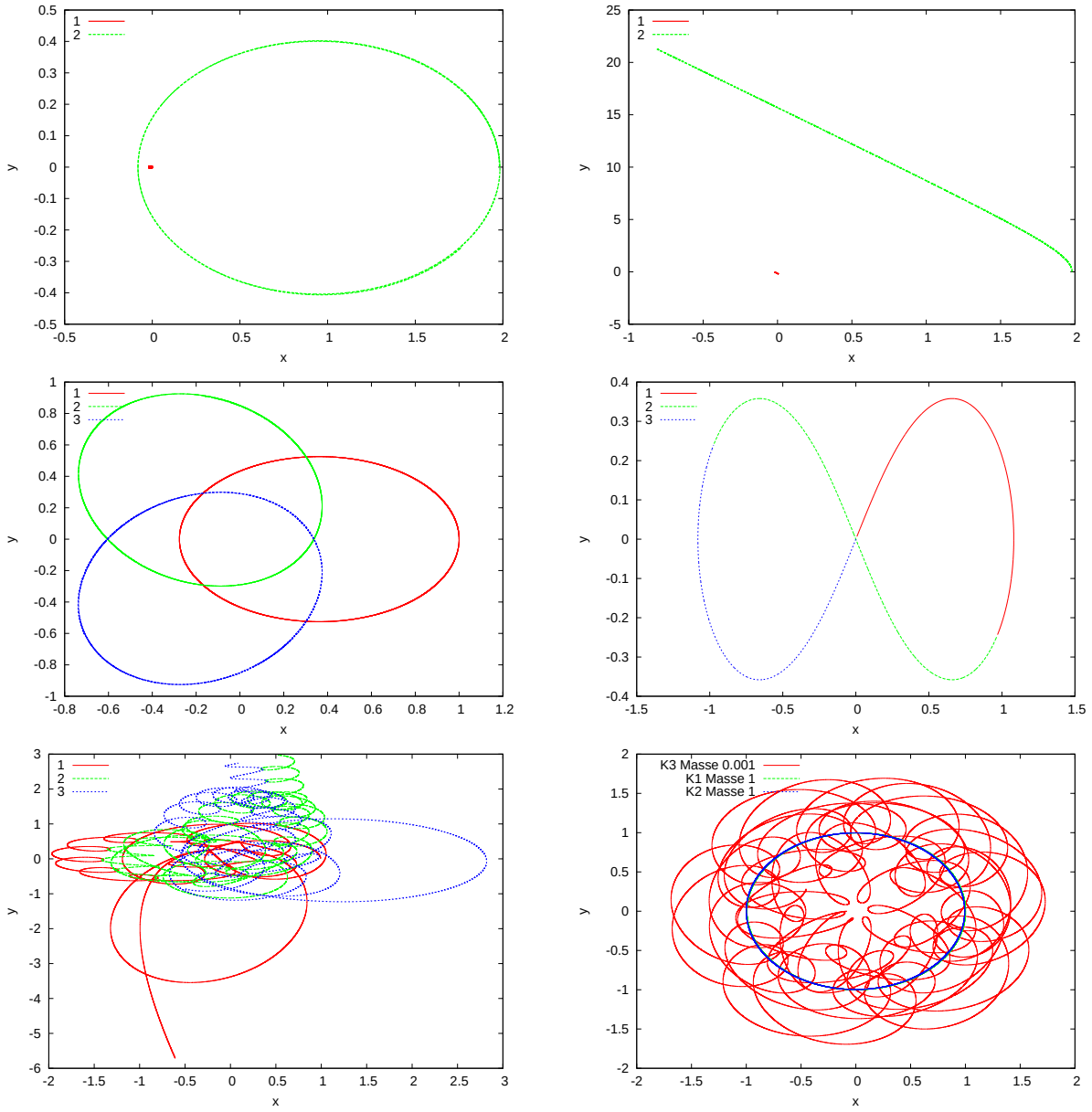


Abbildung 26: Einige Lösungen des N-Körper-Problems. Von links oben nach rechts unten: 2 Körper umkreisen sich, 2 ungebundene Körper, stabile Lösung des 3-Körper-Problems, die Figur-8-Lösung von 1993, instabiles 3-Körper-Problem und Lösung des eingeschränkten 3-Körper-Problems.

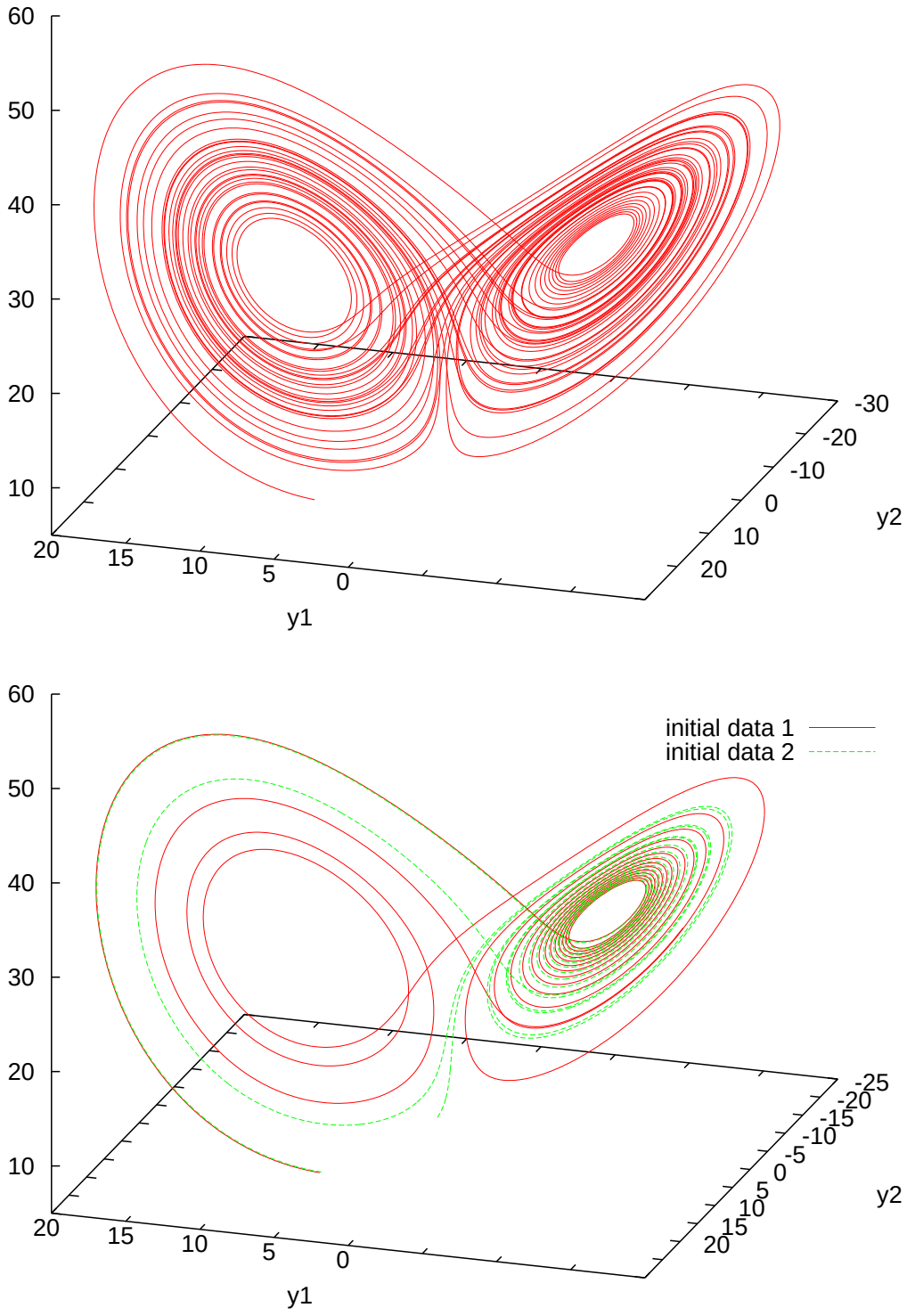


Abbildung 27: Lösung des Lorenz-Systems. Oben: Eine Lösungstrajektorie bis $T = 50$. Unten: Lösungen zu den Startwerten $Y = (1, 2, 3)^T$ und $Z = (1.5, 2, 3)^T$ bis $T = 12.5$.

- Ein konvergentes Verfahren liefert für $h \to 0$ immer bessere Approximationen an die Lösung der Differentialgleichung.
- Bei manchen Verfahren muss die Schrittweite h genügend klein sein, damit sie nicht offensichtlich falsche Ergebnisse liefern.
- Bei steifen Problemen sind solche Verfahren in der Regel ungeeignet, da sie dann sehr ineffizient werden.
- Schließlich haben wir noch einige Beispiele für dynamische Systeme betrachtet. Inhärent instabile Systeme lassen sich nur über einen begrenzten Zeitraum mit ausreichender Genauigkeit vorhersagen.

17 Einführung in die Wahrscheinlichkeitstheorie

Hinweis: Dieser Teil der Vorlesung stützt sich, im Gegensatz zum Numerikteil, stark auf ein einzelnes Buch, nämlich

Schickinger, T. und A. Steger: *Diskrete Strukturen 2.* Springer, 2002.

Daneben kann man auch noch heranziehen:

Teschl, G. und S. Teschl: *Mathematik für Informatiker, Band 2: Analysis und Statistik.* Springer, 2. Auflage, 2007.

17.1 Determinismus und Zufall

Bis jetzt haben wir ausschließlich deterministische Modelle betrachtet.

Beispiele waren: Pendel, N-Körper-Problem. Die gewöhnliche DGL hat genau eine Lösung! Ob wir diese numerische genau genug berechnen können, ist eine andere Sache.

Bei (sehr) vielen praktischen Anwendungen ist ein solches Vorgehen jedoch nicht möglich.

Wir wollen nun verschiedene Arten des Zufalls beschreiben.

Beispiel 17.1. Beobachte *ein* Atom eines radioaktiven Elements und stelle fest, ob es innerhalb der nächsten Minute zerfällt.

Die Physik sagt: Man kann den Ausgang dieses Experiments nicht hundertprozentig vorhersagen, da sich die Anfangsbedingungen nicht exakt messen lassen (Unschärferelation).

Besser ist es mit Wahrscheinlichkeiten zu operieren.

Man nennt dies „natürlichen Zufall". □

Bemerkung 17.2. Normalerweise beobachtet man *viele* Atome gleichzeitig. Dies ist ein völlig anderes Problem!

Das Verhalten vieler Atome *im Mittel* lässt sich sehr gut mit deterministischen Methoden vorhersagen, wie die statistische Physik lehrt.

In diesem Fall ist das die lineare Differentialgleichung $y' = \lambda y$.

Weitere Beispiele für natürlichen Zufall sind die inhärent instabilen Probleme, etwa chaotische Systeme. Hier verhindern unvermeidbare Ungenauigkeiten in der Erfassung der Anfangsbedingungen eine zuverlässige Vorhersage. □

Eine andere Art von Zufall steckt in folgendem Beispiel.

Beispiel 17.3. Untersuche ob Prof. Bastian in der Klausur etwas zu Lagrange-Interpolation frägt.

Hier kann es sein, dass Prof. Bastian dieses schon weiss, aber den Studenten nützt dies nichts, da sie es ja nicht wissen.

Dies ist „Zufall aus Unsicherheit bzw. mangelndem Wissen". □

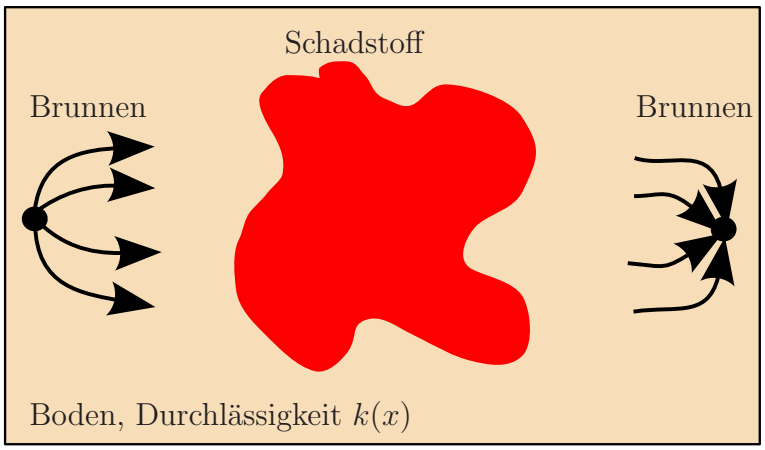
Von dieser Art sind viele Probleme in der Informatik

- Antwortzeit eines Großrechners oder Servers?
- Steigert eine Programmmodifikation den Durchsatz eines Routers?
- Wie ist ein Rechnersystem fehlertolerant auszulegen?

Das Problem liegt hier, wie oben, an mangelndem Wissen, welche Situationen auf das Rechensystem zukommen werden.

Kombination deterministischer und stochastischer Aspekte.

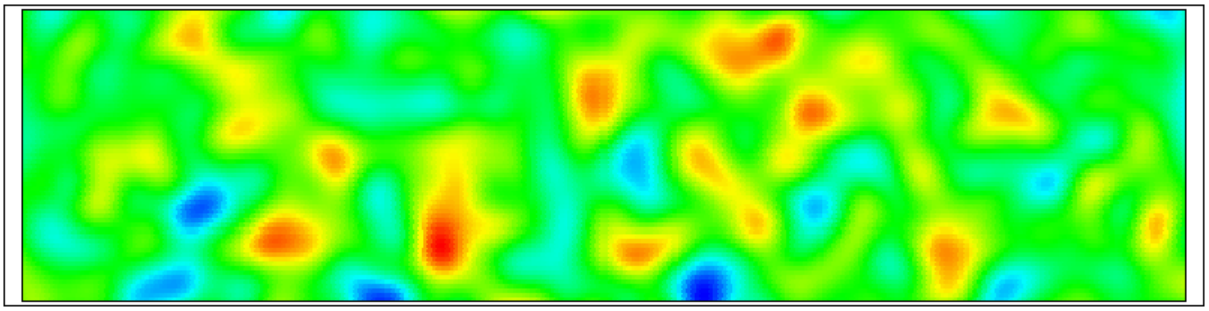
Beispiel 17.4 (Reinigung eines kontaminierten Grundwasserleiters)**.** Verunreinigter Boden kann in manchen Fällen mittels der „Pump and Treat“ Strategie gereinigt werden. Dazu pumpt man (oft viele Jahre!) Wasser durch den Boden, welches anschließend gereinigt wird.



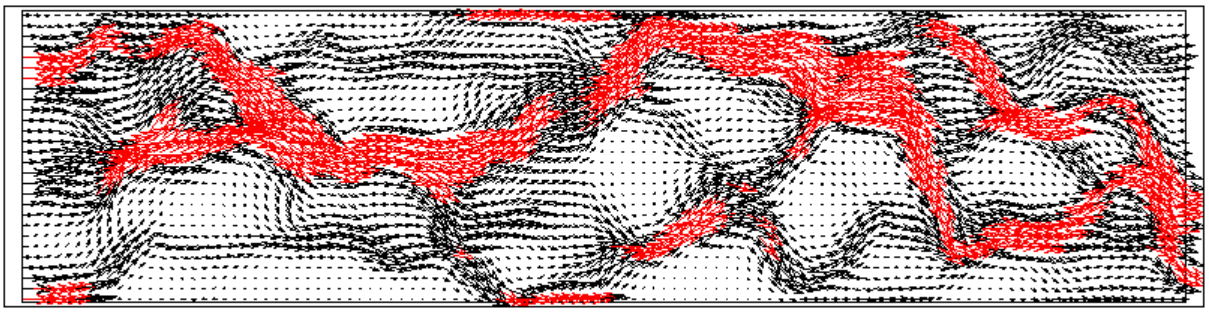
Die Strömung durch den Boden lässt sich recht gut mit einem deterministischen Modell, einer partiellen Differentialgleichung beschreiben.

Allerdings enthält dieses Modell als Parameter die *Durchlässigkeit* $k(x)$ des Bodens an der Stelle x.

Für eine Realisierung des Permeabilitätsfeldes



lässt sich die Grundwasserströmung berechnen:



Will man etwa die Frage beantworten

> Wie groß ist die Wahrscheinlichkeit, dass 90% des Schadstoffes nach 1 Jahr ausgewaschen sind?

so kann man

- eine Reihe möglicher (d. h. physikalisch sinnvoller) Durchlässigkeitsfelder erzeugen
- sowie für jede dieser Realisierungen die Schadstoffauswaschung berechnen.
- Hiermit kann man dann obige Frage beantworten.

Dies nennt man auch „Monte-Carlo-Methode“. □

Ein andere Anwendung von Zufall in der Informatik sind *Randomisierte Algorithmen*.

Hier wird der Ablauf eines Algorithmus durch den Zufall gesteuert.

Beispiele sind:

- Quicksort: Hier wird durch die Wahl des Pivotelements eine Menge von Zahlen in zwei Teile zerlegt, die dann rekursiv sortiert werden.

  Zur Analyse des Algorithmus muss man mit Wahrscheinlichkeiten operieren.
- Kollisionsauflösung im ursprünglichen Ethernet.

In manchen Fällen sind solche Algorithmen einfacher und/oder besser als deterministische Algorithmen.

Stochastik, Statistik, Wahrscheinlichkeitsrechnung: Oft werden diese Begriffe durcheinander gebraucht. Hier der Versuch einer Erklärung.

Wahrscheinlichkeitstheorie: Teilgebiet der Mathematik, formale Untersuchung zufallsbeeinflusster Vorgänge. Zufall im Sinne von nicht vorhersagbar.

Statistik: Beobachtend, beschreibend

deskriptive Statistik: Verständliche Aufbereitung großer Datenmengen, Mittelwerte, Streuung, Diagramme, ...

induktive Statistik Schließen von „Stichproben“(beobachteten Ausprägungen von Zufallsgrößen) auf zugrundeliegende Gesetzmäßigkeiten (z.B. Parameter von Verteilungen).

Stochastik:

- nicht einheitlich
- oft synonym zu Wahrscheinlichkeitstheorie

- Hübner: *neuere* Bezeichnung der Wahrscheinlichkeitstheorie
- Oberbegriff von Wahrscheinlichkeitstheorie und Statistik

17.2 Zufallsexperiment und Wahrscheinlichkeitsraum

Dieser Abschnitt beschreibt die grundsätzliche Herangehensweise bei der Modellierung unsicherer Situationen mittels stochastischer Methoden.

Der Gegenstand der Untersuchung ist ein sog. „Zufallsexperiment".

Definition 17.5 (Zufallsexperiment)**.** Ein Zufallsexperiment ist ein Vorgang, der

- beliebig oft unter gleichen Bedingungen wiederholt werden kann und
- dessen Ergebnis nicht mit Sicherheit vorhergesagt werden kann. □

Die mathematische Modellierung formalisiert das Zufallsexperiment mittels eines „Wahrscheinlichkeitsraumes" $(\Omega, \mathcal{A}, \Pr)$ aus drei Komponenten:

1. Die möglichen Experimentausgänge Ω.
2. Die zu untersuchende Fragenstellung (Ereignis) $\mathcal{A}$.
3. Die zugehörigen Wahrscheinlichkeiten Pr.

Diese Komponenten führen wir nun der Reihe nach ein.

Die Ergebnismenge Ω beschreibt alle *möglichen Ausgänge* eines Zufallsexperiments.

Am besten erklären wir das mit einem

Beispiel 17.6. Einige mögliche Ergebnismengen sind:

- Wetter zu einem gegebenen Zeitpunkt: $\Omega = \{\text{sonnig}, \text{bewölkt}, \text{regnerisch}\}$.
- Einmaliges Werfen einer Münze: $\Omega = \{\text{Kopf}, \text{Zahl}\}$.
- Einmaliges Werfen eines sechsseitigen Würfels: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Antwortzeit eines Servers in Sekunden. Hier gäbe es verschiedene Möglichkeiten:
  - $\Omega = \mathbb{N}$, oder
  - $\Omega = \{1, 2, \ldots 100\}$, „100" bedeutet die Antwortzeit ist größer 100 Sekunden, oder
  - $\Omega = \mathbb{R}_+$.

□

Man unterscheidet prinzipiell

- endliche,
- abzählbar unendliche und
- überabzählbare

Ergebnismengen.

Endliche und abzählbar unendliche Ergebnismengen bezeichnet man auch als „diskret". Die darauf basierenden Wahrscheinlichkeitsräume heißen dann diskret.

Überabzählbare Ergebnismengen bezeichnet man auch als „kontinuierlich“. Die darauf basierenden Wahrscheinlichkeitsräume heißen dann kontinuierlich.

Wir werden in dieser Vorlesung vorwiegend diskrete Wahrscheinlichkeitsräume betrachten.

Jedes „Ereignis“ A ist eine Teilmenge der Ergebnismenge Ω:

$$A \subset \Omega.$$

Man sagt, „das Ereignis A tritt ein“, falls das Zufallsexperiment ein Ergebnis

$$a \in A$$

liefert.

Beispiel 17.7. Passend zu unserem obigen Beispiel hätten wir

- $A = \{\text{sonnig}, \text{bewölkt}\}$. Tritt A ein, so braucht man keinen Schirm auf den Spaziergang mitzunehmen.
- $A = \{1, 3, 5\}$. Der Würfel zeigt eine ungerade Augenzahl an.
- $A = \{10, 11, 12, 13, 14\}$. Die Antwortzeit des Servers liegt zwischen 10 und 14 Sekunden. Alternativ wäre $A = [10, 14]$ möglich, falls $\Omega = \mathbb{R}_+$ ist.

□

Einige Ereignisse haben spezielle Namen:

- $A = \emptyset$ heißt das „unmögliche“Ereignis. Es tritt nie ein.
- $A = \Omega$ heißt das „sichere“ Ereignis. Es tritt immer ein.
- $A = \{\omega\}$ für ein $\omega \in \Omega$ heißt „Elementarereignis“.

Da Ereignisse Mengen sind, kann man Ereignisse mittels Mengen-Operatoren miteinander verknüpfen:

- $A \cup B$: A oder B (oder beide) treten ein.
- $A \cap B$: A und B treten ein.
- $\bar{A} = \Omega \setminus A$: A tritt nicht ein, man sagt auch $\bar{A}$ ist das Komplementärereignis.
- $A \setminus B$: A aber nicht B tritt ein.
- $\bigcup_{i=1}^{\infty} A_i$: Mindestens eines der Ereignisse A_i tritt ein.
- $\bigcap_{i=1}^{\infty} A_i$: Alle Ereignisse A_i treten ein.

Was verstehen wir nun unter dem Ereignissystem $\mathcal{A}$?

Das Ereignissystem $\mathcal{A}$ ist *die Menge aller möglichen Ereignisse.*

Eine mögliche Wahl für das Ereignissystem ist

$$\mathcal{A} = \mathcal{P}(\Omega) \qquad \text{(Potenzmenge)}.$$

Für *diskrete* Ergebnismengen Ω ist das die natürliche Wahl.

Bei *überabzählbarer* Ergebnismenge Ω gibt es allerdings technische Schwierigkeiten: $\mathcal{P}(\Omega)$ ist „zu groß“.

In diesem Fall wählt man $\mathcal{A}$ als sog. Borel'sche Mengen. Da wir uns vorerst auf diskrete Wahrscheinlichkeitsräume beschränken davon erst später mehr.

Den einzelnen Ereignissen sollen nun Wahrscheinlichkeiten zugeordnet werden.

Mittels der Abbildung (Pr steht für „probability“)

$$\Pr : \mathcal{A} \to [0, 1]$$

quantifiziert man die Wahrscheinlichkeit des Eintretens des Ereignisses A.

Dabei heißt

- $\Pr[A] = 0$: A tritt nie ein und
- $\Pr[A] = 1$: A tritt sicher ein.

Eine mögliche Art, die Abbildung Pr festzulegen, ist mittels sog. „relativer Häufigkeiten“ bei der Durchführung vieler Versuche:

$$\text{relative Häufigkeit von } A = \frac{\text{Anzahl Eintreten von A}}{\text{Anzahl aller Versuche}}.$$

Offensichtlich setzt dies das vielmalige Wiederholen des Zufallsexperimentes voraus. Dies ist laut dessen Definition möglich.

Das sog. „Gesetz der großen Zahlen“ sichert, dass sich die relative Häufigkeit der gesuchten Wahrscheinlichkeit immer mehr nähert (dies werden wir später beweisen).

Bei der Festlegung des Wahrscheinlichkeitsmaßes komplexer Ereignisse wie etwa

„Die Rakete erreicht das Weltall“,

ist es schwierig, genügend viele (wieviele sind das?) Versuche durchzuführen.

In diesem Fall betrachtet man zunächst die *Elementarereignisse*, aus denen sich das Ereignis zusammensetzt („Bauteil 1 geht“, „Bauteil 2 geht“, ...), und bestimmt deren Wahrscheinlichkeiten.

Dann *berechnet* man die Wahrscheinlichkeit des zusammengesetzten Ereignisses (indem man alle Kombinationen der Bauteile betrachtet, die funktionieren müssen damit die Rakete das Weltall erreicht).

Es ist also sinnvoll, zunächst den Elementarereignissen Wahrscheinlichkeiten zuzuordnen.

Beispiel 17.8. Aus langjährigen Wetterdaten erhält man für Stuttgart die folgenden Wahrscheinlichkeiten für die Elementarereignisse $\{\text{sonnig}, \text{bewölkt}, \text{regnerisch}\}$:

$$\Pr[\{\text{sonnig}\}] = 0.3, \quad \Pr[\{\text{bewölkt}\}] = 0.45, \quad \Pr[\{\text{regnerisch}\}] = 0.25.$$

Für zusammengesetze Ereignisse erwarten wir, dass

$$\Pr[\{\text{sonnig}, \text{bewölkt}\}] = 0.3 + 0.45 = 0.75$$

da die Ergebnisse nicht zugleich eintreten können (man sagt die Ereignisse sind *disjunkt*). □

Um Schreibarbeit zu sparen, erlauben wir Pr gleichzeitig als Abbildung $\Pr : \Omega \to [0,1]$ zu betrachten, die mit dem obigen Pr auf den Elementarereignissen übereinstimmt.

Wegen diesem „Overloading“ verwendet man eckige Klammern.

Damit haben wir nun alle Komponenten eines Wahrscheinlichkeitsraumes eingeführt.

Definition 17.9 (Diskreter Wahrscheinlichkeitsraum)**.** Ein diskreter Wahrscheinlichkeitsraum besteht aus

- Einer diskreten Ergebnismenge $\Omega = \{\omega_1, \omega_2, \ldots\}$ von Elementarereignissen.
- Einer Funktion $\Pr : \Omega \to \mathbb{R}$ derart, dass

$$0 \leq \Pr[\omega_i] \leq 1 \qquad \text{und} \qquad \sum_{\omega \in \Omega} \Pr[\omega] = 1.$$

- Die Funktion Pr hat eine natürliche Erweiterung $\Pr : \mathcal{A} \to \mathbb{R}$, $\mathcal{A} = \mathcal{P}(\Omega)$, mittels

$$\Pr[A] = \sum_{\omega \in A} \Pr[\omega].$$

- $\Pr : \mathcal{A} \to \mathbb{R}$ heißt Wahrscheinlichkeitsmaß. □

Betrachten wir einige Beispiele für diskrete Wahrscheinlichkeitsräume.

Beispiel 17.10 (Wurf mit dem sechsseitigen Würfel)**.** Hier ist

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

und

$$\Pr[\omega] = \frac{1}{6},$$

d. h. alle Elementarereignisse sind gleich wahrscheinlich, der Würfel ist nicht gezinkt.

Dies bezeichnet man auch als Laplace'sches[38] Prinzip. □

Beispiel 17.11 (Ein mehrstufiges Experiment)**.** Hier ein Beispiel für einen etwas komplizierteren Wahrscheinlichkeitsraum mit Bezug zur Informatik.

