Numerische Mathematik

Patrick Erik Bradley

Skript zur Vorlesung
in den Masterstudiengängen
Geodäsie und Geoinformatik
sowie
Remote Sensing and Geoinformatics

Karlsruhe, 2019

Vorwort

Der vorliegende Text wurde als Begleitlektüre zur Vorlesung „Numerische Mathematik“ für die Masterstudiengänge *Geodäsie und Geoinformatik* und *Remote Sensing and Geoinformatics* konzipiert. Die behandelten Themen sind üblich in einer Numerik-Vorlesung und werden durch Anwendungsbeispiele aus der Welt der Geodäsie und (Geo-)Informatik ergänzt, wobei das Thema Internet auch eine Rolle spielt (PageRank von Google und Kryptographie). Außerdem wird, inspiriert durch das Buch *Homers letzter Satz — Die Simpsons und die Mathematik*[1] von Simon Singh, auch der Bezug zu mathematischen Themen hergestellt, die in den Zeichentrick-Sitcoms *The Simpsons* und *Futurama* eine Rolle spielen. Besonderen Dank möchte ich Vater Nil von der Skite des Heiligen Spiridon in Geilnau aussprechen, der mir bei den geistlichen Beispielen wertvolle Hilfestellung gab.

Hinweise auf Druck- und andere Fehler werden gerne entgegengenommen, was zu einer verbesserten weiteren Auflage führen kann.

Karlsruhe, den 5. Februar 2019 P.E. Bradley

[1] Originaltitel: *The Simpsons and their Mathematical Secrets*

Inhaltsverzeichnis

Kapitel 1

Gleitkomma-Arithmetik

1.1 Gleitkommazahlen

Eine Gleitkommazahl

$$a = m \cdot \beta^e$$

besteht aus einer *Mantisse* m, einer *Basis* β und einem *Exponenten* e. Sie ist *normalisiert*, falls

$$\beta^{-1} \leq m < 1$$

d.h. falls

$$m = 0.\, x_1 x_2 \ldots$$

mit $x_1 \neq 0$ ist.

Bemerkung 1.1.1. *Auch andere Normalisierungen sind möglich, z.B.*

$$2.597\,\mathrm{E} - 03 = 2.597 \cdot 10^{-3}$$

statt $0.2597 \cdot 10^{-2}$.

IEEE-Standard, double precision

Eine *double precision* Zahl im *IEEE-Standard* lässt sich als 64-Bit Wort über dem Alphabet $\{0, 1\}$ darstellen:

$$\underbrace{\sigma}_{\text{Vorzeichen}} \quad \underbrace{a_1 \ldots a_{52}}_{\text{Mantisse}} \quad \underbrace{e_0 \ldots e_{10}}_{\text{Exponent}}$$

Der Wert, der einem solchen Wort zugeordnet wird, ist

$$x = (-1)^\sigma \left(1 + \sum_{i=1}^{51} 2^{-i} a_{52-i}\right) \cdot 2^{e-1023}$$

mit

$$e = \sum_{i=0}^{10} e_i 2^i$$

1.2 Überlauf und Unterlauf

Nicht alle reellen Zahlen können durch ein Wort endlicher Länge dargestellt werden. Da die Mantisse nur endliche Länge hat, treten Rundungsfehler auf. *Überlauf* und *Unterlauf* entstehen auf Grund der endlichen Länge des Exponenten.

Überlauf

Überlauf bedeutet, dass eine arithmetische Operation eine Zahl mit zu großem Exponenten produziert.

Unterlauf

Unterlauf bedeutet, dass eine arithmetische Operation eine Zahl mit zu kleinem Exponenten produziert.

Bemerkung 1.2.1. *Überlauf führt stets zu einer Fehlermeldung, d.h. ist fatal.*
Unterlauf führt zu einer Zahl, die fast Null ist. Dies bedeutet, dass wenn die Zahl gleich Null gesetzt wird, weiter gerechnet werden kann.

Oft kann Überlauf auf Kosten harmloser Unterläufe eliminiert werden.

Beispiel 1.2.2. *Es sei*

$$c = \sqrt{a^2 + b^2}$$

mit $a = 10^{60}$ *und* $b = 1$ *in einem Dezimalsystem mit 2-stelligem Exponenten zu berechnen. Darin führt* a^2 *zu Überlauf. Dieses kann folgendermaßen eliminiert werden:*

$$c = s\sqrt{\left(\frac{a}{s}\right)^2 + \left(\frac{b}{s}\right)^2}, \quad s = \max\{|a|, |b|\} = 10^{60}$$

ergibt

$$c = 10^{60}\sqrt{1^2 + \left(\frac{1}{10^{60}}\right)^2}$$

mit dem Unterlauf

$$\left(\frac{1}{10^{60}}\right)^2$$

Setzen wir dieses gleich Null, ergibt sich $c = 10^{60}$.

1.3 Rundungsfehler

Nicht jede reelle Zahl kann im Rechner exakt dargestellt werden. Beispielsweise

$$\sqrt{7} = 2.6457513\ldots$$

Auf einem fünfstelligen Dezimalrechner müssen die hinteren Stellen wegfallen. Es gibt zwei Möglichkeiten:

1. Rundung:
$$\sqrt{7} \approx 2.6458$$
2. Abschneiden:
$$\sqrt{7} \approx 2.6457$$

Fehlerschranken

Runde auf 5 Stellen. Dann wird aus

$$a = X.XXXXY$$

die Rechner-Zahl

$$b = X.XXXZ$$

Runde dabei auf, falls $Y \geq 5$ und runde ab, falls $Y < 5$ ist. Für den Fehler gilt:

$$|b - a| \leq 5 \cdot 10^{-5}$$

Ist die führende Stelle $\neq 0$, d.h. $|a| \geq 1$, so ist

$$\frac{|b-a|}{|a|} \leq 5 \cdot 10^{-5} = \frac{1}{2} \cdot 10^{-4}$$

Allgemein gilt:

1. Beim Runden auf t Dezimalstellen ist der relative Fehler
$$\frac{|b-a|}{|a|} \leq \frac{1}{2} \cdot 10^{-t+1}$$
2. Beim Abschneiden auf t Dezimalstellen gilt:
$$\frac{|b-a|}{|a|} \leq 10^{-t+1}$$
Für Binärzahlen gilt:
$$\frac{|b-a|}{|a|} \leq \begin{cases} 2^{-t} & \text{(Runden)} \\ 2^{-t+1} & \text{(Abschneiden)} \end{cases}$$

1.4 Maschinen-Epsilon

Sei $b = \mathrm{fl}(a)$ die Maschinendarstellung einer reellen Zahl $a \in \mathbb{R}$. ϵ_M sei die kleinste obere Schranke für den relativen Fehler:

$$\epsilon = \frac{b-a}{a} \quad \text{und} \quad |\epsilon| \leq \epsilon_M$$

Mit anderen Worten:

$$\mathrm{fl}(a) = a(1+\epsilon), \quad |\epsilon| \leq \epsilon_M$$

Die Zahl ϵ_M heißt *Maschinen-Epsilon* und ist eine Charakteristik der Gleitkomma-Arithmetik auf einer gegebenen Maschine.

Bemerkung 1.4.1. *ϵ_M ist etwas größer als die größte Zahl x, für die gilt:*

$$\mathrm{fl}(1+x) = 1$$

Beispiel 1.4.2. *Bei einer 6-stelligen Binärarithmetik liefert $x = 2^{-7}$:*

$$\mathrm{fl}(1+x) = 1$$

Konsequenz. *Eine Approximation an ϵ_M ist durch folgenden Algorithmus gegeben:*
Start. $x_0 = 1$
Schritt n. *Falls* $\mathrm{fl}(1 + x_{n-1}) \neq 1$, *setze* $x_n := \frac{x_{n-1}}{2}$.

Oder in Pseudocode:

```
x=1;
while (1 + x != 1)
    x = x/2;
```

1.5 Arithmetik

Das Ergebnis einer arithmetischen Operation kann auf einem Rechner nur approximativ dargestellt werden: Z.B. hat das Produkt zweier n-stelliger Zahlen bis zu $2n$ Stellen.

Idealerweise ist das Ergebnis einer Gleitkomma-Rechenoperation die korrekte Rundung des exakten Ergebnisses. D.h. für die Operation $\square$ soll gelten:

$$\mathrm{fl}(a \square b) = (a \square b)(1+\epsilon), \quad |\epsilon| \leq \epsilon_M$$

Im IEEE-Standard ist dies realisiert, solange kein Überlauf oder Unterlauf vorkommt.

Bei der Berechnung von Differenzen können große relative Fehler auftreten.

Beispiel 1.5.1. *Es sei die Differenz* $1 - 0.999999$ *in einer sechsstelligen Dezimalarithmetik zu berechnen. Falls* 7 *Stellen möglich wären, käme das korrekte Ergebnis* $0.100000 \cdot 10^{-6}$ *heraus. Bei nur* 6 *Stellen ergibt sich*

$$\begin{array}{r} 1.00000 \\ -0.99999 \\ \hline 0.00001 \end{array} = 0.10000 \cdot 10^{-5}$$

Der relative Fehler ist

$$\frac{\left|0.1 \cdot 10^{-5} - 0.1 \cdot 10^{-6}\right|}{0.1 \cdot 10^{-6}} = 9.9$$

und damit recht groß. Durch eine siebte Stelle für die Berechnung (Schutzstelle) *hätte der Fehler vermieden werden können.*

Beispiel 1.5.2. *Sei eine* 4*-stellige Dezimalarithmetik gegeben. Darin möchten wir die Zahlen* 10.90 *und* 0.009 *addieren. Hier ist das Ergebnis* 10.90*, weil die größere Zahl zwei Stellen vor dem Komma hat, was dazu führt, dass bei der Addition die kleinere Zahl nach der zweiten Nachkommastelle abgeschnitten wird.*

1.5.1 Berechnung von Summen

Verallgemeinern wir die Gleichung

$$\mathrm{fl}(a+b) = (a+b)(1+\epsilon), \quad |\epsilon| \leq \epsilon_M$$

auf Summen der Form

$$S_n = \mathrm{fl}(x_1 + x_2 + \cdots + x_n)$$

Hierbei kommt es auf die Summationsreihenfolge an:

$$\mathrm{fl}(x_1 + x_2 + \cdots + x_n) := \mathrm{fl}(\ldots(\mathrm{fl}(\mathrm{fl}(x_1 + x_2) + x_3)\ldots) + x_n)$$

Es gilt:

$$
\begin{aligned}
S_2 &= \mathrm{fl}(x_1 + x_2) = (x_1 + x_2)(1 + \epsilon_1) = x_1(1 + \epsilon_1) + x_2(1 + \epsilon_1), \quad |\epsilon_1| \leq \epsilon_M \\
S_3 &= \mathrm{fl}(S_2 + x_3) = (S_2 + x_3)(1 + \epsilon_2) \\
&= x_1(1 + \epsilon_1)(1 + \epsilon_2) \\
&+ x_2(1 + \epsilon_1)(1 + \epsilon_2) \\
&+ x_3(1 + \epsilon_2) \\
S_n &= \mathrm{fl}(S_{n-1} + x_n) = (S_{n-1} + x_n)(1 + \epsilon_{n-1}) \\
&= x_1(1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&+ x_2(1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&+ x_3(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&\ldots \\
&+ x_{n-1}(1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) \\
&+ x_n(1 + \epsilon_{n-1}), \qquad\qquad |\epsilon_i| \leq \epsilon_M, \quad i = 1, \ldots, n-1
\end{aligned}
$$

Definiere η_i durch

$$
\begin{aligned}
1 + \eta_1 &= (1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
1 + \eta_2 &= (1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
1 + \eta_3 &= (1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&\ldots \\
1 + \eta_{n-1} &= (1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) \\
1 + \eta_n &= 1 + \epsilon_{n-1}
\end{aligned}
$$

Dann gilt:

$$
S_n = \sum_{i=1}^{n} x_i(1 + \eta_i)
$$

Untersuchung von $1 + \eta_i$

$$
\begin{aligned}
1 + \eta_{n-1} &= (1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) = 1 + \epsilon_{n-2} + \epsilon_{n-1} + \epsilon_{n-2}\epsilon_{n-1} \\
&\simeq 1 + \epsilon_{n-2} + \epsilon_{n-1} \qquad \text{(in 1. Näherung)}
\end{aligned}
$$

In der Tat können wegen

$$
|\epsilon_{n-2}\epsilon_{n-1}| \leq \epsilon_M^2
$$

Terme in ϵ_i höherer Ordnung vernachlässigt werden. Es folgt:

$$
\eta_{n-1} \simeq \epsilon_{n-2} + \epsilon_{n-1} \qquad \text{(in 1. Näherung)}
$$

bzw.

$$
|\eta_{n-1}| \lesssim |\epsilon_{n-2}| + |\epsilon_{n-1}| \leq 2\epsilon_M \qquad \text{(in 1. Näherung)}
$$

Allgemein gilt:

$$
\begin{aligned}
|\eta_1| &\lesssim (n-1)\epsilon_M \\
|\eta_i| &\lesssim (n-i+1)\epsilon_M, \qquad\qquad i = 2, \ldots, n
\end{aligned}
$$

Es gilt:

Satz 1.5.3. *Falls* $n\epsilon_M \leq 0.1$ *und* $|\epsilon_i| \leq \epsilon_M$ *für* $i = 1, \ldots, n$*, so gilt*

$$(1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_n) = 1 + \eta$$

mit

$$\eta \leq 1.06 \cdot n\epsilon_M$$

Setze also

$$\epsilon'_M := 1.06 \cdot \epsilon_M$$

dann werden die approximativen Schranken rigoros:

$$|\eta_1| \leq (n-1)\epsilon'_M$$
$$|\eta_i| \leq (n-i+1)\epsilon'_M, i = 2, \ldots, n$$

Beispiel 1.5.4. *Die Bedingung* $n\epsilon_m \leq 0.1$ *für* $\epsilon_M = 10^{-15}$ *bedeutet* $n \leq 10^{14}$*. Ein Rechner, der pro Addition* $1\mu s = 10^{-6}s$ *braucht, benötigt zur Addition von* 10^{14} *Zahlen*

$$10^8 s = 3.2 \, Jahre$$

D.h. Satz 1.5.3 ist für praktische Zwecke anwendbar.

1.6 Großer Fermatscher Satz und knappe Verfehlungen

Der griechische Mathematiker Diophantos von Alexandria schrieb im dritten Jahrhundert n. Chr. ein mathematisches Werk namens *Arithmetika*. Es ist eine Sammlung von 300 algebraischen Gleichungen zusammen mit Methoden zum Finden von Lösungen. Eine Gleichung, die darin auftaucht ist

$$x^2 + y^2 = z^2$$

wobei x, y, z positive natürlichen Zahlen sein. Die Lösungen sind *pythagoräische Tripel*, und davon gibt es unendlich viele.

Fermat schreibt an dieser Stelle in seinem Exemplar des Werkes an den Rand:[1]

„Es ist jedoch nicht möglich, einen Kubus in 2 Kuben, oder ein Biquadrat in 2 Biquadrate und allgemein eine Potenz, höher als die zweite, in 2 Potenzen mit ebendemselben Exponenten zu zerlegen: Ich habe hierfür einen wahrhaft wunderbaren Beweis entdeckt, doch ist dieser Rand hier zu schmal, um ihn zu fassen."

Mit anderen Worten:

Satz 1.6.1 (Fermat, ca. 1637)**.** *Die Gleichung*

$$x^n + y^n = z^n$$

hat für $n \geq 3$ *und* $x, y, z > 0$ *keine ganzzahlige Lösung.*

Streng genommen, ist Satz 1.6.1 kein Satz, da Fermat keinen Beweis angibt. Daher heißt er auch oft *Fermatsche Vermutung*. Ein erster gültiger Beweis wurde von Andrew Wiles (1995) veröffentlicht. Bis dahin waren Spezialfälle des Satzes bewiesen worden sowie die Tatsache, dass es genügt, ihn für $n = 4$ oder eine ungerade Primzahl zu beweisen.

[1] Im Original ist diese Notiz in Lateinischer Sprache.

Die Methode von Wiles ist wie folgt skizziert: Seit 1990 war bekannt, dass für ein Gegenbeispiel a, b, c mit $a^n + b^n = c^n$ folgt, dass die elliptische Kurve

$$y^2 = x(x - a^n)(x + b^n)$$

(die sog. *Frey-Kurve*) nicht *modular*[2] ist. Andererseits gibt es die *Taniyama-Shimura-Vermutung*, welche besagt, dass alle über $\mathbb{Q}$ definierten elliptischen Kurven modular sind. Andrew Wiles und Richard Taylor bewiesen diese Vermutung für eine große Klasse von elliptischen Kurven, darunter auch für die Frey-Kurve. Dies ergibt einen Widerspruch. Daher kann es kein Gegenbeispiel zum großen Satz von Fermat geben.

Kurze Zeit später erschien in der TV-Zeichentrick-Sitcom *The Simpsons* folgendes Beispiel:

Beispiel 1.6.2 (Homer Simpson, 1995)**.**

$$1782^{12} + 1841^{12} = 1922^{12}$$

Genau genommen erscheint u.A. diese Gleichung in der Szene *Homer*[3] (Homer hoch Drei) der Episode *Treehouse of Horror VI* im Hintergrund, als Homer in die dritte Dimension gerät. Einer Überprüfung auf einem gewöhnlichen Taschenrechner hält dieses „Gegenbeispiel“ stand: die zwölfte Wurzel aus der linken Seite ergibt dort 1922. Der Grund ist die lediglich 10-stellige Gleitkomma-Arithmetik auf dem Taschenrechner. Mit mehr Stellen sieht man, dass diese zwölfte Wurzel ein winziges bisschen größer ist als die ganze Zahl 1922. Der Drehbuchautor Cohen erzeugte diese knappe Verfehlung der Fermat-Gleichung mithilfe eines Computer-Programms. Auch ohne Taschenrechner sieht man, dass dieses Beispiel falsch sein muss: die linke Seite ist die Summe einer geraden mit einer ungeraden Zahl, also ungerade, während die rechte Seite gerade ist, ein Widerspruch.

Drei Jahre später schreibt Homer in *The Wizard of Evergreen Terrace* neben einer Vorhersage der Masse des Higgs-Bosons (14 Jahre vor dessen Entdeckung), der Dichte des Universums sowie einer eigenartigen topologischen Transformation eines Donuts in eine Sphäre, eine weitere knappe Verfehlung eines Gegenbeispiels zum Satz von Fermat an die Tafel:

Beispiel 1.6.3 (Homer Simpson, 1998)**.**

$$3987^{12} + 4365^{12} = 4472^{12}$$

Auf dem Taschenrechner lässt sich auch dieses Beispiel „verifizieren“. Auch hier lässt sich schnell einsehen, dass diese Gleichung nicht stimmen kann: die rechte Seite ist die Summe zweier Zahlen, die durch drei teilbar sind (Quersumme!), während es die rechte Seite nicht ist: diese ist nämlich

$$4472^{12} \equiv 2^{12} \equiv 4^6 \equiv 1 \not\equiv 0 \mod 3$$

[2]Dieser Rand ist zu schmal, um eine Erklärung dieses Begriffs zu fassen...