- Die Operationen eines Prozessors werden in zwei Typen I/O (Ein/Ausgabe) und CPU (Rechenoperation) unterteilt.
- Die auf dem Prozessor laufenden Prozesse werden ebenfalls in die zwei Typen I/O (Ein-/Ausgabe lastig) und CPU (rechenlastig) unterteilt.

Als Ergebnismenge wählen wir somit

$$\Omega = \{(p, o) \,|\, p \in \{\text{I/O}, \text{CPU}\} \wedge o \in \{\text{I/O}, \text{CPU}\}\}.$$

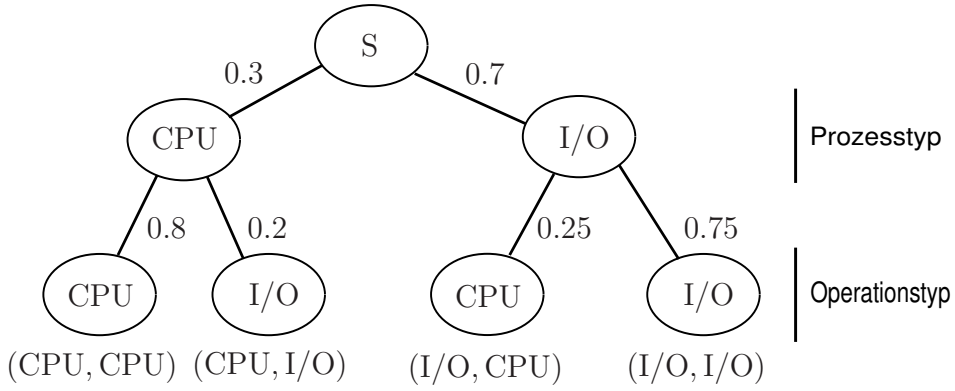
Dabei steht die erste Komponente p für den Prozesstyp und die zweite Komponente o für den Operationstyp.

[38] Pierre-Simon (Marquis de) Laplace, 1749-1827, frz. Mathematiker.

Es seien weiterhin folgende Wahrscheinlichkeiten bekannt:

- Ein Prozess gehört mit Wahrscheinlichkeit 0.7 zur Gruppe I/O.
- Für I/O-Prozesse beträgt die Wahrscheinlichkeit eine Rechenoperation auszuführen 0.25.
- Für CPU-Prozesse beträgt die Wahrscheinlichkeit eine Rechenoperation auszuführen 0.8.

Die Wahrscheinlichkeiten für die Elementarereignisse bestimmt man jetzt mit Hilfe eines sog. *Entscheidungsbaumes*.



Damit erhält man folgende Wahrscheinlichkeiten für die Elementarereignisse

$$\Pr[(\text{CPU}, \text{CPU})] = 0.3 \cdot 0.8 = 0.24, \qquad \Pr[(\text{CPU}, \text{I/O})] = 0.3 \cdot 0.2 = 0.06,$$
$$\Pr[(\text{I/O}, \text{CPU})] = 0.7 \cdot 0.25 = 0.175, \qquad \Pr[(\text{I/O}, \text{I/O})] = 0.7 \cdot 0.75 = 0.525.$$

Damit lassen sich nun auch Wahrscheinlichkeit von zusammengesetzten Ereignissen berechnen.

Die Wahrscheinlichkeit, dass eine Rechenoperation ausgeführt wird, ist

$$\Pr[\{(\text{CPU}, \text{CPU}), (\text{I/O}, \text{CPU})\}] = 0.24 + 0.175 = 0.415.$$

□

Diese sog. mehrstufigen Experimente treten in der Praxis häufig auf.

Man kann per Induktion zeigen dass die Wahrscheinlichkeiten für die so entstehenden Elementarereignisse (die den Blättern entsprechen) sich immer zu 1 summieren, falls sich die Wahrscheinlichkeiten in jedem inneren Knoten zu eins summieren.

Ein Beispiel mit unendlicher Ergebnismenge.

Beispiel 17.12. Von Rechner P zu Rechner Q wird ein Paket mit der Wahrscheinlichkeit p erfolgreich übertragen.

Wir betrachten das Zufallsexperiment: Übertrage ein Paket solange, bis es erfolgreich übertragen wird. Als Ergebnismenge wählen wir

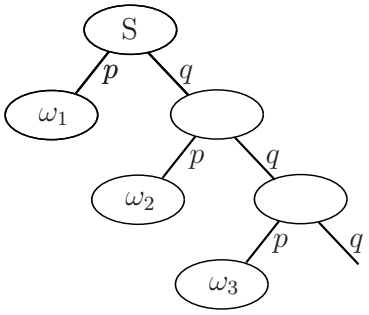
$$\Omega = \{\omega_i | i \in \mathbb{N}\}.$$

Dabei soll ω_i das Ereignis bedeuten, dass $i \in \mathbb{N}$ Versuche bis zur *erstmaligen* erfolgreichen Übertragung notwendig sind.

Die Anzahl der Versuche ist unbeschränkt, wir haben es mit einer abzählbar unendlichen Ergebnismenge zu tun.

Die Zuordnung der Wahrscheinlichkeiten zu den ω_i gelingt wieder mit Hilfe eines Entscheidungsbaumes.

$q = 1 - p$ bezeichne die Wahrscheinlichkeit einer fehlerhaften Übertragung.



Erfolgreiche Übertragung im Schritt i bedeutet, dass $i-1$ mal fehlerhaft und einmal erfolgreich übertragen wurde, also

$$\Pr[\omega_i] = pq^{i-1}.$$

Die Summationsbedingung prüfen wir explizit:

$$\sum_{\omega\in\Omega} \Pr[\omega] = \sum_{i=1}^{\infty} pq^{i-1} = p\sum_{i=0}^{\infty} q^i = p\frac{1}{1-q} = 1.$$

□

17.3 Gesetzmäßigkeiten für Wahrscheinlichkeitsmaße

Bei völlig unbekanntem Verhalten eines Systems muss man die Wahrscheinlichkeiten der Elementarereignisse mittels relativer Häufigkeiten bestimmen.

Oft kann man jedoch generelle Gesetzmäßigkeiten heranziehen.

Prinzip von Laplace Oben haben wir bereits einmal das Prinzip von Laplace herangezogen. Es lautet

> Falls nichts dagegen spricht, sind alle Elementarereignisse gleich wahrscheinlich.

Formal heißt das

$$\Pr[\omega] = \frac{1}{|\Omega|},$$
$$\Pr[A] = \frac{|A|}{|\Omega|}.$$

Informationstheoretisch kann man das Prinzip von Laplace als Wahrscheinlichkeitsraum mit größtmöglicher Entropie (Unordnung) deuten.

Abweichung von der Gleichwahrscheinlichkeit ist nur dann sinnvoll, wenn entsprechende zusätzliche Information über das Problem vorliegt (z.B. aus Beobachtung von relativen Häufigkeiten).

Das Prinzip von Laplace kann man offensichtlich nur bei *endlicher* Ergebnismenge anwenden.

Newcomb-Benford'sches Gesetz Für Datensätze „natürlichen" Ursprungs, etwa

- Aktienkurse verschiedener Unternehmen,
- Umsatzzahlen von Unternehmen,
- Einwohnerzahlen von Städten,
- Flächeninhalten von Inseln, oder
- Durchlässigkeiten von Böden

beobachtet man häufig das folgende Phänomen:

Betrachtet man die *erste* Ziffer der Zahl in der Darstellung zur Basis 10, so beträgt die Wahrscheinlichkeit für die Ziffer $n \in \{1, 2, \ldots, 9\}$

$$f_n = \log_{10}\left(1 + \frac{1}{n}\right), \qquad f_1 \approx 0.3, f_9 \approx 0.05.$$

Dies bezeichnet man als Newcomb[39]-Benford'sches[40] Gesetz.

Eine Folge des Newcomb-Benford'schen Gesetzes ist, dass die *Logarithmen* der Zahlenwerte gleiche Wahrscheinlichkeiten haben.

Man sagt die Zahlenwerte seien log-normalverteilt. Die Bodendurchlässigkeiten in obigem Beispiel etwa wurden log-normalverteilt gewählt.

Auch bei den Flächeninhalten von Inseln ist unmittelbar einsichtig, dass es sehr viel mehr kleine Inseln wie große Inseln gibt.

Eine Erklärung für dieses Gesetz basiert darauf, dass Zahlenwerte natürlichen Ursprungs *skaleninvariant* sein müssen, d. h. multipliziert man die Werte mit einer Konstanten (oder: man misst in einer anderen Einheit, die war ja willkührlich vom Menschen festgelegt) so dürfen sich die relativen Häufigkeiten der ersten Ziffer nicht ändern.

Das Newcomb-Benford'sche Gesetz kann man dazu verwenden, Unregelmäßigkeiten in Daten, etwa Bilanzen oder Steuererklärungen, aufzuspüren.

[39] Simon Newcomb, 1835-1909, amerik. Astronom.
[40] Frank Benford, 1883-1948, amerik. Physiker.

17.4 Zusammenfassung

- Stochastische Methoden verwendet man, um Systeme zu modellieren, deren Verhalten sich nicht mit Sicherheit vorhersagen lässt, die vom Zufall beeinflusst sind. Dabei kann der Zufall natürlichen Ursprungs oder auf mangelndem Wissen begründet sein.
- Formal erfolgt diese Beschreibung durch einen Wahrscheinlichkeitsraum, der aus Ergebnismenge, Ereignissystem und Wahrscheinlichkeitsmaß besteht.

18 Bedingte Wahrscheinlichkeiten

18.1 Rechnen mit Wahrscheinlichkeiten

Aus der Definition 17.9 des diskreten Wahrscheinlichkeitsraumes erschließen wir einige elementare Rechenregeln für Wahrscheinlichkeiten.

Lemma 18.1. 1. $\Pr[\emptyset] = 0, \quad \Pr[\Omega] = 1$

2. $0 \leq \Pr[A] \leq 1$ für jedes $A \in \mathcal{A}$
3. $\Pr[\bar{A}] = 1 - \Pr[A]$
4. $A \subseteq B \Rightarrow \Pr[A] \leq \Pr[B]$
5. Additionssatz: Sind die Ereignisse $A_1, \ldots, A_n$ *paarweise disjunkt* so gilt

$$Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \Pr[A_i]$$

Für eine unendliche Menge von paarweise disjunkten Ereignissen gilt analog

$$Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} Pr\left[A_i\right]$$

Beweis:

1) folgt direkt aus der Definition $\sum_{\omega\in\emptyset} \Pr[\omega] = 0, \quad \sum_{\omega\in\Omega} \Pr[\omega] = 1$

5) Direkt aus der Definition $A = A_1 \cup \ldots \cup A_n$

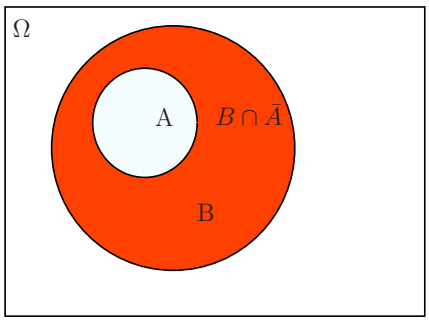
$$\Pr[A] = \sum_{\omega\in A} \Pr[\omega] = \sum_{i=1}^{n} \sum_{\omega\in A_i} \Pr[\omega] = \sum_{i=1}^{n} \Pr[A_i].$$

Genauso für die unendliche Summe.

2) $\Pr[A] \geq 0$, klar, da $\Pr[\omega] \geq 0$ und $A \subseteq \Omega$

3) $1 = \Pr[\Omega] = \Pr[A \cup \bar{A}] = \Pr[A] + \Pr[\bar{A}]$ zeigt 3)

4) $\Pr[B] = \Pr[B \cap \Omega] = \Pr[B \cap (A \cup \bar{A})] = \Pr[(B \cap A) \cup (B \cap \bar{A})] = \Pr[A] + \Pr[B \cap \bar{A}] \Rightarrow$ $\Pr[A] = \Pr[B] - \Pr[B \cap \bar{A}] \leq \Pr[B]$



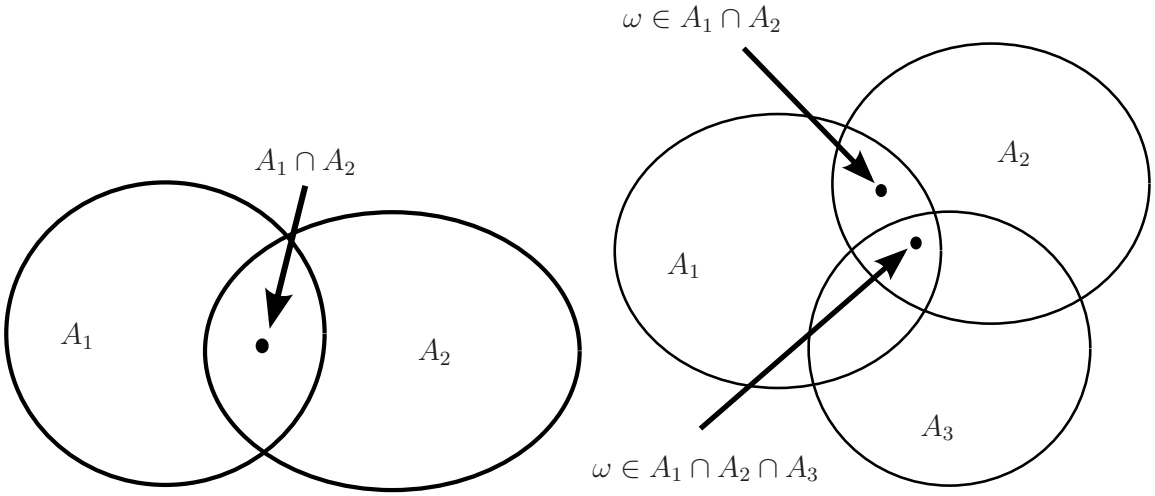
2) Setze $B = \Omega$ in 4) so gilt $\Pr[A] \leq \Pr[\Omega] = 1$ □

Hat man allgemeine, nicht unbedingt disjunkte Ereignisse $A_1, \ldots A_n$ so gilt der folgende

Satz 18.2 (Siebformel)**.** Für beliebige Ereignisse $A_1, \ldots A_n, (n \geq 2)$ gilt

$$\begin{aligned}\Pr\left[\bigcup_{i=1}^{n} A_i\right] = &\sum_{i=1}^{n} \Pr[A_i] - \sum_{1\leq i_1<i_2\leq n} \Pr[A_{i_1} \cap A_{i_2}] \\ &+ \sum_{1\leq i_1<i_2<i_3\leq n} \Pr[A_{i_1} \cap A_{i_2} \cap A_{i_3}] \\ &+ (-1)^{l-1} \sum_{1\leq i_1<\ldots<i_l\leq n} \Pr[A_{i_1} \cap \ldots \cap A_{i_l}] \\ &+ (-1)^{n-1} \Pr[A_{i_1} \cap \ldots \cap A_{i_n}].\end{aligned}$$

Beweis: Graphisch. Wir zeigen nur $n = 2, 3$.



$\omega \in A_1 \cap A_2$ wird in der ersten Summe zweimal gezählt, muss also einmal abgezogen werden.

I	$\Pr[A_1] + \Pr[A_2] + \Pr[A_3]$
II	$- \Pr[A_1 \cap A_2] - \Pr[A_1 \cap A_3] - \Pr[A_2 \cap A_3]$
III	$+ \Pr[A_1 \cap A_2 \cap A_3]$

In II gibt es zwei Fälle: ω kommt in genau 2 A_i vor und ω kommt in genau 3 A_i vor ($\rightarrow$ III).

□

Folgende Aussage erlaubt eine Abschätzung der Wahrscheinlichkeit zusammengesetzter Ereignisse:

Satz 18.3 (Bool'sche[41] Ungleichung)**.** Für Ereignisse $A_1, \ldots, A_n$ gilt

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i].$$

[41] George Bool, 1815-1864, brit. Mathematiker.

Die Erweiterung auf den unendlichen Fall ist möglich.

Beweis: Links steht

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{\omega \in \bigcup_{i=1}^{n} A_i} \Pr[\omega].$$

Rechts steht:

$$\sum_{i=1}^{n} \Pr[A_i] = \sum_{i=1}^{n} \sum_{\omega \in A_i} \Pr[\omega] \quad \Rightarrow \quad \text{jedes } \omega \in \bigcup_{i=1}^{n} A_i \text{ kommt } \geq 1 \text{ mal vor}$$

Da links jedes $\omega \in \bigcup_{i=1}^{n} A_i$ genau einmal gezählt wird, ist die linke Seite kleiner gleich der rechten Seite.

Man beachte aber, dass im allgemeinen

$$\Rightarrow \sum_{i=1}^{n} \Pr[A_i] \not\leq 1!$$

□

18.2 Bedingte Wahrscheinlichkeiten

Bekanntwerden zusätzlicher Informationen verändert möglicherweise die Wahrscheinlichkeit von Ereignissen. Dies wollen wir zunächst an einem Beispiel demonstrieren.

Beispiel 18.4 (Würfeln mit idealem, sechsseitigem Würfel)**.** Angenommen, nach dem Würfeln erfahren wir *zunächst* nur, ob das Ereignis $B = \{2, 4, 6\}$, d. h. „Augenzahl gerade“, eingetreten ist.

Nun ist klar, dass *unter dieser Bedingung* die Wahrscheinlichkeit für $\omega \in \{1, 3, 5\}$ Null und für $\omega = \{6\}$ gleich $1/3$ ist (da nur noch unter 3 Möglichkeiten gewählt werden kann). □

Wir führen die folgenden Bezeichnungen ein:

$A|B$ bezeichnet das Ereignis, dass A eintritt, wenn wir *wissen*, dass B auf jeden Fall eintritt.

Kurz: „A unter der Bedingung B“ oder noch kürzer„A gegeben B“.

Man sagt $A|B$ ist ein bedingtes Ereignis.

$\Pr[A|B]$ ist die entsprechende Wahrscheinlichkeit dieses bedingten Ereignisses.

$\Pr[A|B]$ heißt „bedingte“ Wahrscheinlichkeit.

Definition 18.5 (Bedingte Wahrscheinlichkeit)**.** A und B seien Ereignisse und es sei $\Pr[B] \geq 0$. Dann setzen wir

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

□

Diese Definition ist motiviert durch folgende Argumente:

- $\Pr[B|B] = 1$. Wenn B schon eingetreten ist, ist die Wahrscheinlichkeit für das Eintreten von B eins.
- $\Pr[A|\Omega] = \Pr[A]$. Eintreten von Ω bringt *keine* zusätzliche Information. Die Wahrscheinlichkeit für A unter dieser Bedingung ist unverändert.
- Ist B eingetreten, so kann A nur dann eintreten, wenn zugleich auch $A \cap B$ eintritt. Damit sollte $\Pr[A|B]$ proportional zu $\Pr[A \cap B]$ sein für *festes* B (Verdoppelt sich $\Pr[A \cap B]$ so auch $\Pr[A|B]$).

Beispiel 18.6 (Poker)**.** Wir verwenden zum Pokerspiel ein französisches Blatt mit den Werten

$$2, 3, \ldots, 9, 10, \text{Bube}, \text{Dame}, \text{König}, \text{Ass} = 13 \text{ Stück}$$

sowie den Farben

$$\text{Kreuz, Pik, Herz, Karo.}$$

Damit gibt es $\Rightarrow 4 \cdot 13 = 52$ Karten.

Wir betrachten zwei Spieler A und B.

A habe vier Asse und die Herz 2.

B kann dies nur mit einem Straight Flush, d.h. 5 Karten *einer* Farbe in aufsteigender Reihenfolge überbieten.

Betrachte jetzt das Ereignis $F =$ „Spieler B hat einen Straight Flush“. Dessen Wahrscheinlichkeit ist

$$\Pr[F] = \frac{|F|}{|\Omega|} = \frac{3 \cdot 8 + 7}{\binom{52-5}{5}} = \frac{31}{1533939} = 2.02 \cdot 10^{-5}.$$

Wie kommt man drauf?

Für eine Farbe≠Herz hat B die Möglichkeiten $2, 3, 4, 5, 6$ oder $3, 4, 5, 6, 7$ oder $\ldots, 9, 10, B, D, K$. Das sind 8 Stück bei 3 Farben. Bei Herz fällt die Möglichkeit $2, 3, 4, 5, 6$ weg, damit $|F| = 3 \cdot 8 + 7 = 31$.

Spieler B bekommt $k = 5$ aus $n = 52 - 5$ Karten, ohne Zurücklegen und ungeordnet gibt es dafür

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Möglichkeiten.

A hat allerdings die Karten gezinkt und *weiß*, dass B nur Kreuz auf der Hand hält, wir untersuchen also das Ereignis $F' =$ „B hat Straight Flush der Farbe Kreuz“.

Die Wahrscheinlichkeit dieses Ereignisses können wir direkt berechnen:

$$\Pr[F'] = \frac{|F'|}{|\Omega'|} = \frac{8}{\binom{12}{5}} = \frac{8}{792} \approx 0.01 \qquad \Rightarrow \text{ Faktor 500(!) größer.}$$

Das Wissen, dass B nur Kreuz hat äußert sich darin, dass wir im Nenner nur noch die verschiedenen Möglichkeiten 5 Kreuzkarten zu ziehen berücksichtigen.

Das Vorwissen verändert also die Wahrscheinlichkeiten drastisch.

Beachte, dass das Ereignis $F \cap K$, „B hat Straight Flush und nur Kreuz“, folgende Wahrscheinlichkeit hat:

$$\Pr[F \cap K] = \frac{|F \cap K|}{|\Omega|} = \frac{8}{\binom{47}{5}} = \frac{8}{1533939} \approx 5.2 \cdot 10^{-6}.$$

Versuchen wir dieses Resultat mit unserer Definition der bedingten Wahrscheinlichkeit zu erhalten.

Die *Bedingung* ist das Ereignis $K =$ „B hat nur Kreuz“ mit

$$\Pr[K] = \frac{|K|}{|\Omega|} = \frac{\binom{12}{5}}{\binom{52-5}{5}},$$

da nur noch 12 Karten von Kreuz übrig sind.

Nun untersuchen wir das Ereignis

$$F|K = \text{„B hat Straigh Flush unter der Bedingung Kreuz“}$$

$$\Pr[F|K] = \frac{\Pr[F \cap K]}{\Pr[K]} = \frac{\frac{|F \cap K|}{|\Omega|}}{\frac{|K|}{|\Omega|}} = \frac{|F \cap K|}{|K|} = \frac{8}{792} = \Pr[F']$$

Dies ist also dasselbe wie F'!

$F|K$ und $F \cap K$ bezeichnen das gleiche Ereignis. *Aber* sie haben *nicht* die gleiche Wahrscheinlichkeit, da sie in verschiedenen Wahrscheinlichkeitsräumen definiert sind:

$$\Pr[F|K] = \frac{|F \cap K|}{|K|}, \qquad \Pr[F \cap K] = \frac{|F \cap K|}{|\Omega|}.$$

$F|K$ bezieht sich auf die Ergebnismenge $\Omega' = K$, also 5 aus 12 übrigen Kreuzkarten und $F \cap K$ auf die Ergebnismenge Ω, also 5 aus allen übrigen 47 Karten.

Eine beliebige Bedingung $B \subseteq \Omega$, $B \neq \emptyset$ erzeugt aus einem WR mit Ergebnismenge Ω einen neuen WR mit der Ergebnismenge B.

Es gilt nämlich 17.9 wegen

$$\sum_{\omega \in \Omega} \Pr\left[\{\omega\}|B\right] = \sum_{\omega \in \Omega} \frac{\Pr\left[\{\omega\} \cap B\right]}{\Pr[B]} = \sum_{\omega \in B} \frac{\Pr[\omega]}{\Pr[B]} = \frac{\Pr[B]}{\Pr[B]} = 1.$$

Alle Elementarereignisse $\omega \notin B$ haben die Wahrscheinlichkeit Null:

$$Pr\left[\{\omega\}|B\right] = \frac{Pr\left[\{\omega\} \cap B\right]}{\Pr[B]} = \frac{|\emptyset|}{|B|} = 0.$$

Wir können Definition 18.5 auch anders schreiben:

$$\Pr[A \cap B] = \Pr[B|A] \cdot P[A] = \Pr[A|B] \cdot P[B].$$

Dies folgt aus der Definition

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} \qquad \text{bzw.} \qquad \Pr[B|A] = \frac{\Pr[B \cap A]}{\Pr[A]}$$

und Auflösen nach $\Pr[A \cap B]$.

In Worten bedeutet dies

$$\Pr[A \cap B] = \underbrace{\Pr[A]}_{\text{„Wahrscheinlichkeit von } A\text{“}} \cdot \underbrace{\Pr[B|A]}_{\substack{\text{„ Wahrscheinlichkeit, dass } B \\ \text{eintritt, wenn } A \text{ schon} \\ \text{eingetreten ist“} \\ B \text{ alleine genügt nicht, da} \\ \text{ja } A \text{ und } B \text{ eintreten müssen.}}}$$

Dies führt zum folgenden Satz.

Satz 18.7 (Multiplikationssatz)**.** Seien die Ereignisse $A_1, \ldots, A_n$ gegeben. Falls $\Pr[A_1 \cap \ldots \cap A_n] > 0$ so gilt

$$\Pr[A_1 \cap \ldots \cap A_n] = \Pr[A_1] \cdot \Pr[A_2|A_1] \cdot \Pr[A_3|A_1 \cap A_2] \cdot \ldots \cdot \Pr[A_n|A_1 \cap \ldots \cap A_{n-1}].$$

Beweis: Auf der rechten Seite steht

$$\frac{\Pr[A_1]}{1} \cdot \frac{\Pr[A_1 \cap A_2]}{\Pr[A_1]} \cdot \frac{\Pr[A_1 \cap A_2 \cap A_3]}{\Pr[A_1 \cap A_2]} \cdot \ldots \cdot \frac{\Pr[A_1 \cap \ldots \cap A_n]}{\Pr[A_1 \cap \ldots A_{n-1}]}$$

und nach Kürzen folgt die Behauptung.

Wegen Definition 17.9(4) gilt $0 < \Pr[A_1 \cap \ldots \cap A_n] \leq \Pr[A_1 \cap \ldots \cap A_{n-1}] \leq \ldots \leq \Pr[A_1 \cap A_2] \leq \Pr[A_1]$ und alle Nenner sind positiv. □

Beispiel 18.8 (Geburtstagsproblem)**.** Wie groß ist die Wahrscheinlichkeit, dass in einer m-köpfigen Gruppe *mindestens* zwei Personen am gleichen Tag Geburtstag haben.

Eine abstraktere Formulierung des Problems lautet:

- Verteile m Bälle zufällig auf $n \geq m$ Körbe.
- Werfe dabei einen Ball nach dem anderen.
- Definiere Ereignis A_i: „Ball $\#i$ landet in einem leeren Korb “. (Dies bedeutet *nicht* automatisch, dass alle vorherigen Bälle jeweils in einem leeren Korb gelandet sind).
- Definiere Ereignis A: „Alle Bälle liegen alleine in einem Korb “.

Mit dem Multiplikationssatz 18.7 gilt dann :

$$\Pr[A] = \Pr\left[\bigcap_{i=1}^{m} A_i\right] = \Pr[A_1] \cdot \Pr[A_2|A_1] \cdot \ldots \cdot \Pr\left[A_m \Big| \bigcap_{i=1}^{m-1} A_i\right]$$

Das Ereignis $A_j | \bigcap_{i=1}^{j-1} A_i$ bedeutet, dass der j-te Ball in einem leeren Korb landet unter der Bedingung, dass alle vorherigen Bälle auch jeweils in einem leeren Korb gelandet sind.

Für dessen Wahrscheinlichkeit erhalten wir

$$\Pr\left[A_j \middle| \bigcap_{i=1}^{j-1} A_i\right] = \frac{n-(j-1)}{n} = 1 - \frac{j-1}{n}$$

(für den j-ten Ball sind $n-(j-1)$ von n Körben möglich) und somit

$$\Pr[A] = \prod_{j=1}^{m}\left(1-\frac{j-1}{n}\right) \leq \prod_{j=2}^{m} e^{-\frac{j-1}{n}} = e^{-(1/n)\sum_{j=1}^{m-1} j} = e^{-\frac{m(m-1)}{2n}}.$$

Hierbei haben wir die Abschätzung

$$1 - x \leq e^{-x}$$

verwendet. Diese zeigt man folgendermaßen.