Kapitel 2

Nichtlineare Gleichungen

2.1 Die Grundaufgabe

Hier geht es um die Grundaufgabe, eine Gleichung

$$f(x) = 0$$

in einer Unbekannten x zu lösen. Offen gehalten ist dabei, ob etwa alle Lösungen gesucht werden, oder ob es Einschränkungen für den Lösungsraum geben soll.

Beispiel 2.1.1. *Betrachten wir die Gleichung $f = 0$ mit*

$$f = x^2 - 9$$

Diese Gleichung ist über den ganzen Zahlen $\mathbb{Z}$ lösbar. Eine Lösungsmöglichkeit existiert mithilfe der Primfaktorzerlegung:

$$9 = 3^2 \quad \Rightarrow \quad x = \pm 3$$

Verwendet wurde im Beispiel:

Satz 2.1.2 (Hauptsatz der Arithmetik)**.** *Jede natürliche Zahl lässt sich in eindeutiger Weise als Produkt von Primzahlen darstellen.*

Hierbei handelt es sich um einen reinen Existenzsatz. Es ist ein bis heute ungelöstes Problem, ob die Primfaktorzerlegung in *polynomieller Zeit* möglich ist.

Beispiel 2.1.3. *Sei*

$$f = x^2 - 2$$

Hier hat $f(x) = 0$ in den rationalen Zahlen $\mathbb{Q}$ keine Lösung. Dies liegt daran, dass $\sqrt{2}$ irrational ist. Wir können aber $\sqrt{2}$ durch Rationalzahlen approximieren. Es ist

$$f(1) = -1 < 0 \quad \textit{und} \quad f(2) = 2 > 0$$

Also hat f nach dem Zwischenwertsatz eine Nullstelle $x \in [1, 2]$. Weiter ist

$$f(3/2) = \frac{1}{4} > 0$$

Also hat f eine Nullstelle $x \in [1, 3/2]$. Dies lässt sich beliebig fortführen, wobei in jedem Schritt das Intervall, welches eine Nullstelle x enthält, halbiert wird.

Verwendet wurde hierbei:

Satz 2.1.4 (Zwischenwertsatz)**.** *Sei $f\colon [a,b] \to \mathbb{R}$ stetig. Dann existiert zu jedem u zwischen $f(a)$ und $f(b)$ ein $c \in [a,b]$ mit $f(c) = u$.*

Beispiel 2.1.5. *Sei*

$$f = x^2 + 1$$

Hier hat $f(x) = 0$ in den reellen Zahlen $\mathbb{R}$ keine Lösung. Jedoch eine Erweiterung des Zahlenraums $\mathbb{Q}$ führt zu einer Lösung: Sei $i := \sqrt{-1}$ (symbolisch). Dann sei

$$\mathbb{Q}(i) := \mathbb{Q} \oplus \mathbb{Q}i$$

die Menge der Zahlen der Form

$$z = x + yi \quad \textit{(symbolisch)}$$

Gerechnet wird mit den üblichen Rechenregeln von Addition und Multiplikation unter Beachtung, dass $i^2 = -1$ gilt. Die Gleichung $f(z) = 0$ ist in $\mathbb{Q}(i)$ exakt lösbar:

$$z = \pm i$$

Bemerkung 2.1.6. *Die Methode aus Beispiel 2.1.5 funktioniert auch mit $\sqrt{2}$:*

$$\mathbb{Q}(\sqrt{2}) := \mathbb{Q} \oplus \mathbb{Q}\sqrt{2}$$

Dann ist $z^2 - 2 = 0$ mit $z = \pm\sqrt{2}$ exakt lösbar.

Häufig wird als Lösungsraum gewählt:

1. $\mathbb{R}$ (reelle Zahlen)
2. $\mathbb{C} = \mathbb{R} \oplus \mathbb{R}i$ (komplexe Zahlen)

Der wichtigste Grund für die Wahl von $\mathbb{C}$ als Lösungsraum ist

Satz 2.1.7 (Hauptsatz der Algebra)**.** *Jedes Polynom*

$$f(X) = a_0 + a_1 X + \cdots + a_n X^n$$

mit Koeffizienten $a_0, \ldots, a_n \in \mathbb{C}$ hat (mindestens) eine Nullstelle in $\mathbb{C}$.

In der Regel gibt es noch weitere Einschränkungen an den Lösungsraum. Im Fall $\mathbb{R}$ soll eine Lösung beispielsweise in einem vorgegebenen Intervall $[a,b]$ gesucht werden. Bei $\mathbb{C}$ könnte der Lösungsraum etwa ein Gebiet oder eine Kreisscheibe sein.

Auf dem Rechner werden noch weitere Einschränkungen gemacht. So ist bei reellem Lösungsraum die Lösung häufig durch endliche Dezimalzahlen zu approximieren. Im Komplexen wird häufig Real- und Imaginärteil durch endliche Dezimalzahlen approximiert.

2.2 Bisektionsmethode

Hier geht es um die Aufgabe, die Gleichung

$$f(x) = 0$$

für eine stetige Funktion f auf einem Intervall $[a, b]$ zu lösen.

Die Voraussetzung sei $f(a) \cdot f(b) < 0$. Dann können wir die *Binärsuche* ansetzen. Setze dazu

$$x_1 := \frac{a+b}{2}, \quad I_0 := [a, b]$$

Es gibt 3 Möglichkeiten:

1. Falls $f(a)f(x_1) < 0$, so setze $I_1 := [a, x_1]$.
2. Falls $f(x_1)f(b) < 0$, so setze $I_1 := [x_1, b]$.
3. Falls $f(x_1) = 0$, so sind wir fertig: Die Lösung ist $x = x_1$.

Führen wir dies fort, so erhalten wir eine Folge x_n als Mittelpunkt des Intervalls I_{n-1}.

Satz 2.2.1. *Die Folge x_n konvergiert gegen*

$$x := \lim_{n\to\infty} x_n$$

und es gilt:

$$f(x) = 0$$

Der Approximationsfehler für x_n beträgt

$$\epsilon_n := |x_n - x| \leq \frac{b-a}{2^n}$$

und die Konvergenz ist linear.

Wir benötigen noch eine Definition. Sei $x_n \to x$ eine konvergente Folge.

Definition 2.2.2. *x_n konvergiert* mit Ordnung q, *falls ein $\rho > 0$ existiert, sodass für alle n gilt:*

$$|x_{n+1} - x| \leq \rho |x_n - x|^q$$

ρ heißt die Konvergenzrate.

Definition 2.2.3. *x_n heißt R-*linear konvergent, *falls es eine Folge $\alpha_n > 0$ gibt, die mit Ordnung 1 und Konvergenzrate $\rho \in (0, 1)$ gegen 0 konvergiert, sodass für alle n gilt:*

$$|x_n - x| \leq \alpha_n$$

Beweis von Satz 2.2.1. Nach dem Zwischenwertsatz (Satz 2.1.4) enthält jedes der Intervalle I_n eine Nullstelle von f. Da der Durchschnitt aller dieser Intervalle ein Punkt ξ ist, muss ξ zugleich Nullstelle und Grenzwert der Folge x_n sein.

Die Konvergenz ist R-linear wegen

$$\epsilon_n \leq \kappa_n := \frac{b-a}{2^n} \quad \Rightarrow \quad \lim_{n\to\infty} \kappa_n = 0 \quad \text{und} \quad \kappa_{n+1} \leq \frac{1}{2}\kappa_n$$

□

2.3 Fixpunktmethoden

Wir wollen die Gleichung

$$f(x) = 0$$

mit $f = x^3 - x - 1$ iterativ mit

$$x_n = \phi(x_{n-1})$$

lösen. Hier sei

1. $\phi(x) := (x+1)^{\frac{1}{3}}$
2. $\phi(x) := x^3 - 1$

Die Lösung ist jeweils als *Fixpunkt*

$$\phi(x) = x$$

gegeben. Anhand von Abbildung 2.1 nehmen wir als Startwert $x_0 = 1.5$. Die Graphen der Iteratoren ϕ sind in Abbildung 2.2 gegeben.

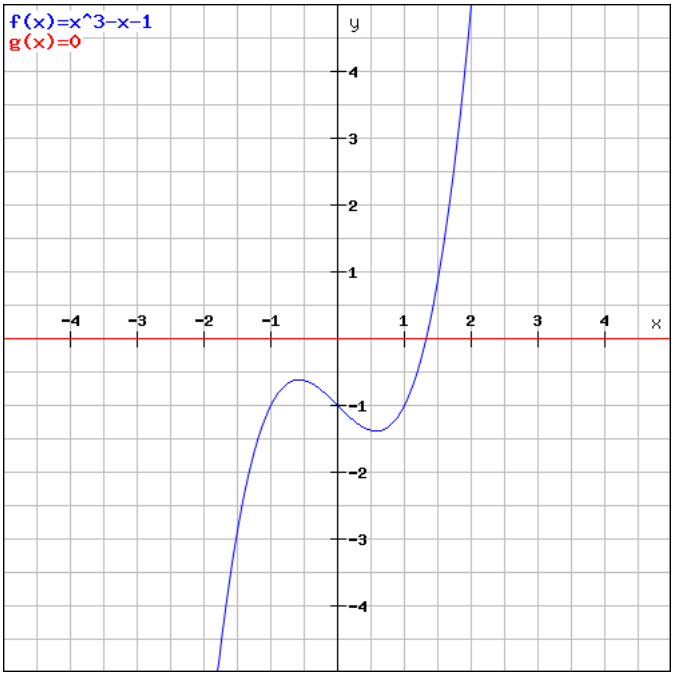


Abbildung 2.1: Der Graph von $x^3 - x - 1$.

Berechnen wir die ersten paar Iterationen:

1. $\phi(x) = (x+1)^{\frac{1}{3}}$.

$$x_1 = 1.3572,\ x_2 = 1.3309,\ x_3 = 1.3259,\ x_4 = 1.3249$$

2. $\phi(x) = x^3 - 1$.

$$x_1 = 2.375,\ x_2 = 12.396,\ x_3 = 1904.003,\ x_4 = 6.902 \cdot 10^9$$

Die Folge in 1. könnte wohl konvergieren, während die Folge in 2. eher divergiert. Ein Hilfsmittel, um dies zu entscheiden, ist

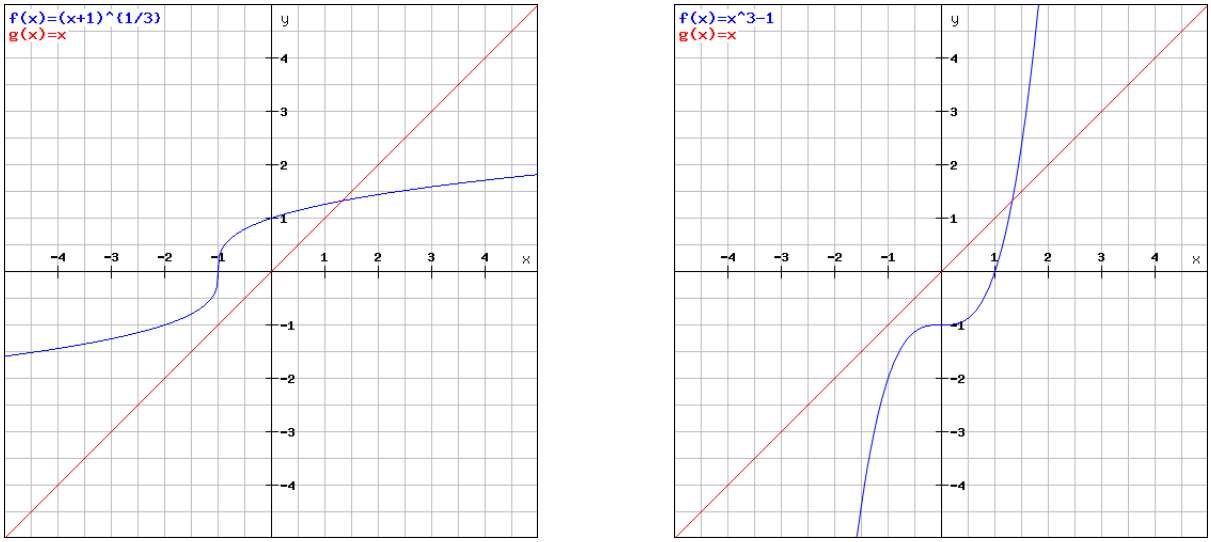


Abbildung 2.2: Die Graphen der Iteratoren ϕ.

Satz 2.3.1 (Banachscher Fixpunktsatz)**.** *Sei* (X, d) *ein nichtleerer* metrischer Raum. *Ist* (X, d) vollständig *und*

$$T\colon X \to X$$

eine Kontraktion, *dann besitzt* T *genau einen Fixpunkt in* X.

Wir werden sogleich die hervorgehobenen Begriffe definieren.

Definition 2.3.2. *Eine Abbildung*

$$d\colon X \times X \to \mathbb{R}$$

ist eine Metrik *auf* X, *falls* $d(x, y) \geq 0$ *für alle* $x, y \in X$, *und für alle* $x, y, z \in X$ *gilt:*

1. $d(x, y) = 0$ *genau dann, wenn* $x = y$. (Definitheit)
2. $d(x, y) = d(y, x)$ (Symmetrie)
3. $d(x, y) \leq d(x, z) + d(z, y)$ (Dreiecksungleichung)

Definition 2.3.3. *Ein metrischer Raum heißt* vollständig, *wenn jede* Cauchy-Folge *in* X *konvergiert.*

Definition 2.3.4. *Eine Folge* (x_n) *heißt* Cauchy-Folge, *wenn zu jedem* $\epsilon > 0$ *ein* $N \in \mathbb{N}$ *existiert, sodass für alle* $m, n > N$ *gilt:*

$$d(x_m, x_n) < \epsilon$$

Definition 2.3.5. *Eine Abbildung* $T\colon X \to X$ *heißt* Kontraktion, *falls eine nichtnegative reelle Zahl* $L < 1$ *existiert mit*

$$d(T(x), T(y)) \leq L \cdot d(x, y)$$

Der banachsche Fixpunktsatz (Satz 2.3.1) gibt eine konstruktive Formulierung her:

Satz 2.3.6. *Ist* $T\colon X \to X$ *eine Kontraktion in einem vollständigen metrischen Raum. so konvergiert die Folge*

$$x_{n+1} = T(x_n)$$

für jeden Startwert $x_0 \in X$ *gegen den Fixpunkt von* T.

Der Grenzwert $x := \lim x_n$ bei einer Kontraktion T ist tatsächlich Fixpunkt.

Beweis.

$$\lim x_n = \lim T(x_{n-1}) \overset{(*)}{=} T(\lim x_{n-1})$$

$(*)$ gilt, da Kontraktionen stetig sind. Also ist $x = T(x)$. □

Die Zahl L bei Kontraktionen T heißt *Lipschitz-Konstante*.

Für stetig differenzierbare reelle Funktionen auf Intervallen gibt es ein Kriterium für das Vorliegen einer Kontraktion.

Satz 2.3.7. *Ist $I \subset \mathbb{R}$ ein abgeschlossenes Intervall und $\phi\colon I \to \mathbb{R}$ stetig differenzierbar mit $\phi(I) \subseteq I$ und*

$$\left|\phi'(x)\right| \leq L < 1$$

für alle $x \in I$, so ist ϕ eine Kontraktion.

Beweis. Nach dem Mittelwertsatz der Differentialrechnung (Satz 2.3.8) existiert für alle $x < y$ in I ein $\xi \in (x, y)$ mit

$$\frac{|\phi(y) - \phi(x)|}{|y - x|} = \left|\phi'(\xi)\right| \leq L < 1$$

Also ist ϕ Kontraktion. □

Verwendet wurde:

Satz 2.3.8 (Mittelwertsatz der Differentialrechnung)**.** *Sei $f\colon [a, b] \to \mathbb{R}$ eine stetige Funktion, die im offenen Intervall (a, b) differenzierbar sei. Dann gibt es ein $\xi \in (a, b)$ mit*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}$$

Beispiel 2.3.9. *Nehmen wir $\phi(x) = (x + 1)^{\frac{1}{3}}$. Dann ist*

$$\phi'(x) = \frac{1}{3}(x + 1)^{-\frac{2}{3}}$$

Für $I = [1, 2]$ gilt:

$$\left|\phi'(x)\right| < 0.21 =: L < 1$$

für $x \in I$ (vgl. Abbildung 2.3).

Beispiel 2.3.10. *Für $\phi(x) = x^3 - 1$ ist $\phi'(x) = 3x^2$. Mit $x_0 = 1.5$ ergibt sich*

$$\phi'(x_0) = 6.75 > 1$$

Also ist ϕ auf keinem Intervall, das x_0 enthält, eine Kontraktion.

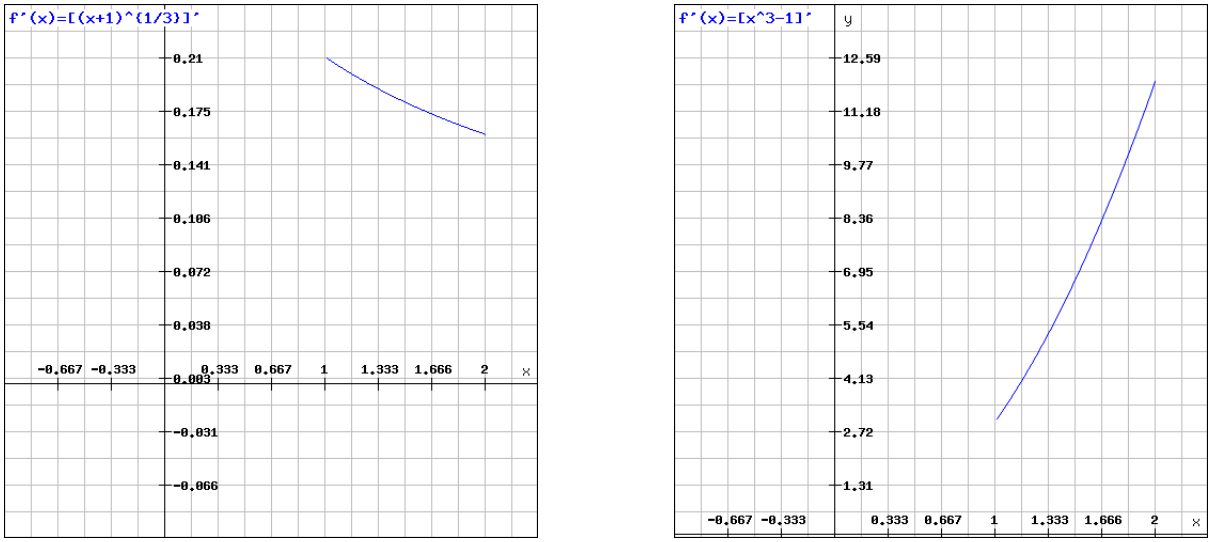


Abbildung 2.3: Die Ableitungen der Iteratoren auf dem Intervall $[1, 2]$.

2.3.1 Fehlerschranken für Kontraktionen

A priori

Sei $T\colon X \to X$ Kontraktion mit Fixpunkt $x \in X$. Für $x_n = T(x_{n-1})$ ist der Fehler

$$\epsilon_n := d(x_n, x)$$

Es gilt:

$$d(x_k, x_{k-1}) \leq L \cdot d(x_{k-1}, x_{k-2}) \leq \cdots \leq L^{k-1} \cdot d(x_1, x_0)$$

wobei L die Lipschitz-Konstante von T sei. Weiter gilt:

$$\begin{aligned} d(x_{m+n}, x_n) &\leq d(x_{m+n}, x_{m+n-1}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (L^{m+n-1} + \cdots + L^n) \cdot d(x_1, x_0) \end{aligned}$$

Mit $m \to \infty$ ergibt sich als *a priori Fehlerschranke*:

$$\epsilon_n \leq \frac{L^n}{1-L} \cdot d(x_1, x_0)$$

Hier wurde die *geometrische Reihe*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad \text{falls} \quad |x| < 1$$

verwendet.

Aus der a priori Fehlerschranke lässt sich die Anzahl der Iterationen abschätzen, wenn der Fehler höchstens $\epsilon > 0$ betragen darf:

$$\epsilon_n \leq \frac{L^n}{1-L} \cdot d(x_1, x_0) \leq \epsilon \quad \Rightarrow \quad n \geq \frac{\log \frac{\epsilon \cdot (1-L)}{d(x_1, x_0)}}{\log L}$$

A posteriori

Für eine Kontraktion T gilt mit $x_n = T(x_{n-1})$:

$$d(x_{n+k}, x_{n+k-1}) \leq L^k \cdot d(x_n, x_{n-1})$$

Daraus ergibt sich:

$$\begin{aligned} d(x_{n+m}, x_n) &\leq d(x_{n+m}, x_{n+m-1}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (L^m + \cdots + L) \cdot d(x_n, x_{n-1}) \end{aligned}$$

Mit $m \to \infty$ folgt die *a posteriori Fehlerschranke*:

$$\epsilon_n \leq \frac{L}{1-L} \cdot d(x_n, x_{n-1})$$

A priori vs. a posteriori

Die a priori Fehlerschranke

$$\tilde{\epsilon}_n := \frac{L^n}{1-L} \cdot d(x_1, x_0)$$

ist vorab ermittelbar, während die a posteriori Fehlerschranke

$$\hat{\epsilon}_n := \frac{L}{1-L} \cdot d(x_m, x_{n-1})$$

erst dann ermittelbar ist, wenn x_n bekannt ist.

Satz 2.3.11. *Bei Kontraktionen ist die a posteriori Fehlerschranke schärfer als die a priori Fehlerschranke.*

Beweis. Da T Kontraktion ist, gilt:

$$\epsilon_n \leq \hat{\epsilon} = \frac{L}{1-L} \cdot d(x_n, x_{n-1}) \leq \frac{L}{1-L} \cdot L^{n-1} d(x_1, x_0) = \tilde{\epsilon}_n$$

□

2.4 Newton-Verfahren

Wir setzen hier voraus, dass $f\colon I \to \mathbb{R}$ zweimal stetig differenzierbar sei und in $\xi \in I$ eine *einfache* Nullstelle habe. Dies bedeutet:

$$f(\xi) = 0 \quad \text{und} \quad f'(\xi) \neq 0$$

Entwickeln von f in eine Taylor-Reihe in $x_0 \in I$ ergibt:

$$f(\xi) = f(x_0) + f'(x_0)(\xi - x_0) + R(\xi, x_0)$$

mit

$$R(\xi, x_0) = f''(\alpha)\frac{(\xi - x_0)^2}{2}$$

wobei α echt zwischen ξ und x_0 ist. Für $|\xi - x_0| \to 0$ gilt:

$$R(\xi, x_0) \to 0$$

also

$$0 \approx f(x_0) + f'(x_0)(\xi - x_0)$$

d.h.

$$\xi \approx x_0 - \frac{f(x_0)}{f'(x_0)}$$

wobei $f'(x_0) \neq 0$ für $|\xi - x_0|$ hinreichend klein. Hieraus wird der Iterator

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

mit Startwert x_0 in der Nähe der gesuchten Nullstelle ξ.

Lemma 2.4.1. *Es gibt in I eine Umgebung $U(\xi) = [\xi - h, \xi + h]$ von ξ, in welcher f' keine Nullstelle hat und*

$$\left|\phi'(x)\right| < 1$$

gilt.

Beweis. Da f' stetig ist, gibt es eine Umgebung $V(\xi)$ von ξ ohne Nullstelle von f'. Es ist

$$\phi'(x) = \frac{f(x)f''(x)}{f'(x)^2}$$

stetig als Produkt stetiger Funktionen, und es ist

$$\phi'(\xi) = 0$$

Also existiert in I eine Umgebung $W(\xi)$ von ξ mit $|\phi'| < 1$. Eine Umgebung von ξ innerhalb $V(\xi) \cap W(\xi)$ leistet somit das Gewünschte. □

Konsequenz. *Der Iterator $\phi\colon U(\xi) \to \mathbb{R}$ ist eine Kontraktion mit Lipschitz-Konstante*

$$L = \max\left\{\left|\phi'(x)\right| \mid x \in U(\xi)\right\} < 1$$

Auch über die Konvergenz des Newton-Iterators können wir etwas aussagen:

Satz 2.4.2. *Sei $f\colon [a,b] \to \mathbb{R}$ zweimal stetig differenzierbar mit einfacher Nullstelle $\xi \in [a,b]$. Dann gibt es in $[a,b]$ eine Umgebung $U(\xi)$, sodass der Iterator ϕ für jeden Startwert $x_0 \in U(\xi)$ quadratisch (d.h. mit Ordnung 2) konvergiert.*

Beweis. Die Taylor-Reihe um $x_n \in U(\xi)$ ist

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2}f''(\alpha_n)(\xi - x_n)^2$$

Wegen

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

folgt:

$$\xi - x_{n+1} = \frac{f(x_n)}{f'(x_n)} + (\xi - x_n) = \frac{-f''(\alpha_n)}{2f'(x_n)}(\xi - x_n)^2$$

Also gilt für den Fehler $\epsilon_n = |\xi - x_n|$:

$$\epsilon_{n+1} = \underbrace{\left|\frac{-f''(\alpha_n)}{2f'(x_n)}\right|}_{=:C_n}\epsilon_n^2$$

mit

$$\lim_{n\to\infty} C_n = \left|\frac{f''(\xi)}{2f'(\xi)}\right|$$

Also ist C_n beschränkt. Mit $\rho = \min\{C_n\}$ ist

$$\epsilon_{n+1} \leq \rho\epsilon_n^2$$

also liegt Konvergenzordnung 2 vor. □

Wir fassen zusammen, dass das Newton-Verfahren quadratisch konvergiert, aber nur *lokal*, d.h. in einer Umgebung der gesuchten Nullstelle.

Ein geeigneter Startwert kann z.B. mit der Bisektionsmethode gesucht werden.

2.4.1 Zwei Anwendungen

Optimierung

Sei $f\colon I \to \mathbb{R}$ dreimal stetig differenzierbar. Gesucht sind Maximal- und Minimalstellen von f. Diese sind Nullstellen von der Ableitung f'. Dies führt auf den Newton-Iterator

$$\phi(x) = x - \frac{f'(x)}{f''(x)}$$

Newton-Raphson-Division

Hier ist $\frac{1}{D}$ für $D \neq 0$ numerisch zu berechnen. Löse dazu die Gleichung

$$f(x) = 0$$

mit $f = \frac{1}{x} - D$ und $x \neq 0$. Der Iterator ist

$$\phi(x) = x - \frac{f(x)}{f'(x)} = x(2 - Dx)$$

und verwendet nur Multiplikation und Subtraktion.

2.5 Sekantenverfahren

Das Newton-Verfahren

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

benötigt in jedem Schritt die Ableitung der Funktion f an der Stelle x_n. Ist diese nicht bekannt oder schwierig zu berechnen, kann der Differentialquotient f' durch den Differenzenquotient approximiert werden:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Dies liefert die Iteration

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

Der Vorteil liegt darin, dass nur Funktionswerte $f(x)$ benötigt werden. Nachteilig ist, dass die Konvergenz langsamer ist. Für den Fehler gilt:

$$\epsilon_{n+1} \approx C \cdot \epsilon_n^\alpha$$

mit

$$\alpha = \frac{1+\sqrt{5}}{2} \approx 1.618 < 2$$

2.6 Newton-Fraktal

Wir beschreiben nun den Newton-Iterator im Komplexen. Dazu sei p eine nicht-konstante *meromorphe* Funktion auf $\mathbb{C}$, d.h.

$$p = \frac{f}{g}$$

mit f, g *holomorph*. Letzteres bedeutet, dass die Funktionen sich in jedem Punkt in $\mathbb{C}$ lokal (d.h. in einer Umgebung) in eine Taylor-Reihe entwickeln lassen. Dann lautet der komplexe Newton-Iterator für die Gleichung $p = 0$:

$$\phi(z) = z - \frac{p(z)}{p'(z)}$$

Die Resultate für den reellen Newton-Iterator übertragen sich auf $\phi(z)$. Die Konvergenz ist abhängig vom Startwert $z_0 \in \mathbb{C}$.

Für die folgende Betrachtung führen wir folgende Bezeichnung ein:

$$\phi^n := \underbrace{\phi \circ \cdots \circ \phi}_{n\text{-mal}}$$

D.h.

$$\phi^n(z) = \underbrace{\phi(\phi(\ldots(\phi}_{n\text{-mal}}(z))\ldots)$$

Für $z \in \mathbb{C}$ betrachten wir die Folge

$$F_z(w)\colon |z - w|, |\phi(z) - \phi(w)|, \left|\phi^2(z) - \phi^2(w)\right|, \ldots$$

Für $F_z(w)_n = |\phi^n(z) - \phi^n(w)|$ gibt es nun zwei Möglichkeiten:

1. Es gibt eine Umgebung $U(z)$ von z, sodass für alle $w \in U(z)$:

$$\lim_{n\to\infty} F_z(w)_n = 0$$

2. In jeder Umgebung $U(z)$ gibt es ein w, sodass

$$F_z(w)_n \not\to 0 \qquad (n \to \infty)$$

Definition 2.6.1. *Die Menge*

$$\mathcal{F}(\phi) := \{z \in \mathbb{C} \mid 1. \text{ gilt für } z\}$$

heißt die Fatou-Menge *von ϕ. Die Menge*

$$\mathcal{J}(\phi) := \{z \in \mathbb{C} \mid 2. \text{ gilt für } z\}$$

heißt die Julia-Menge *oder auch das* Newton-Fraktal *von ϕ.*

Satz 2.6.2. *Sei ϕ ein komplexer Newton-Iterator. Ist der Startwert $z_0 \in \mathcal{F}(\phi)$, so konvergiert $\phi^n(z_0)$ gegen einen periodischen Zyklus endlicher Länge. Ist $z_0 \in \mathbb{J}(\phi)$, so konvergiert $\phi^n(z_0)$ nicht.*

Ist speziell im ersten Fall die Periodenlänge gleich 1, so konvergiert der Newton-Iterator für den Startwert z_0.

Beispiel 2.6.3. *Sei $p = z^3 - 1$. Dies ergibt den Newton-Iterator*

$$\phi = \frac{2z^3 + 1}{3z^2}$$

Wir stellen zunächst fest, dass die Lösungen von $p = 0$ die dritten Einheitswurzeln sind:

$$\zeta = e^{\frac{2\pi i}{3}},\ \zeta^2 = e^{\frac{4\pi i}{3}},\ \zeta^3 = 1$$

mit $i = \sqrt{-1} \in \mathbb{C}$.

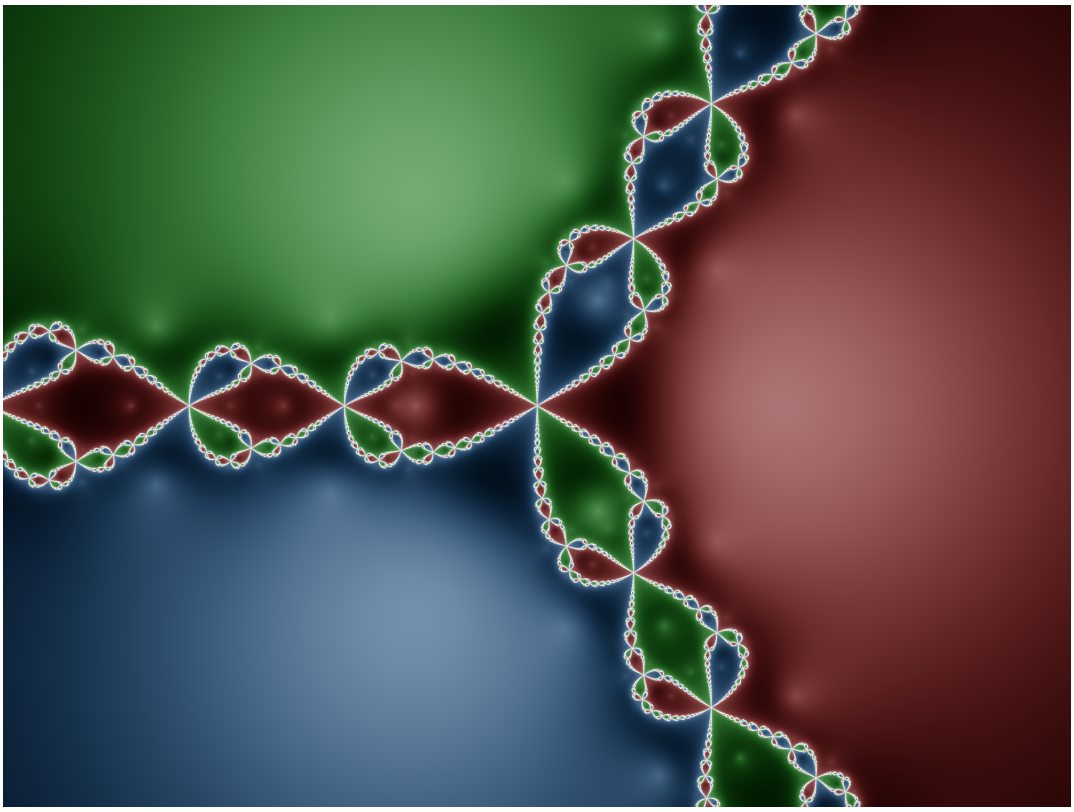


Abbildung 2.4: Newton-Fraktal und Fatou-Menge für $z^3 - 1 = 0$ (Quelle: Wikipedia, Author: Georg-Johann Lay).

In Abbildung 2.4 ist die Julia-Menge weiß. Rot bedeutet Konvergenz gegen 1, grün Konvergenz gegen ζ und blau Konvergenz gegen ζ^2. Die Helligkeit der farbigen Punkte gibt die Konvergenzgeschwindigkeit wieder: hell bedeutet rasch, dunkel bedeutet langsam.

Für das nächste Beispiel benötigen wir noch

Lemma 2.6.4. *Jedes Polynom f mit reellen Koeffizienten und ungeradem Grad hat stets eine reelle Nullstelle.*

Beweis. Fasse f als komplexes Polynom auf. Nach dem Hauptsatz der Algebra (Satz 2.1.7) hat f eine komplexe Nullstelle. Später (in Abschnitt 3.2) werden wir sehen, dass f eine Darstellung hat:

$$f(X) = \alpha \cdot \prod_{\mu} (X - \alpha_\mu), \quad \alpha, \alpha_\nu \in \mathbb{C} \tag{2.1}$$

mit $\alpha_\mu \in \mathbb{C}$. Da f reell ist, ist mit jeder Nullstelle $\xi \in \mathbb{C}$ auch die komplex-konjugierte $\bar{\xi}$ eine Nullstelle von f, denn:

$$f(\bar{\xi}) = \sum_\nu a_\nu \bar{\xi}^\nu = \sum_\nu \bar{a}_\nu \overline{\xi^\nu} = \overline{\sum_\nu a_\nu \xi^\nu} = \bar{0} = 0$$

wenn $f(X) = \sum_\nu a_\nu X^\nu$ ist. Es folgt wegen (2.1), dass bei ungeradem Grad für mindestens eine der Nullstellen ξ gelten muss, dass

$$\bar{\xi} = \xi$$

also ξ reell sein muss. □

Beispiel 2.6.5. *Sei $p = z^3 - 2z + 2$. Dann ist der Newton-Iterator*

$$\phi = \frac{2z^3 - 2}{3z^2 - 2}$$

p hat mindestens eine und höchstens drei reelle Nullstellen. Die kritischen Stellen sind gegeben durch:

$$p'(x) = 0 \quad \Leftrightarrow \quad x = \pm\sqrt{\frac{2}{3}}$$

Wegen

$$0 < p\left(\sqrt{\frac{2}{3}}\right) = 2 - \frac{4}{3}\sqrt{\frac{2}{3}} < p\left(-\sqrt{\frac{2}{3}}\right) = 2 + \frac{4}{3}\sqrt{\frac{2}{3}}$$

ist $-\sqrt{\frac{2}{3}}$ lokales Maximum, $\sqrt{\frac{2}{3}}$ lokales Minimum, und dazwischen gibt es keine Nullstelle auf der reellen Achse. Also haben wir die Zerlegung

$$p(X) = (Z - \alpha)(X - \beta)(X - \bar{\beta})$$

mit $\alpha < 0$ und $\beta \in \mathbb{C} \setminus \mathbb{R}$. Ein reeller Plot des Polynoms p findet sich in Abbildung 2.5.