Zunächst gilt die Potenzreihenentwicklung

$$e^{-x} = 1 - \frac{x}{1} + \underbrace{\frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!}\ldots}_{=:S_2}.$$

Wir zeigen nun, dass $S_2 \geq 0$. Daraus folgt dann wegen $e^{-x} = 1-x+S_2 \geq 1-x$ die Behauptung.

Betrachte zwei aufeinanderfolgende Folgenglieder k und $k+1$ mit k gerade:

$$\frac{x^k}{k!} - \frac{x^{k+1}}{(k+1)!} = \frac{x^k(k+1)-x^{k+1}}{(k+1)!} = \frac{x^k(k+1-x)}{(k+1)!} \geq 0 \Leftrightarrow x \leq k+1.$$

Für $x = (j-1)/n < 1$ und $k \geq 2$ gilt also $S_2 \geq 0$.

Nun zurück zum Geburtstagsproblem.

Das Ereignis A bezeichnet den Fall, dass alle Personen an *verschiedenen* Tagen Geburtstag haben. Die Lösung des Geburtstagsproblems ist also

$$\Pr[\bar{A}] = 1 - \Pr[A] \geq 1 - e^{-\frac{m(m-1)}{2n}}.$$

Zahlenbeispiel: Setze $n = 365$ und $m = 50$, dann gilt : $1 - \Pr[A] \geq 1 - e^{-\frac{m(m-1)}{2n}} \approx 95\%$.

Anwendung in der Informatik: Kollisionen in Hashtabellen. □

Mit dem Multiplikationssatz ließ sich dieses Problem recht elegant lösen.

Ein anderer Satz, der das Rechnen mit Wahrscheinlichkeiten in der Praxis vereinfachen kann, ist der folgende

Satz 18.9 (Totale Wahrscheinlichkeit)**.** Die Ereignisse $A_1, \ldots A_n$ seien paarweise disjunkt und es sei B ein Ereignis mit $B \subseteq A_1 \cup \ldots \cup A_n$.

Dann kann man die Wahrscheinlichkeit von B wie folgt berechnen

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B|A_i] \cdot \Pr[A_i].$$

Man kann also von den bedingten Wahrscheinlichkeiten bezüglich der A_i auf die totale Wahrscheinlichkeit von B zurückschließen.

Beweis: Nach Voraussetzung gilt

$$A = \bigcup_{i=1}^{n} A_i \qquad \text{und} \qquad B = B \cap A \text{ wegen } B \subseteq A$$

und damit

$$B = B \cap \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} B \cap A_i.$$

Man zerlegt also B in Teilereignisse $B \cap A_i$. Da die A_i paarweise disjunkt sind, sind auch die $B \cap A_i$ paarweise disjunkt.

Damit gilt der Additionssatz aus 18.1:

$$\Pr[B] = \Pr\left[\bigcup_{i=1}^{n} B \cap A_i \right] = \sum_{i=1}^{n} \Pr[B \cap A_i] = \sum_{i=1}^{n} \Pr[B|A_i] \cdot \Pr[A_i]$$

□

Der Satz erlaubt oft eine einfachere Berechnung komplexer Wahrscheinlichkeiten.

Beispiel 18.10 (Ziegenproblem)**.** Die Kandidatin einer Fernsehshow darf eine von drei Türen wählen. Hinter genau einer der Türen ist der Hauptgewinn (ein Auto), hinter den anderen beiden ist als Trostpreis je eine Ziege versteckt.

Nachdem die Kandidatin gewählt hat, öffnet der Moderator eine der beiden übrigen Türen hinter der sich (natürlich) eine Ziege verbirgt. Dann bekommt die Kandidatin die Möglichkeit, die Türe zu wechseln. Frage: Sollte die Kandidatin das Angebot annehmen?

Wir betrachten folgende Ereignisse:

- $A =$ „Kandidatin hat bei der ersten Wahl das Auto gewählt“.
- $G =$ „Kandidatin gewinnt *nach Wechseln der Tür* “.

Gesucht ist demnach die Wahrscheinlichkeit $\Pr[G]$.

Als disjunkte Ereignisse verwenden wir A und $\bar{A}$. Wir erhalten

- $\Pr[G|A] = 0$, denn wenn die Kandidatin anfangs das Auto gewählt hat, ist nach dem Wechsel immer eine Ziege hinter der Tür.
- $\Pr[G|\bar{A}] = 1$, denn bei der ersten Wahl hat die Kandidatin eine der beiden Ziegen erwischt und der Moderator musste die andere Ziege aufdecken, also ist unter der verbleibenden Türe das Auto.

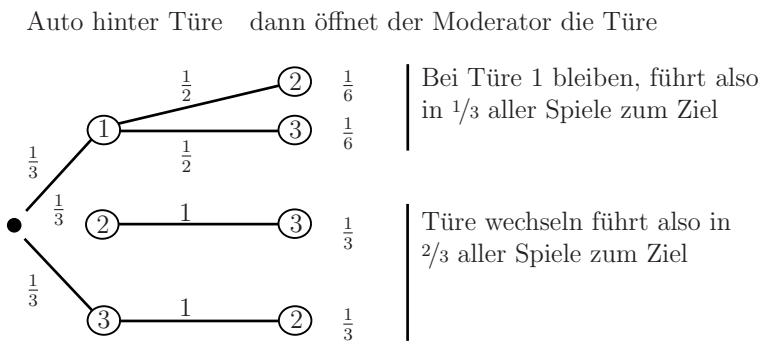
Darauf wenden wir nun den Satz 18.9 an:

$$\begin{array}{rlll} \Pr[G] = & \Pr[G|A] \cdot \Pr[A] & +\Pr[G|\bar{A}] \cdot \Pr[\bar{A}] & \\ = & 0 \cdot \frac{1}{3} & +1 \cdot \frac{2}{3} & = \frac{2}{3} \end{array}$$

Es ist also besser zu wechseln!

Man kann dieses Problem auch direkt mit Hilfe eines Entscheidungsbaumes lösen.

O.B.d.A. zeige die Kandidatin auf Türe 1. Dann gibt es folgende Fälle



□

Eine Folge des Satzes von der totalen Wahrscheinlichkeit ist der

Satz 18.11 (Satz von Bayes[42])**.** Die Ereignisse $A_1, \ldots, A_n$ seien wieder disjunkt und $B \subseteq A_1 \cup \ldots \cup A_n$ sei ein Ereignis mit $\Pr[B] > 0$. Dann gilt für jedes $i \in 1, \ldots, n$

$$\Pr[A_i|B] = \frac{\Pr[A_i \cap B]}{\Pr[B]} = \frac{\Pr[B|A_i] \cdot \Pr[A_i]}{\sum_{j=1}^{n} \Pr[B|A_j] \cdot \Pr[A_j]}.$$

Eine Erweiterung auf unendlich viele Ereignisse A_j ist möglich. □

Mit dem Satz von Bayes kann man die Richtung der Bedingung umdrehen. Das $\Pr[A_i|B]$ wird aus den $\Pr[B|A_j]$ errechnet.

Machen wir zum Abschluss noch ein Beispiel, das alle behandelten Konzepte nochmals illustriert.

Beispiel 18.12 (Fehlerhafter Übertragungskanal)**.** Über einen Übertragungskanal werden die Bits $0, 1$ übertragen.

Wir definieren die Ereignisse

- $S_i =$ „Das Zeichen i wird gesendet“, für $i = 0, 1$.
- $R_i =$ „Das Zeichen i wird empfangen“, für $i = 0, 1$.

Die Trennung der Ereignisse S_i und R_i erlaubt es, Übertragungsfehler zu modellieren.

Als Wahrscheinlichkeiten geben wir vor

$$\Pr[S_0] = 0.3, \qquad \Pr[S_1] = 0.7,$$

es werden also mehr Einsen als Nullen gesendet.

Schließlich sei noch Information über die Übertragungsfehler vorhanden:

$$\Pr[R_1|S_0] = 0.3, \qquad \Pr[R_0|S_1] = 0.1,$$

d.h. 30% aller Nullen und 10% aller Einsen werden falsch übertragen.

[42] Thomas Bayes, 1702-1762, brit. Theologe.

Dies ist ein typisches Beispiel für eine sehr informelle Definition eines Wahrscheinlichkeitsraumes.

Die Ergebnismenge Ω wird gar nicht angeben.

Eine formale Definition des Wahrscheinlichkeitsraumes wäre

$$\Omega = \{(s,r) \mid s,r \in \{0,1\}\}, \qquad \text{also } |\Omega| = 4.$$

Das Ereignis (s,r) bedeutet, dass s gesendet und r empfangen wurde.

Die S_0, S_1 sind dann

$$S_0 = \{(0,0),(0,1)\}, \qquad S_1 = \{(1,0),(1,1)\}.$$

Wie groß ist die Wahrscheinlichkeit für einen Übertragungsfehler?

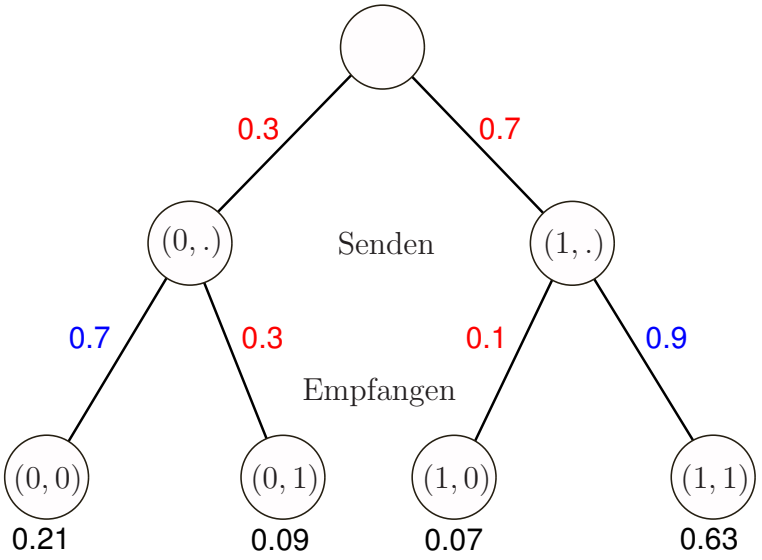
Wir nutzen den Additionssatz für disjunkte Ereignisse sowie die Definition bedingter Wahrscheinlichkeiten:

$$\begin{aligned}\Pr[\text{„Übertragungsfehler“}] &= \Pr[S_0 \cap R_1] + \Pr[S_1 \cap R_0] \\ &= \Pr[R_1|S_0] \cdot \Pr[S_0] + \Pr[R_0|S_1] \cdot \Pr[S_1] \\ &= 0.3 \cdot 0.3 + 0.1 \cdot 0.7 = 0.16.\end{aligned}$$

Mit dem Satz von der totalen Wahrscheinlichkeit berechnen wir $\Pr[R_1]$:

$$\begin{aligned}\Pr[R_1] &= \Pr[R_1|S_0] \cdot \Pr[S_0] + \Pr[R_1|S_1] \cdot \Pr[S_1] \\ &= \Pr[R_1|S_0] \cdot \Pr[S_0] + (1 - \Pr[R_0|S_1]) \cdot \Pr[S_1] \\ &= 0.3 \cdot 0.3 + 0.9 \cdot 0.7 = 0.72.\end{aligned}$$

Das hätte man nach Berechnen der fehlenden (blauen) Wahrscheinlichkeiten auch mit einem Entscheidungsbaum bekommen:



Mit dem Satz von Bayes berechnen wir $\Pr[S_0|R_0]$, also die Wahrscheinlichkeit, dass eine Null gesendet wurde, wenn man eine Null empfangen hat:

$$\begin{aligned}\Pr[S_0|R_0] &= \frac{\Pr[R_0|S_0] \cdot \Pr[S_0]}{\Pr[R_0|S_0] \cdot \Pr[S_0] + \Pr[R_0|S_1] \cdot \Pr[S_1]} = \frac{0.7 \cdot 0.3}{0.7 \cdot 0.3 + 0.1 \cdot 0.7} \\ &= \frac{0.21}{0.21 + 0.07} = 0.75.\end{aligned}$$

□

18.3 Zusammenfassung

- In diesem Abschnitt wurde der Begriff der bedingten Wahrscheinlichkeit eingeführt. Dabei geht es um die Wahrscheinlichkeit von Ereignissen bei Bekanntwerden zusätzlicher Information.
- Mittels dem Multiplikationssatz, dem Satz von der totalen Wahrscheinlichkeit sowie dem Satz von Bayes kann man komplexere Probleme einfacher lösen.

19 Unabhängigkeit von Ereignissen

19.1 Unabhängigkeit zweier Ereignisse

Wir untersuchen die Frage: Wann beeinflusst ein Ereignis die Wahrscheinlichkeit eines anderen Ereignisses (nicht)?

$\Pr[A|B]$ bedeutet die Wahrscheinlichkeit, dass A eintritt, wenn man weiss, dass B eingetreten ist.

Hat das Eintreten von B keinen Einfluss auf die Wahrscheinlichkeit von A, so bezeichnet man diese als unabhängig. Dies motiviert folgende Definition:

Definition 19.1 (Unabhängige Ereignisse)**.** Die Ereignisse A und B heißen *unabhängig*, falls gilt

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B].$$

□

Denn, falls $\Pr[B] \neq 0$, gilt dann

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A] \cdot \Pr[B]}{\Pr[B]} = \Pr[A].$$

Beispiel 19.2. Wir untersuchen das Zufallsexperiment „Zweimaliges Würfeln mit sechseckigem Würfel“:

$$\Omega = \{(i,j) | 1 \leq i,j \leq 6\}, \qquad \Pr[(i,j)] = 1/36.$$

Wir betrachten folgende Ereignisse:

- $A =$ „Augenzahl im ersten Wurf gerade“.
- $B =$ „Augenzahl im zweiten Wurf gerade“.
- $C =$ „Summe beider Augenzahlen ist 7“.

Nun untersuchen wir deren Unabhängigkeit.

Intuitiv würde man erwarten, dass A und B unabhängig sind, da schon in der Definition des Zufallsexperimentes gefordert wurde, dass der erste den zweiten Wurf nicht beeinflusst.

Prüfen wir dies formal mit unserer Definition. Es ist

$$\Pr[A] = \frac{18}{36} = \frac{1}{2}, \qquad \Pr[B] = \frac{18}{36} = \frac{1}{2}.$$

Sowie

$$|A \cap B| = |\{(i,j) \mid i,j \in \{2,4,6\}\}| = 9.$$

Also

$$\Pr[A \cap B] = \frac{9}{36} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \Pr[A] \cdot \Pr[B].$$

Die formale Definition liefert ebenfalls die Unabhängigkeit!

Untersuchen wir nun die Unabhängigkeit von A und C.

Hier ist die Sache nicht so klar. Die Tatsache, dass im ersten Wurf eine gerade Zahl gewürfelt wurde könnte ja durchaus einen Einfluss auf die „Summe der Zahlen ist 7“ haben.

Rechnen wir nach:

$$|C| = 6, \qquad |A \cap C| = |\{(2,5),(4,3),(6,1)\}| = 3,$$

also

$$\Pr[A \cap C] = \frac{3}{36} = \frac{1}{12} = \frac{1}{2} \cdot \frac{1}{6} = \Pr[A] \cdot \Pr[C].$$

Die Ereignisse A und C sind also laut Definition unabhängig.

Betrachten wir nun, das Ereignis

$$C' = \text{„Summe der Augenzahlen ist 2“}.$$

Wegen

$$C' = \{(1,1)\}$$

ist

$$\Pr[C' \cap A] = \frac{|\emptyset|}{36} = 0 \neq \Pr[C'] \cdot \Pr[A] = \frac{1}{36} \cdot \frac{1}{2}.$$

Mit der Summe 7 hatten wir also Glück. □

Trenne klar die Begriffe unabhängig und disjunkt! Für zwei *disjunkte* Ereignisse A und B gilt

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] \qquad \text{(Additionssatz!)}.$$

Für zwei *unabhängige* Ereignisse gilt

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B].$$

Für zwei *disjunkte* Ereignisse A, B mit $\Pr[A], \Pr[B] > 0$ gilt

$$\Pr[A] \cdot \Pr[B] > 0 = \Pr[A \cap B]$$

Zwei unabhängige Ereignisse A, B mit $\Pr[A], \Pr[B] > 0$ sind nie disjunkt! Denn wären sie disjunkt, so wäre

$$\Pr[A \cap B] = 0 < \Pr[A] \cdot \Pr[B].$$

Umgekehrt können zwei disjunkte Ereignisse A und B mit $\Pr[A], \Pr[B] > 0$ nie unabhängig sein, denn dann wäre

$$\Pr[A] \cdot \Pr[B] > 0 = \Pr[A \cap B].$$

Insbesondere sind also Elementarereignisse ω_1, ω_2 mit $\Pr[\omega_1] \cdot \Pr[\omega_2] > 0$ nie unabhängig.

19.2 Unabhängigkeit von mehr als zwei Ereignissen

Der Begriff der Unabhängigkeit kann auf mehr als zwei Ereignisse erweitert werden. Hierzu das

Beispiel 19.3. Fortsetzung von Beispiel 19.2. Die Ereignisse A und B, A und C sowie B und C (nicht gezeigt, geht aber genauso) sind jeweils unabhängig, d. h. A, B und C sind paarweise unabhängig.

Betrachten wir aber

$$\Pr[C|A \cap B] = \frac{\Pr[C \cap A \cap B]}{\Pr[A \cap B]} = \frac{|\emptyset|}{\Pr[A] \cdot \Pr[B]} = 0 < \Pr[C] \cdot \Pr[A] \cdot \Pr[B]$$

so sind also die Ereignisse C und $A \cap B$ *nicht* unabhängig! □

Für die Unabhängigkeit von mehr als zwei Ereigeignissen trifft man daher die folgende

Definition 19.4. Die Ereignisse $A_1, \ldots, A_n$ heißen unabhängig wenn für *alle Teilmengen von Indizes* $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ gilt, dass

$$\Pr[A_{i_1} \cap \ldots \cap A_{i_k}] = \Pr[A_{i_1}] \cdot \ldots \cdot \Pr[A_{i_k}]. \tag{19.1}$$

Eine unendliche Menge von Ereignissen heißt unabhängig, wenn (19.1) für jede endliche Teilmenge gilt. □

Man beachte: Für den Additionssatz genügt die paarweise Disjunktheit der Ereignisse, für die Unabhängigkeit muss jede endliche Teilmenge die Multiplikationsformel erfüllen.

Eine äquivalente Charakterisierung mehrerer unabhängiger Ereignisse, die manchmal leichter zu prüfen ist, gibt das folgende Lemma.

Lemma 19.5. Die Ereignisse $A_1, \ldots, A_n$ sind genau dann unabhängig, wenn für alle $(s_1, \ldots, s_n) \in \{0,1\}^n$ gilt, dass

$$\Pr[A_1^{s_1} \cap \ldots \cap A_n^{s_n}] = \Pr[A_1^{s_1}] \cdot \ldots \cdot \Pr[A_n^{s_n}]$$

wobei $A_i^0 = \bar{A}_i$ und $A_i^1 = A_i$ sein soll. Man muss also 2^n Kombinationen von Ereignissen und Komplementärereignissen durchsehen.

Beweis: Siehe [SS02, S.24]. □

Eine direkete Konsequenz aus der letzten Charakterisierung ist:

Sind A und B unabhängige Ereignisse, so auch $\bar{A}$ und B, A und $\bar{B}$, sowie $\bar{A}$ und $\bar{B}$. (Dies ist ja gerade die Aussage).

Außerdem zeigt man

Lemma 19.6. Sind A, B und C unabhängige Ereignisse, so sind auch $A \cap B$ und C bzw. $A \cup B$ und C unabhängig.

Beweis: wir rechnen nach

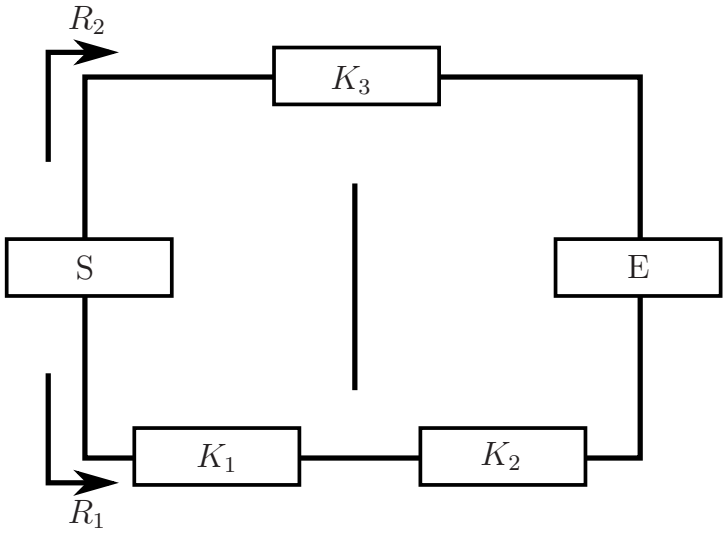
$$\begin{aligned}\Pr[(A \cap B) \cap C] &= \Pr[A \cap B \cap C] \\ &= \Pr[A] \cdot \Pr[B] \cdot \Pr[C] = \Pr[A \cap B] \cdot \Pr[C],\end{aligned}$$

$$\begin{aligned}\Pr[(A \cup B) \cap C] &= \Pr[(A \cap C) \cup (B \cap C)] \\ &= \Pr[A \cap C] + \Pr[B \cap C] - \Pr[A \cap B \cap C] \\ &= \Pr[A] \cdot \Pr[C] + \Pr[B] \cdot \Pr[C] - \Pr[A \cap B] \cdot \Pr[C] \\ &= \Pr[C] \cdot (\Pr[A] + \Pr[B] - \Pr[A \cap B]) = \Pr[C] \cdot \Pr[A \cup B].\end{aligned}$$

□

Üben wir den Begriff der Unabhängigkeit nochmal an einem größeren Beispiel.

Beispiel 19.7 (Rechnernetz)**.** Vom Sender S zum Empfänger E gebe es zwei Routen R_1 und R_2. R_1 besteht aus Knotenrechnern K_1 und K_2 sowie R_2 aus dem Knotenrechner K_3. Die Knotenrechner K_i sind *unabhängig* und mit der Wahrscheinlichkeit p intakt. Weiter nehmen wir an, dass die Verbindungen nicht ausfallen können.



Wir definieren folgende Ereignisse:

- K_i = „Knotenrechner i is intakt“.
- R_i = „Route i ist verfügbar“.
- A = „Es gibt eine intakte Route von S nach E“.

Hinweis: Überlegen Sie sich, wie der Wahrscheinlichkeitsraum, insbesondere Ω, aussieht!

Wegen $R_2 = K_3$ gilt natürlich $\Pr[R_2] = p$.

Für R_1 erhalten wir mit Hilfe der Unabhängigkeit:

$$\Pr[R_1] = \Pr[K_1 \cap K_2] = \Pr[K_1] \cdot \Pr[K_2] = p^2.$$

Schließlich das Ereignis A:

$$\begin{aligned}\Pr[A] &= \Pr[R_1 \cup R_2] = 1 - \Pr[\overline{R_1 \cup R_2}] \\ &= 1 - \Pr[\bar{R}_1 \cap \bar{R}_2] \\ &= 1 - \Pr[\bar{R}_1] \cdot \Pr[\bar{R}_2] \\ &= 1 - (1-p^2)(1-p) = 1 - (1 - p - p^2 + p^3) = p + p^2 - p^3.\end{aligned}$$

Hier haben wir benutzt:

- Komplementärereignis und De-Morgan-Regeln.
- $R_1 = K_1 \cap K_2$ und $R_2 = K_3$ sind unabhängig da K_1, K_2 und K_3 unabhängig sind und obiges Lemma gilt.
- Aus der Unabhängigkeit von R_1 und R_2 folgt die Unabhängigkeit von $\bar{R}_1$ und $\bar{R}_2$. □

19.3 Zusammenfassung

- Wir haben den Begriff der Unabhängigkeit eingeführt: Zwei Ereignisse sind unabhängig, wenn $\Pr[A|B]$ nicht von B abhängt, also $\Pr[A|B] = \Pr[A]$.
- Für unabhängige Ereignisse ergibt sich die Wahrscheinlichkeit, dass beide eintreten, durch Multiplikation: $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$.
- Bei der Unabhängigkeit von mehr als zwei Ereignissen muss diese Multiplikationsformel entsprechend erweitert für alle Teilmengen von Ereignissen gelten. Die paarweise Unabhängigkeit ist nicht ausreichend!

20 Zufallsvariablen

20.1 Einführung des Begriffes

Zufallsvariablen sind zunächst ein geschickter Weg zur Definition von Ereignissen zu einer Ergebnismenge Ω.

Oft möchte man einem Experimentausgang einen Zahlenwert zuordnen, etwa den Gewinn bei einem Spiel. Deshalb definiert man:

Definition 20.1 (Zufallsvariable)**.** Gegeben sei ein Wahrscheinlichkeitsraum mit Ergebnismenge Ω. Eine Abbildung

$$X : \Omega \to \mathbb{R}$$

heißt (numerische) Zufallsvariable (Abkürzung: ZV). Eine Zufallsvariable X über einer endlichen oder abzählbar unendlichen Menge Ω heißt diskret. Zunächst behandeln wir nur diskrete ZV. □

Weiterhin beschränken wir uns auf *numerische* Zufallsvariablen. Diese ordnen einem Experimentausgang eine Zahl aus $\mathbb{R}$ zu.

Die Abbildung X hat natürlich einen Wertebereich:

$$W_X = \{x \in \mathbb{R} \,|\, \exists \omega \in \Omega \text{ mit } X(\omega) = x\}.$$

Ist Ω diskret, dann ist auch W_X diskret, d.h. die Elemente sind abzählbar:

$$W_X = \{x_1, x_2, \ldots\}.$$

Die Elemente in W_X definieren auf kanonische Weise Ereignisse über Ω. Zu jedem $x_i \in W_X$ erhalten wir das Ereignis

$$A_i = \{\omega \in \Omega \,|\, X(\omega) = x_i\} =: X^{-1}(x_i) \subseteq \Omega.$$

Dieses Ereignis A_i besitzt die Wahrscheinlichkeit $\Pr[A_i]$. Jedem Element des Wertebereiches W_X ist auf diese Weise eine Wahrscheinlichkeit zugeordnet.

Beachte: Zwei Ereignisse A_i, A_j zu $x_i \neq x_j$ sind offensichtlich disjunkt.

Man definiert die folgende Schreibweise für Ereignisse, die über Zufallsvariablen definiert sind:

$$\text{„}X = x_i\text{“ entspricht } A_i.$$

Für die Wahrscheinlichkeiten schreibt man

$$\Pr[\text{„}X = x_i\text{“}] = \Pr[A_i] = \sum_{\omega \in A_i} \Pr[\omega].$$

Da die Wertemenge numerisch ist, ist auch folgendes Ereignis sinnvoll:

$$\text{„}X \leq x_i\text{“} = \bigcup_{x \in W_X \,:\, x \leq x_i} \text{„}X = x\text{“} = \{\omega \in \Omega \,|\, X(\omega) \leq x_i\}$$

mit der Wahrscheinlichkeit

$$\Pr[„X \leq x_i“] = \sum_{x \in W_X \,:\, x \leq x_i} \Pr[„X = x“] = \sum_{\omega \in \Omega : X(\omega) \leq x_i} \Pr[\omega].$$

(Hier haben wir benutzt, dass die Ereignisse „$X = x_i$“ disjunkt sind).

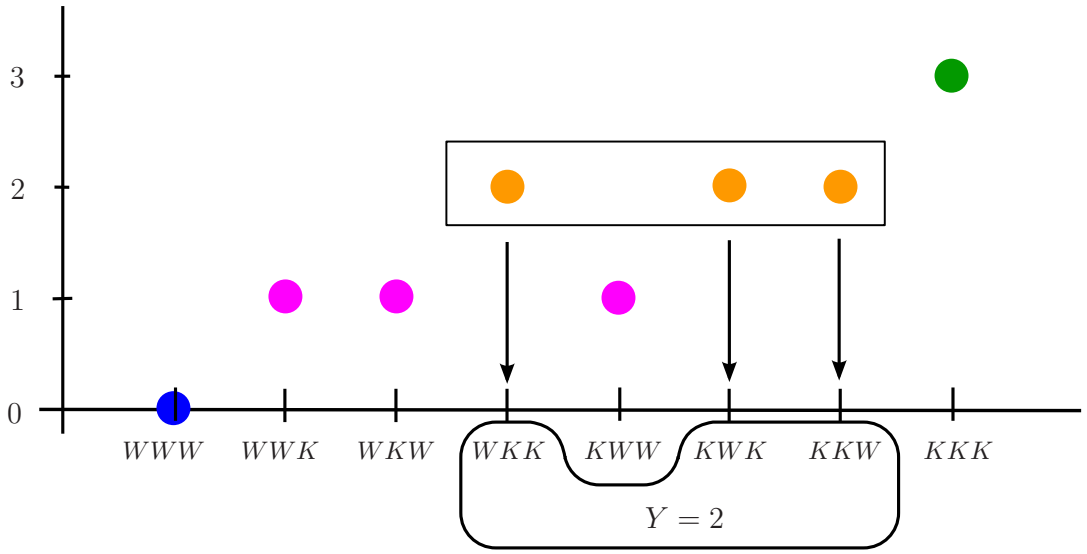
Analog definiert man auch „$X \geq x_i$“ oder $\Pr[„2 \leq X \leq 7„]$.

In Zukunft werden wir die Anführungszeichen weglassen: $\Pr[X = x_i]$, $\Pr[X \leq x_i]$. Selbst $\Pr[X^2 = x_i]$ ist möglich!

Beispiel 20.2 (Dreimaliges Werfen einer Münze)**.** Hier ist $\Omega = \{W, K\}^3$ mit W=Wappen und K=Kopf.

Betrachte die ZV Y : Gesamtzahl der Ergebnisse mit Kopf, also z.B. $Y(WWW) = 0, Y(KWK) = 2$.

Die Wertemenge ist also $W_Y = \{0, 1, 2, 3\}$.



□

Formal wird über diese Konstruktion jedem $x \in W_X$ über das zugehörige Ereignis „$X = x$“ eine Wahrscheinlichkeit $\Pr[X = x] \in [0, 1]$ zugeordnet.

Die so definierte Funktion nennt man *(diskrete) Dichtefunktion* von X und schreibt:

$$f_X : \mathbb{R} \to [0, 1], \qquad f_X(x) = \Pr[X = x].$$

Für alle $x \notin W_X$ ist $f_X(x) = 0$, da es sich um das leere Ereignis handelt.

Eine weitere Funktion ist die *Verteilungsfunktion*. Diese ist definiert als:

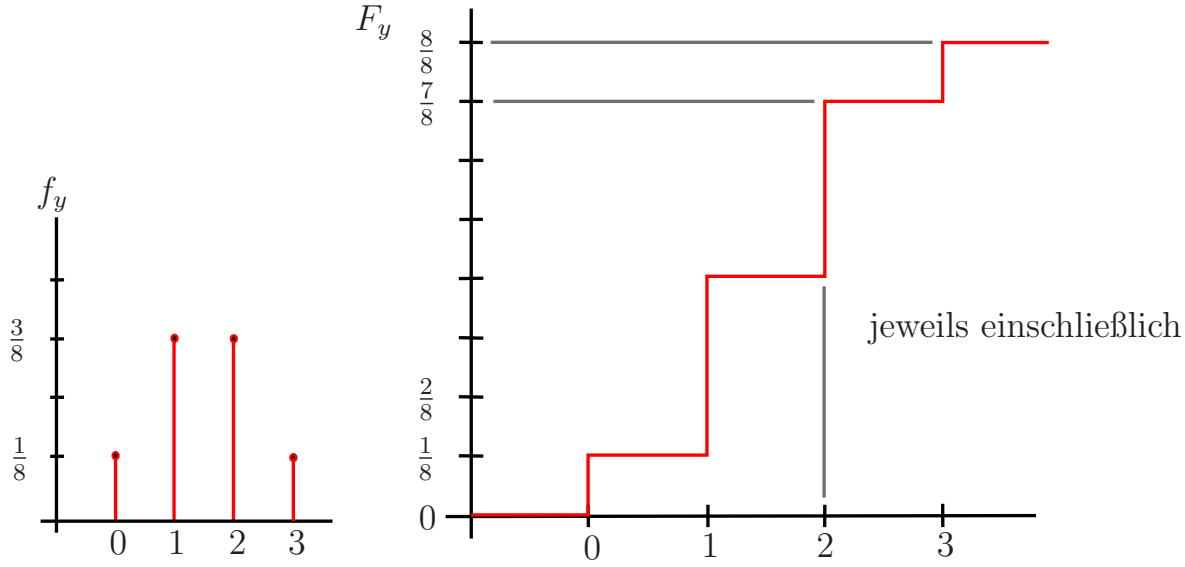
$$F_X : \mathbb{R} \to [0, 1], \qquad F_X(x) = \Pr[X \leq x] = \sum_{x' \in W_X \,:\, x' \leq x} f_X(x').$$

Die Verteilungsfunktion ist monoton steigend und bei diskreten ZV treppenförmig.

Für $x < \min W_X$ gilt $F_X(x) = 0$.

Für $x \geq \max W_X$ gilt $F_X(x) = 1$.

Beispiel 20.3. Wir setzen das letzte Beispiel fort.



□

W_X ist eine Menge und f_X ist ein Wahrscheinlichkeitsmaß auf dieser Menge. Die $x_i \in W_X$ kann man als neue Elementarereignisse auffassen.

Offensichtlich kann man dann

$$(W_X, f_X)$$

als einen neuen Wahrscheinlichkeitsraum auffassen.

Oft spielt das zugrundeliegende Ω gar keine Rolle mehr, da die weiteren Berechnungen nicht mehr davon abhängen. Dichte- oder Verteilungsfunktion genügen im folgenden, um Aussagen über die ZV zu machen.

Man kann also sagen: Eine ZV X erzeugt aus $(\Omega, \Pr)$ einen neuen Wahrscheinlichkeitsraum (W_X, f_X).

20.2 Erwartungswert

Oft interessiert man sich dafür, welche Werte eine Zufallsvariable im Mittel annimmt. Insbesondere dann, wenn X z.B. den Gewinn in einem Spiel angibt.

Angenommen beim Werfen eines sechsseitigen Würfels gewinnt jede 6 einen Euro, dann erwarten wir nach n Würfen einen Gewinn von $n/6$ Euro, da im Mittel $n/6$ 6en geworfen wurden.

Der *mittlere Gewinn pro Spiel* ist also 1/6 Euro.

Gewinnt jede gerade Augenzahl 10 Euro, so erwarten wir $(n/2)\cdot 10$ Euro Gewinn nach n Würfen, also 5 Euro je Spiel im Mittel.

Definition 20.4 (Erwartungswert)**.** Zu einer ZV X definiert man den Erwartungswert

$$\mathbb{E}[X] := \sum_{x\in W_X} x \cdot \Pr[X = x] = \sum_{x\in W_X} x \cdot f_X(x),$$

sofern $\sum_{x\in W_X} |x| \cdot \Pr[X = x]$ konvergiert. □

Beispiel 20.5. Für Beispiel 20.2 auf Seite 234 gilt

$$\mathbb{E}[Y] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Also erscheinen im Mittel 1.5 Köpfe je Spiel. □

Der Zusatz „$\sum_{x\in W_X} |x|\, f_X(x)$“ konvergiert, ist für abzählbar unendliche Zufallsvariablen notwendig.

Dann ist der Erwartungswert über eine unendliche Reihe definiert und diese muss nicht unbedingt konvergieren.

Der Zusatz in der Definition bedeutet, dass diese Reihe absolut konvergiert (man darf die Glieder beliebig zusammenfassen und umsortieren).

Wenn man den der Zufallsvariablen zugrundeliegenden Wahrscheinlichkeitsraum betrachtet erhält man folgende Formel für den Erwartungswert:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in W_X} x \cdot \Pr[X = x] = \sum_{x\in W_X} x \left(\sum_{\omega\in\Omega\,:\,X(\omega)=x} \Pr[\omega] \right) \\ &= \sum_{\omega\in\Omega} X(\omega) \cdot \Pr[\omega].\end{aligned}$$

Bei abzählbar unendlichem Ω ist analog zu fordern, dass die Reihe absolut konvergiert.

Damit folgt der

Satz 20.6 (Monotonie des Erwartungswertes)**.** Sind X, Y Zufallsvariablen über Ω mit $X(\omega) \le Y(\omega)$, für alle $\omega \in \Omega$ so gilt

$$\mathbb{E}[X] \le \mathbb{E}[Y].$$

Beweis: Nutze die alternative Formel des Erwartungswertes:

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)\Pr[\omega] \le \sum_{\omega\in\Omega} Y(\omega)\Pr[\omega] = \mathbb{E}[Y].$$

□

Ein Folgerung hieraus ist:

$$a \leq \mathbb{E}[X] \leq b \qquad \text{falls} \qquad a \leq X(\omega) \leq b.$$

Das zeigt man mittels der beiden ZV $Y_a(x) = a$ und $Y_b(x) = b$ für die $Y_a \leq X \leq Y_b$ gilt.

Wir zeigen nun eine Reihe weiterer Rechenregeln. Dabei wird immer angenommen, dass die Erwartungswerte wohldefiniert sind.

Beobachtung 20.7. Hat man eine beliebige Funktion

$$f : \mathbb{D} \to \mathbb{R}$$

mit

$$W_X \subset \mathbb{D}$$

so bildet die Funktion $Y = f \circ X$ eine neue Zufallsvariable

$$Y : \Omega \to \mathbb{R}.$$

(Hintereinanderschaltung von Abbildungen). □

Für diese neue ZV erhält man die Dichtefunktion

$$\Pr[Y = y] = \Pr\left[\bigcup_{x \in W_X \,:\, f(x)=y} \text{„X=x“}\right] = \sum_{x \in W_X \,:\, f(x)=y} \Pr[X = x].$$

Hier haben wir wieder die Disjunktheit der Ereignisse $X = x$ verwendet.

Für den Erwartungswert der neuen ZV gilt

$$\begin{aligned}\mathbb{E}[f(X)] = \mathbb{E}[Y] &= \sum_{y \in W_Y} y \cdot \Pr[Y = y] = \sum_{y \in W_Y} y \cdot \sum_{x \,:\, f(x)=y} \Pr[X = x] \\ &= \sum_{x \in W_X} f(x) \cdot \Pr[X = x].\end{aligned}$$

Alternativ kann man noch auf Ω zurückgehen und erhält

$$\mathbb{E}[f(X)] = \sum_{\omega \in \Omega} f(X(\omega)) \cdot \Pr[\omega].$$

Damit zeigt man nun

Satz 20.8 (Linearität des Erwartungswertes, einfache Version)**.** Für $f(X) = aX + b$ mit $a, b \in \mathbb{R}$, d. h. einer linearen Transformation der ZV gilt

$$\mathbb{E}[a \cdot X + b] = a \cdot \mathbb{E}[X] + b.$$

Beweis: Wir wenden obige Formel an:

$$\begin{aligned}\mathbb{E}[a \cdot X + b] &= \sum_{x \in W_X} (ax + b) \cdot \Pr[X = x] \\ &= a \cdot \sum_{x \in W_X} x \cdot \Pr[X = x] + b \cdot \sum_{x \in W_X} \Pr[X = x] \\ &= a\mathbb{E}[X] + b.\end{aligned}$$

□

Da ZV nur eine andere Schreibweise für Ereignisse sind, lassen sich auch bedingte Ereignisse auf ZV übertragen.

Definition 20.9. Sei X eine Zufallsvariable und A ein Ereignis mit $\Pr[A] > 0$. Die *bedingte Zufallsvariable* $X|A$ besitzt die Dichte

$$f_{X|A}(x) := \Pr[X = x \mid A] = \frac{\Pr[„X = x“ \cap A]}{\Pr[A]} = \frac{\sum\limits_{\omega \in \Omega \,:\, \omega \in A \wedge X(\omega) = x} \Pr[\omega]}{\Pr[A]}.$$

$X|A$ ist also eine neue ZV mit der Dichtefunktion $f_{X|A}$. □

Dass dies wirklich eine Dichte ist, zeigt

$$\sum_{x \in W_X} f_{X|A}(x) = \sum_{x \in W_X} \sum_{\omega \in \Omega \,:\, \omega \in A \wedge X(\omega) = x} \frac{\Pr[\omega]}{\Pr[A]} = \frac{\sum_{\omega \in A} \Pr[\omega]}{\Pr[A]} = 1.$$

Für den Erwartungswert von $X|A$ gilt:

$$\mathbb{E}[X|A] = \sum_{x \in W_X} x \cdot f_{X|A}(x).$$

20.3 Varianz

Zwei Zufallsvariablen können gleiche Erwartungswerte haben aber trotzdem sehr verschieden sein. Deshalb sucht man nach weiteren Maßen, um Zufallsvariablen charakterisieren zu können.

Betrachte etwa $W_X = \{-\alpha, \alpha\}$ und $\Pr[X = \pm\alpha] = 1/2$ so gilt $\mathbb{E}[X] = 0$ unabhängig von α.

Idee: Man möchte die „Abweichung vom Mittelwert“ quantifizieren.

Eine Möglichkeit wäre: Erwartete Abweichung vom Erwartungswert, etwa $\mathbb{E}[|X - \mathbb{E}[X]|]$. Das ist ungünstig wegen der Fallunterscheidung in der Betragsfunktion (bzw. fehlende Differenzierbarkeit im kontinuierlichen Fall).

Definition 20.10 (Varianz)**.** Für eine ZV X mit Erwartungswert $\mu = \mathbb{E}[X]$ definiert man die *Varianz*

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mu)^2] = \sum_{x \in W_X} (x - \mu)^2 \cdot \Pr[X = x].$$

Also die erwartete quadratische Abweichung vom Mittelwert. Die Größe $\sigma := \sqrt{\mathrm{Var}[X]}$ heißt *Standardabweichung* von X. □

Die Varianz existiert im abzählbar unendlichen Fall wenn der Erwartungswert existiert.

Diese Konstruktion lässt sich verallgemeinern.

Definition 20.11 (Momente)**.** Allgemein bezeichnet man

$$m_k(X) = \mathbb{E}[X^k] = \sum_{x \in W_X} x^k \cdot f_X(x)$$

als *k-tes Moment* und

$$c_k(X) = \mathbb{E}[(X-\mu)^k] = \sum_{x \in W_X} (x-\mu)^k f_X(x)$$

als *k-tes zentrales Moment.* □

$\mathbb{E}[X]$ ist also das 1. Moment und $\mathrm{Var}[X]$ das 2. zentrale Moment.

Auch für die Varianz werden wir nun einige Rechenregeln herleiten.

Satz 20.12 (Alternative Berechnung der Varianz)**.** Man kann die Varianz oft effizienter mit folgender Formel berechnen:

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Beweis: Sei $\mu = \mathbb{E}[X]$. Dann gilt

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] \\
&= \sum_{x \in W_X} (x^2 - 2\mu x + \mu^2) \cdot \Pr[X = x] \\
&= \sum_{x \in W_X} x^2 \cdot \Pr[X = x] - 2\mu \underbrace{\sum_{x \in W_X} x \cdot \Pr[X = x]}_{\mu} + \mu^2 \sum_{x \in W_X} \Pr[X = x].
\end{aligned}$$

□

Für die Varianz einer linear transformierten Zufallsvariablen gilt:

Satz 20.13. Für eine ZV X und $a, b \in \mathbb{R}$ gilt

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X].$$

Beweis: Zunächst nur die Verschiebung:

$$\begin{aligned}
\mathrm{Var}[X + b] &= \mathbb{E}[(X + b - \mathbb{E}[X + b])^2] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}[X].
\end{aligned}$$

Hier haben wir die Linearität des Erwartungswertes benutzt.

Eine Verschiebung ändert die Varianz nicht. Deshalb genügt es nur die Skalierung mit a zu betrachten:

$$\begin{aligned}
\mathrm{Var}[aX] &= \mathbb{E}[(aX)^2] - \mathbb{E}[aX]^2 = a^2\mathbb{E}[X^2] - a^2\mathbb{E}[X]^2 \\
&= a^2(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = a^2 \mathrm{Var}[X].
\end{aligned}$$

Hier haben wir die zweite Formel für die Varianz und die Linearität des Erwartungswertes benutzt. □

Bei einer einfachen Skalierung $Y = aX$ gilt also $\mathbb{E}[Y] = a\mathbb{E}[X]$ und $\mathrm{Var}[Y] = a^2\mathrm{Var}[X]$.

20.4 Mehrere Zufallsvariablen

Oft hat man nicht nur eine, sondern mehrere ZV über dem selben Wahrscheinlichkeitsraum.

Einfachster Fall sind zwei ZV, d.h.:

$$X : \Omega \to \mathbb{R}, \qquad Y : \Omega \to \mathbb{R}.$$

Beispiel 20.14. Betrachte das zweimalige Werfen eines Würfels: $\Omega = \{1, 2, 3, 4, 5, 6\}^2$.

Als Zufallsvariable könnten wir definieren

$$\begin{aligned} X &= \text{„Anzahl Würfe mit gerader Augenzahl“}, & W_X &= \{0, 1, 2\}, \\ Y &= \text{„Summe der Augenzahlen“}, & W_Y &= \{2, 3, \ldots, 12\}. \end{aligned}$$

Die Zufallsvariablen X und Y beeinflussen sich gegenseitig: Z. B. folgt aus $X = 2$, dass Y nur gerade Werte annehmen kann. □

Es geht nun darum, wie man mit mehreren ZV rechnet.

ZV sind ja eine andere Schreibweise für Ereignisse. Wir vereinbaren die Schreibweise:

$$X = x, Y = y \quad = \quad \text{„}X = x\text{“} \cap \text{„}Y = y\text{“} \quad = \quad \{\omega \mid X(\omega) = x \wedge Y(\omega) = y\}.$$

Dies ist also die Und-Verknüpfung bzw. der Schnitt von Ereignissen.

Entsprechend schreibt man für die Wahrscheinlichkeiten:

$$\Pr[X = x, Y = y] = \Pr[\text{„}X = x\text{“} \cap \text{„}Y = y\text{“}].$$

Auch kompliziertere Ereignisse sind möglich:

$$\Pr[X = x, Y \leq y], \qquad \Pr[X \geq x, \sqrt{Y} = y, Z \leq z].$$

Beispiel 20.15. Fortsetzung von oben Zweimaliges Werfen eines Würfels $\Omega = \{1, \ldots, 6\}^2$ mit den ZV:

$$\begin{aligned} X &= \text{„Anzahl Würfe mit gerader Augenzahl“}, & W_X &= \{0, 1, 2\}, \\ Y &= \text{„Summe der Augenzahlen“}, & W_Y &= \{2, 3, \ldots, 12\}. \end{aligned}$$

Wir erhalten folgende Wahrscheinlichkeiten:

- $\Pr[X = 0, Y = 5] = 0$ da Summe zweier ungerader Zahlen gerade.
- $\Pr[X = 1, Y = 5] = \frac{|\{(2,3),(3,2),(4,1),(1,4)\}|}{36} = \frac{4}{36} = \frac{1}{9}$. □

Jedem Ereignis $(x, y) \in W_X \times W_Y$ wird mittels $\Pr[X = x, Y = y]$ eine Wahrscheinlichkeit zugeordnet.

Dies bezeichnet man als *gemeinsame Dichtefunktion*:

$$f_{X,Y} : W_X \times W_Y \to \mathbb{R}, \qquad f_{X,Y}(x, y) := \Pr[X = x, Y = y].$$

Aus einer gegebenen gemeinsamen Dichte erhält man die Dichten der einzelnen ZV zurück:

$$\begin{aligned}\sum_{y \in W_Y} f_{X,Y}(x, y) &= \sum_{y \in W_Y} \Pr[X = x, Y = y] = \sum_{\omega \in \Omega \,:\, X(\omega) = x \wedge Y(\omega) \in W_Y} \Pr[\omega] \\ &= \sum_{\omega \in \Omega \,:\, X(\omega) = x} \Pr[\omega] = \Pr[X = x] = f_X(x).\end{aligned}$$

Analog erhält man $f_Y(y) = \sum_{x \in W_X} f_{X,Y}(x, y)$.

f_X und f_Y heißen *Randdichten*.

Man überzeuge sich auch davon, dass $\sum_{x \in W_X} \sum_{y \in W_Y} f_{X,Y}(x, y) = 1$.

Analog kann man auch die gemeinsame Verteilungsfunktion einführen:

$$\begin{aligned}F_{X,Y}(x, y) &:= \Pr[X \le x, Y \le y] \\ &= \Pr[\{\omega \mid X(\omega) \le x \wedge Y(\omega) \le y\}] \\ &= \sum_{x' \le x} \sum_{y' \le y} f_{X,Y}(x', y').\end{aligned}$$

Entsprechend bildet man wieder Randverteilungen:

$$\begin{aligned}F_X(x) &= \sum_{x' \le x} f_X(x') = \sum_{x' \le x} \sum_{y' \in W_Y} f_{X,Y}(x', y') = F_{X,Y}(x, \max W_Y), \\ F_Y(x) &= \sum_{y' \le y} f_Y(y') = \sum_{y' \le y} \sum_{x' \in W_X} f_{X,Y}(x', y') = F_{X,Y}(\max W_X, y).\end{aligned}$$

F_X und F_Y heißen *Randverteilungen*.

Unabhängige *Ereignisse* beeinflussen sich gegenseitig nicht.

Analog kann man auch danach fragen ob zwei oder mehrere Zufallsvariablen sich gegenseitig nicht beeinflussen, also unabhängig sind.

Definition 20.16 (Unabhängigkeit von Zufallsvariablen)**.** Die ZV $X_1, \ldots, X_n$ heißen *unabhängig*, wenn für *alle* $(x_1, \ldots, x_n) \in W_{X_1} \times \ldots \times W_{X_n}$ gilt

$$\Pr[X_1 = x_1, \ldots, X_n = x_n] = \Pr[X_1 = x_1] \cdot \ldots \cdot \Pr[X_n = x_n].$$

$A_i =$ „$X_i = x_i$“ bezeichnet ein Ereignis und $X_1 = x_1, \ldots, X_n = x_n$ ist der Schnitt all dieser Ereignisse, also $\bigcap_{i=1}^n A_i$. Man fordert also, dass für alle möglichen Schnitte die Produktformel $\Pr[\bigcap_{i=1}^n A_i] = \prod_{i=1}^n \Pr[A_i]$ gilt. □

Der Zusammenhang mit der Definition unabhängiger Ereignisse wird nach dem folgenden Satz noch klarer.

Satz 20.17. $X_1, \ldots, X_n$ seien unabhängige ZV und $S_1 \subset W_{X_1}, \ldots, S_n \subset W_{X_n}$ beliebige Mengen. Dann sind die Ereignisse „$X_1 \in S_1$“, …, „$X_n \in S_n$“ unabhängig.

Beweis: Man rechnet nach

$$\begin{aligned}
\Pr[X_1 \in S_1, \ldots, X_n \in S_n] = & \\
&= \sum_{x_1 \in S_1} \ldots \sum_{x_n \in S_n} \Pr[X_1 = x_1, \ldots, X_n = x_n] \\
&= \sum_{x_1 \in S_1} \ldots \sum_{x_n \in S_n} \Pr[X_1 = x_1] \cdot \ldots \cdot \Pr[X_n = x_n] \\
&= \left(\sum_{x_1 \in S_1} \Pr[X_1 = x_1] \right) \cdot \ldots \cdot \left(\sum_{x_n \in S_n} \Pr[X_n = x_n] \right) \\
&= \Pr[X_1 \in S_1] \cdot \ldots \cdot \Pr[X_n \in S_n].
\end{aligned}$$

Dies gilt für beliebige Mengen, also darf man insbesondere beliebige S_i durch $\bar{S}_i$ ersetzen und damit sind die Ereignisse $X \in S_i$ auch nach Lemma 19.5 auf Seite 229 unabhängig. □

20.5 Zusammengesetzte Zufallsvariablen

Hat man n ZV $X_1, \ldots, X_n$ so kann man daraus mittels einer Funktion $g : \mathbb{R}^n \to \mathbb{R}$ eine neue ZV $Y : \Omega \to \mathbb{R}$ definieren mittels

$$Y(\omega) = g(X_1(\omega), \ldots, X_n(\omega)).$$

Dahinter stecken wieder Ereignisse:

$$„Y = y“ = \{\omega \in \Omega | g(X_1(\omega), \ldots, X_n(\omega)) = y\}.$$

Man nennt Y eine *zusammengesetzte* ZV.

Beispiel 20.18. Zweimaliges Werfen eines Würfels. Sei X die Augenzahl im ersten und Y die Augenzahl im zweiten Wurf. Dann bezeichnet $Z = X + Y$ die Summe der Augenzahlen. □

Wir leiten nun Rechenregeln für zusammengesetzte ZV her.

Satz 20.19. Seien X und Y *unabhängige* Zufallsvariablen. Weiter sei $Z = X + Y$. Dann gilt

$$f_Z(z) = \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x).$$

Beweis:

$$\begin{aligned}
f_Z(z) = \Pr[Z = z] &= \Pr[X + Y = z] = \Pr[\{\omega \mid X(\omega) + Y(\omega) = z\}] \\
&= \Pr\left[\bigcup_{x \in W_X} \{\omega \mid X(\omega) = x \wedge Y(\omega) = z - x\}\right] \\
&= \sum_{x \in W_X} \Pr[\{\omega \mid X(\omega) = x \wedge Y(\omega) = z - x\}] \\
&= \sum_{x \in W_X} \Pr[X = x, Y = z - x] \\
&= \sum_{x \in W_X} \Pr[X = x] \cdot \Pr[Y = z - x] = \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x).
\end{aligned}$$

Hier haben wir die Disjunktheit der Ereignisse $\{\omega \mid X(\omega) = x \wedge Y(\omega) = z - x\}$, den Additionssatz und die Unabhängigkeit der ZV benutzt. □

Diese Formel erinnert an eine sog. „Faltung“.

Auch über Erwartungswert und Varianz zusammengesetzer Zufallsvariablen lässt sich etwas sagen.

Satz 20.20 (Linearität des Erwartungswertes)**.** Für Zufallsvariablen $X_1, \ldots, X_n$ und $X = a_1X_1 + \ldots + a_nX_n$ wobei $a_1, \ldots, a_n \in \mathbb{R}$, gilt

$$\mathbb{E}[X] = a_1\mathbb{E}[X_1] + \ldots + a_n\mathbb{E}[X_n].$$

Beweis: Einfach einsetzen liefert

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} (a_1 \cdot X_1(\omega) + \ldots a_n \cdot X_n(\omega)) \cdot \Pr[\omega] \\
&= a_1 \cdot \left(\sum_{\omega \in \Omega} X_1(\omega) \cdot \Pr[\omega]\right) + \ldots + a_n \cdot \left(\sum_{\omega \in \Omega} X_n(\omega) \cdot \Pr[\omega]\right) \\
&= a_1\mathbb{E}[X_1] + \ldots + a_n\mathbb{E}[X_n].
\end{aligned}$$

□

Wichtig ist, dass keine Voraussetzungen an die ZV notwendig sind. Das werden wir noch ausnutzen.

Beispiel 20.21 (Seemannsproblem)**.** n betrunkene Seeleute torkeln nach dem Landgang zurück in ihre Kojen. In jeder Koje kommt genau ein Seeman zum liegen. Jede Zuordnung sei gleichwahrscheinlich. Wieviele Seeleute liegen im Mittel im richtigen Bett?