Die nächste Vorüberlegung ist, dass es hier außerhalb der Julia-Menge nicht immer Konvergenz gibt. Hier ist nämlich

$$\phi(0) = 1, \quad \phi(1) = 0$$

d.h. es liegt ein Zyklus der Länge 2 vor. Es gibt also Startwerte, die gegen diesen Zyklus konvergieren. In Abbildung 2.6 ist die Julia-Menge weiß; rot bedeutet Konvergenz zum Zyklus $\{0, 1\}$, beige Konvergenz zur Nullstelle α, grün Konvergenz zur Nullstelle β und blau Konvergenz zur Nullstelle $\bar{\beta}$.

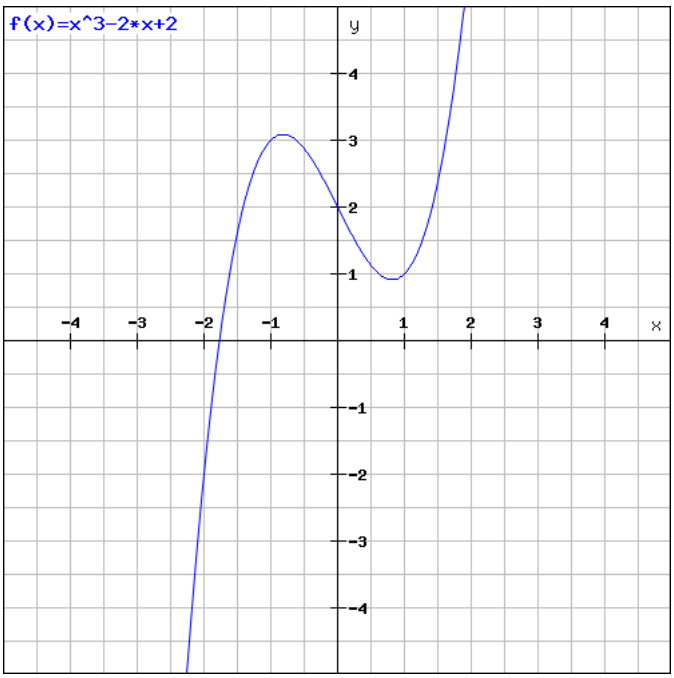


Abbildung 2.5: Plot von $x^3 - 2x + 2$

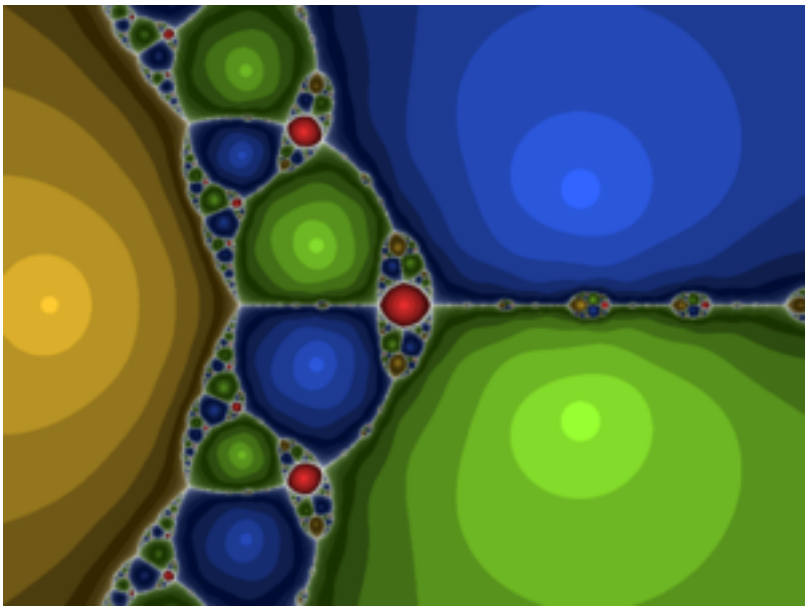


Abbildung 2.6: Das Newton-Fraktal und Fatou-Menge zu $z^3 - 2z + 2 = 0$ (Quelle: Wikipedia, Author: Georg-Johann Lay).

Kapitel 3

Polynome

Es bezeichne $K[X]$ die Menge aller Polynome mit Koeffizienten aus K, wobei $K = \mathbb{Q}$, $\mathbb{R}$ oder $\mathbb{C}$ der Körper der rationalen, reellen oder komplexen Zahlen sei.

Sei $f = \sum_{\nu \in \mathbb{N}} a_\nu X^\nu \in K[X]$ ein Polynom. Für $f \neq 0$ existiert die Zahl

$$\deg(f) := \max\{\nu \mid a_\nu \neq 0\}$$

Diese Zahl heißt der *Grad* von f. Weiter definieren wir

$$\deg(0) := -\infty$$

3.1 Euklidischer Algorithmus

Der Euklidische Algorithmus beruht auf der *Division mit Rest*:

Division mit Rest

Seien $f, g \in K[X]$ mit $g \neq 0$. Dann existieren $q, r \in K[X]$ mit $\deg(r) < \deg(g)$, sodass

$$f(X) = q(X) \cdot g(X) + r(X)$$

Falls bei der Division mit Rest $r = 0$ ist, schreiben wir

$$g \mid f$$

(„g teilt f“). Der *größte gemeinsame Teiler* $\mathrm{ggT}(f, g)$ ist ein Polynom $d \in K[X]$ mit:

1. $d \mid f$ und $d \mid g$ (d.h. d ist gemeinsamer Teiler)
2. $e \mid f$ und $e \mid g \Rightarrow e \mid d$ (d.h. d ist maximal als gemeinsamer Teiler)

Beachte, dass der größte gemeinsame Teiler nicht eindeutig bestimmt ist. Aber es gilt:

Lemma 3.1.1. *Seien $f, g \neq 0$. Sind d_1 und d_2 größte gemeinsame Teiler von f und g, so gilt:*

$$d_1 = c_2 \cdot d_2$$

mit $\deg(c_2) = 0$.

Beweis. Da d_1, d_2 beide größte gemeinsame Teiler sind, gilt:

$$d_1 \mid d_2 \quad \text{und} \quad d_2 \mid d_1$$

Dies bedeutet:

$$d_2 = c_1 \cdot d_1 \quad \text{und} \quad d_1 = c_2 \cdot d_2$$

Da $d_1 \neq 0$ und $d_2 \neq 0$, folgt:

$$\begin{aligned} &\deg(d_2) = \deg(c_1) + \deg(d_1) \quad \text{und} \quad \deg(d_1) = \deg(c_2) + \deg(d_2) \\ \Rightarrow \quad &\deg(d_2) = \deg(c_1) + \deg(c_2) + \deg(d_2) \\ \Rightarrow \quad &0 = \deg(c_1) + \deg(c_2) \end{aligned}$$

Da $\deg(c_1)$ und $\deg(c_2)$ natürliche Zahlen sind, folgt:

$$\deg(c_1) = \deg(c_2) = 0$$

□

Der euklidische Algorithmus

„[The Euclidean Algorithm] is the granddaddy of all algorithms, because it is the oldest nontrivial algorithm that has survived to the present day."

(Donald Knuth, *The Art of Computer Programming*, Vol. 2: *Seminumerical Algorithms*, 2nd edition (1981), p. 318.)

Algorithm 3.1.2 (Euklid)**.** Eingabe: *Polynome* $a, b \in K[X] \setminus \{0\}$.

Führe solange aus, bis sich ein Rest $r_{N+1} = 0$ *ergibt:*

$$\begin{aligned} a &= q_0 \cdot b + r_0, && \deg(r_0) < \deg(b) \\ b &= q_1 \cdot r_0 + r_1, && \deg(r_1) < \deg(r_0) \\ r_0 &= q_2 \cdot r_1 + r_2, && \deg(r_2) < \deg(r_1) \\ &\vdots \\ r_{N-1} &= q_{N+1} \cdot r_N && (r_{N+1} = 0) \end{aligned}$$

Ausgabe: r_N.

Satz 3.1.3 (Euklid)**.** *Der euklidische Algorithmus terminiert. Der letzte Rest* $r_N \neq 0$ *ist der größte gemeinsame Teiler von* a *und* b.

Beweis. Die Folge $\deg(r_n) \in \mathbb{N} \cup \{-\infty\}$ ist echt absteigend. Also existiert ein kleinstes N, sodass $r_{N+1} = 0$. Dann ist $d := r_N \neq 0$ der letzte Rest. Zeigen wir nun, dass r_N die Eigenschaften des ggT erfüllt.

- $r_N \mid a$ und $r_N \mid b$.

  $$\begin{aligned} r_{N-1} &= q_{N+1} \cdot r_N \quad &&\Rightarrow r_N \mid r_{N-1} \\ r_{N-2} &= q_N \cdot r_{N-1} + r_N \quad &&\Rightarrow r_N \mid r_{N-2} \end{aligned}$$

  Aus jeder vorangehenden Gleichung in Algorithmus 3.1.2 folgt: $r_N \mid r_n$ für alle n. Insbesondere gilt:

  $$r_N \mid a \quad \text{und} \quad r_N \mid b$$

- $e \mid a$ und $e \mid b \;\Rightarrow\; e \mid r_N$. Aus

$$e \mid a = q_0 \cdot b + r_0 \quad \text{und} \quad e \mid b$$

folgt $e \mid r_0$. Aus der nächsten Gleichung von Algorithmus 3.1.2 folgt dann $e \mid r_1$ usw. bis schließlich $e \mid r_N$ folgt.

Dies zeigt die Behauptung. □

Ganz allgemein funktioniert der euklidische Algorithmus für so-genannte *euklidische Ringe*, in denen es eine Division mit Rest gibt.

Beispiel 3.1.4. *Der Ring $\mathbb{Z}$ der ganzen Zahlen ist ein euklidischer Ring. Die Rolle der Grad-Funktion* deg *übernimmt hier der Absolutbetrag $|\cdot|$:*

$$a = q \cdot b + r, \quad |r| < |b|$$

ist hier die Division mit Rest.

Bedeutung für Euklid

Euklid selbst verwendet seinen Algorithmus, um den Hauptsatz der Arithmetik (Satz 2.1.2) zu zeigen. Über den ganzen Zahlen formuliert er sich so: Jede Zahl $n \in \mathbb{Z} \setminus \{0\}$ hat eine Darstellung

$$n = \pm p_1 \cdots p_r$$

mit eindeutig bestimmten Primzahlen p_i bis auf die Reihenfolge der Faktoren.

3.2 Sturmsche Kette

Möchte man die Nullstellen eines Polynoms bestimmen, so kann man bei kleinem Grad die Nullstellen durch Radikale ausdrücken, wodurch sie explizit gegeben sind. Ein Beispiel ist die bekannte Lösungsformel für die allgemeine quadratische Gleichung. Auch für Polynome dritten und vierten Grades gibt es Lösungsformeln (die *Cardano-Formeln*) für deren Nullstellen. Ab Grad fünf gilt:

Satz 3.2.1 (Abel-Ruffini)**.** *Die allgemeine Polynom-Gleichung von Grad fünf oder höher hat keine Lösung un Radikalen.*

Dies bedeutet, dass die Nullstellen des allgemeinen Polynoms sich nicht durch Radikale ausdrücken lassen. Also ist man für Polynome höheren Grades auf numerische Methoden zu deren Berechnung angewiesen.

In diesem Abschnitt gilt es, die Aufgabe zu lösen, wieviele Nullstellen ein reelles Polynom $f \in \mathbb{R}[X]$ in einem vorgegebenen Intervall $[a, b]$ hat.

Zunächst ein paar allgemeine Aussagen für $K = \mathbb{Q}, \mathbb{R}$ oder $\mathbb{C}$:

Lemma 3.2.2. *Sei $f \in K[X] \setminus \{0\}$. Genau dann hat $f(X)$ in $\xi \in K$ eine Nullstelle, wenn*

$$(X - \xi) \mid f$$

gilt.

Beweis. $\Rightarrow$: Sei $\xi \in K$ Nullstelle von f. Division mit Rest ergibt:

$$f(X) = q(X)(X - \xi) + r(X)$$

mit $\deg(r) < 1$. Es ist

$$0 = f(\xi) = r(\xi)$$

Wegen $\deg(r) < 1$ folgt $r = 0$. Also $(X - \xi) \mid f$.

$\Leftarrow$: Gelte $(X - \xi) \mid f$, also

$$f(X) = q(X) \cdot (X - \xi)$$

Dann ist $f(\xi) = 0$. □

Als Konsequenz ergibt sich, dass ein Polynom $f \in K[X] \setminus \{0\}$ höchstens $\deg(f)$ Nullstellen in K hat.

Beweis. Sei $\xi \in K$ Nullstelle von f. Aus

$$f = q \cdot (X - \xi)$$

folgt:

$$\deg(q) = \deg(f) - 1$$

und nach spätestens $\deg(f)$ Nullstellen ist Schluss. □

Für $K = \mathbb{C}$ ergibt sich, dass ein nichtkonstantes Polynom f eine Darstellung

$$f(X) = \alpha \cdot \prod_{\mu} (X - \alpha_\mu), \quad \alpha, \alpha_\nu \in \mathbb{C}$$

hat.

Definition 3.2.3. *Ein Polynom $f \in K[X]$ heißt* quadratfrei*, falls für kein nichtkonstantes Polynom $g \in K[X]$ gilt: $g^2 \mid f$.*

Lemma 3.2.4. *Quadratfreie Polynome haben nur einfache Nullstellen.*

Beweis. Sei $\xi \in K$ Nullstelle von $f \in K[X]$. Dann gilt:

$$f = q \cdot (X - \xi)$$

mit $q(\xi) \neq 0$, da f quadratfrei ist. Es ist zu zeigen, dass $f'(\xi) \neq 0$. Es gilt:

$$f' = q' \cdot (X - \xi) + q \quad \Rightarrow \quad f'(\xi) = q(\xi) \neq 0$$

□

Jetzt kommen wir zur sturmschen Kette.

Definition 3.2.5. *Sei $f \in \mathbb{R}[X]$ ein Polynom. Dann heißt die Folge $p_0 := f$, $p_1 := f'$, $p_2, \ldots, p_N$ mit*

$$\begin{aligned}
f &= q_1 \cdot f' - p_2, && \deg(p_2) < \deg(f') \\
f' &= q_2 \cdot p_2 - p_3, && \deg(p_3) < \deg(p_2) \\
p_2 &= q_3 \cdot p_3 - p_4, && \deg(p_4) < \deg(p_3) \\
&\vdots \\
p_{N-1} &= q_N \cdot p_N && (p_{N+1} = 0)
\end{aligned}$$

eine sturmsche Kette *zu f.*

Die sturmsche Kette gemäß Definition sind, bis auf das Vorzeichen, die Reste beim euklidischen Algorithmus zur Berechnung von $\mathrm{ggT}(f, f')$.

Satz 3.2.6 (Sturm)**.** *Sei $f \in \mathbb{R}[X]$ quadratfrei und $p_0, \ldots, p_N$ eine sturmsche Kette zu f. Dann ist für $a < b$ die Anzahl der Nullstellen im Intervall $(a, b]$ gleich*

$$\sigma(a) - \sigma(b)$$

wobei $\sigma(\xi)$ die Anzahl der Vorzeichenwechsel in der Folge

$$p_0(\xi), \ldots, p_N(\xi)$$

sei.

Bemerkung 3.2.7. *Der Satz von Sturm gilt auch, wenn eine sturmsche Kette $p_0, \ldots, p_N$ durch*

$$\alpha_0 \cdot p_0, \ldots, \alpha_N \cdot p_N \tag{3.1}$$

mit $\alpha_0, \ldots, \alpha_N > 0$ ersetzt wird. Die Folge (3.1) heißt ebenfalls eine sturmsche Kette *zu f. Mithilfe der α_i können auftretende Nenner in Brüchen entfernt werden.*

An einer sturmschen Kette kann erkannt werden, ob f quadratfrei ist, ob sich also die Nullstellen in Intervallen zählen lassen:

Lemma 3.2.8. *$f \in K[X]$ ist genau dann quadratfrei, wenn*

$$\mathrm{ggT}(f, f') = 1$$

gilt.

Beweis. $\Rightarrow$: Falls $d(X) = \mathrm{ggT}(f, f')$ nicht konstant ist, so hat $d(X)$ nach dem Hauptsatz der Algebra (Satz 2.1.7) eine Nullstelle $\xi \in \mathbb{C}$. Da d Teiler von f und f' ist, folgt:

$$f(\xi) = f'(\xi) = 0$$

Also ist die Nullstelle ξ nicht einfach, folglich f nach Lemma 3.2.4 nicht quadratrei.

$\Leftarrow$: Sei $\mathrm{ggT}(f, f') = 1$. Ist $f = g^2 \cdot q$, so gilt:

$$f' = 2gg'q + g^2q' = g \cdot (2g'q + gq')$$

Also gilt $g \mid \mathrm{ggT}(f, f') = 1$, d.h. g ist konstant. Somit ist f quadratfrei. □

Was kann getan werden, wenn f nicht quadratfrei ist?

Satz 3.2.9. *Sei $f \in K[X] \setminus \{0\}$. Dann ist*

$$g := \frac{f}{\mathrm{ggT}(f, f')}$$

quadratfrei und hat dieselben Nullstellen wie f.

Beweis. Sei

$$f = (X - \xi)^k \cdot h$$

mit $h(\xi) \neq 0$. Dann ist

$$f' = k \cdot (X - \xi)^{k-1} \cdot h + (X - \xi)^k \cdot h'$$

Also $(X - \xi)^{k-1} \mid f'$ und $(X - \xi)^k \nmid f'$, da sonst

$$(X - \xi)^k \mid f' - (X - \xi)^k \cdot h' = k \cdot (X - \xi)^{k-1}$$

gilt, was nicht geht. Es folgt:

$$(X - \xi)^{k-1} \mid \mathrm{ggT}(f, f') \quad \text{und} \quad (X - \xi)^k \nmid \mathrm{ggT}(f, f')$$

D.h. die Multiplizität der Nullstelle ξ in $\mathrm{ggT}(f, f')$ ist Eins weniger als in f. Also ist g quadratfrei und hat dieselben Nullstellen wie f. □

Beispiel 3.2.10. *Sei $f(X) = X^3 - 2X^2$. Wir möchten die reellen Nullstellen von f zählen. Sturms Methode ergibt die sturmsche Kette*

$$f, f' = 3X^2 - 4X, X$$

und wir sehen, dass $\mathrm{ggT}(f, f') = X$ ist, also f nicht quadratfrei ist. Also nehmen wir

$$g := \frac{f}{\mathrm{ggT}(f, f')} = X^2 - 2X$$

und erhalten die sturmsche Kette

$$g, g' = 2X - 2, p_2 = 1$$

Die folgende Vorzeichentabelle:

	$a >> 0$	$-a << 0$
g	$+$	$+$
g'	$+$	$-$
p_2	$+$	$+$
σ	0	2

ergibt

$$\sigma(-a) - \sigma(a) = 2$$

als Anzahl der reellen Nullstellen von f (ohne Vielfachheiten).