Kombinatorischer Lösungsansatz: Betrachte die $n!$ Permutationen $(x_1, \ldots, x_n)$, $x_i \in \{1, \ldots, n\}$, $x_i \neq x_j$. Für wieviele x_i gilt im Mittel $x_i = i$? Das ist sehr aufwändig.

Alternativ betrachte die n Zufallsvariablen

$$X_i = \begin{cases} 1 & \text{Seemann } i \text{ liegt im richtigen Bett,} \\ 0 & \text{sonst.} \end{cases}$$

und die Zufallsvariable

$$X = \text{„Anzahl Seeleute im richtigen Bett“}.$$

Es gilt $X = X_1 + \ldots + X_n$ und wir sind interessiert an $\mathbb{E}[X]$.

Nun gilt

$$\Pr[X_i = 1] = \frac{|\{(*, \ldots, *, i, *, \ldots, *)\}|}{|\Omega|} = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

und damit

$$\mathbb{E}[X_i] = 0 \cdot \Pr[X_i = 0] + 1 \cdot \Pr[X_i = 1] = \frac{1}{n}.$$

Weiter gilt dann

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot \frac{1}{n} = 1.$$

Also im Mittel liegt nur einer im richtigen Bett! □

Satz 20.22 (Multiplikativität des Erwartungswertes)**.** Für *unabhängige* ZV $X_1, \ldots, X_n$ gilt

$$\mathbb{E}[X_1 \cdot \ldots \cdot X_n] = \mathbb{E}[X_1] \cdot \ldots \cdot \mathbb{E}[X_n].$$

Beweis: Im Fall $n = 2$ gilt:

$$\begin{aligned}
\mathbb{E}[X \cdot Y] &= \sum_{x \in W_X} \sum_{y \in W_Y} xy \cdot \Pr[X = x, Y = y] \\
&= \sum_{x \in W_X} \sum_{y \in W_Y} xy \cdot \Pr[X = x] \cdot \Pr[Y = y] \\
&= \left(\sum_{x \in W_X} x \cdot \Pr[X = x] \right) \left(\sum_{y \in W_Y} y \cdot \Pr[Y = y] \right) = \mathbb{E}[X] \cdot \mathbb{E}[Y].
\end{aligned}$$

Der Fall $n > 2$ geht analog. □

Definition 20.23 (Indikatorvariable)**.** Sei $A \subseteq \Omega$ ein Ereignis. Eine ZV der Form

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \text{ (A tritt ein)}, \\ 0 & \text{sonst} \end{cases}$$

heißt *Indikatorvariable* (zum Ereignis A). □

Offenbar gilt

$$\begin{aligned}
\mathbb{E}[I_A] &= 0 \cdot \Pr[\bar{A}] + 1 \cdot \Pr[A] = \Pr[A], \\
\mathrm{Var}[I_A] &= \mathbb{E}[I_A^2] - \mathbb{E}[I_A]^2 = \Pr[A] - \Pr[A]^2 = \Pr[A](1 - \Pr[A]), \\
\mathbb{E}[I_{A_1} \cdot \ldots \cdot I_{A_n}] &= \Pr[A_1 \cap \ldots \cap A_n].
\end{aligned}$$

Letzteres folgt, weil das Produkt eins ist, wenn jede Indikatorvariable eins ist. □

Beispiel 20.24 (Fortsetzung des Seemannproblems)**.** Im Seemannproblem haben wir schon gesehen wie Indikatorvariablen die Berechnung des Erwartungswertes vereinfachen.

Jetzt wollen die Varianz $\mathrm{Var}[X]$ ausrechnen. Mit A_i dem Ereignis, dass Seemann i im richtigen Bett ist, gilt

$$\mathbb{E}[X_i X_j] = \mathbb{E}[I_{A_i} I_{A_j}] = \Pr[A_i \cap A_j] = \frac{1}{n(n-1)}.$$

(Überlegt man wie oben das $1/n$).

Weiter ist

$$\mathbb{E}[X_i^2] = 0^2 \cdot \Pr[\bar{A}] + 1^2 \cdot \Pr[A] = \frac{1}{n}.$$

Dann folgt für $X = X_1 + \ldots + X_n$:

$$\mathbb{E}[X^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} X_i X_j\right] = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} = 2.$$

Und schließlich

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2 - 1^2 = 1.$$

□

Für Erwartungswert von Summen und Produkten von ZV gelten einfache Formeln. Schließlich untersuchen wir noch die Varianz.

Satz 20.25. Für *unabhängige* ZV $X_1, \ldots, X_n$ und $X = X_1 + \ldots + X_n$ gilt

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n].$$

Beweis: Für $n = 2$, d.h. die ZV X und Y, erhalten wir:

$$\begin{aligned}\mathbb{E}[(X+Y)^2] &= \mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] \\ \mathbb{E}[X+Y]^2 &= (\mathbb{E}[X] + \mathbb{E}[Y])^2 = \mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2.\end{aligned}$$

Die erste Beziehung benötigt die Unabhängigkeit für den gemischten Term. Schließlich gilt

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathrm{Var}[X] + \mathrm{Var}[Y].\end{aligned}$$

□

20.6 Zusammenfassung

- Zufallsvariable erlauben eine kompakte Definition von Ereignissen.
- Dichte- und Verteilungsfunktion.
- Erwartungswert und Varianz.
- Mehrere und zusammengesetzte Zufallsvariablen und ihre Eigenschaften.

21 Diskrete Verteilungen

Eine ZV $X : \Omega \to \mathbb{R}$ definiert einen neuen Wahrscheinlichkeitsraum (W_X, f_X) mit der Wertemenge W_X und der Dichtefunktion f_X.

In der Praxis führen unterschiedliche Anwendungen auf gleiche Verteilungen. Daher macht es Sinn diese Verteilungen zu untersuchen ohne eine konkrete Anwendung zu betrachten.

Oftmals genügt es dann sich zu überzeugen, dass für eine konkrete Anwendung ein bestimmte Verteilung anwendbar ist.

Die hier vorgestellten Verteilungen hängen von Parametern ab, die dann entsprechend anzupassen sind.

21.1 Bernoulli-Verteilung

Ist ganz simpel

$$W_X = \{0, 1\} \qquad \text{und } f_X(x) = \begin{cases} p & \text{für } x = 1 \\ 1 - p & \text{für } x > 0 \end{cases}$$

Der Parameter p heißt *Erfolgswahrscheinlichkeit.*

Die Bernoulli-Verteilung modelliert ein Bernoulli-Experiment. Das ist ein Experiment bei dem es nur zwei Ausgänge $\{0, 1\}$ gibt. Der Ausgang 1 wird mit Wahrscheinlichkeit p angenommen.

Es gilt ausserdem gilt mit $q = 1 - p$:

$$\mathbb{E}[X] = p, \qquad \mathrm{Var}[X] = pq$$

wegen $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p) = pq$.

Die oben eingeführten *Indikatorvariablen* sind Bernoulli-verteilt.

21.2 Binomial-Verteilung

Die n-malige Durchführung eines Bernoulli-Experiments heißt *Bernoulli*-Kette.

Beispiel 21.1. Hier einige Anwendungen:

1. n-maliges Würfeln mit sechseitigem Würfel. $X =$ „Anzahl Einsen“.
2. n-maliges Werfen einer Münze. $X =$ „Anzahl Kopf“.
3. Parallelrechner mit n Rechnern. Mit Wahrscheinlichkeit p ist ein Rechner intakt. $X =$ „Anzahl der intakten Rechner“.
4. Ein Druckerserver prüft n mal in einer Stunde ob ein neuer Druckauftrag vorliegt. n sei so groß, dass bei einer Überprüfung höchstens ein neuer Auftrag vorliegt. Die Wahrscheinlichkeit dafür sei p. $X =$ „Anzahl Aufträge in einer Stunde“. □

Kennzeichen dieser Experimente ist, dass die Erfolgswahrscheinlichkeit in jedem Teilschritt gleich ist. Die Teilexperimente sind unabhängig.

Man sagt eine ZV ist binomialverteilt, wenn gilt

$$W_X = \{0, \ldots, n\}, \qquad f_X(x) = \binom{n}{x} p^x q^{n-x} =: b(x; n, p).$$

Die Binomial-Verteilung hat also zwei Parameter: $n \in \mathbb{N}$ und $p \in (0, 1)$.

Man sagt in diesem Fall auch $X \sim \text{Bin}(n, p)$.

Wegen $X = X_1 + \ldots + X_n$, mit X_i Bernoulli-verteilt, folgt aus den Sätzen über Summen unabhängiger ZV für Erwartungswert und Varianz:

$$\mathbb{E}[X] = n \cdot \mathbb{E}[X_i] = n \cdot p, \qquad \text{Var}[X] = n \cdot p \cdot q.$$

Herleitung der Dichtefunktion über das Zufallsexperiment:

Die Ergebnismenge ist $\Omega = \{0, 1\}^n$.

Die Wahrscheinlichkeit eines Elementarereignisses ist $\Pr[\omega \in \Omega] = p^x(1-p)^{n-x}$ wobei x die Anzahl der Erfolge in ω ist.

Pr ist ein Wahrscheinlichkeitsmaß (Betrachte den zugehörigen Entscheidungsbaum).

Betrachte Ereignisse $\Omega \supseteq A_x = \{\omega \in \Omega \,|\, \omega \text{ enthält } x \text{ Mal Erfolg}\}$ dann gilt $|A_x| = \binom{n}{x}$ also

$$\Pr[A_x] = f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Mit der binomischen Formel folgt auch sofort:

$$\sum_{x=0}^{n} f_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1^n = 1.$$

Die Abbildungen 28 und 29 zeigen die Binomial-Verteilung für festes n und verschiedene Werte von p.

Die Abbildungen 30 und 31 zeigen die Binomial-Verteilung für $p = 0.5$ und verschiedene Werte von n. Man kann zeigen, dass das Maximum wie $O(1/\sqrt{n})$ fällt.

Satz 21.2. Sei $X \sim \text{Bin}(n_x, p)$ und $Y \sim \text{Bin}(n_y, p)$ und $Z = X + Y$. Dann gilt $Z \sim \text{Bin}(n_x + n_y, p)$.

Dies ist klar, da X und Y jeweils die Summe aus n_x bzw. n_y Bernoulli-verteilten ZV sind und damit Z die Summe aus $n_x + n_y$ Bernoulli-verteilten ZV ist. X und Y müssen Bernoulli-verteilt mit dem selben p sein! □

Typische Situationen, in denen die Binomial-Verteilung Anwendung findet:

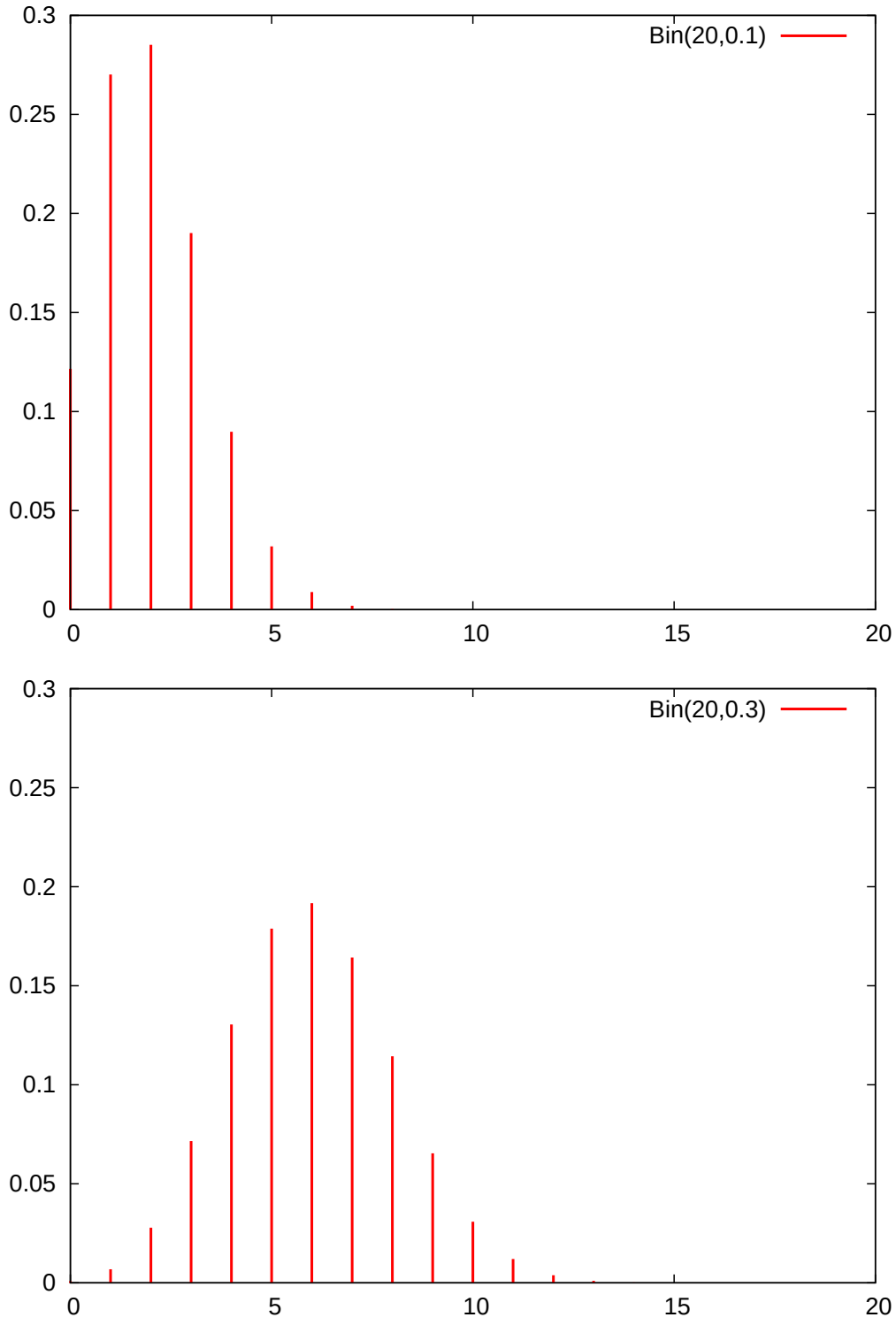


Abbildung 28: Binomial-Verteilung für $n = 20$ und $p = 0.1, 0.3$.

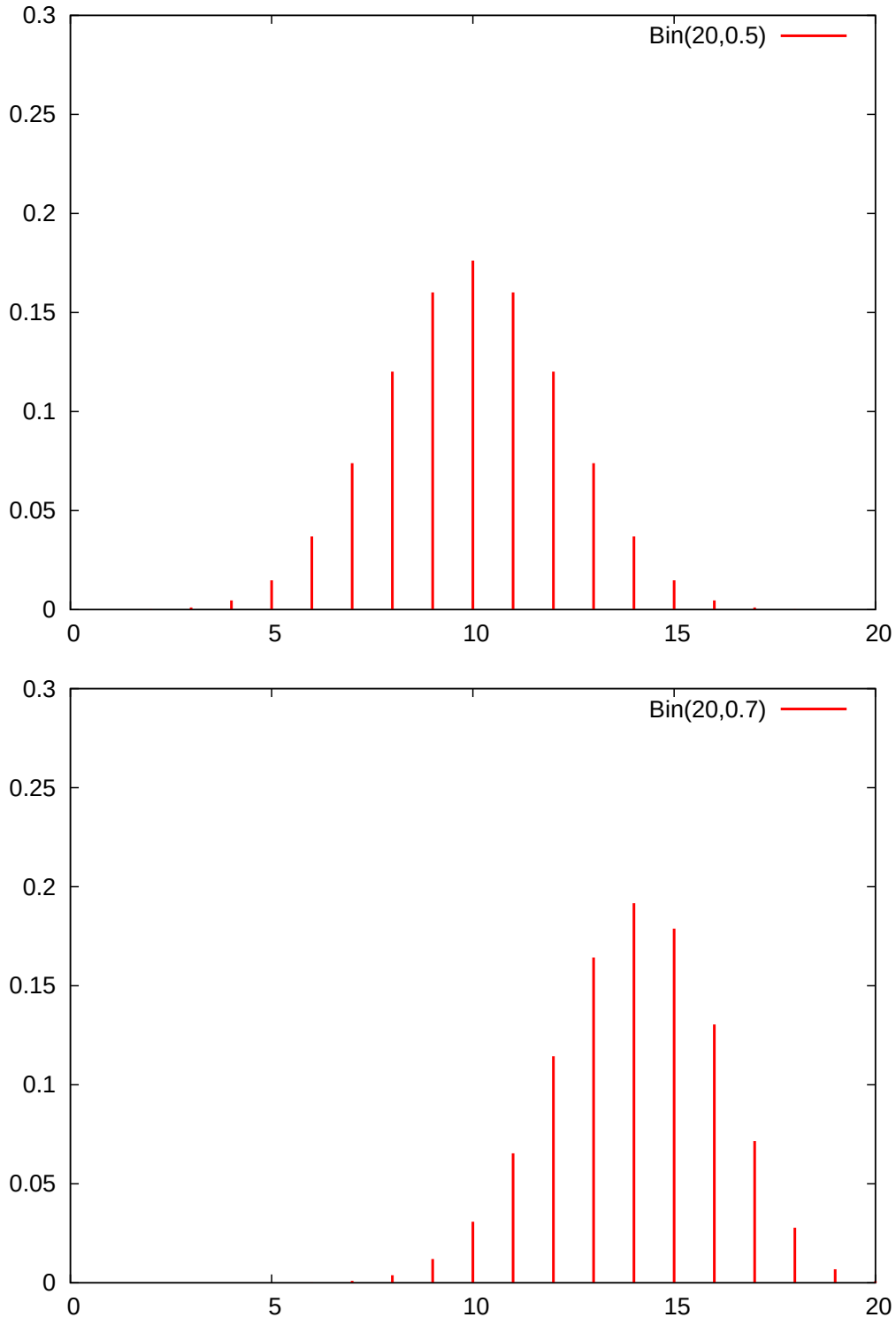


Abbildung 29: Binomial-Verteilung für $n = 20$ und $p = 0.5, 0.7$.

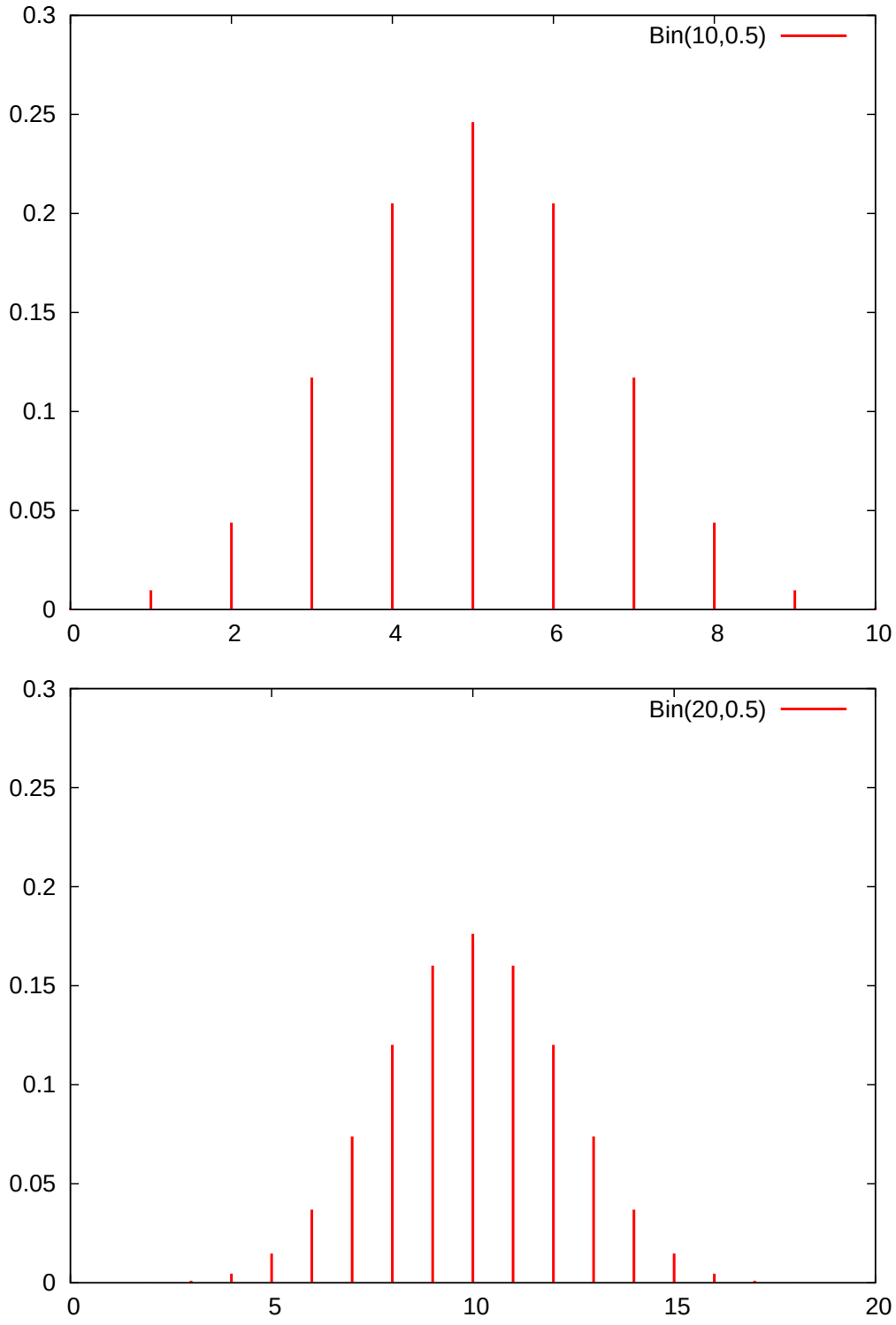


Abbildung 30: Binomial-Verteilung für $p = 0.5$ und $n = 10, 20$.

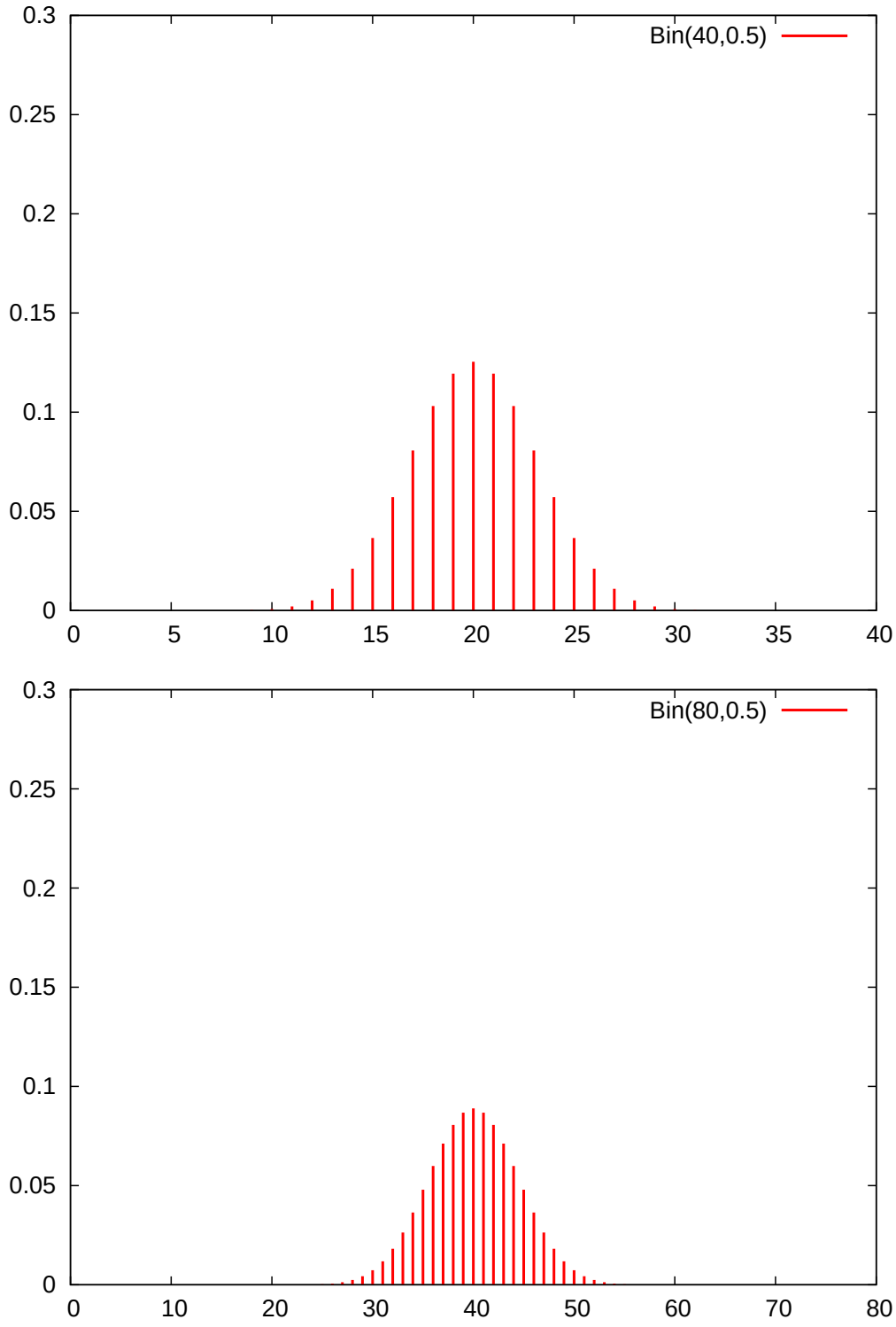


Abbildung 31: Binomial-Verteilung für $p = 0.5$ und $n = 40, 80$.

- Unabhängige, n-malige Wiederholung eines Zufallsxperiments (Würfel, Münze).
- n-maliges Ziehen mit zurücklegen aus einer endlichen Grundgesamtheit. X bezeichnet wie oft ein Element mit einer bestimmten Eigenschaft gezogen wurde.
- n-maliges Ziehen ohne Zurücklegen aus einer unendlichen Grundgesamtheit. Das Ziehen beeinflusst die Wahrscheinlichkeit für die folgenden Ziehungen nicht. Anwendung: Stichprobe aus einer laufenden Produktion.
- n-maliges Ziehen ohne Zurücklegen aus einer endlichen, aber sehr großen Grundgesamtheit mit N Elementen. Hier stellt die Binomial-Verteilung aber nur eine Näherung (der hypergeometrischen Verteilung dar). Es gibt verschiedene Faustregeln: etwa $n \leq N/20$, siehe [TT07].

21.3 Geometrische Verteilung

Diese baut auch auf der Durchführung von Bernoulli-Experimenten auf.

Man führt Bernoulli-Experimente solange durch *bis zum ersten Mal Erfolg eintritt.*

Die ZV X bezeichne die Anzahl der Versuche bis zum Erfolg.

Damit gilt mit $q = 1 - p$:

$$W_X = \mathbb{N} \text{ also ohne } 0, \qquad f_X(i) = pq^{i-1}.$$

$f_X(i)$ ist die Wahrscheinlichkeit, dass der erste Erfolg im i-ten Versuch eintritt.

Man rechnet nach (entsprechende Reihen):

$$\mathbb{E}[X] = \frac{1}{p} \qquad \mathrm{Var}[X] = \frac{q}{p^2}.$$

Die Abbildungen 32 und 33 zeigen die geometrische Verteilung für verschiedene Werte von p.

Dichtefunktion ist streng monoton fallend und „erinnert“ an eine (abgetastete) Exponentialfunktion. Darauf werden wir im Kapitel über kontinuierliche Wahrscheinlichkeitsräume wieder zurückkommen.

Wir betrachten nun eine wichtige Eigenschaft der geometrischen Verteilung.

Die geometrische Verteilung basiert auf der Annahme, dass die einzelnen Versuche unabhängig voneinander sind.

Wir untersuchen die Wahrscheinlichkeit $\Pr[X > y + x \,|\, X > x]$.

Wenn wir *wissen*, dass x Versuche erfolglos waren dann können wir uns vorstellen erst mit dem $(x+1)$-ten Versuch zu beginnen. Ab da ist das Eintreten des Erfolges wieder geometrisch verteilt, da die Vorgeschichte irrelevant ist.

Also erwarten wir

$$\Pr[X > y + x \,|\, X > x] = \Pr[X > y].$$

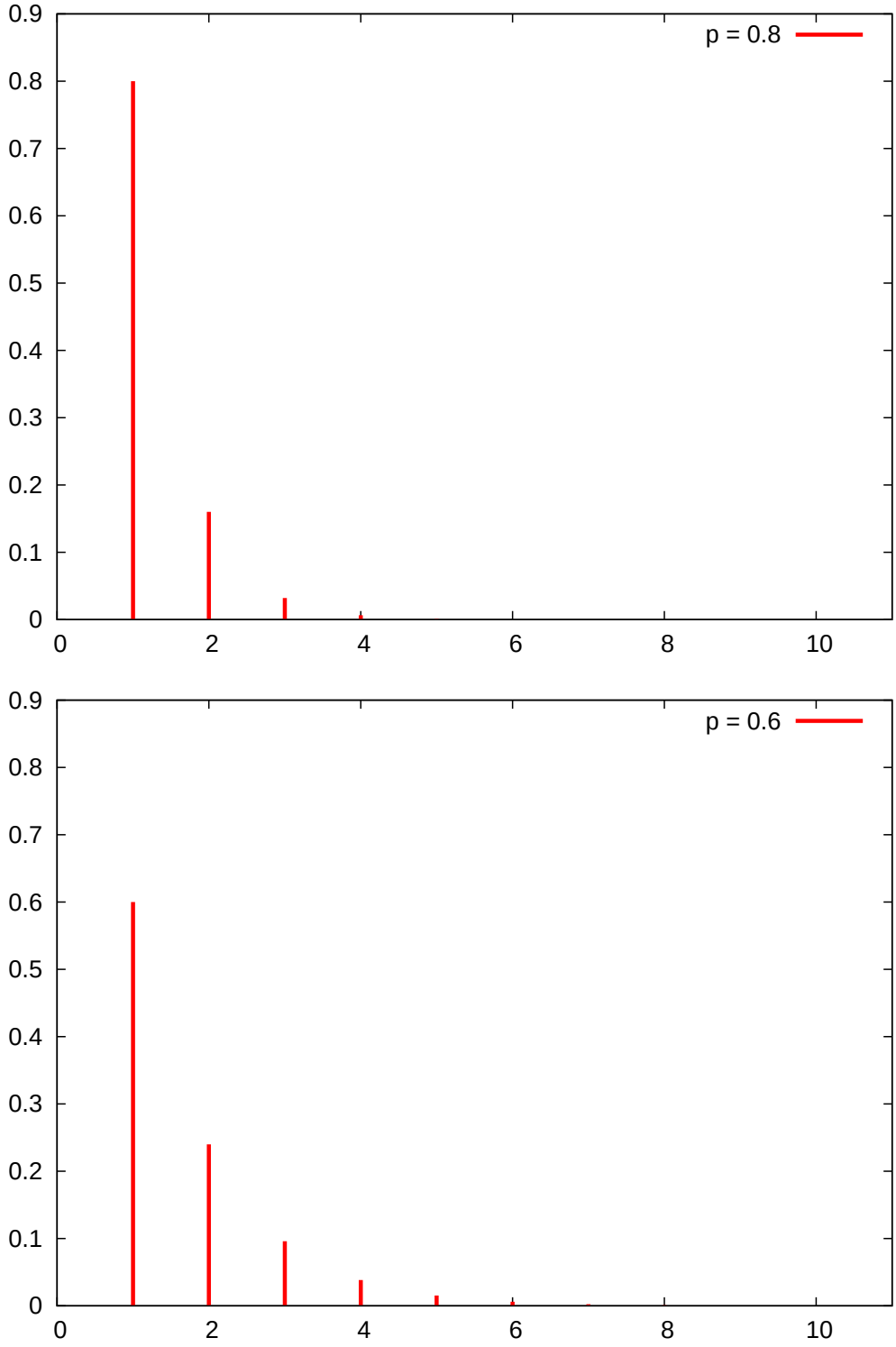


Abbildung 32: Geometrische Verteilung für $p = 0.8, 0.6$.

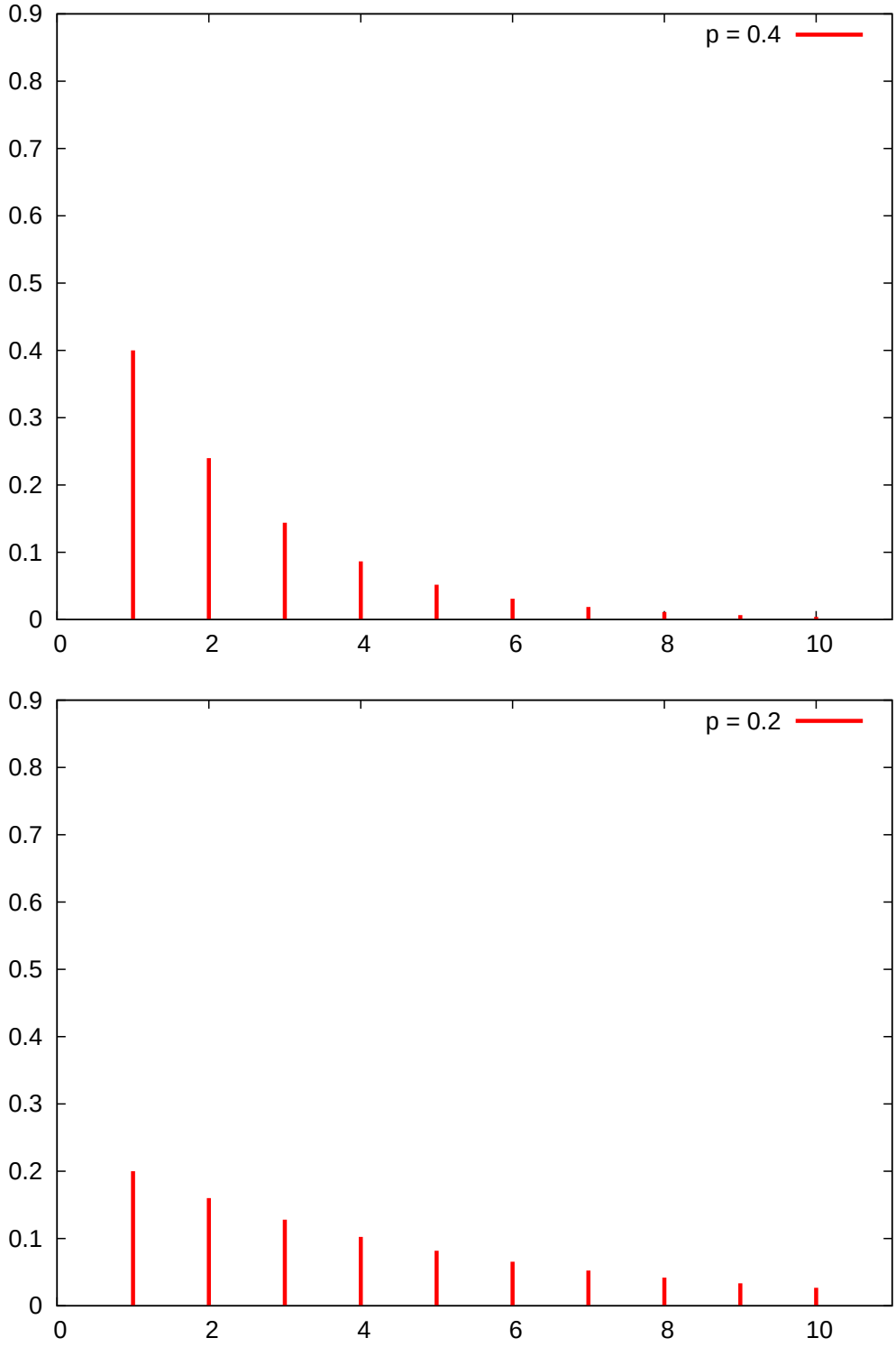


Abbildung 33: Geometrische Verteilung für $p = 0.4, 0.2$.

Dies nennt man *Gedächtnislosigkeit*.

Das rechnen wir jetzt formal nach. Erst die Bedingung:

$$\begin{aligned}\Pr[X > x] &= \sum_{i=x+1}^{\infty} p(1-p)^{i-1} = (1-p)^x p \sum_{i=0}^{\infty} (1-p)^i \\ &= (1-p)^x p \cdot \frac{1}{1-(1-p)} = (1-p)^x.\end{aligned}$$

Das ist klar: Das Ereignis mindestens $x+1$ Versuche zu brauchen tritt genau ein wenn man x mal Misserfolg hatte.

Damit gilt dann

$$\begin{aligned}\Pr[X > y + x | X > x] &= \frac{\Pr[„X > y + x“ \cap „X > x“]}{\Pr[X > x]} = \frac{\Pr[X > y + x]}{\Pr[X > x]} \\ &= (1-p)^{y+x} \cdot (1-p)^{-x} = (1-p)^y = \Pr[X > y].\end{aligned}$$

Dabei benutzten wir, dass „$X > y + x$“ $\subseteq$ „$X > x$“.

Für die Verteilungsfunktion der geometrischen Verteilung erhalten wir mit obigem Resultat $F_X(x) = \Pr[X \leq x] = 1 - \Pr[X > x] = 1 - (1-p)^x$.

Die Abbildungen 34 und 34 zeigen die Verteilungsfunktion für die geometrische Verteilung für verschiedene Werte von p.

Bemerkung 21.3 (Geometrische Verteilung und Binomialverteilung)**.** Der Erwartungswert der geometrischen Verteilung $\mathbb{E}[X] = 1/p$ gibt die mittlere Zahl von Versuchen bis zum Erfolg an.

Für $n = 1/p$ hat die Binomialverteilung den Erwartungswert $\mathbb{E}[X] = np = p/p = 1$, d. h. wir erwarten im Mittel einen Erfolg bei $1/p$ Versuchen.

Beide Verteilungen bauen auf dem Bernoulli-Experiment auf, bieten aber unterschiedliche Sichtweisen:

- Die Anzahl der Erfolge (bei bekannter Zahl von Versuchen) ist binomialverteilt.
- Der Abstand zwischen den Erfolgen ist geometrisch verteilt.

□

Die geometrische Verteilung beschreibt das klassischen Warteproblem. Ein komplizierteres Warteproblem beschreibt das folgende Beispiel.

Beispiel 21.4 (Hanuta-Problem)**.** Jeder Hanuta-Packung liegt zufällig eines von n Abziehbildchen bei. Wieviele Hanuta muss man im Mittel kaufen (und essen!), bis man die komplette Sammlung besitzt?

Definiere Zufallsvariable $X =$ „Anzahl Käufe bis man alle hat“.

Teile den gesamten Ablauf in Phasen ein: Phase i sind die Versuche vom Erwerb des $(i-1)$-ten Bildes (ausschließlich) bis zum Erwerb des i-ten Bildes (einschließlich).

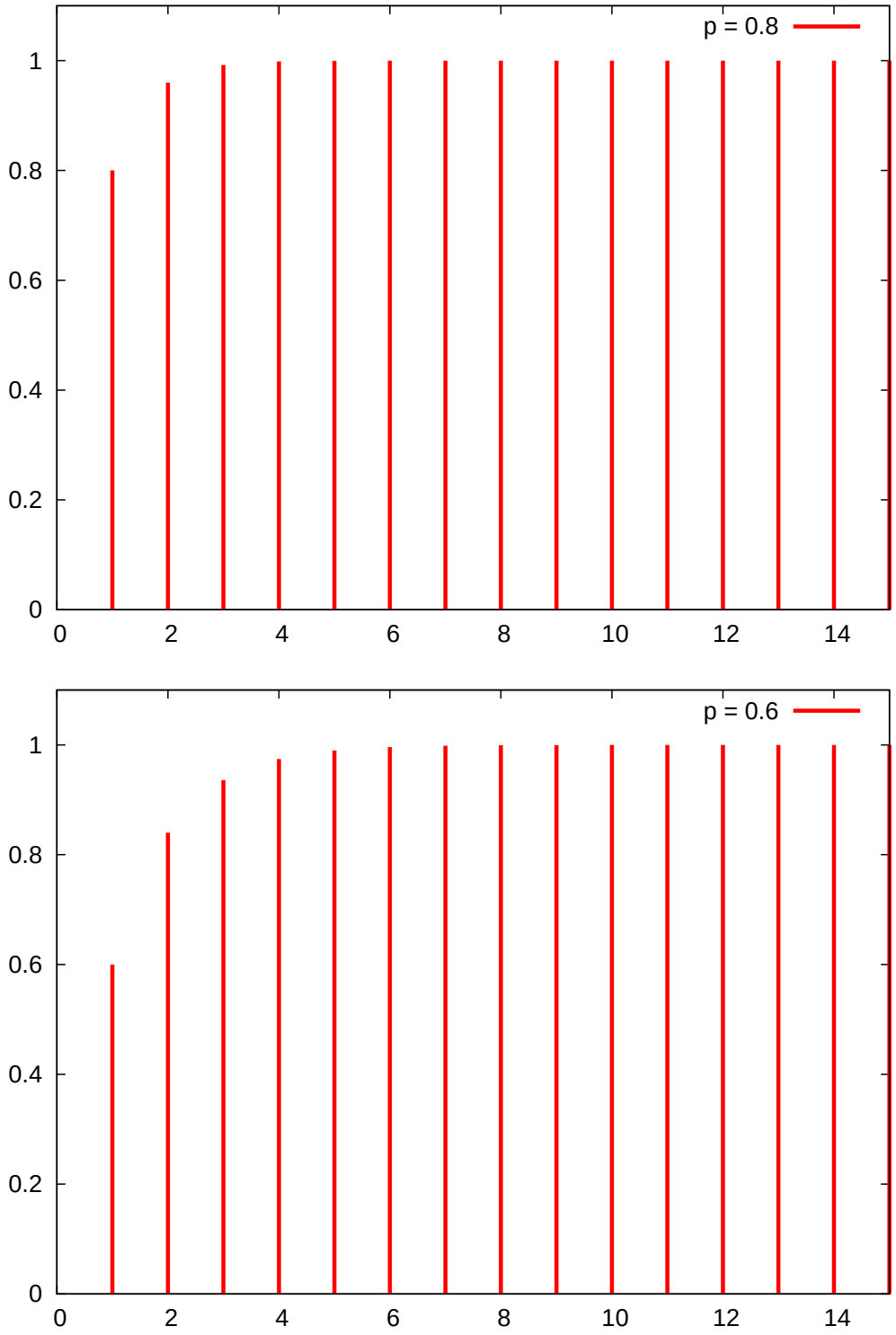


Abbildung 34: F_X der geometrischen Verteilung für $p = 0.8, 0.6$.

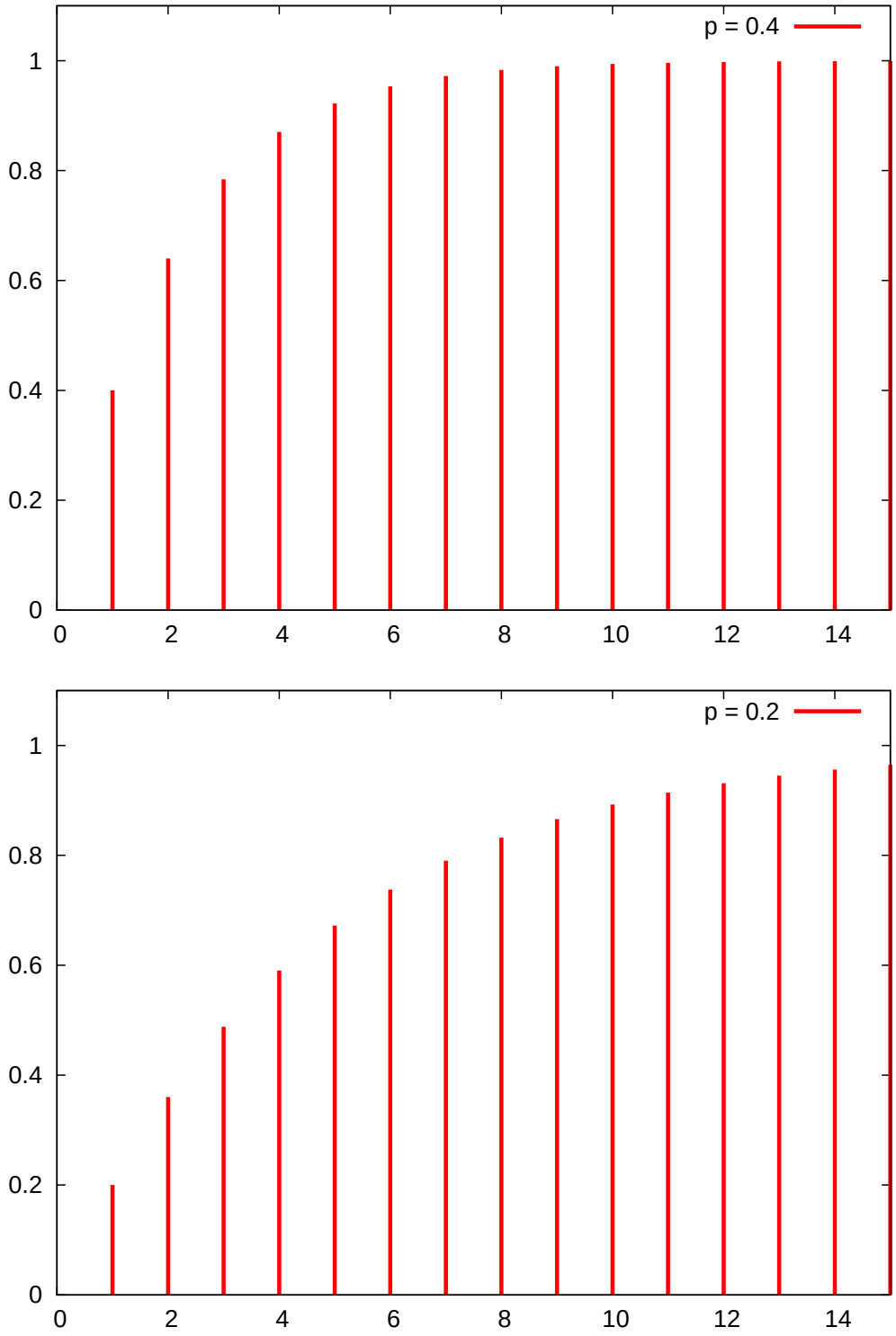


Abbildung 35: F_X der geometrischen Verteilung für $p = 0.4, 0.2$.

Beispiel im Beispiel. Ein möglicher Ablauf bei $n = 4$ wäre:

$$\underbrace{\boxed{3}}_{\substack{\text{Phase 1}\\ \text{(immer Länge 1)}}} \quad \underbrace{,3,3,\boxed{2}}_{\text{Phase 2}}, \quad \underbrace{2,3,3,2,\boxed{4}}_{\text{Phase 3}}, \quad \underbrace{2,4,2,2,3,4,\boxed{1}}_{\text{Phase 4}}.$$

Definiere nun die ZV $X_i =$ „Anzahl Versuche in Phase i“.

X_i ist geometrisch verteilt mit $p = \frac{n-i+1}{n}$, da man in Phase i noch auf eines von $n - (i-1)$ übrigen Bildchen von n Bildchen insgesamt wartet.

Damit ist $\mathbb{E}[X_i] = \frac{1}{p} = \frac{n}{n-i+1}$.

Wegen und $X = X_1 + \ldots + X_n$ und der Linearität des Erwartungswertes gilt

$$\mathbb{E}[X] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \cdot \sum_{i=1}^{n} \frac{1}{i} = n \cdot H_n.$$

H_n ist die n-te harmonische Zahl. Für diese gilt $H_n = \ln n + O(1)$, also

$$\mathbb{E}[X] = n \ln n + O(n).$$

Man muss also im Mittel nur $n \ln n$ Hanuta kaufen. Eigentlich ziemlich wenig. □

Informatikbezug dieses Beispiels: Stelle alle Benutzer eines Netzwerkes durch Abhören fest.

21.4 Poisson-Verteilung

Die *Poisson-Verteilung*[43] lautet:

$$W_X = \mathbb{N}_0, \qquad f_X(i) = \frac{e^{-\lambda}\lambda^i}{i!} \text{ für } i \in \mathbb{N}_0.$$

Dass f_X eine zulässige Dichte ist zeigt

$$\sum_{i=0}^{\infty} f_X(i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Ist die ZV X Poisson-verteilt mit Parameter λ so schreibt man kurz $f_X(x) \sim \mathrm{Po}(\lambda)$.

Berechnen wir den Erwartungswert:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i \cdot \frac{e^{-\lambda}\lambda^i}{i!} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

[43] Siméon Denis Poisson, 1781-1840, frz. Mathematiker.

Mit dem Zwischenergebnis

$$\mathbb{E}[X(X-1)] = \sum_{i=0}^{\infty} i(i-1) \cdot \frac{e^{-\lambda}\lambda^i}{i!} = \lambda^2 e^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} = \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2$$

erhält man die Varianz:

$$\begin{aligned} \mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X] + \mathbb{E}[X] - \mathbb{E}[X]^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

Die Abbildungen 36 und 37 zeigen die Poisson-Verteilung für verschiedene Werte von λ.

Beispiel 21.5 (Druckerserver)**.** Benutzer einer Rechenanlage legen ihre Druckaufträge in einer Warteschlange ab. Der Druckerserver prüft periodisch ob Aufträge vorliegen und führt diese gegebenenfalls aus.

Angenommen es kommen im Mittel λ Aufträge pro Zeiteinheit an, dann ist die Wahrscheinlichkeit, dass in einer Zeiteinheit genau k Aufträge eingehen Poisson-verteilt mit Parameter λ. □

Die Bilder zur Poisson-Verteilung erinnern stark an die Binomial-Verteilung. Dasss hier ein sehr enger Zusammenhang besteht erläutert die nun folgende Überlegung.

Beispiel 21.6 (Fortsetzung des Druckerserverbeispiels)**.** Wir spezifizieren die Arbeitsweise des Druckerservers noch etwas genauer.

Die Zeiteinheit werde in $n \in \mathbb{N}$ gleichgroße Teilintervalle geteilt und der Druckerserver prüft jeweils am Ende des abgelaufenen Intervalls ob ein Auftrag eingegangen ist (also n mal pro Zeiteinheit).

Dabei sei n so groß, dass in einem Teilintervall nur höchstens ein Auftrag eingehen kann.

Das Nachsehen am Ende des Teilintervalles können wir als Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p_n auffassen. Die Größe $X_n =$ „Anzahl Aufträge pro Zeiteinheit“ ist binomialverteilt mit $X_n \sim \mathrm{Bin}(n, p_n)$.

Die Erfolgswahrscheinlichkeit p_n ist eine Funktion von n: Je mehr Teilintervalle n man bildet, desto kleiner ist die Wahrscheinlichkeit p_n dass in einem Intervall ein Auftrag ankommt.

Man bestimmt p_n mittels

$$\mathbb{E}[X_n] = np_n = \lambda \quad \Rightarrow \quad p_n = \lambda/n.$$

□

Die Poisson-Verteilung kann als Grenzwert der Binomial-Verteilung für $p_n = \lambda/n$ und $n \to \infty$

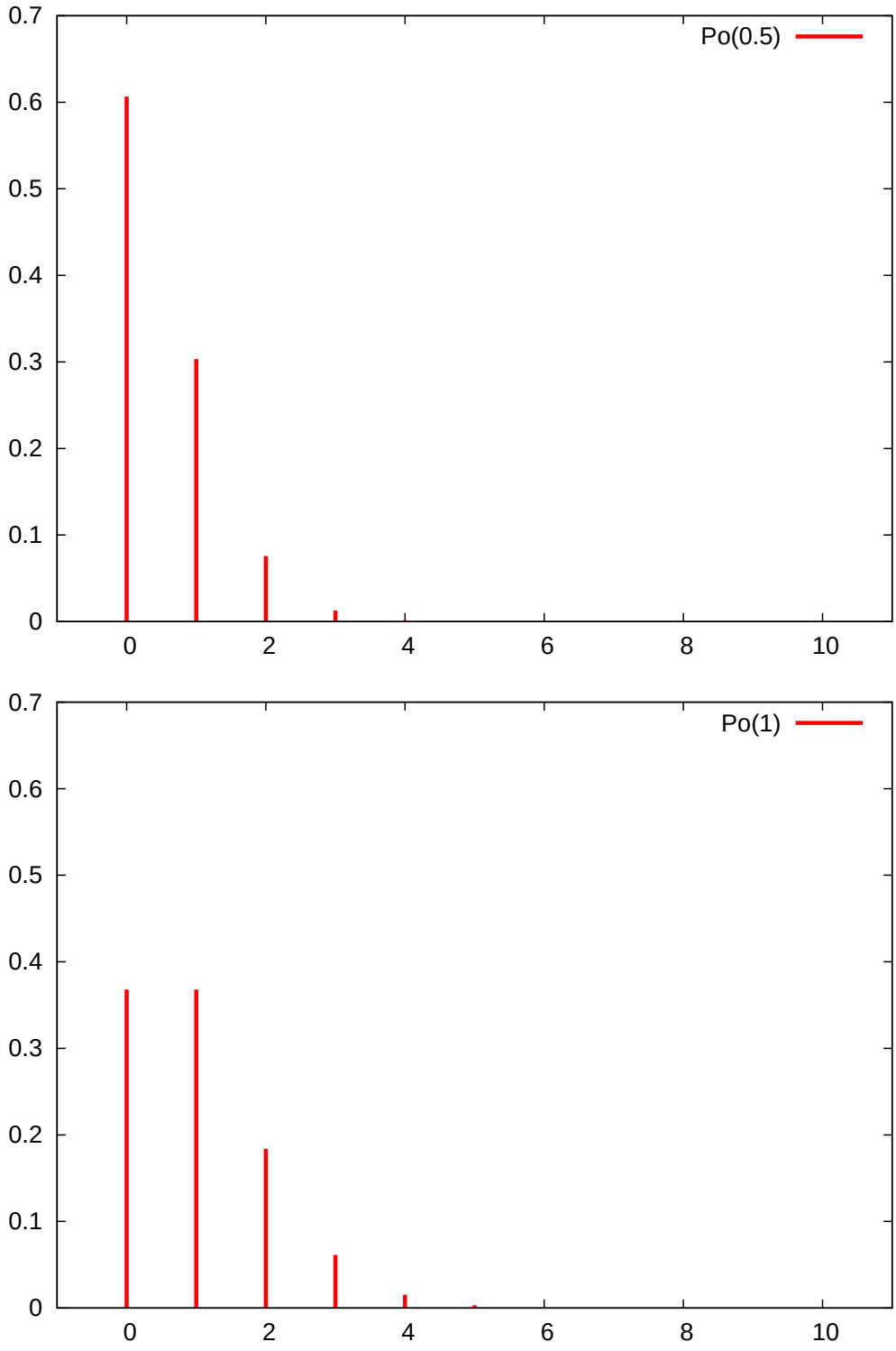


Abbildung 36: Poisson-Verteilung für $\lambda = 0.5, 1$.