3.3 Primzahlen, vollkommene und narzisstische Zahlen

In der Episode *Homerun für die Liebe*[1] erscheint im Baseball-Stadion, als Tabitha ihre Liebeserklärung an ihren Mann abgeben will, auf der Großbildleinwand eine Aufforderung an das Publikum, die Anzahl der Zuschauer zu schätzen:

a) 8191

b) 8128

[1] Originaltitel: *Marge and Homer Turn a Couple Play*

c) 8208

d) No way to tell

Die erste Zahl ist eine Primzahl. Sie ist sogar eine *Mersenne-Primzahl*. Solche sind Primzahlen der Form $2^p - 1$, wobei p selbst eine Primzahl ist. In der Tat ist

$$8191 = 2^{13} - 1$$

Mersenne-Primzahlen sind Rekordhalter: die zehn größten von ihnen sind die größten Primzahlen, die je entdeckt wurden. Die größte bekannte Primzahl ist $2^{74207281} - 1$, eine Mersenne-Primzahl, die im Jahr 2016 entdeckt wurde.

Die zweite Zahl ist eine *vollkommene Zahl*, d.h. eine Zahl, die gleich der Summe ihrer echten Teiler ist. Die kleinste vollkommene Zahl ist

$$6 = 1 + 2 + 3$$

Die nächste ist

$$28 = 1 + 2 + 4 + 7 + 14$$

Diese ist gefolgt von 496 und von 8128. Es ist nicht bekannt, ob es unendlich viele vollkommene Zahlen gibt oder nicht. Auch bisher unbekannt ist, ob alle vollkommenen Zahlen gerade sind.

Die dritte Zahl ist eine *narzisstische Zahl*. Diese haben die Eigenschaft, dass die Summe ihrer Ziffern, jeweils mit der Anzahl der Stellen potenziert, die Zahl selbst ergibt. In der Tat ist

$$8208 = 8^4 + 2^4 + 0^4 + 8^4$$

Es wurde bewiesen, dass es nur endlich viele narzisstische Zahlen gibt. Es sind 88 Stück, und die größte ist

$$115\,132\,219\,018\,763\,992\,565\,095\,597\,973\,971\,522\,401$$

Kapitel 4

Interpolation

Sei $K = \mathbb{Q}, \mathbb{R}$ oder $\mathbb{C}$. Seien $n+1$ Paare $(x_i, f_i) \in K^2$ und eine Funktion $\Phi(x, a_0, \ldots, a_n)$ gegeben. Die Aufgabe besteht nun darin, die Parameter $a_0, \ldots, a_n$ so zu wählen, dass

$$\Phi(x_i, a_0, \ldots, a_n) = f_i$$

ist.

Wir betrachten das *lineare Interpolationsproblem*:

Lineares Interpolationsproblem

Hier ist

$$\Phi(x, a_0, \ldots, a_n) = a_0 \cdot \Phi_0(x) + \cdots + a_n \cdot \Phi(x)$$

mit linear unabhängigen Funktionen $\Phi_0, \ldots, \Phi_n$.

4.1 Polynominterpolation

Hier besteht die Aufgabe darin, das Polynom P von Grad $\leq n$ zu finden, welches an $n+1$ verschiedenen Stellen α_i die Werte $f(\alpha_i)$ annimmt.

Der Raum, in dem gesucht wird, ist

$$K[X]_n := \{\text{Polynome vom Grad} \leq n\}$$

Es handelt sich dabei um ein lineares Interpolationsproblem: Löse das lineare Gleichungssystem

$$P(\alpha_i) = f(\alpha_i)$$

abhängig von einer Basis $b_i(X)$ des Vektorraums $K[X]_n$. Es ist

$$P(X) = \sum_{i=0}^{n} \rho_i b_i(X) =: \Phi(X, \rho_0, \ldots, \rho_n)$$

4.1.1 Standardbasis

Die Standardbasis für $K[X]_n$ ist

$$b_i(X) = X^i, \quad i = 0, \ldots, n$$

Bezüglich dieser hat das Polynom die übliche Darstellung

$$P(X) = \sum_{i=0}^{n} \rho_i X^i$$

Das Interpolationsproblem führt auf das lineare Gleichungssystem

$$\underbrace{\begin{pmatrix} 1 & \alpha_0 & \ldots & \alpha_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \ldots & \alpha_n^n \end{pmatrix}}_{=V(\alpha_0,\ldots,\alpha_n)} \begin{pmatrix} \rho_0 \\ \vdots \\ \rho_n \end{pmatrix} = \begin{pmatrix} f(\alpha_0) \\ \vdots \\ f(\alpha_n) \end{pmatrix}$$

Definition 4.1.1. *Die Matrix $V(\alpha_0, \ldots, \alpha_n)$ heißt* Vandermonde-Matrix.

Lemma 4.1.2. *Die Vandermonde-Matrix ist genau dann regulär, wenn die α_i paarweise verschieden sind.*

Beweisskizze. Es gilt:

$$\det V(\alpha_0, \ldots, \alpha_n) = \prod_{0 \le j < k \le n} (\alpha_j - \alpha_k)$$

Diese Determinante ist genau dann von Null verschieden, wenn die α_i paarweise verschieden sind. □

Als Konsequenz ergibt sich daraus:

Satz 4.1.3. *Das Polynominterpolationsproblem ist eindeutig lösbar.*

Beweis. Es ist nach Lemma 4.1.2 für die Standardbasis von $K[X]_n$ eindeutig lösbar. Somit ist es auch für eine beliebige Basis $b_0(X), \ldots, b_n(X)$ eindeutig lösbar. □

Bemerkung 4.1.4. *Lemma 4.1.2 ist der Grund dafür, dass beim Polynominterpolationsproblem für $n+1$ verschiedene Punkte verlangt wird, dass der Polynomgrad $\le n$ ist.*

Beispiel 4.1.5. *Durch 2 verschiedene Punkte in der euklidischen Ebene ist eine Gerade eindeutig bestimmt, aber Parabeln gibt es unendlich viele, welche durch diese Punkte verlaufen.*

Die Methode, die sich bei der Polynominterpolation für die Standardbasis anbietet, ist der Gauß-Algorithmus. Dessen Komplexität ist jedoch mit $O(n^3)$ recht hoch.

4.1.2 Lagrange-Polynome

Die *Lagrange-Polynome*

$$\ell_i(X) = \prod_{\substack{j=0 \\ j \ne i}}^{n} \frac{X - \alpha_j}{\alpha_i - \alpha_j}$$

erfüllen die Eigenschaft

$$\ell_i(\alpha_j) = \delta_{ij} := \begin{cases} 1, & i = j \\ 0, & i \ne j \end{cases}$$

δ_{ij} ist das *Kronecker-Delta*. Die Lagrange-Polynome sind eine Basis von $K[X]_n$.

Beweis. Sei

$$B(X) := \sum_{\nu=0}^{n} \beta_\nu \ell_\nu(X) = 0$$

Dann ist

$$0 = B(\alpha_\mu) = \sum_{\nu=0}^{n} \beta_\nu \delta_{\nu\mu} = \beta_\mu$$

Also sind die $\ell_0, \ldots, \ell_n$ linear unabhängig. □

Die Lösung des Interpolationsproblems ist gegeben durch:

$$P(X) = \sum_{i=0}^{n} f(\alpha_i)\ell_i(X)$$

Beweis.

$$P(\alpha_\mu) = \sum_{\nu=0}^{n} f(\alpha_\nu)\ell_\nu(\alpha_\mu) = \sum_{\nu=0}^{n} f(\alpha_\nu)\delta_{\nu\mu} = f(\alpha_\mu)$$

□

Die Koeffizienten ρ_i sind also einfach die Funktionswerte:

$$\rho_i = f(\alpha_i)$$

Ein Nachteil der Lagrange-Polynome ist, dass sie sich bei Hinzunahme einer weiteren Stützstelle α_{n+1} allesamt ändern.

Beispiel 4.1.6. *Wir interpolieren mit Lagrange-Polynomen:*

$$f(0) = 3,\ f(1) = 2,\ f(3) = 1,\ f(4) = 0 \tag{4.1}$$

und schätzen damit den Wert $f(2.5)$. Es ist

$$\begin{aligned}
\ell_0(X) &= -\frac{1}{12}(X-1)(X-3)(X-4) \\
\ell_1(X) &= \frac{1}{6}X(X-3)(X-4) \\
\ell_2(X) &= -\frac{1}{6}X(X-1)(X-4)
\end{aligned}$$

Das Interpolationspolynom ist

$$P(X) = 3 \cdot \ell_0 + 2 \cdot \ell_1 + 1 \cdot \ell_2 + 0 \cdot \ell_3 = \frac{1}{12}\left(-X^3 + 6X^2 - 17X + 36\right)$$

und $f(2.5) = P(2.5) = 1.28125$. Das Polynom und die Werte sind in Abbildung 4.1 abgebildet.

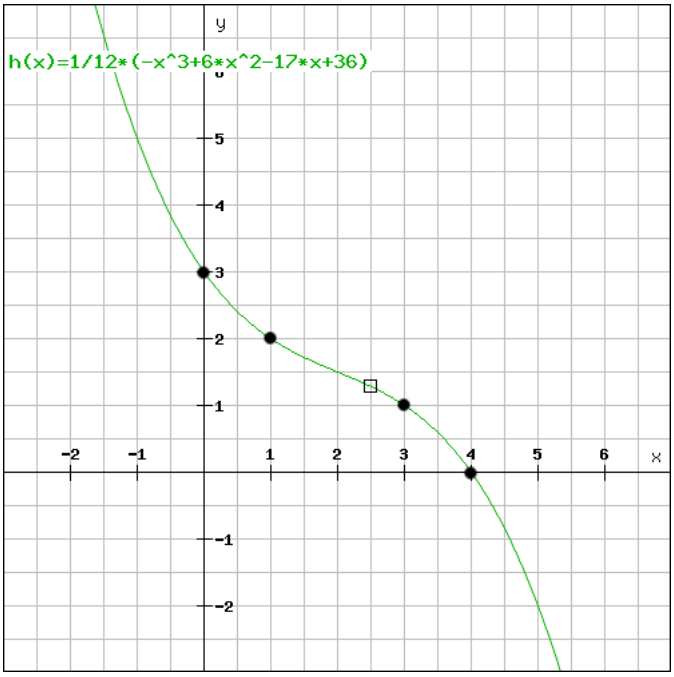


Abbildung 4.1: Das Interpolationspolynom zu den Werten in (4.1).

4.1.3 Newton-Polynome

Seien $\alpha_0, \ldots, \alpha_n \in K$ paarweise verschieden. Die *Newton-Polynome* sind

$$N_i(X) = \prod_{k=0}^{i-1}(X - \alpha_k), \quad i = 0, \ldots, n$$

Es ist für $i > 0$:

$$N_i(\alpha_j) = \begin{cases} \prod_{k=0}^{i-1}(\alpha_j - \alpha_i), & j \geq i \\ 0, & j < i \end{cases}$$

Daher führt der Ansatz

$$P(X) = \sum_{i=0}^{n} \rho_i N_i(X)$$

auf das lineare Gleichungssystem

$$\text{(4.2)} \qquad \begin{pmatrix} 1 & & & 0 \\ 1 & (\alpha_1 - \alpha_0) & & \\ 1 & (\alpha_2 - \alpha_0) & (\alpha_2 - \alpha_0)(\alpha_2 - \alpha_1) & \\ \vdots & \vdots & \ddots & \\ 1 & (\alpha_n - \alpha_0) & \cdots & \prod_{i=0}^{n-1}(\alpha_n - \alpha_i) \end{pmatrix} \begin{pmatrix} \rho_0 \\ \vdots \\ \rho_n \end{pmatrix} = \begin{pmatrix} f(\alpha_0) \\ \vdots \\ f(\alpha_n) \end{pmatrix}$$

denn es ist

$$P(\alpha_j) = \sum_{i=0}^{j} \rho_i N_i(\alpha_j)$$

Da die Koeffizientenmatrix in (4.2) invertierbar ist, folgt dass die Newton-Polynome $N_0(X)$ bis $N_n(X)$ eine Basis von $K[X]_n$ sind.

Da die Koeffizientenmatrix eine obere Dreiecksmatrix ist, kann die Lösung durch *Vorwärtssubstitution* gefunden werden, indem mit der obersten Gleichung

$$\rho_0 = f(\alpha_0)$$

begonnen wird, und diese in die zweite Gleichung eingesetzt wird:

$$f(\alpha_0) + (\alpha_1 - \alpha_0)\rho_1 = f(\alpha_1)$$

Dies ist eine lineare Gleichung mit einer Unbekannten, usw.

4.1.4 Interpolationsfehler

Eine typische Aufgabe ist es, eine stetige Funktion durch ein Interpolationspolynom zu approximieren. Grundlage hierfür ist:

Satz 4.1.7 (Weierstraß)**.** *Sei $f\colon [a,b] \to \mathbb{R}$ stetig. Dann gibt es für jedes $\epsilon > 0$ ein Polynom $P(X) \in \mathbb{R}[X]$, sodass*

$$\|f - P\|_\infty := \max\{|f(t) - P(t)| \mid t \in [a,b]\} < \epsilon$$

gilt.

Dies ergibt Interpolationsfehler. Hierbei gilt folgendes:

Satz 4.1.8. *Sei $f\colon [a,b] \to \mathbb{R}$ $(n+1)$-fach stetig differenzierbar. Dann existiert zu jedem $t \in [a,b]$ ein $\xi \in I_t$, sodass*

$$f(t) - P(t) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot N_{n+1}(t)$$

gilt. Hierbei sei $P(X)$ das Interpolationspolynom für $n+1$ Stützstellen $\alpha_0, \ldots, \alpha_n$.

Im Satz sei $N_{n+1}(X)$ das $(n+1)$-te Newton-Polynom und I_t das kleinste Intervall, das die Stützstellen $\alpha_0, \ldots, \alpha_n$ und $t \in \mathbb{R}$ enthält.

Konsequenz. *Es gilt die Fehlerabschätzung*

$$|f(t) - P(t)| \leq \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!} \prod_{\nu=0}^{n} |t - \alpha_\nu|$$

wobei

$$\|\cdot\|_\infty\colon C[a,b] \to \mathbb{R}, \quad g \mapsto \max\{|g(t)| \mid t \in [a,b]\}$$

die Maximum-Norm *auf dem Vektorraum $C[a,b]$ der stetigen Funktionen $[a,b] \to \mathbb{R}$ sei.*

Example 4.1.9. *Sei $f(X) = \frac{1}{12}(-X^3 + 6X^2 - 17X + 36)$. Wir möchten diese Funktion in $(0,3),(1,2),(3,1)$ interpolieren. Das interpolierende Polynom ist*

$$P_2(X) = 3 - \frac{7}{6}X + \frac{1}{6}X^2$$

Der Interpolationsfehler ist in Abbildung 4.2 dargestellt. Er stimmt mit der Fehlerabschätzung überein, da f ein Polynom ist.

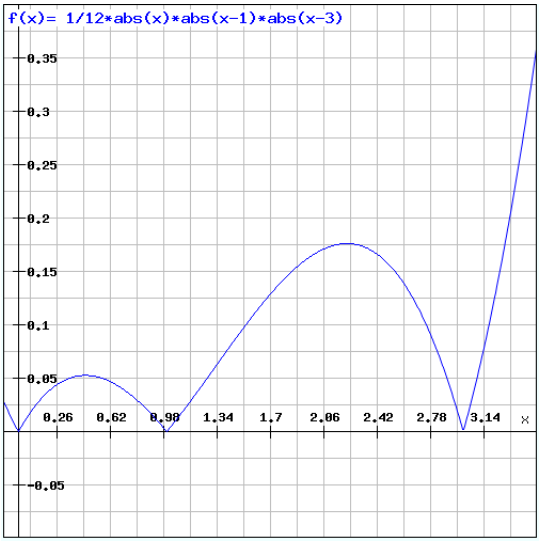


Abbildung 4.2: Der Interpolationsfehler stimmt in diesem Beispiel mit der Fehlerabschätzung überein.

4.1.5 Runges Phänomen

Da für Polynome $P \in \mathbb{R}[X]$ gilt:

$$\lim_{t \to \pm\infty} P(t) = \pm\infty$$

ist es zweckmäßig, nur Werte von Funktionen mit demselben Grenzverhalten zu interpolieren. Andernfalls treten in Randnähe, insbesondere bei äquidistanten Stützstellen, starke Oszillationen auf.

Runges Beispiel

Runge betrachtet

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

auf dem Intervall $[-5, 5]$. Siehe Abbildung 4.3.

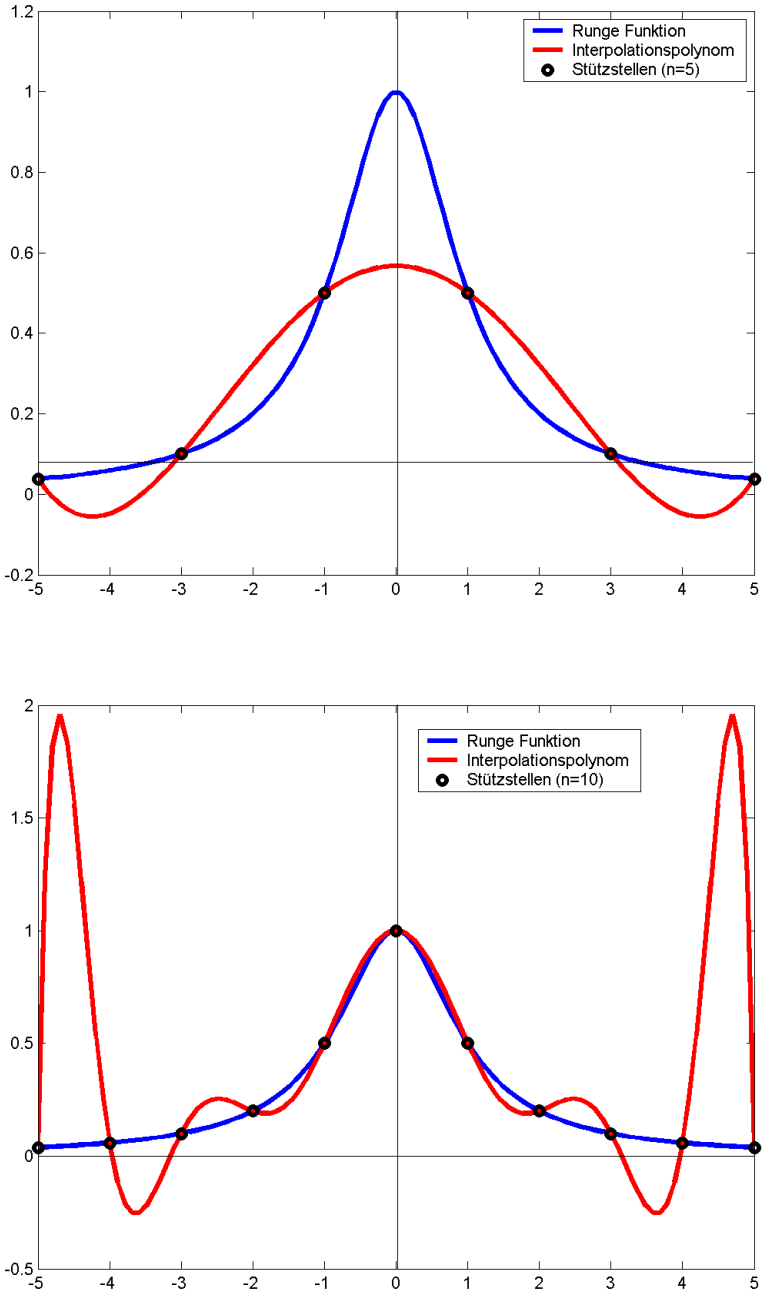


Abbildung 4.3: Interpolation von f mit 5 bzw. 10 äquidistanten Stützstellen (Quelle: Wikipedia, Autor: Mártin Pieper).

4.2 Spline-Interpolation

Eine *Straklatte* (engl. *spline*) ist eine elastische Latte (s. Abbildung 4.4). Sie wird im Schiffsbau verwendet für Linien ohne plötzliche Änderung des Krümmungsradius.

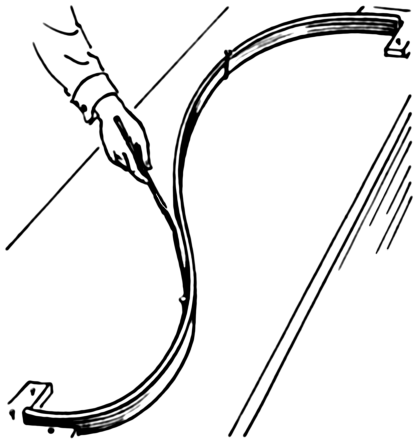


Abbildung 4.4: Eine Straklatte (Quelle: Wikipedia)

4.2.1 Streckenzug

Für reelle Stützstellen $x_0 < \cdots < x_n$ und Werte $f(x_0), \ldots, f(x_n)$ gibt es die *Knotenbasis* $\phi_i(x)$, bei der stückweise linear interpoliert wird gemäß Abbildung 4.5, sodass

$$\phi_i(x_j) = \delta_{ij}$$

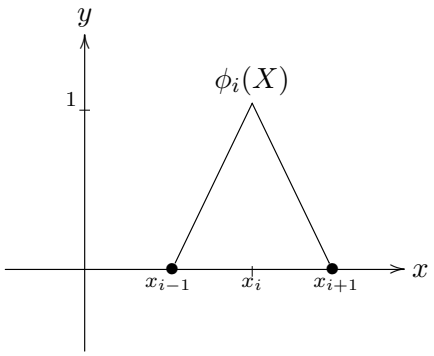


Abbildung 4.5: Eine Knotenbasisfunktion.

Die interpolierende Funktion ist

$$P(x) = \sum_{i=0}^{n} f(x_i)\phi_i(x)$$

4.2.2 Spline-Räume

Gegeben sei das Intervall $[a, b]$ mit Stützstellen

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

Dies führt zur *Zerlegung* von $[a, b]$

$$\mathcal{Z} := \{I_i = [x_{i-1}, x_i] \mid i = 1, \ldots, n\}$$

Die *Feinheit* der Zerlegung $\mathcal{Z}$ ist

$$h_{\mathcal{Z}} = \max\{x_i - x_{i-1} \mid i = 1, \ldots, n\}$$

Dies führt auf den *Spline-Raum* für $\mathbb{Z}$:

$$S_{\mathcal{Z}}^{(k,r)}[a,b] := \{P \in C^r[a,b] \colon P|_{I_i} \in \mathbb{R}[X]_k,\ i = 1, \ldots, n\}$$

wobei

$$C^r[a,b] = \{f \colon [a,b] \to \mathbb{R} \mid f \text{ ist } r\text{-fach stetig differenzierbar}\}$$

und $P|_{I_i}$ die Einschränkung von P auf I_i ist:

$$P|_{I_i} \colon I_i \to \mathbb{R}, \quad t \mapsto P(t)$$

Bemerkung 4.2.1. *Der Spline-Raum $S_{\mathcal{Z}}^{(k,r)}$ ist ein Vektorraum.*

Interpolationsfehler

Sei q_i ein interpolierendes Polynom auf I_i. Aus Abschnitt 4.1.4 erinnern wir uns, dass

$$|f(t) - q_i(t)| \leq \frac{\left\|f^{(r+1)}\right\|_\infty}{(r+1)!} \prod_{\nu=0}^{k} |t - \alpha_i|$$

gilt, wobei $\alpha_0 = x_{i-1}, \ldots, \alpha_k = x_i \in I_i = [x_{i-1}, x_i]$ weitere Stützstellen sind. Wegen

$$|t - \alpha_i| \leq |x_i - x_{i-1}| \leq h_{\mathcal{Z}}$$

folgt:

$$|f(t) - P(t)| \leq \frac{\left\|f^{(r+1)}\right\|_\infty}{(r+1)!} \cdot h_{\mathcal{Z}}^{k+1}$$

Wir haben also:

Satz 4.2.2. *Der Interpolationsfehler bei Interpolation von f mit $P \in S_{\mathcal{Z}}^{(k,r)}$ ist durch*

$$|f(t) - P(t)| \leq \frac{\left\|f^{(r+1)}\right\|_\infty}{(r+1)!} \cdot h_{\mathcal{Z}}^{k+1}$$

gegeben.

Beispiel 4.2.3. *Bei Interpolation mit einem Streckenzug haben wir*

$$P \in S_{\mathcal{Z}}^{(1,0)}$$

Also gilt für den Interpolationsfehler

$$|f(t) - P(t)| \leq \left\|f'\right\|_\infty \cdot h_{\mathcal{Z}}^2$$

4.2.3 Kubische Splines

Ein Spline $P \in S_{\mathcal{Z}}^{(3,2)}$ heißt *kubischer Spline*.

Satz 4.2.4. *Der interpolierende kubische Spline P existiert und ist durch die zusätzliche Vorgabe von $P''(a)$ und $P''(b)$ eindeutig bestimmt.*

Beweis. Existenz. Jedes Polynom $q_i(X) = P|_{I_i}$ hat 4 Koeffizienten. Das macht zusammen $4n$ Parameter. In diesen gibt es

- $2n$ lineare Gleichungen $q_i(x_i) = f(x_i)$, $q_{i+1}(x_i) = f(x_i)$
- $n-1$ lineare Gleichungen für P' stetig
- $n-1$ lineare Gleichungne für P'' stetig
- 2 lineare Zusatzgleichungen durch Vorgabe von $P''(a), P''(b)$

Also sind dies $4n$ lineare Gleichungen in $4n$ Unbekannten. D.h. im Falle der Eindeutigkeit ist das Gleichungssystem auch lösbar.

Eindeutigkeit. Seien $P, Q \in S_{\mathcal{Z}}^{(3,2)}[a,b]$ mit denselben zusätzlichen Vorgaben. Dann ist

$$s := P - Q \in \big\{ w \in C^2[a,b] \mid w(x_i) = 0,\ i = 0, \dots, n \big\} =: N \tag{4.3}$$

Genauer ist $s \in N \cap S_{\mathcal{Z}}^{(3,2)}[a,b]$. Für $w \in N$ beliebig gilt:

$$\int_a^b s''(x) w''(x)\, dx = 0 \tag{4.4}$$

denn:

$$\begin{aligned}
\int_a^b s''(x)w''(x)\,dx &= \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} s''(x)w''(x)\,dx = \sum_{i=0}^{n} s''(x)w'(x)\Big|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} s'''(x)w'(x)\,dx \\
&= \sum_{i=0}^{n} s''(x)w'(x)\Big|_{x_i}^{x_{i+1}} - \underbrace{s'''(x)w(x)\Big|_{x_i}^{x_{i+1}}}_{=0} + \int_{x_i}^{x_{i+1}} \underbrace{s''''(x)}_{=0} w(x)\,dx \\
&= \underbrace{s''(b)}_{=0} w'(b) - \underbrace{s''(a)}_{=0} w'(a) = 0
\end{aligned}$$

Mit $w = s$ folgt für die Krümmung

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

Also ist s linear. Da aber $s \in N$, folgt: $s = 0$. Also ist $P = Q$. □

Definition 4.2.5. *Der kubische Spline P mit $P''(a) = P''(b) = 0$ heißt* natürlich.

Bemerkung 4.2.6. *Der natürliche kubische Spline minimiert die Gesamtkrümmung*

$$\int_a^b \left|f''(x)\right|^2 dx$$

unter den interpolierenden Funktionen $f \in C^2[a,b]$.

Beweis. Sei $P \in S_{\mathfrak{Z}}^{(3,2)}$. Dann ist $w := f - P \in N$ (wie in (4.3) definiert), und es gilt:

$$\begin{aligned}\int_a^b \left|f''(x)\right|^2 dx &= \int_a^b \left|P''(x) + w''(x)\right|^2 dx \\ &= \int_a^b \left|P''(x)\right|^2 dx + 2\underbrace{\int_a^b P''(x)w''(x)\,dx}_{=0} + \underbrace{\int_a^b \left|w''(x)\right|^2 dx}_{\geq 0}\end{aligned}$$

(Das mittlere Integral verschwindet nach (4.4)). Also ist $\int_a^b |P''(x)|^2\,dx$ minimal. □

Berechnung der natürlichen kubischen Splines

Wir stellen die $4n$ linearen Gleichungen auf. Es sei

$$q_i(X) = a_0^{(i)} + a_1^{(i)}(X - x_i) + a_2^{(i)}(X - x_i)^2 + a_3^{(i)}(X - x_i)^3, \quad i = 1, \ldots, n$$

Die Bedingung $q_i(x_i) = f(x_i)$ führt auf:

$$a_0^{(i)} = f(x_i), \quad i = 1, \ldots, n \tag{4.5}$$

Mit $h_i := x_{i-1} - x_i$ wird

$$q_i(x_{i-1}) = a_0^{(i)} + a_1^{(i)}h_i + a_2^{(i)}h_i^2 + a_3^{(i)}h_i^3$$

Dann führt die Bedingung $q_i(x_{i-1}) = f(x_{i-1})$ auf:

$$f(x_{i-1}) - f(x_i) = a_1^{(i)}h_i + a_2^{(i)}h_i^2 + a_3^{(i)}h_i^3, \quad i = 1, \ldots, n \tag{4.6}$$

Die Bedingung $q_1''(x_0) = q_n''(x_n) = 0$ führt auf:

$$a_2^{(1)} + 3a_3^{(1)}h_1 = 0, \quad a_2^{(n)} = 0 \tag{4.7}$$

Die Bedingung $q_i'(x_i) = q_{i+1}'(x_i)$ führt auf:

$$a_1^{(i)} = a_1^{(i+1)} + 2a_2^{(i+1)}h_{i+1} + 3a_3^{(i+1)}h_{i+1}^2, \quad i = 1, \ldots, n-1 \tag{4.8}$$

Schließlich führt die Bedingung $q_i''(x_i) = q_{i+1}''(x_i)$ auf:

$$a_2^{(i)} = a_2^{(i+1)} + 3a_3^{(i+1)}h_{i+1}, \quad i = 1, \ldots, n-1 \tag{4.9}$$

(4.5) bis (4.9) vereinfacht sich zu

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1}) a_2^{(i)} + h_{i+1} a_2^{(i+1)} = 3 \left(\frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1})}{h_i} \right),$$

mit $i = 1, \ldots, n-1$. Dies ist ein lineares Gleichungssystem, dessen Koeffizientenmatrix folgende Geststalt hat:

$$A = \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & & & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & & & & \\ & h_3 & \ddots & \ddots & & & \\ & & \ddots & & & & \\ & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} & \\ 0 & & & & h_{n-1} & 2(h_{n-1} + h_n) & \end{pmatrix}$$

Diese ist regulär, wie wir aus der Eindeutigkeit wissen.

Kapitel 5

Numerische Lineare Algebra

5.1 Die Potenzmethode zur Bestimmung von Eigenvektoren am Beispiel von PageRank

Beim PageRank von Google lernen wir einerseits wichtige Eigenschaften stochastischer Matrizen kennen, andererseits lernen wir die Potenzmethode zur Berechnung von Eigenvektoren kennen.

Die Idee hinter dem PageRank ist, dass die Wichtigkeit einer Webseite von der Zahl der Seiten bestimmt ist, die auf diese verweisen und deren Wichtigkeiten.

Die Seite P_j habe ℓ_j Links auf andere Seiten. Gibt es einen Verweis auf die Seite P_i (wir schreiben dies hier als $P_j \to P_i$), so erhält P_i von P_j $\frac{1}{\ell_j}$ seiner Wichtigkeit. Der *Wichtigkeitsrang* von P_i ist die Summe aller Beiträge von Seiten, die auf P_i verweisen:

$$I(P_i) = \sum_{P_j \to P_i} \frac{I(P_j)}{\ell_j} \tag{5.1}$$

Um (5.1) besser zu verstehen, betrachten wir die *Hyperlinkmatrix* $H = (H_{ij})$ mit

$$H_{ij} = \begin{cases} \frac{1}{\ell_j}, & P_j \to P_i \\ 0 & \text{sonst} \end{cases}$$

Diese hat die Eigenschaften:

- Alle Einträge sind nicht negativ.
- Die Spaltensumme ist stets 1 oder 0.

Definition 5.1.1. *Eine* stochastische Matrix *ist eine quadratische Matrix mit nicht negativen reellen Einträgen, und deren Spaltensumme stets* 1 *ist.*

Mit dem Vektor $I = (I(P_i))$ schreibt sich (5.1) als:

$$I = H \cdot I$$

D.h. I ist *Eigenvektor* von H zum *Eigenwert* 1. So etwas nennt man einen *stationären Vektor* von H.

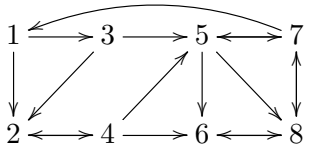


Abbildung 5.1: Ein Mini-Internet.

Beispiel 5.1.2. *Der Graph in Abbildung 5.1 hat Hyperlinkmatrix*

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{pmatrix}$$

Ein stationärer Vektor ist

$$I = \begin{pmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{pmatrix}$$

Der wichtigste Knoten ist also 8*.*

5.1.1 Die Potenzmethode

Zur Berechnung des stationären Vektors I der Hyperlinkmatrix H des Internets sei gesagt:

- H hat ca. 25 Mrd. Zeilen und Spalten.
- Die meisten Einträge von H sind Null.
- Im Durchschnitt gibt es ca. 10 Einträge pro Spalte.

Deshalb ist eine möglichst schnelle Methode zur Berechnung von I interessant. Diese ist die *Potenzmethode*:

- Startvektor $I^0 \neq 0$.
- $I^{k+1} := H \cdot I^k$

Das Prinzip lautet $I^k \to I$ für $k \to \infty$.

Im Beispiel 5.1.2 ist bereits $I^{60} = I$.

Es ergeben sich die Fragen:

- Konvergiert I^k stets?

- Ist der Limes-Vektor unabhängig vom Startvektor?
- Enthalten die Wichtigkeitsränge die gewünschte Information?

Die Antworten sind dreimal: *Nein.*

Der Ausweg ist eine Modifikation der Hyperlinkmatrix.

Beispiel 5.1.3. *Gegeben sei der Graph* $1 \to 2$*. Dessen Hyperlinkmatrix ist*

$$H = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

Für den Startvektor $I^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ *ist* $I^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ *und* $I^2 = I = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$*. Der tiefere Grund hierfür ist, dass Knoten* 2 *keine Links hat. Solch ein Knoten heißt* anhängender Knoten.

Interpretieren wir den PageRank $I(P_i)$ als Anteil der Zeit, die ein zufälliger Surfer auf einer Seite verbringt, so können wir verlangen, dass die Spaltensumme von I gleich Eins sein soll. Bei einem anhängenden Knoten soll der Surfer einfach mit gleicher Wahrscheinlichkeit auf irgendeine Seite springen.

Beispiel 5.1.4. *Gegeben sei wieder der Graph* $1 \to 2$*. Dann springt der zufällige Surfer gemäß Matrix*

$$S = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{pmatrix}$$

Ein stationärer Vektor hiervon ist $I = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}$*. Der Knoten* 2 *ist nun doppelt so wichtig wie Knoten* 1*, was auch der Intuition entspricht.*

Wir ersetzen die Hyperlinkmatrix H nun durch

$$S = H + A$$

wobei A zu jedem Knoten ohne Links die Spalte $\begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}$ und sonst Nullen hat.

Ein *Eigenwert* einer quadratischen Matrix S ist eine Zahl λ, für die es einen Vektor $v \neq 0$ gibt, sodass

$$S \cdot v = \lambda \cdot v \tag{5.2}$$

Der Vektor v heißt *Eigenvektor* von S zum Eigenwert λ. Die Menge aller Vektoren, die (5.2) erfüllen, bildet einen Vektorraum und heißt *Eigenraum* von S zum Eigenwert λ.

Für eine stochastische Matrix ist 1 der betragsgrößte Eigenwert.

Wir nehmen nun an, dass für die Eigenwerte λ_i der $n \times n$-Matrix S gilt:

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

Weiter nehmen wir an, dass es eine Basis v_j des $\mathbb{R}^n$ aus Eigenvektoren von S gibt. Dann ist

$$\begin{aligned} I^0 &= c_1v_1 + c_2v_2 + \cdots + c_nv_n \\ I^1 = SI^0 &= c_1v_1 + c_2\lambda_2v_2 + \cdots + c_n\lambda_nv_n \\ I^2 = SI^1 &= c_1v_1 + c_2\lambda_2^2v_2 + \cdots + c_n\lambda_n^2v_n \\ &\vdots \\ I^k = SI^{k-1} &= c_1v_1 + c_2\lambda_2^kv_2 + \cdots + c_n\lambda_n^kv_n \end{aligned}$$

Da $\lambda_j^k \to 0$ für $k \to \infty$ und $j \geq 2$, folgt:

$$I^k \to I = c_1v_1, \quad \text{ein stationärer Vektor} \quad (k \to \infty) \tag{5.3}$$

Ist $|\lambda_2|$ sehr klein, so ist die Konvergenz (5.3) sehr schnell.