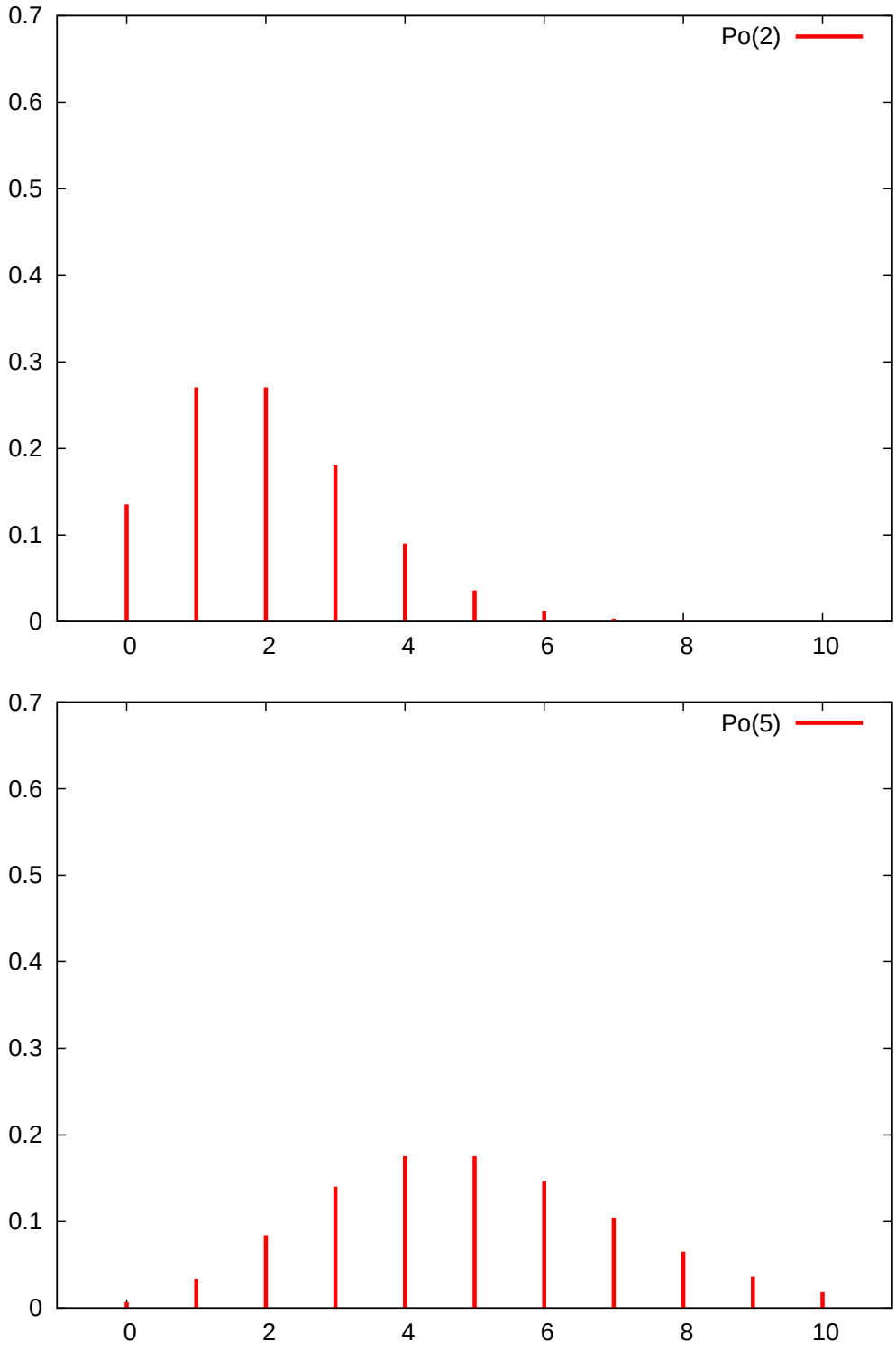


Abbildung 37: Poisson-Verteilung für $\lambda = 2,5$.

aufgefasst werden:

$$\begin{aligned} b(k;n,p_n) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n!}{k!\cdot(n-k)!} p_n^k (1-p_n)^{n-k} \\ &= \frac{(np_n)^k}{k!} \cdot \frac{n!}{n^k \cdot (n-k)!} \cdot (1-p_n)^{-k} \cdot (1-p_n)^n \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n!}{n^k \cdot (n-k)!}}_{\to 1} \cdot \underbrace{(1-\frac{\lambda}{n})^{-k}}_{\to 1} \cdot \underbrace{(1-\frac{\lambda}{n})^n}_{\to e^{-\lambda}} \\ &= \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

Der Vorteil ist, dass die im letzten Beispiel eingeführte künstliche Zeitdiskretisierung nicht mehr notwendig ist.

Für hinreichend kleines p kann man die Poisson-Verteilung als Näherung für die Binomial-Verteilung verwenden.

Deswegen sagt man auch *Gesetz seltener Ereignisse*.

Die Poisson-Verteilung modelliert wie oft ein Ereignis innerhalb einer festen Zeitspanne eintritt.

Aufgrund der Herleitung als Grenzübergang der Binomial-Verteilung müssen folgende Voraussetzungen für ihre Anwendung erfüllt sein:

- Ereignisse treten nie gleichzeitig auf, denn für n hinreichend groß, d. h. p_n hinreichend klein haben wir angenommen dass in jedem Teilintervall höchstens ein Ereignis eintritt.
- Die Wahrscheinlichkeit dass ein Ereignis in einem kleinen Zeitintervall Δt auftritt ist proportional zu dessen Länge, denn bei $\Delta t = T/n$ ist die Wahrscheinlichkeit $p_n = \lambda/n = \lambda \Delta t/T$.
- Die Anzahl der Ereignisse in einem festen Zeitintervall hängt nur von der Länge aber nicht von der absoluten Zeit ab, denn die W. des Eintretens ist in jedem Δt gleich.
- In zwei disjunkten Zeitintervallen sind die Anzahlen des Eintretens der Ereignisse unabhängig voneinander. Auch dies ist klar, wenn man jedes Intervall als Bernoulli-Kette approximiert.

21.5 Zusammenfassung

- In diesem Abschnitt wurden Binomial-, geometrische und Poisson-Verteilung eingeführt.
- Die Binomial-Verteilung $\mathrm{Bin}(n,p)$ gibt an wie wahrscheinlich es ist genau k-mal Erfolg zu haben bei n gleichen Versuch mit der Erfolgswahrscheinlichkeit p.
- Die geometrische Verteilung gibt an wie wahrscheinlich es ist genau im i-ten Versuch zum ersten Mal erfolg zu haben wenn die Erfolgswahrscheinlichkeit in jedem Versuch p ist.
- Die Poisson-Verteilung $\mathrm{Po}(\lambda)$ gibt an wie wahrscheinlich es ist, dass innerhalb einer Zeitspanne genau k Aufträge ankommen, wenn im Mittel λ Aufträge pro Zeitspanne ankommen.

22 Asymptotik

22.1 Ungleichungen von Markov und Chebyshev

Wir lernen nun Methoden kennen mit denen man Wahrscheinlichkeiten gewisser Ereignisse abschätzen kann *ohne* die Dichtefunktion explizit zu können.

Stattdessen versucht man nur mit Erwartungswert und Varianz auszukommen.

Ein erstes Resultat ist der

Satz 22.1 (Ungleichung von Markov[44])**.** Sei X eine ZV, die nur *nichtnegative* Werte annimmt. Dann gilt für alle $t \in \mathbb{R}$ mit $t > 0$:

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Dazu äquivalent ist die Aussage

$$\Pr\Big[x \geq t \cdot \mathbb{E}[X]\Big] \leq \frac{1}{t}.$$

Beweis: Rechne

$$\mathbb{E}[X] = \sum_{x \in W_x} x \cdot \Pr[X = x] \geq \sum_{x \in W_x : x \geq t} x \cdot \Pr[X = x] \geq t \sum_{x \in W_x : x \geq t} \Pr[X = x] = t \cdot \Pr[X \geq t].$$

Die zweite Aussage ergibt sich indem man $t = s \cdot \mathbb{E}[X]$ einsetzt. □

Der folgende Satz ist eine Anwendung hiervon.

Satz 22.2 (Ungleichung von Chebyshev[45])**.** Sei X eine ZV und $\mathbb{R} \ni t > 0$, dann gilt

$$\Pr\Big[|X - \mathbb{E}[X]| \geq t\Big] \leq \frac{\mathrm{Var}[X]}{t^2}.$$

Äquivalent dazu ist die Aussage

$$\Pr\Big[|X - \mathbb{E}[X]| \geq t \cdot \sqrt{\mathrm{Var}[X]}\Big] \leq \frac{1}{t^2}$$

Beweis: Quadrieren der Bedingung ändert nichts an dem Ereignis und seiner Wahrscheinlichkeit, d. h.

$$\Pr\Big[|X - \mathbb{E}[X]| \geq t\Big] = \Pr\Big[(X - \mathbb{E}[X])^2 \geq t^2\Big].$$

Nun gilt nach Definition der Varianz:

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}[X]$$

und damit nach der Ungleichung von Markov, denn $(X - \mathbb{E}[X])^2$ hat nur nichtnegative Werte,

$$\Pr\Big[|X - \mathbb{E}[X]| \geq t\Big] = \Pr\Big[(X - \mathbb{E}[X])^2 \geq t^2\Big] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}.$$

Die zweite Aussage ergibt sich wieder durch die Substitution $t^2 = s^2 \cdot \mathrm{Var}[X]$. □

[44] Andrey Andreyevich Markov, 1856-1922, russ. Mathematiker.
[45] Pavnuty Lvovich Chebyshev, 1821-1894, russ. Mathematiker.

Beispiel 22.3. Betrachte das n-malige Werfen eines Würfels. Die ZV $X_n = \#$ Einsen ist binomialverteilt wie $\mathrm{Bin}(n,p)$, $p = 1/6$, mit Erwartungswert np und Varianz $np(1-p)$.

Wir sind nun interessiert an der Wahrscheinlichkeit, dass die Anzahl der gewürfelten Einsen *um 10% größer als der Erwartungswert ist.*

Mit der Ungleichung von Chebyshev erhalten wir:

$$\begin{aligned}\Pr[X_n \geq \mathbb{E}[X_n] + 0.1 \cdot \mathbb{E}[X_n]] &= \Pr[X_n - \mathbb{E}[X_n] \geq 0.1 \cdot \mathbb{E}[X_n]] \\ &\leq \Pr[|X_n - \mathbb{E}[X_n]| \geq 0.1 \cdot \mathbb{E}[X_n]] \\ &\leq \frac{\mathrm{Var}[X_n]}{0.1^2 \cdot \mathbb{E}[X_n]^2} = \frac{np(1-p)}{0.1^2 n^2 p^2} \\ &= \frac{1-p}{0.1^2 p} \cdot \frac{1}{n}.\end{aligned}$$

Diese Wahrscheinlichkeit fällt also mit steigendem n und wir können sogar abschätzen wie schnell. □

22.2 Gesetz der großen Zahlen

Bei der Einführung des Wahrscheinlichkeitsraumes haben wir den Begriff der Wahrscheinlichkeit mit der „relativen Häufigkeit“ motiviert.

Mit den hergeleiteten Abschätzungen können wir das nun formal fassen.

Satz 22.4 (Gesetz der großen Zahlen)**.** Sei X eine ZV, $\varepsilon, \delta > 0$ beliebige aber fest gewählte Zahlen sowie $n \in \mathbb{N}$ mit $n \geq \frac{\mathrm{Var}[X]}{\varepsilon \cdot \delta^2}$.

Sind $X_1, \ldots X_n$ *unabhängige* ZV mit der selben Verteilung wie X und

$$Z := \frac{X_1 + \ldots + X_n}{n}$$

das arithmetische Mittel, so gilt

$$\Pr\Big[|Z - \mathbb{E}[X]| \geq \delta\Big] \leq \varepsilon.$$

Beweis: Berechne zunächst den Erwartungswert von Z:

$$\mathbb{E}[Z] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n \cdot \mathbb{E}[X] = \mathbb{E}[X]$$

sowie die Varianz (unabhänge ZV):

$$\mathrm{Var}[Z] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{n^2} \cdot n \cdot \mathrm{Var}[X] = \frac{\mathrm{Var}[X]}{n}.$$

Nun die Ungleichung von Chebyshev:

$$\Pr\Big[|Z - \mathbb{E}[X]| \geq \delta\Big] = \Pr\Big[|Z - \mathbb{E}[Z]| \geq \delta\Big] \leq \frac{\mathrm{Var}[Z]}{\delta^2} = \frac{\mathrm{Var}[X]}{n \cdot \delta^2} \leq \varepsilon.$$

□

Die Konvergenz der relativen Häufigkeit gegen die Wahrscheinlichkeit des Ereignisses ergibt sich nun als Speziallfall.

Sei $X = I_A$ eine Indikatorvariable für das Ereignis A mit $\Pr[A] = p$.

X ist Bernoulli-verteilt und es gilt $\mathbb{E}[X] = p$.

Nun wird das Experiment n-mal durchgeführt. Dann ist

$$Z = \frac{1}{n}(X_1 + \ldots + X_n)$$

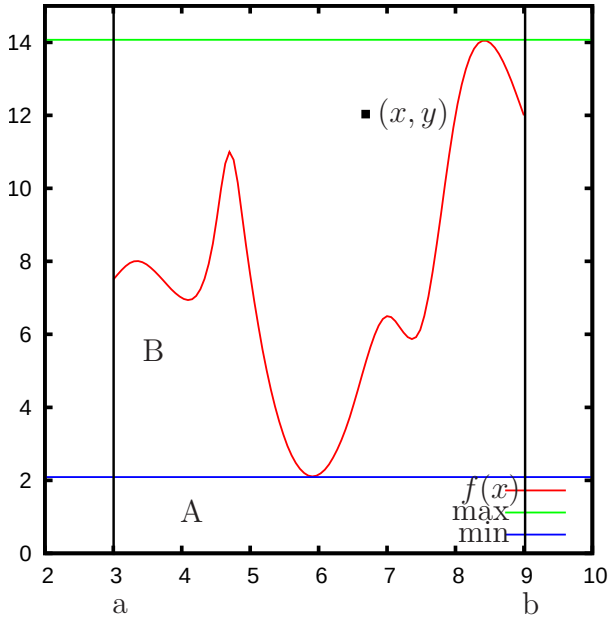
die relative Häufigkeit mit der A bei n-maliger Wiederholung eintritt.

Mit dem Satz gilt dann

$$\Pr[|Z - p| \geq \delta] \leq \varepsilon \qquad \text{für } n \geq \frac{p(1-p)}{\varepsilon \cdot \delta^2}.$$

Dies kann man auch algorithmisch ausnutzen.

Beispiel 22.5 (Numerische Integration mit der Monte-Carlo Methode)**.** Das Integral $\int_a^b f(x)\,dx$ soll numerische berechnet werden.



Setze $F = (b - a) \cdot (\max - \min)$.

Sei $(x, y) \in [a, b] \times [\min, \max]$ zufällig gewählt so ist intuitiv klar (kontinuierlicher Wahrscheinlichkeitsraum), dass

$$\Pr[„(x, y) \in B“] = \frac{B}{F}.$$

Dabei ist B die Fläche unter dem Graph von f im Quadrat F.

Also gilt $B = F \cdot \Pr[„(x, y) \in B“]$ und damit

$$\int_a^b f(x)\,dx = A + B = (b-a) \cdot \min + F \cdot \Pr[„(x, y) \in B“].$$

$\Pr[„(x, y) \in B“]$ bestimmen wir über die relative Häufigkeit, d. h. wir würfeln n Punkte (x_i, y_i) und bestimmen wie oft $(x_i, y_i) \in B$ (d. h. $f(x_i) \leq y_i$).

Für $n \to \infty$ konvergiert der so bestimmte Wert gegen das Integral.

Praktisch verwendet wird diese Methode für hochdimensionale Integrale. □

22.3 Zusammenfassung

- Mit der Ungleichung von Chebyshev kann man die Wahrscheinlichkeit der Abweichung vom Erwartungswert abschätzen.
- Als Anwendung hiervon kann man zeigen, dass sich die relative Häufigkeit bei vielen Versuchen der Wahrscheinlichkeit des Ereignisses immer besser nähert.

23 Kontinuierliche Wahrscheinlichkeitsräume

23.1 Einführung in kontinuierliche Wahrscheinlichkeitsräume

Viele Anwendungen führen auf Ergebnismengen oder ZV mit überabzählbaren Wertebereichen insbesondere physikalische Messgrößen wie Druck, Temperatur, Position, Geschwindigkeit, Zeit,...

Eine Diskretisierung dieser Wertebereiche ist möglich aber unnatürlich.

Beispiel 23.1. In eine Warteschlange werden Druckaufträge eingereiht. Betrachte das Ereignis A=„Innerhalb der Zeit τ kommt ein Auftrag an“.

Es erscheint sinnvoll $\tau \in \mathbb{R}$ zu wählen, aber was ist $\Pr[A]$?

Wie bei dem Übergang von Binomial- zu Poissonverteilung könnte man „diskretisieren“.

Wir teilen die Wartezeit τ in n gleiche Intervalle $\Delta t_n = \tau/n$. Am Ende jedes Intervalles testen wir ob ein Auftrag angekommen ist. Für jedes $n \in \mathbb{N}$ sei X_n geometrisch verteilt (mit einem noch zu bestimmenden p_n).

Wir nehmen nun an, dass

$$\Pr[A] = \lim_{n\to\infty} \Pr[X_n \leq n].$$

Die mittlere Zeit zwischen zwei Aufträgen sei $T \in \mathbb{R}$.

Da $\mathbb{E}[X_n] = 1/p_n$ die mittlere Zahl von Versuchen bis zum ersten Erfolg ist muss gelten

$$T = \mathbb{E}[X_n] \cdot \Delta t_n = \frac{1}{p_n} \cdot \frac{\tau}{n} \qquad \Rightarrow \qquad p_n = \frac{\tau}{nT}.$$

Mit der geometrischen Verteilung gilt

$$\Pr[A] = \lim_{n\to\infty} \Pr[X_n \leq n] = \lim_{n\to\infty} 1 - (1-p_n)^n = 1 - \left(1 - \frac{\tau}{nT}\right)^n = 1 - e^{-\tau/T}.$$

Dies ist die *Exponentialverteilung*, unser erstes Beispiel für ein kontinuierliches Wahrscheinlichkeitsmaß. □

Wir haben gesehen: Eine ZV X definiert mittels (X, f_x) einen Wahrscheinlichkeitsraum.

Im folgenden beginnen wir gleich mit Zufallsvariablen und nehmen implizit an, dass für die zugrundeliegende Ergebnismenge $\Omega = \mathbb{R}$ gilt.

Dies motiviert folgende Definition.

Definition 23.2 (Kontinuierliche ZV)**.** Eine *kontinuierliche* (oder auch stetige) ZV $X : \mathbb{R} \to \mathbb{R}$ (d. h. $\Omega = W_X = \mathbb{R}$) ist definiert durch eine integrierbare (Wahrscheinlichkeits-)Dichtefunktion $f_X : \mathbb{R} \to \mathbb{R}_0^+$ mit der Eigenschaft

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

Eine Menge $A \subseteq \mathbb{R}$, die durch Vereinigung $A = \bigcup_k I_k$ *abzählbar vieler paarweise disjunkter Intervalle* beliebiger Art (offen, geschlossen, halboffen, einseitig unendlich) gebildet werden kann heißt Ereignis.

Das Ereignis A tritt ein, falls X einen Wert aus A annimmt. Für die Wahrscheinlichkeit gilt:

$$\Pr[A] = \int_A f_X(x)\,dx = \sum_k \int_{I_k} f_X(x)\,dx.$$

□

Bemerkung 23.3. Zwei Anmerkungen hierzu:

1. Diese Definition mutet zunächst etwas umständlich an. Warum lässt man nicht beliebige $A \subseteq \mathbb{R}$ zu?

   Das Problem ist, dass das Integral nicht für beliebige solche Teilmengen erklärt ist. Für obige Wahl lässt sich jedoch ein sinnvoller Integralbegriff erklären.

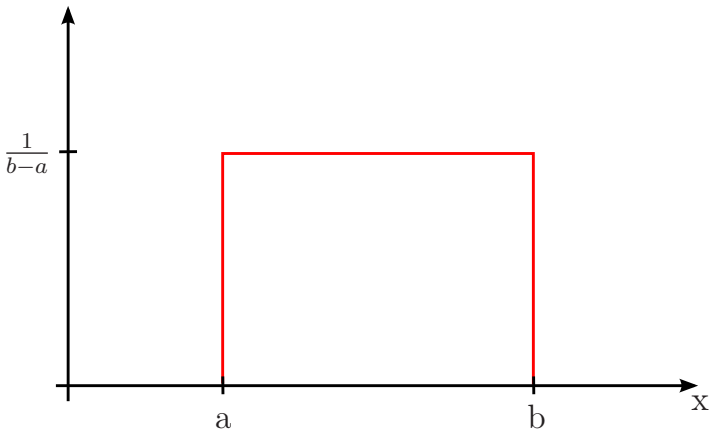
2. Die Ereignisse sind hier als Teilmenge von W_X definiert! Im diskreten Fall waren Ereignisse Teilmengen von Ω. Dies liegt daran, dass die Ergebnismenge Ω hier immer $\mathbb{R}$ ist und somit eine Trennung von Ω und W_X nicht notwendig ist. □

Wir lernen nun eine weitere kontinuierliche Dichtefunktion kennen.

Beispiel 23.4 (Gleichverteilung)**.** Diese ist für zwei Parameter $a, b \in \mathbb{R}$, $a < b$, gegeben durch die Dichtefunktion

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a,b] \\ 0 & \text{sonst} \end{cases}$$

Hier ein Bild:



□

Auch im kontinuierlichen Fall definieren wir eine Verteilungsfunktion:

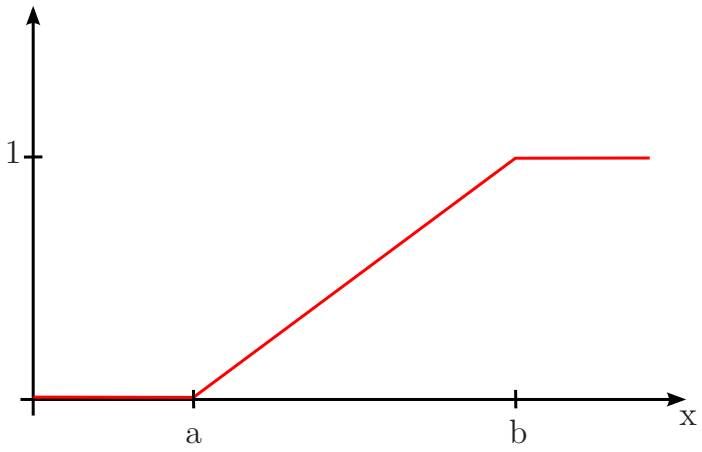
$$F_X(x) := \Pr[X \leq x] = \Pr[X \in A = (-\infty, x)] = \int_{-\infty}^{x} f_X(t)\,dt.$$

(Achtung: $A \subseteq W_X$!).

Beispiel 23.5 (Gleichverteilung (Forts.))**.** Für die Gleichverteilung gilt

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

Im Bild:



□

Die Verteilungsfunktion hat folgende Eigenschaften:

a) F_X ist stetig da Integral über f_X. Man spricht deshalb auch von stetiger Zufallsvariable.

b) F_X ist monoton steigend (da f_X nicht negativ).

c) $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$.

Für das Ereignis $a < X \leq b$ erhalten wir einfach die Wahrscheinlichkeit:

Aus

$$\Pr[X \leq b] = \Pr[„X \leq a“ \cup „a < X \leq b“] = \Pr[X \leq a] + \Pr[a < X \leq b]$$

und damit

$$\Pr[a < X \leq b] = \Pr[X \leq b] - \Pr[X \leq a] = F_X(b) - F_X(a).$$

Wegen

$$\int_{(a,b)} \ldots = \int_{[a,b)} \ldots = \int_{(a,b]} \ldots = \int_{[a,b]} \ldots$$

ist es egal ob man $<$ oder $\leq$ schreibt.

Bemerkung 23.6. Die formal korrekte Einführung kontinuierlicher Wahrscheinlichkeitsräume, insbesondere der Ereignismenge $\mathcal{A}$, der Dichte f_X und des zugehörigen Intervallbegriffs erfordert einige Analysis. Wir verweisen auf [SS02, Abschnitt 2.1.3].

Macht man alles richtig, so gilt Lemma 18.1 auf Seite 215 entsprechend, d. h. $\Pr[\bar{A}] = 1 - \Pr[A]$, $A \subseteq B \Rightarrow \Pr[A] \leq \Pr[B]$, Additionssatz für abzählbare Mengen von Ereignissen, usw. □

23.2 Rechnen mit kontinuierlichen ZV

Wie im diskreten Fall erhält man durch Anwenden einer Funktion auf eine ZV eine neue ZV:

$$Y := g(X) \quad \text{oder} \quad Y = g \circ X.$$

Für die Verteilung von Y erhält man dann

$$\begin{aligned} F_Y(y) = \Pr[Y \leq y] = \Pr[g(X) \leq y] = \Pr[X \in \underbrace{\{t \in \mathbb{R} | g(t) \leq y\}}_{:=C}] \\ = \int_C f_X(t)\, dt. \end{aligned}$$

C muss dabei ein zulässiges Ereignis sein.

Beispiel 23.7. Sei X gleichverteilt auf $]0,1[$. Für $\lambda > 0$ betrachten wir die ZV $Y := -(1/\lambda) \ln X$. Bestimme Verteilung und Dichte von Y.

Für die Menge C von oben gilt in diesem Fall:

$$C = \{t \in \mathbb{R} \mid -(1/\lambda) \ln t \leq y\} = \{t \in \mathbb{R} \mid \ln t \geq -\lambda y\} \qquad = \{t \in \mathbb{R} \mid t \geq e^{-\lambda y}\}.$$

Damit erhalten wir

$$\begin{aligned} F_Y(y) &= \int_{e^{-\lambda y}}^{\infty} f_X(t)\, dt = 1 - \int_{-\infty}^{e^{-\lambda y}} f_X(t)\, dt = 1 - F_X(e^{-\lambda y}) \\ &= \begin{cases} 1 - e^{-\lambda y} & y \geq 0 \\ 0 & \text{sonst} \end{cases}. \end{aligned}$$

Die Dichte erhalten wir durch differenzieren:

$$f_Y(y) = \begin{array}{ll} \lambda e^{-\lambda y} & y \geq 0 \\ 0 & \text{sonst} \end{array}$$

□

23.3 Simulation von ZV

Im letzten Beispiel haben wir mittels Transformation aus einer gleichverteilten ZV eine exponentiell verteilte ZV gemacht (siehe erstes Beispiel in diesem Kapitel).

Das kann man zu einer Methode verallgemeinern um (nahezu) beliebig verteilte ZV aus der Gleichverteilung zu erzeugen.

Sei also X gleichverteilt auf $]0,1[$ und F_Y die gewünschte stetige und *streng* monoton steigende Verteilung.

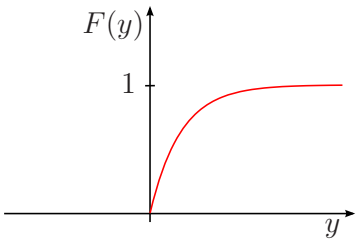
Idee ist wie im obigen Beispiel: Auf X eine Funktion g anwenden, so dass $Y = g(X)$ die Verteilung F_Y hat.

Mit der Wahl $g(t) = F_Y^{-1}(t)$ (die Umkehrfunktion existiert wg. der Voraussetzung streng monoton) gilt dann

$$\begin{aligned}\Pr[g(X) \le y] &= \Pr[F_Y^{-1}(X) \le y] = \Pr[X \le F_Y(y)] \\ &= F_X(\underbrace{F_Y(y)}_{\in[0,1]}) = F_Y(y) \quad \text{wie gewünscht.}\end{aligned}$$

Dies können wir natürlich formal auch auf Beispiel 23.7 auf der vorherigen Seite anwenden:

gewünscht $\quad F_Y(y) = \begin{cases} 1 - e^{-\lambda y} & \boxed{y \ge 0} \\ 0 & \text{sonst} \end{cases}$



Umkehrfunktion lautet

$$\begin{aligned}F_Y(y) = 1 - e^{-\lambda y} = x \Leftrightarrow 1 - x = e^{-\lambda y} \Leftrightarrow -\ln(1-x) = \lambda y \\ \Leftrightarrow F_Y^{-1}(x) = -1/\lambda \ln(1-x) =: g(x).\end{aligned}$$

Oben hatten wir $g(x) = -1/\lambda \ln x$ verwendet, aber mit x ist auch $1 - x$ gleichverteilt.