Allerdings gilt nicht immer $1 = \lambda_1 > |\lambda_2|$.

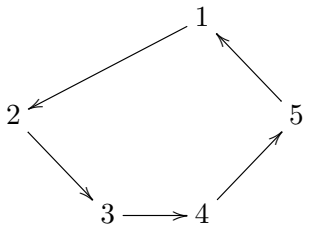


Abbildung 5.2: Ein zyklisches Internet.

Beispiel 5.1.5. *Im Graph in Abbildung 5.2 ist*

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Hier ist, mit $I^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $I^5 = I^0$, *d.h.* I^k *konvergiert nicht. Der Grund ist, dass* $|\lambda_2| = 1$ *ist.*

Definition 5.1.6. *Die Matrix* S *heißt* primitiv, *wenn für ein* m S^m *nur positive Einträge hat.*

Die Bedeutung von S primitiv ist, dass zu je zwei Seiten A, B es eine Folge von Links $A \to \cdots \to B$ gibt.

Definition 5.1.7. *Ein Graph, für den zu je zwei Knoten* A, B *ein gerichteter Weg* $A \to \cdots \to B$ *existiert, heißt* stark zusammenhängend.

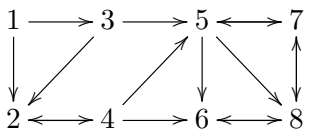


Abbildung 5.3: Ein Netz im Netz: in $\{5, 6, 7, 8\}$ kommt man hinein, aber nicht wieder heraus.

Beispiel 5.1.8. *Ein stationärer Vektor zu Abbildung 5.3 ist*

$$I = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.12 \\ 0.24 \\ 0.24 \\ 0.4 \end{pmatrix}$$

Hier sind die Nulleinträge problematisch. Der Grund dafür ist, dass ein Netz im Netz existiert: in die Knotenmenge $\{5, 6, 7, 8\}$ kommt man über Links hinein, aber nicht wieder heraus. Die zugehörige Matrix ist

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{2} & 0 \end{pmatrix}$$

Diese Matrix ist reduzibel.

Definition 5.1.9. *Eine quadratische Matrix S heißt* reduzibel*, wenn sie nach geeigneter Permutation von Zeilen und Spalten die Gestalt*

$$S = \begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$$

hat. Andernfalls heißt S irreduzibel*.*

Lemma 5.1.10. *Ist S irreduzibel, so existiert ein stationärer Vektor mit lauter positiven Einträgen.*

Ersetzt man in S alle von Null verschiedenen Einträge durch Eins, so erhält man die *Adjazenzmatrix* des Netzes.

Lemma 5.1.11. *Genau dann ist ein Graph stark zusammenhängend, wenn seine Adjazenzmatrix irreduzibel ist.*

Lemma 5.1.12. *Eine primitive Matrix ist irreduzibel.*

Der Satz von Perron-Frobenius liefert nun, was wir brauchen:

Satz 5.1.13 (Perron-Frobenius)**.** *Sei S eine primitive stochastische Matrix. Dann gilt:*

1. *1 ist Eigenwert von S von Multiplizität 1 (d.h. der Eigenraum zu 1 ist eindimensional).*
2. *1 ist betragsgrößter Eigenwert von S, alle anderen Eigenwerte haben kleineren Betrag.*
3. *Die Eigenvektoren zum Eigenwert 1 haben nur positive oder nur negative Einträge. Insbesondere existiert ein Eigenvektor zum Eigenwert 1, bei dem die Summe aller Einträge gleich Eins ist.*

Endgültige Modifikation

Die endgültige Modifikation geschieht mit einem Parameter $\alpha \in (0,1)$. Mit Wahrscheinlichkeit α folgt ein zufälliger Surfer der Matrix S und mit Wahrscheinlichkeit $1-\alpha$ geht er/sie auf irgendeine Seite.

Sei

$$\mathbb{1} := \begin{pmatrix} 1 & 1 & \dots \\ 1 & 1 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Dann ist die *Google-Matrix*

$$G = \alpha S + (1-\alpha)\frac{1}{n}\mathbb{1}$$

Zuweilen wird α auch *Dämpfungskonstante* genannt.

Bemerkung 5.1.14. *Die Google-Matrix G ist stochastisch und primitiv.*

Konsequenz. *G hat einen stationären Vektor mit positiven Einträgen, der mit der Potenzmethode gefunden werden kann.*

$\alpha = 1$ liefert das originale Netz, und bei $\alpha = 0$ haben wir den vollständigen Graph: jeder Knoten ist gleich wahrscheinlich. Für die Netzstruktur ist also α nahe 1 zu bevorzugen.

Bemerkung 5.1.15. *Für den zweiten Eigenwert λ_2 von G gilt:*

$$|\lambda_2| = \alpha$$

Dies bedeutet, dass α nicht zu nahe an 1 sein darf. Sergey Brin und Larry Page wählen $\alpha = 0.85$.

Anwendung auf das Internet

Die Google-Matrix hat die Struktur

$$G = \alpha H + \alpha A + \frac{1-\alpha}{n}\mathbb{1}$$

Bei der Potenzmethode ist

$$GI^k = \alpha H I^k + \alpha A I^k + \frac{1-\alpha}{n}\mathbb{1} I^k$$

Bei den drei Summanden ist zu bemerken, dass H dünn besetzt ist, und bei A und $\mathbb{1}$ alle Zeilen identisch sind. Zur Berechnung der letzten beiden Summanden, sind die Wichtigkeitsränge der anhängenden Knoten bzw. aller Knoten zu addieren. Dies muss nur einmal gemacht werden.

Angeblich benötigt es 50-100 Iterationen, um I hinreichend gut zu approximieren, was ein paar Tage benötigt. Gerüchteweise wird der PageRank I etwa jeden Monat aktualisiert.

5.1.2 Etwas Topologie

„Mmm... donuts."
H. Simpson

Inspiriert durch den Internetgraph des vorigen Abschnitts halten wir fest, dass die Graphentheorie als Teilgebiet der Topologie angesehen werden kann. Die Topologie beschäftigt sich mit Eigenschaften von Räumen, die unverändert bleiben unter stetigen Verformungen wie Dehnen und Biegen, aber nicht z.B. Auseindanderreißen. Ziel dieses Abschnittes ist es, Homers letzte Zeile auf der Tafel in *The Wizard of Evergreen Terrace* zu verstehen (s. Abschnitt 1.6). Diese Zeile ist in Abbildung 5.4 nachgebildet.

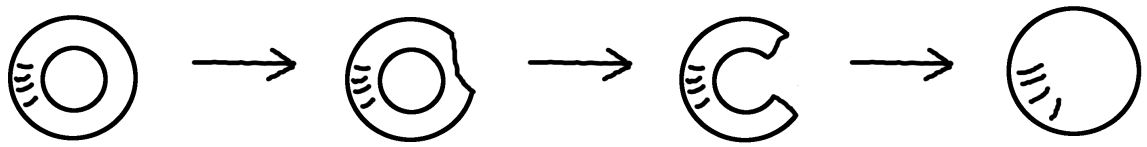


Abbildung 5.4: Anknabbern (gefolgt von einer topologischen Deformation) ist eine bei Homer Simpson erlaubte topologische Transformation, um einen Donut in eine Sphäre zu verwandeln.

Zunächst zurück zu den Graphen. Ein Graph besteht aus Knoten, von denen einige mit einigen anderen jeweils durch Kanten verbunden sind. Zwischen einem Knotenpaar gibt es höchstens eine Kante. Sei b_0 die Anzahl der Zusammenhangskomponenten und b_1 die Anzahl der „Löcher", dann gibt es eine Beziehung zu der Anzahl V an Knoten und der Anzahl E an Kanten:

Lemma 5.1.16. *Für endliche Graphen gilt:*

$$V - E = b_0 - b_1$$

Beweis. Sowohl die rechte als auch die linke Seite ändert sich nicht, wenn eine Kante kontrahiert wird: dabei wird eine Kante mit zwei verschiedenen Endknoten durch einen Knoten ersetzt. Führt man solange Kantenkontraktionen durch bis nur noch Kanten übrigbleiben, die jeweils einen einzigen Endknoten haben, so ist jede Zusammenhangskomponente ein einpunktiger Graph. Kanten sehen dabei aus wie „Blütenblätter". Jeder Kante entspricht genau ein Loch und umgekehrt. Daher ist dann $V = b_0$ und $E = b_1$, und die Gleichung stimmt. □

Die rechte Seite der Gleichung in Lemma 5.1.16 heißt *Euler-Charakteristik*, und b_i heißt *i-te Bettizahl*.

Bettizahlen gibt es auch in höherer Dimension: b_2 ist die Anzahl der Hohlkörper einer geschlossenen orientierbaren Fläche (oder auch in höherer Dimension). *Orientierbar* heißt dabei, dass die Fläche zwei verschiedene Seiten hat: ein „Innen" und ein „Außen". Die Bettizahlen sind topologische Invarianten, in dem Sinn, dass sie sich unter stetigen Verformungen nicht ändern. Man spricht dabei von einem *Homöomorphismus*, wenn es eine stetige Verformung zwischen den beiden Objekten gibt. So ist beispielsweise ein Quadrat homöomorph zu einem Kreis oder eine Kaffeetasse homöomorph zu einem Torus (für Homer Simpson: einem Donut).

Bei einer Sphäre ist die erste Bettizahl gleich Null, es gibt kein „Loch" oder „Tunnel". Bei einem Donut ist die erste Bettizahl gleich Zwei, denn einerseits gibt es das äussere Loch in der Mitte und andererseits den Tunnel im Innern. Bei einer Sphäre mit g Henkeln ist $b_1 = 2g$, denn jeder Henkel liefert einen Beitrag von 2: einmal ein äusseres Loch und einmal ein innerer Tunnel.

Für geschlossene orientierbare Flächen gibt es folgende Klassifikation:

Satz 5.1.17. *Jede geschlossene orientierbare Fläche gibt es ein $g \geq 0$, sodass sie homöomorph zu einer Sphäre mit g Henkeln ist.*

Die Zahl g heißt das *Geschlecht* der Fläche. Da auch das Geschlecht eine topologische Invariante ist, folgt:

Konsequenz. *Ein Donut und eine Sphäre sind nicht homöomorph.*

Beweis. Ein Donut hat Geschlecht 1, während eine Sphäre Geschlecht 0 hat. Da das Geschlecht sich unter einem Homöomorphismus nicht ändert, können daher Donut und Sphäre nicht homöomorph sein. □

Da Homer Simpson Donuts sehr liebt, erlaubt er nicht nur stetige Verformungen, sondern auch eine Transformation namens „Anknabbern“. Dadurch kann er einen Donut in eine Sphäre verwandeln. Es bietet sich dazu der Ausdruck „Homeromorphismus“ an.

Satz 5.1.18 (Homer Simpson, 1998)**.** *Donut und Sphäre sind „homeromorph“.*

Beweis. Transformiere den Donut durch Anknabbern solange, bis er homöomorph zum dritten Objekt von links in Abbildung 5.4 ist. Führe anschließend eine stetige Verformung durch. □

Es gibt auch nicht-orientierbare Flächen. In der Episode *Möbius Dick* der Zeichentrick-Sitcom *Futurama*[1] fliegt das Raumschiff *Planet Express* durch den *Bermuda-Tetraeder*. Ähnlich wie beim Bermuda-Dreieck auf der Erde finden sich dort haufenweise verschollene Raumschiffe. Dort erscheint ein vierdimensionaler Wal. Eine Anspielung auf diesen ist der Titel dieser Episode. Weiter spielt „Möbius“ auf das *Möbiusband* an, einer nicht-orientierbaren Fläche, die folgendermaßen konstruiert werden kann:

Bei einem (schmalen) Rechteck werden die beiden kurzen Seiten nach einem halben Twist identifiziert. Das Ergebnis kann in Abbildung 5.5 betrachtet werden.

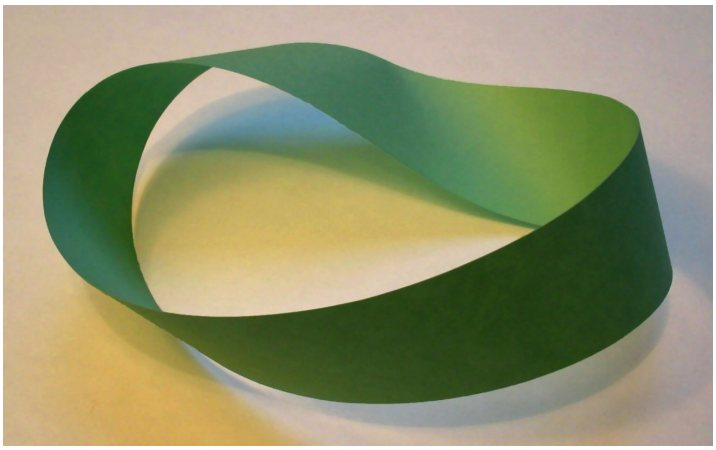


Abbildung 5.5: Ein Möbiusband aus Papier (Quelle: Wikipedia, Autor: David Benbennick).

Startet man im Innern der Fläche von einem Punkt aus und läuft parallel zum Rand, so erreicht man nach einem Umlauf den Ausgangspunkt, aber auf der anderen „Seite“. Nach einem zweiten Umlauf ist man wieder am Ausgangspunkt auf der „ursprünglichen“ Seite angekommen. Dies zeigt, dass die Fläche nur eine Seite hat: sie ist daher nicht orientierbar.

Was für eine Fläche ergibt sich, wenn man ein Möbiusband parallel zum Rand aufschlitzt?

[1] erschaffen von Matt Groening, dem Erfinder der *Simpsons*

5.1.3 Alexandroff-Topologien

Definition 5.1.19. *Sei $R \subset X \times X$ eine Relation. R heißt* reflexiv*, wenn für alle $x \in X$ gilt:*

$$(x, x) \in R$$

R heißt transitiv*, wenn für alle $x, y, z \in X$ gilt:*

$$(x, y) \in R \text{ und } (y, z) \in R \Rightarrow (x, z) \in R$$

Gerne schreiben wir auch

$$xRy$$

statt $(x, y) \in R$.

Pavel Alexandroff (1896–1982) entdeckte, dass eine reflexive und transitive Relation R auf einer Menge X eine Topologie definiert, bei der jeder Punkt $x \in X$ eine *minimale Umgebung*

$$U(x) = \{y \in X \mid (x, y) \in R\}$$

hat. Interessant ist der Fall einer *partiellen Ordnung* $\leq$ auf X. Dies ist eine reflexive und transitive Relation, die zusätzlich noch *antisymmetrisch* ist:

$$x \leq y \text{ und } y \leq x \Rightarrow x = y$$

für alle $x, y \in X$. Der zugehörige topologische Raum heißt T_0*-Raum.*

Definition 5.1.20. *Sei $R \subset X \times X$ eine Relation. Dann ist*

$$\begin{aligned} R^0 &= \{(x, x) \in X \times X\} \\ R^2 = R \circ R &= \{(x, z) \in X \times X \mid \text{es existiert } y \in X, \text{ sodass } (x, y) \in R \text{ und } (y, z) \in R\} \\ R^{n+1} &= R^n \circ R \end{aligned}$$

Die Relation

$$R^* = \bigcup_{n \in \mathbb{N}} R^n$$

heißt die reflexive und transitive Hülle *von R.*

Beispiel 5.1.21. *In der Geoinformatik gibt es die Relation* ist-berandet-von*, bei der z.B. ein Raum in einem Gebäude von einer Wand berandet ist, diese von einer Kante und diese wiederum von einer Ecke berandet ist. Die sogenannte* Inzidenztopologie *ist die reflexive und transitive Hülle von* ist-berandet-von*. Diese definiert einen T_0-Raum, dessen Punkte die Volumenkörper, Flächen, Liniensegmente und Punkte sind, die als Gebäudeteile aufgefasst werden.*

Beispiel 5.1.22. *Ein weiteres Beispiel ist die Relation*

$$\preceq = \textit{ ist-Voraussetzung-zur-Erlangung-von}$$

auf der Menge

$$M = \{\textit{Gebet, Erleuchtung des Verstandes, Aufmerksamkeit, Selbstbeobachtung, Selbsterkenntnis, Reue, Demut, Gnade Gottes}\}$$

Dabei meint Gnade*: Gabe bzw. unverdientes Geschenk. Es gilt:*

$$\begin{aligned} &\textit{Gebet} \preceq \textit{Erleuchtung des Verstandes} \preceq \textit{Aufmerksamkeit} \preceq \textit{Selbstbeobachtung} \\ &\quad \preceq \textit{Selbsterkenntnis} \preceq \textit{Reue} \preceq \textit{Demut} \preceq \textit{Gnade Gottes} \end{aligned}$$

Außerdem ist die Relation $\preceq$ eine partielle Ordnung. Dieser Spezialfall heißt totale Ordnung*. Dies hat zur Folge, dass wenn bei einem Menschen ein Element der Menge M fehlt, dann ruht auf diesem die Gnade Gottes nicht:*

> *Denn wenn jemand das ganze Gesetz hält und sündigt gegen ein einziges Gebot, der ist am ganzen Gesetz schuldig.* *Jak 2, 10*

Insbesondere gibt es ohne Demut keine Gnade Gottes! Es ist allerdings $\preceq$ nicht die einzig mögliche Topologie auf M. Diese ist eher als eine Anleitung zur Erlangung göttlicher Gnade zu lesen. Was allerdings auf jeden Fall gilt, ist

$$\begin{aligned} \textit{Demut} &\preceq \textit{Gnade Gottes} \\ \textit{Reue} &\preceq \textit{Gnade Gottes} \end{aligned}$$

Die restlichen Elemente von M dienen der Erlangung von Demut und Reue. Beispiele sind das Gleichnis vom Zöllner und dem Pharisäer, sowie die beiden mit Jesus gekreuzigten Räuber. Weiter schreibt der Heilige Johannes Klimakos in seiner Leiter göttlichen Aufstiegs und geistiger Vervollkommnung*:*

> *Aus einigen zogen sich die Leidenschaften zurück, nicht nur Gläubigen, sondern auch Ungläubigen, mit Ausnahme von einer [dem Stolz]. Diese als einzige blieb, um die Stelle aller anderen einzunehmen, da sie die erste unter den Übeln ist und derartiges Unheil anrichtet, dass sie sogar aus dem Himmel stürzen lässt.*
> *(26.1, 39)*

Denn erst wenn wir demütig sind, ähneln wir Gott:

> *Nehmet auf euch mein Joch und lernet von mir; denn ich bin sanftmütig und von Herzen demütig; so werdet ihr Ruhe finden für eure Seelen.* (*Mt 11, 29*)

5.2 Eine Gleichung in einer Variablen

Die Aufgabe in diesem Abschnitt lautet: Löse die Gleichung

$$a \cdot X = b$$

mit gegebenen a, b.