Praktisch heißt das: Erzeuge gleichverteilte Zufallszahlen x, dann sind die Zahlen $g(x) = -1/\lambda \ln(1-x)$ exponentialverteilt.

23.4 Erwartungswert und Varianz

Die Begriffe können völlig analog zum diskreten Fall eingeführt werden. Ersetze einfach Summen durch Integrale.

Definition 23.8. Der Erwartungswert einer kontinuierlichen ZV is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t) dt$$

dabei setzen wir voraus, dass $\int_{-\infty}^{\infty} |t| f_X(t)\, dt$ existiert.

Entsprechend für die Varianz:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 f_X(t)\, dt$$

wenn das Integral existiert. □

Die Sätze über Rechenregeln für Erwartungswert und Varianz lassen sich übertragen, insbesondere

$$Y = g(X) \quad \Rightarrow \quad \mathbb{E}[Y] = \int_{-\infty}^{\infty} g(t) F_X(t) dt$$
$$Y = aX + b \quad \Rightarrow \quad \mathbb{E}[Y] = a \cdot \mathbb{E}[X] + b$$

oder

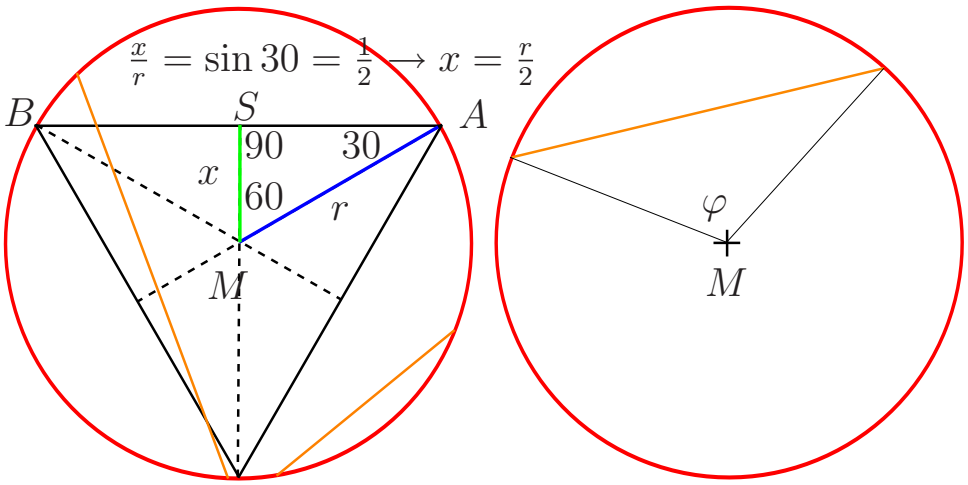
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Ebenso lassen sich auch wieder mehrere ZV gleichzeitig betrachten, etwa $Y = X_1 + X_2 + \ldots + X_n$.

23.5 Bertrand'sches Paradoxon

Im diskreten Fall gab es das Prinzip von Laplace um den Elementarereignissen Wahrscheinlichkeiten zuzuordnen. Im kontinuierlichen Fall ist das nicht so einfach.

Beispiel 23.9 (Bertrand'sches Paradoxon)**.** Betrachte einen Kreis mit eingeschriebenem gleichseitigen Dreieck sowie das Ereignis $A =$ „Die Länge einer beliebig gewählten Sehne des Kreises übersteigt die Seitenlänge des Dreiecks".



Wie modelliert man dieses Problem?

Wähle einen Parameterraum Ω. Die ZV $X(\omega)$ messe dann die Länge der Sehne. Für Ω hat man verschiedene Möglichkeiten:

- Wähle den Abstand d des Sehnenmittelpunktes S vom Kreismittelpunkt M. $d \in [0, r]$ sei gleichverteilt.

  Sei s eine Sehne und S ihr Mittelpunkt. Die Strecke $\overline{SM}$ steht senkrecht auf s und d ist die Länge dieser Strecke.

  A tritt genau dann ein, wenn $d < r/2$ wie man im linken Bild sieht. Dann ist also $\Pr[A] = 1/2$.

- Wähle den Sehnenmittelpunkt S gleichverteilt innerhalb der *Kreisfläche*. A tritt ein, wenn S innerhalb eines Kreises mit Radius $r/2$.

  Dieser hat die Fläche $(\frac{r}{2})^2\pi = \frac{1}{4}r^2\pi$, also $\Pr[A] = \frac{1/4 r^2\pi}{r^2\pi} = \frac{1}{4}$.

- Wähle Betrag des Winkels ϕ am Punkt M des Dreiecks BMA.

  Jedes $\phi \in [0, 180]$ beschreibt bis auf Rotation eine Kreissehne. ϕ sei ausserdem gleichverteilt.

  A tritt ein, wenn $\varphi > 120$ (das ist der Winkel BMA im linken Bild). Also $\Pr[A] = \frac{60}{180} = \frac{1}{3}$.

Welche Wahrscheinlichkeit ist nun richtig?

Das kann man so nicht sagen, es kommt eben genau darauf an *wie* modelliert wird. Man muss also das Zufallsexperiment in der Aufgabe genauer festlegen! (Frage: Wann hat man die Aufgabenstellung genau genug festgelegt?)

23.6 Gleichverteilung

Wir behandeln nun einige wichtige Verteilungen. Der Vollständigkeit wegen beginnen wir mit der

Gleichverteilung auf dem Intervall $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{sonst} \end{cases} \qquad F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathrm{Var}[X] = \frac{(a-b)^2}{12}.$$

23.7 Normalverteilung; Zentraler Grenzwertsatz

Eine herausragende Stellung nimmt die sogenannte *Normalverteilung* ein:

Definition 23.10 (Normalverteilung)**.** Eine ZV X mit Wertebereich $W_X = \mathbb{R}$ heißt normalverteilt mit den Parametern $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$ wenn sie die Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) =: \varphi(x; \mu, \sigma)$$

hat.

Als Abkürzung schreibt man $X \sim \mathcal{N}(\mu, \sigma^2)$. $\mathcal{N}(0,1)$ heißt Standardnormalverteilung und man setzt $\varphi(x) = \varphi(x; 0, 1)$.

Die Verteilungsfunktion ist dann

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right) dt =: \Phi(x; \mu, \sigma).$$

Diese ist nicht geschlossen angebbar und wird daher numerisch berechnet und tabelliert. Man kürzt ab: $\Phi(x) = \Phi(x; 1, 0)$. □

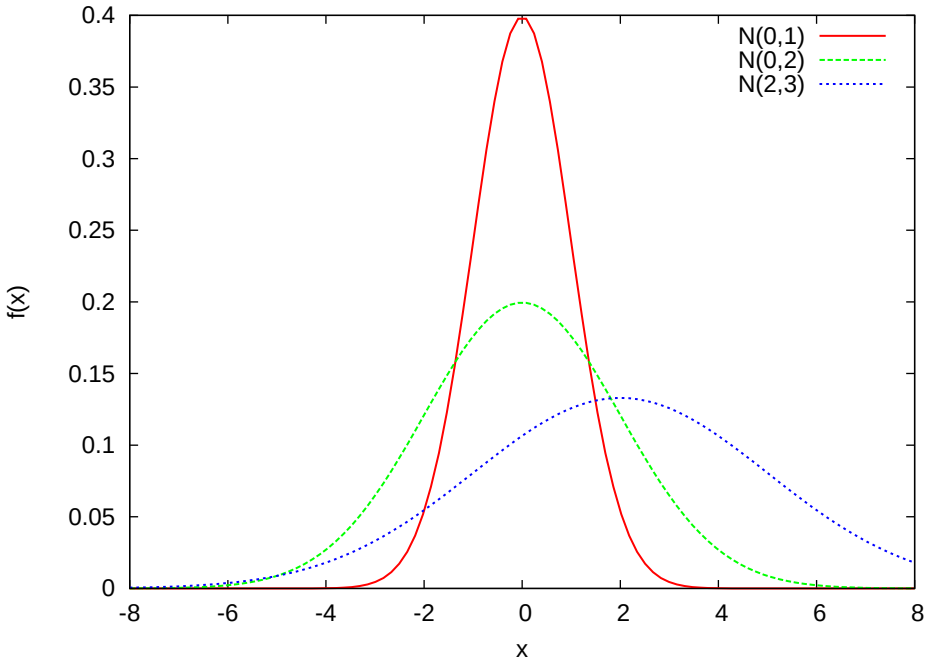


Abbildung 38: Die Dichte der Normalverteilung für verschiedene Werte von μ und σ.

Abbildung 38 zeigt die Dichte der Normalverteilung für verschiedene Werte von μ und σ.

Wir geben noch zwei wichtige Eigenschaften ohne Beweis an.

Satz 23.11. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ dann gilt für $a \in \mathbb{R} \setminus \{0\}$, dass $Y = aX + b$ (lineare Transformation) wieder normalverteilt ist mit $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Beweis: [SS02, Satz 2.20, S. 107]. □

Anwendung dieses Satzes: Aus $X \sim \mathcal{N}(0, 1)$ lässt sich durch lineare Transformation jede Normalverteilung ZV erzeugen.

Benötigt man μ, σ^2, dann setzte $a = \sigma$ und $b = \mu$.

Umgedreht kann man eine mit $X \sim \mathcal{N}(\mu, \sigma)$ normalverteilte ZV durch die lineare Transformation $Y = \frac{X-\mu}{\sigma}$ in eine standardnormalverteilte ZV verwandeln.

Satz 23.12. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, dann gilt

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Beweis: [SS02, Satz 2.22, S. 109]. □

Beispiel 23.13. Die Normalverteilung tritt in Anwendungen oft auf.

Bei der mehrfachen Messung physikalischer Größen kann man oft die Messwerte, bzw. den Fehler in den Messwerten, als normalverteilt annehmen (z.B. Messung der Bahnparameter eines Asteroiden).

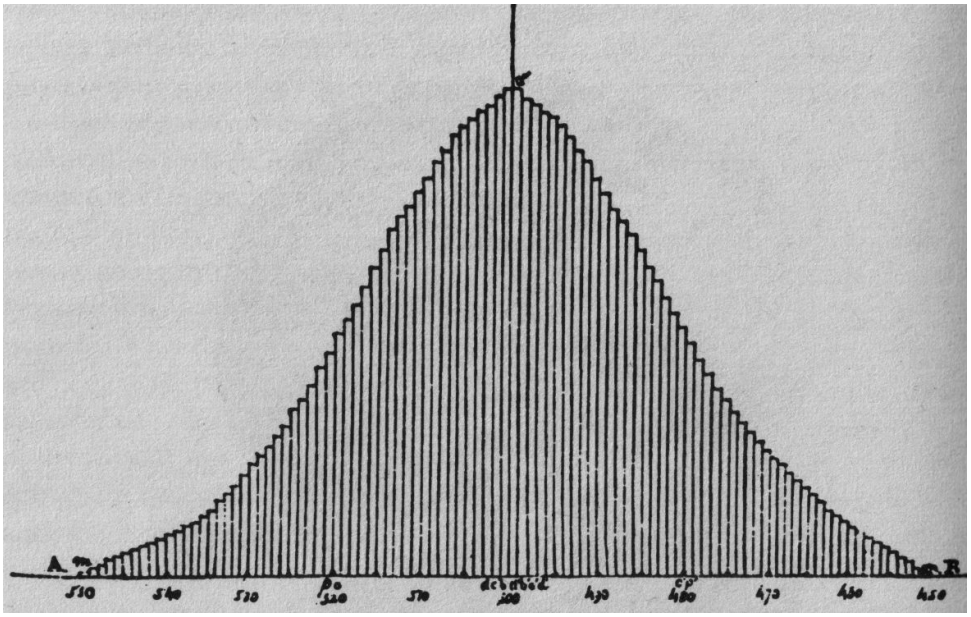


Abbildung 39: Vermessung der Körpergröße von Menschen als ein praktisches Beispiel für die Annäherung der Normalverteilung. Aus [Ste97].

Auch beim Menschen kann man die Normalverteilung beobachten: Messe die Körpergröße von n Menschen. Betrachte die Anzahl der Menschen n_i, deren Körpergröße in cm im Intervall $[i - 1/2, i + 1/2)$ liegt. Man erhält annähernd eine Normalverteilung.

Die Zeichnung erinnert stark an die Binomialverteilung. Offensichtlich haben wir es hier wieder mit einem Übergang diskret nach kontinuierlich zu tun. □

Abbildung 39 zeigt die praktische Messung der Körpergröße von Quetelet[47].

Betrachtet man mehrere normalverteilte ZV so gilt:

Satz 23.14 (Additivität der Normalverteilung)**.** Die ZV $X_1, \ldots, X_n$ seien unabhängig und normalverteilt mit den Parametern $\mu_i, \sigma_i (1 \leq i \leq n)$. Dann ist die zusammengesetzte ZV

$$Z = a_1 X_1 + \ldots + a_n X_n$$

normalverteilt mit Erwartungswert

$$\mu = a_1 \mu_1 + \ldots + a_n \mu_n$$

und Varianz

$$\sigma^2 = a_1^2 \sigma_1^2 + \ldots + a_n^2 \sigma_n^2.$$

Beweis: [SS02, Satz 2.38, S. 120]. □

[47] Lambert Adolphe Jacques Quételet, 1796-1874, belg. Astronom und Statistiker.

Bemerkenswert ist hier, dass die einzelnen normalverteilten ZV unterschiedliche Parameter haben dürfen!

Beispiel 23.15. Was bedeutet der letzte Satz?

Angenommen wir haben zwei Populationen: Zwerge und Riesen. Das Merkmal Körpergröße sei in beiden Populationen normalverteilt aber mit unterschiedlichem Erwartungswert und Varianz.

Wählt man nun aus jeder Population zufällig ein Individuum aus, so ist die *Summe ihrer Körpergrößen* wieder normalverteilt, wobei sich Erwartungswert und Varianz exakt angeben lassen. □

Der nun folgende Satz ist von zentraler Bedeutung in der Statistik.

Satz 23.16 (Zentraler Grenzwertsatz)**.** Die ZV $X_1, \ldots, X_n$ besitzen jeweils *dieselbe* Verteilung und seien *unabhängig*. Erwartungswert und Varianz der X_i existieren und seien μ bzw $\sigma^2 \neq 0$.

Dann ist die ZV

$$Z_n = \frac{(X_1 + \ldots + X_n) - n\mu}{\sigma\sqrt{n}}$$

asymptotisch (d. h. für $n \to \infty$) standardnormalverteilt.

Sei F_n die Verteilungsfunktion von Z_n, dann gilt

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

.

Beweis: [SS02, Satz 2.40, S. 123]. Auch dort nur skizziert. □

Dieser Satz belegt die enorme Bedeutung der Normalverteilung, er sagt, dass die Summe *entsprechend vieler unabhängiger aber identisch verteilter* ZV annähernd normalverteilt ist.

Angewandt auf Bernoulli-verteilte ZV ergibt sich:

Satz 23.17 (Spezialfall von DeMoivre[48])**.** Die ZV $X_1, \ldots, X_n$ seien unabhängig und *Bernoulli-verteilt* mit gleicher Erfolgswahrscheinlichkeit p. Bezeichne ihre Summe mit $H_n = X_1 + \ldots + X_n$.

Dann gilt, dass

$$H_n^* = \frac{H_n - np}{\sqrt{np(1-p)}}$$

für $n \to \infty$ standardnormalverteilt ist.

Beweis: Setze $\mu = \mathbb{E}[X_i] = p$ und $\sigma^2 = p(1-p)$ in den Zentralen Grenzwertsatz ein. □

H_n basiert auf einer Bernoullikette der Länge n, ist also Binomial-verteilt. Der Satz sagt, dass wenn man bei festem p und $n \to \infty$ die Binomialverteilung entsprechend verschiebt $(-n \cdot p)$ und skaliert, dann konvergiert die Binomialverteilung als Treppenfunktion gegen die Normalverteilung.

Dies kann zur schnellen Berechnung der Binomialverteilung genutzt werden. Dies ist wichtig, da für große n die Auswertung der Binomialverteilung sehr unpraktisch wird (etwa $b(x; 10^6, p) = \binom{10^6}{x} p^x (1-p)^{10^6 - x}$).

[48] Abraham de Moivre, 1674-1754, frz. Mathematiker.

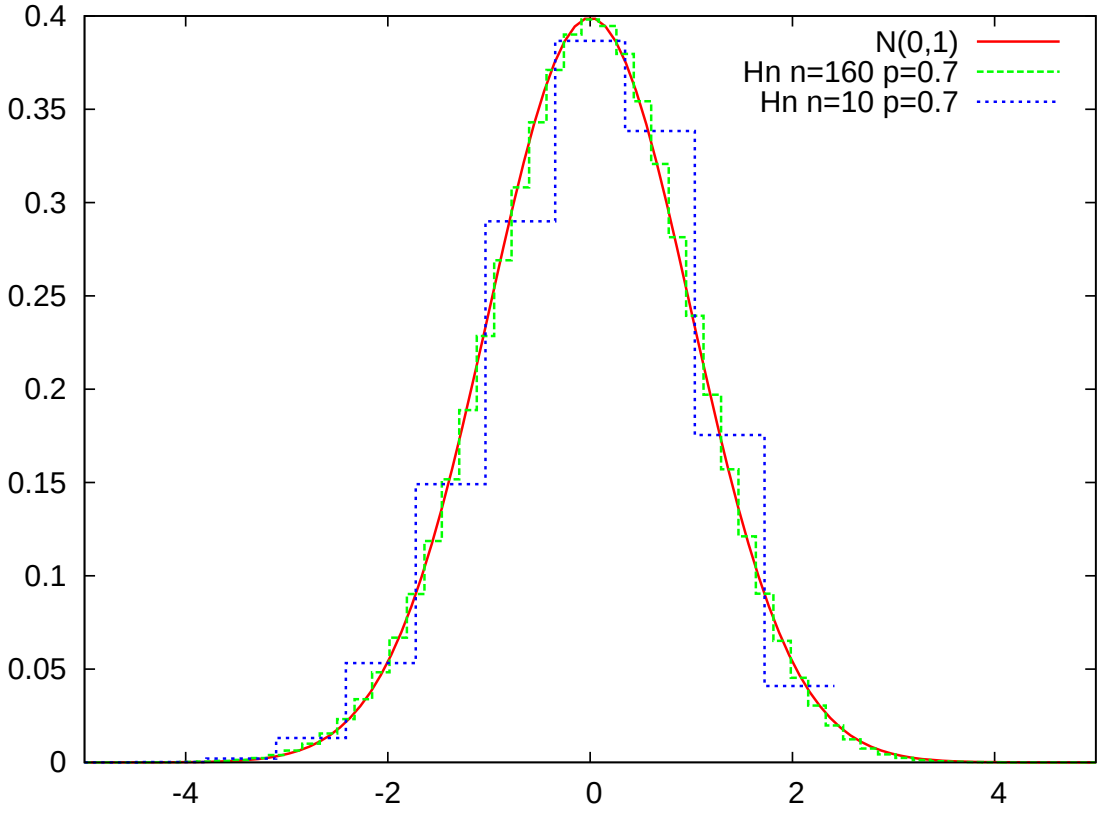


Abbildung 40: Konvergenz der entsprechend skalierten Binomialverteilung gegen die Normalverteilung.

Abbildung 40 zeigt Konvergenz der entsprechend skalierten Binomialverteilung gegen die Normalverteilung wie sie der Satz von DeMoivre behauptet.

23.8 Exponentialverteilung

Die Exponentialverteilung haben wir schon als Grenzwert der geometrischen Verteilung kennengelernt.

Definition 23.18 (Exponentialverteilung)**.** Eine ZV mit der Dichte

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

heißt exponentialverteilt. Die Verteilungsfunktion dazu lautet

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{falls } x \geq 0 \\ 0 & x < 0 \end{cases}$$

Es gilt

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \text{und} \qquad \mathrm{Var}[X] = \frac{1}{\lambda^2}.$$

□

Die Exponentialverteilung hat ihre Anwendung bei Warteprozessen:

Beispiel 23.19. Die ZV X bezeichne die Zeit bis ein *bestimmtes* Atom eines radioaktiven Elements zerfällt. X sei exponentialverteilt mit Parameter λ.

$\Pr[X \leq x]$ ist dann die Wahrscheinlichkeit, dass das Atom innerhalb der nächsten x Zeiteinheiten zerfällt.

Da x die Einheit einer Zeit hat, hat λ die Einheit Zeit^{-1} (Rate).

Der Erwartungswert $\mathbb{E}[X] = \frac{1}{\lambda}$ ist die erwartete Lebensdauer des Atoms, d. h. die mittlere Lebensdauer bei vielen Versuchen. □

Wie die geometrische Verteilung ist die Exponentialverteilung gedächtnislos.

Satz 23.20 (Gedächtnislosigkeit)**.** Für die Exponentialverteilung gilt $\Pr[X > y + x | X > x] = \Pr[X > y]$.

Beweis: Wir rechnen nach:

$$\begin{aligned} \Pr[X > y + x | X > x] &= \frac{\Pr[X > y + x, X > x]}{\Pr[X > x]} = \frac{e^{-\lambda(y+x)}}{e^{-\lambda x}} = e^{-\lambda y} \\ &= \Pr[X > y]. \end{aligned}$$

Hier hat man benutzt, dass $\Pr[X > y + x, X > x] = \Pr[X > y + x]$ und $\Pr[X > x] = 1 - \Pr[X \leq x]$. □

Man kann auch die Umkehrung zeigen: Ist eine stetige Verteilung gedächtnislos, so ist es die Exponentialverteilung.

Die Exponentialverteilung hat ihre Anwendung bei Warteproblemen.

Oben haben wir ein Atom und seine Lebensdauer (d. h.. die Wartezeit bis zum Zerfall) betrachtet. In der Regel betrachten wir aber mehrere Atome, d.h. wir warten auf mehrere Ereignisse gleichzeitig.

Hierüber machen die folgenden Sätze Aussagen.

Satz 23.21. Die ZV $X_1, \ldots, X_n$ seien unabhängig und exponentialverteilt mit den Parametern $\lambda_1, \ldots, \lambda_n$. Dann ist auch $X = \min\{X_1, \ldots, X_n\}$ exponentialverteilt mit dem Parameter $\lambda_1 + \ldots + \lambda_n$.

Beweis: $X > t$ ist das Ereignis, dass *alle* Atome die Zeit t überleben. Dafür erhalten wir

$$\begin{aligned} 1 - F_X(t) &= \Pr[X > t] = \Pr[\min(X_1, \ldots, X_n) > t] \\ &= \Pr[X_1 > t, \ldots, X_n > t] = \Pr[X_1 > t] \cdot \ldots \cdot \Pr[X_n > t] \\ &= e^{-\lambda_1 t} \cdot \ldots \cdot e^{-\lambda_n t} \\ &= e^{-(\lambda_1 + \ldots + \lambda_n)t}, \end{aligned}$$

also

$$F_X(t) = 1 - e^{-(\lambda_1 + \ldots + \lambda_n)t}.$$

□

Für den Spezialfall, dass alle ZV exponentialverteilt mit dem gleichen λ sind erhalten wir:

Korollar 23.22. Die ZV $X_1, \ldots X_n$ seien unabhängig und exponentialverteilt mit Parameter λ. Dann ist $X = \min(X_1, \ldots X_n)$ exponentiell verteilt mit Parameter $n\lambda$. □

Für unsere Atome bedeutet dies

Beispiel 23.23. Die Wahrscheinlichkeit, dass ein Atom die Zeit t überlebt ist $\Pr[X > t] = 1 - F_X(t) = e^{-\lambda t}$.

Die Wahrscheinlichkeit, dass n Atome die Zeit t/n überleben beträgt

$$\begin{aligned} \Pr[X_1 > t/n, \ldots, X_n > t/n] &= \Pr[\min(X_1, \ldots, X_n) > t/n] = \Pr[X > t/n] \\ &= 1 - F_X(t) = e^{-n\lambda t/n} = e^{-\lambda t}, \end{aligned}$$

ist also gleich der Wahrscheinlichkeit, dass ein Atom die Zeit t überlebt. □

Machen wir noch ein praktischeres Beispiel.

Beispiel 23.24. An einem Bahnhof stehen drei Telefonzellen. Es wird überlegt ob und wie gut durch Aufstellen zusätzlicher Telefonzellen die Wartezeit verkürzt werden kann.

Die Dauer eines Telefongesprächs sei exponentialverteilt mit $\lambda = 1/10$, d. h. ein Telefongespräch dauert im Mittel 10 Einheiten (etwa Minuten).

Damit ist $\min(X_1, X_2, X_3)$ ebenfalls exponentialverteilt mit Parameter $3 \cdot 1/10 = 3/10$. Die mittlere Wartezeit bis eine Telefonzelle frei wird beträgt $10/3 = 3.33$ Minuten.

Durch Aufstellen von zwei weiteren Telefonzellen ($n = 5$) würde man die Wartezeit auf $^{10}/_{5} = 2$ Minuten reduzieren.

23.9 Zusammenfassung

- Bei kontinuierlichen Wahrscheinlichkeitsräumen ist die Ergebnismenge überabzählbar. Ereignisse werden durch Vereinigung von Intervallen definiert.
- Die Wahrscheinlichkeit eines Ereignisses ist das Integral über die Dichtefunktion. Die meisten Sätze übertragen sich von den diskreten Wahrscheinlichkeitsräumen indem man Summen durch Integrale ersetzt.
- Wichtige stetige Verteilungen sind die Gleichverteilung (enstpricht dem Laplace-Prinzip aber Vorsicht!), die Exponentialverteilung (entspricht der geometrischen Verteilung) und die Normalverteilung (entspricht der Binomialverteilung).
- Die Normalverteilung hat eine enorme Bedeutung, da die Summe beliebiger, unabhängiger und identisch verteilter Zufallsgrößen immer eine Normalverteilung ergibt (Zentraler Grenzwertsatz).

Lehrbücher Numerik

[DH02] Deuflhard, P. und A. Hohmann: *Numerische Mathematik I, Eine algorithmisch orientierte Einführung*. de Gruyter, 2002.

[GO96] Golub, G. und J. M. Ortega: *Scientific Computing*. Teubner, 1996.

[Ran06] Rannacher, R.: *Einführung in die Numerische Mathematik (Numerik 0)*. `http://numerik.iwr.uni-heidelberg.de/~lehre/notes`, 2006.

[SB05] Stoer, J. und R. Bulirsch: *Numerische Mathematik II*. Springer, 5. Auflage, 2005.

[SK05] Schwarz, H.-R. und N. Köckler: *Numerische Mathematik*. Teubner, 5. Auflage, 2005.

[Sto05] Stoer, J.: *Numerische Mathematik I*. Springer, 9. Auflage, 2005.

Lehrbücher Stochastik

[Hüb03] Hübner, G.: *Stochastik*. Vieweg, 4. Auflage, 2003.

[SS02] Schickinger, T. und A. Steger: *Diskrete Strukturen 2*. Springer, 2002.

[TT07] Teschl, G. und S. Teschl: *Mathematik für Informatiker, Band 2: Analysis und Statistik*. Springer, 2. Auflage, 2007.

Weiterführende Literatur

[Gol91] Goldberg, D.: *What Every Computer Scientist Should Know About Floating Point Arithmetic*. Computing Surveys, 1991. `http://citeseer.ist.psu.edu/goldberg91what.html`.

[Knu98] Knuth, D. E.: *The Art of Computer Programming*, Band 2. Addison-Wesley, 3. Auflage, 1998.

[Ran] Rannacher, Rolf: *Numerische Mathematik 1 (Numerik gewöhnlicher Differentialgleichungen)*. `http://numerik.iwr.uni-heidelberg.de/~lehre/notes`.

[Sim] Simeon, B.: *Numerik gewöhnlicher Differentialgleichungen*. `http://www-m2.ma.tum.de/~simeon/numerik4/skript.html`.

[Ste97] Stewart, I.: *Does God Play Dice*. Penguin, 1997.

[SW70] Schaback, R. und H. Werner: *Praktische Mathematik I/II*. Springer, 1970.