Beispiel 5.2.1. *Über $K = \mathbb{Q}$ oder $\mathbb{R}$ kann a^{-1} z.B. mit der Newton-Raphson-Division (s. Abschnitt 2.4.1) berechnet werden. Dann ist $X = a^{-1} \cdot b$.*

Beispiel 5.2.2. *Löse*

$$2 \cdot X \equiv 1 \mod 3$$

Lösung: $X \equiv 2 \mod 3$, *denn:*

$$2 \cdot 2 \equiv 4 \equiv 1 \mod 3$$

und für $X \equiv 0$ *oder* $1 \mod 3$ *gilt:*

$$2 \cdot X \not\equiv 1 \mod 3$$

Wir sind also durch Ausprobieren auf die Lösung gekommen. Geht dies auch effizienter?

Antwort. *Berechne*

$$1 = \mathrm{ggT}(2,3) = x \cdot 2 + y \cdot 3$$

Dann ist

$$x \cdot 2 \equiv 1 \mod 3$$

Z.B. $x = 2$, $y = -1$ *tun es.*

Satz 5.2.3. *Der größte gemeinsame Teiler* $d = \mathrm{ggT}(a,b)$ *von* $a, b \in \mathbb{Z}$ *hat eine lineare Darstellung*

$$d = x \cdot a + y \cdot b$$

mit $x, y \in \mathbb{Z}$.

Konsequenz. *Sind* a *und* n *teilerfremd, so ist die Kongruenz*

$$a \cdot X \equiv b \mod n$$

eindeutig lösbar.

Beweis des Satzes. Der euklidische Algorithmus liefert

$$\begin{aligned} a &= b \cdot q + r, \quad |r| < |b| \\ b &= r \cdot q_1 + r_1, \quad |r_1| < |r| \\ &\vdots \\ r_{n-2} &= r_{n-1} \cdot q_{n-2} + r_n, \quad |r_n| < |r_{n-1}| \end{aligned}$$

und $r_n = d$. Dies ergibt:

$$\begin{aligned} r &= a - b \cdot q \\ r_1 &= b - r \cdot q_1 = b - (a - b \cdot q) \cdot q_1 = b \cdot (1 + qq_1) - a \cdot q_1 \\ &\vdots \\ r_{n-1} &= r_{n-3} - r_{n-2} \cdot q_{n-3} \\ d &= r_{n-2} - r_{n-1} \cdot q_{n-2} \\ &= \underbrace{r_{n-2} - (r_{n-3} - r_{n-2} \cdot q_{n-3}) \cdot q_{n-2}}_{= r_{n-2} \cdot (\dots) + r_{n-3} \cdot q_{n-2}} = \cdots = a \cdot x + b \cdot y \end{aligned}$$

□

Die Methode im Beweis von Satz 5.2.3 heißt *erweiterter euklidischer Algorithmus.*

Konsequenz. *Ist p eine Primzahl, so ist jede Gleichung*

$$a \cdot X \equiv b \mod p$$

mit $a \not\equiv 0 \mod p$ eindeutig lösbar.

Insbesondere hat jedes $a \not\equiv 0 \mod p$ ein multiplikatives Inverses modulo p. Dies bedeutet, dass

$$\mathbb{F}_p := \{0, \ldots, p-1\}$$

ein *Körper* ist, wenn Addition und Multiplikation modulo p genommen wird. Hierin gibt es dann die *Euklid-Division*.

5.3 Gauß-Algorithmus

Die $n \times n$-Matrizen

$$K^{n\times n} := \{A = (a_{ij}) \mid a_{ij} \in K\}$$

mit Einträgen aus einem Körper K bilden mit der Addition und Matrixmultiplikation einen *Ring mit Eins*, der für $n \geq 2$ nichtkommutativ ist. Dies bedeutet, dass die üblichen Rechengesetze für $+$ und $\cdot$ gelten[2] mit den Ausnahmen, dass nicht jede von Null verschiedene $n \times n$-Matrix invertierbar ist, und dass in der Regel $A \cdot B \neq B \cdot A$ gilt. Z.B. für $n = 2$:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Elementarmatrizen

Sei

$$E_{ij} = (\epsilon_{pq}^{ij})$$

mit

$$\epsilon_{pq}^{ij} = \begin{cases} 1, & (p,q) = (i,j) \\ 0 & \text{sonst} \end{cases}$$

Diese Matrix heißt *Elementarmatrix* und hat eine 1 in der i-ten Zeile und j-ten Spalte, sonst 0. Es gilt:

$$E_{ij} \cdot E_{k\ell} = \begin{cases} E_{i\ell}, & j = k \\ 0 & \text{sonst} \end{cases}$$

Beweis. $E_{ij} \cdot E_{k\ell} = (\gamma_{rs})$ mit

$$\gamma_{rs} = \sum_t \epsilon_{rt}^{ij} \epsilon_{ts}^{k\ell} = \epsilon_{rj}^{ij} \epsilon_{js}^{k\ell} = \begin{cases} 1, & (r,s) = (i,\ell) \quad \text{und} \quad j = k \\ 0 & \text{sonst} \end{cases}$$

□

[2] Die Rolle der Null spielt die Nullmatrix, die Rolle der Eins die Einheitsmatrix.

Aus Elementarmatrizen kann man durch Linearkombination andere Matrizen bauen, z.B.:

$$I := \sum_{i=1}^{n} E_{ii} \qquad \text{(Einheitsmatrix)}$$

$$\text{Diag}(\alpha_1, \ldots, \alpha_n) := \sum_{i=1}^{n} \alpha_i E_{ii} \qquad \text{(Diagonalmatrix)}$$

$$M_i(\alpha) := I + (\alpha - 1) \cdot E_{ii} \qquad \text{(Multiplikationsmatrix)}$$

$$A_{ij}(\alpha) := I + \alpha E_{ij} \qquad (i \neq j) \qquad \text{(Additionsmatrix)}$$

$$V_{ij} := I - E_{ii} - E_{jj} + E_{ij} + E_{ji} \qquad \text{(Vertauschungsmatrix)}$$

Die Bedeutung der letzten drei Matrizen wird durch die nachfolgenden Beispiele in $K^{4\times 4}$ klar.

Beispiel 5.3.1.

$$M_1(\alpha) = \begin{pmatrix} \alpha & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} = I + (\alpha - 1)E_{11}$$

Multiplikation von links und von rechts an eine beliebige 4×4*-Matrix ergibt:*

$$M_1(\alpha) \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} = \begin{pmatrix} \alpha \cdot \alpha_{11} & \alpha \cdot \alpha_{12} & \alpha \cdot \alpha_{13} & \alpha \cdot \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot M_1(\alpha) = \begin{pmatrix} \alpha \cdot \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha \cdot \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha \cdot \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha \cdot \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

Beispiel 5.3.2.

$$A_{23}(\beta) = \begin{pmatrix} 1 & & & \\ & 1 & \beta & \\ & & 1 & \\ & & & 1 \end{pmatrix} = I + \beta \cdot E_{23}$$

$$A_{23} \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} + \beta \cdot \alpha_{31} & \alpha_{22} + \beta \cdot \alpha_{32} & \alpha_{23} + \beta \cdot \alpha_{33} & \alpha_{24} + \beta \cdot \alpha_{34} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot A_{23}(\beta) = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} + \beta \cdot \alpha_{12} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} + \beta \cdot \alpha_{22} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} + \beta \cdot \alpha_{32} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} + \beta \cdot \alpha_{42} & \alpha_{44} \end{pmatrix}$$

Beispiel 5.3.3.

$$V_{24} = \begin{pmatrix} 1 & & & \\ & & & 1 \\ & & 1 & \\ & 1 & & \end{pmatrix} = I - E_{22} - E_{44} + E_{24} + E_{42}$$

$$V_{24} \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot V_{24} = \begin{pmatrix} \alpha_{11} & \alpha_{14} & \alpha_{13} & \alpha_{12} \\ \alpha_{21} & \alpha_{24} & \alpha_{23} & \alpha_{22} \\ \alpha_{31} & \alpha_{34} & \alpha_{33} & \alpha_{32} \\ \alpha_{41} & \alpha_{44} & \alpha_{43} & \alpha_{42} \end{pmatrix}$$

Lemma 5.3.4. *Die Matrix $M_i(\alpha)$ ist für $\alpha \neq 0$ invertierbar. $A_{ij}(\alpha)$ ist für alle $\alpha \in K$ invertierbar. V_{ij} ist invertierbar. Diese Matrizen führe folgende Operationen auf eine gegebene der Größe nach passende Matrix A aus:*

- *$M_i(\alpha)$ von links: multipliziert die i-te Zeile von A mit α.*
- *$A_{ij}(\alpha)$ von links: addiert das α-Fache der j-ten Zeile von A auf die i-te Zeile von A.*
- *V_{ij} von links: vertauscht i-te und j-te Zeile von A.*

- *$M_i(\alpha)$ von rechts: multipliziert die i-te Spalte von A mit α.*
- *$A_{ij}(\alpha)$ von rechts: addiert das α-Fache der i-ten Spalte auf die j-te Spalte von A.*
- *V_{ij} von rechts: vertauscht die i-te und die j-te Spalte von A.*

Die Inversen sind jeweils:

$$M_i(\alpha)^{-1} = M_i(\alpha^{-1})$$
$$A_{ij}(\alpha)^{-1} = A_{ij}(-\alpha)$$
$$V_{ij}^{-1} = V_{ij}$$

Als Konsequenz ergibt sich für das Lösen eines linearen Gleichungssystems

$$A \cdot x = b$$

mit $A \in K^{m \times n}$ und $b \in K^m$: der Gauß-Algorithmus (mit nur Zeilenoperationen): Multiplikation von links mit einer invertierbaren Matrix $B \in K^{m \times m}$ mit Ergebnis die Treppennormalform

$$T = \begin{pmatrix} 1 & * & & & * & * & \\ & & 1 & & * & * & \\ & & & 1 & * & * & \\ & & & & & & 1 \end{pmatrix}$$

Die Treppennormalform liefert folgendermaßen eine Basis des homogenen Lösungsraums

$$Ax = 0$$

Einfügen von Zeilen mit genau einem -1-Eintrag unterhalb einer Nichtstufe liefert die erweiterte Treppe

$$\tilde{T} = \begin{pmatrix} 1 & * & & & * & * \\ & -1 & & & & \\ & & 1 & & * & * \\ & & & 1 & * & * \\ & & & & -1 & \\ & & & & & -1 \\ & & & & & & 1 \end{pmatrix}$$

Die Spalten mit den neuen -1-en sind eine Basis des homogenen linearen Gleichungssystems:

$$\mathbb{L} = \left\langle \begin{pmatrix} * \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} * \\ 0 \\ * \\ * \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} * \\ 0 \\ * \\ * \\ 0 \\ -1 \\ 0 \end{pmatrix} \right\rangle$$

Dies führt zu unserer ersten Zerlegung von A:

$$A = C \cdot T$$

mit C invertierbar und T in Treppennormalform.

5.4 LU-Zerlegung

Sei $A \in K^{n \times n}$.

Definition 5.4.1. *Eine LU-Zerlegung von A ist eine Faktorisierung der Form*

$$A = L \cdot U,$$

wobei L eine untere Dreiecksmatrix, dessen Diagonalelemente gleich Eins sind, und U eine obere Dreiecksmatrix sei.

Doolittle-Algorithmus

Der *Doolittle-Algorithmus* kann unter Umständen eine LU-Zerlegung von A erreichen.

Sei $A_0 := A$. Für $\nu = 1, \ldots, n$ sei $A_\nu := L_\nu \cdot A_{\nu-1} = (\alpha_{ij}^{(\nu)})$ mit

$$L_\nu = A_{n,\nu}(\ell_{n,\nu}) \cdot A_{n-1,\nu}(\ell_{n-1,\nu}) \cdots A_{\nu+1,\nu}(\ell_{\nu+1,\nu}) = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{\nu+1,\nu} & \ddots & \\ & & \vdots & & \ddots \\ & & \ell_{n,\nu} & & & 1 \end{pmatrix}$$

und

$$\ell_{i,\nu} := -\frac{\alpha_{i\nu}^{(\nu-1)}}{\alpha_{\nu\nu}^{(\nu-1)}}, \quad (i = \nu + 1, \ldots, n)$$

Beachte, dass L_ν in der ν-ten Spalte von $A_{\nu-1}$ alles unter der Diagonale eliminiert. Insbesondere ist

$$U := A_{n-1} = L_{n-1} \cdots L_1 \cdot A$$

eine obere Dreiecksmatrix. Dann ist

$$L := L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & 0 \\ -\ell_{2,1} & \ddots & & & \\ & & 1 & & \\ \vdots & & -\ell_{\nu+1,\nu} & \ddots & \\ & & \vdots & & 1 \\ -\ell_{n,1} & & -\ell_{n,\nu} & -\ell_{n,n-1} & 1 \end{pmatrix}$$

eine untere Dreiecksmatrix mit Diagonal-Einsen und

$$A = L \cdot U$$

Allerdings wird in jedem Schritt benötigt, dass $\alpha_{\nu\nu}^{(\nu-1)} \neq 0$ ist. Falls dies nicht der Fall ist, so vertausche die ν-te Zeile mit einer Zeile unterhalb. Dann bekommt man am Ende eine *PLU-Zerlegung*

$$A = P \cdot L \cdot U$$

mit einer *Permutationsmatrix* $P = (\pi_{ij})$, wobei

$$\pi_{ij} = \delta_{i,\sigma(j)}$$

für eine Permutation σ der Zahlen $1, \ldots, n$.

Anwendungen der *LU*-Zerlegung

1. Lösung von $Ax = b$ mit $A \in K^{n\times n}$, $b \in K^n$. Die Zerlegung $A = L \cdot U$ führt auf

$$L\underbrace{(Ux)}_{=y} = b \quad \rightsquigarrow \quad Ux = y$$

   Dies sind zwei lineare Gleichungssysteme. Das erste wird durch *Vorwärtssubstitution*, und das zweite durch *Rückwärtssubstitution* gelöst (s. nachfolgende Paragraphen).

2. Inverse von $A \in K^{n\times n}$. Wende dazu 1. simultan auf jede Spalte von I als rechte Seite b an.

3. Determinante. Es gilt:

$$\det(A) = \det(L) \cdot \det(U) = \det(U)$$

Vorwärtssubstitution

Gegeben sei das dreieckige lineare Gleichungssystem

$$\begin{aligned} \ell_{11}y_1 &= b_1 \\ \ell_{21}y_1 + \ell_{22}y_2 &= b_2 \\ &\vdots \\ \ell_{n1}y_1 + \cdots + \ell_{nn}y_n &= b_n \end{aligned}$$

Dann ist

$$\begin{aligned} y_1 &= \frac{b_1}{\ell_{11}} \\ y_i &= \frac{1}{\ell_{ii}} \left(b_i - \sum_{k=1}^{i-1} \ell_{ik} y_k \right), \qquad i = 2, \ldots, n \end{aligned}$$

die Lösung, falls alle $\ell_{ii} \neq 0$ sind.

Rückwärtssubstitution

Gegeben sei das dreieckige lineare Gleichungssystem

$$\begin{aligned} u_{nn} x_n &= y_n \\ u_{n-1,n-1} x_{n-1} + u_{n-1,n} x_n &= y_{n-1} \\ &\vdots \\ u_{11} x_1 + \cdots + u_{1n} x_n &= y_1 \end{aligned}$$

Dann ist

$$\begin{aligned} x_n &= \frac{y_n}{u_{nn}} \\ x_i &= \frac{1}{u_{ii}} \left(y_i - \sum_{k=i+1}^{n} u_{ik} x_k \right), \qquad i = n-1, \ldots, 1 \end{aligned}$$

die Lösung, falls alle $u_{ii} \neq 0$ sind.

5.4.1 Pfannkuchen sortieren

Permutationen wie aus dem vorigen Abschnitt kommen in *The Simpsons* versteckt vor, und zwar in der Gestalt des Pfannkuchenhauses: In Springfield, Homer Simpsons Heimatstadt, gibt es ein städtisches Pfannkuchenhaus. Dieses ist in der Episode *Marge und das Brezelbacken* (1997) zu sehen. Nehmen wir an, dass dort der Kellner n Pfannkuchen in zufälliger Reihenfolge serviert. Er kann Pfannkuchen auf dem Servierteller wenden, indem er von oben ein paar nimmt und diesen Stapel wendet. Die Frage ist nun, wie oft er sie im schlimmsten Fall auf diese Weise wenden muss, bis sie der Größe nach sortiert sind. Die Anzahl, wie oft gewendet wird, heißt *Pfannkuchenzahl* und sei mit P_n bezeichnet. Gesucht ist eine Formel, die P_n beschreibt.

Informatiker sortieren gerne Daten, und da gibt es Parallelen zu Pfannkuchen. Außerdem ist P_n nur bis $n = 19$ bekannt. Deshalb ist das Pfannkuchen-Sortierproblem von Interesse.

Die Pfannkuchenzahl für die ersten paar Werte von n kann man sich erarbeiten, indem man alle Kombinationen verschieden großer Pfannkuchen vorgibt und für jede von ihnen die Wendezahl bestimmt.

$P_1 = 0$, da der einzige Pfannkuchen schon in der richtigen Reihenfolge ist.

$P_2 = 1$, da schlimmstenfalls der große auf dem kleinen Pfannkuchen liegt und dann einmal gewendet werden muss.

P_3 zu bestimmen ist schon etwas schwieriger. Es gibt die 6 Möglichkeiten $(1,2,3)$, $(1,3,2)$, $(2,3,1)$, $(2,1,3)$, $(3,1,2)$, $(3,2,1)$, wobei die Zahl der Größe und die Position der Zahl von

links nach rechts der Position des Pfannkuchens von oben nach unten entspricht. Die Anzahl an Wendemanövern ist dann in der folgenden Tabelle gegeben:

Permutation	$(1,2,3)$	$(1,3,2)$	$(2,3,1)$	$(2,1,3)$	$(3,1,2)$	$(3,2,1)$
Anzahl Wendemanöver	0	3	3	2	2	1

Somit ergibt sich $P_3 = 3$.

Im Jahr 1979 wurde eine obere Schranke für P_n gefunden. Nämlich:

Satz 5.4.2 (William H. Gates & Christo H. Papadimitriou, 1979)**.** *Es gilt:*

$$P_n \leq \frac{5n+5}{3}$$

William H. Gates ist besser als Bill Gates bekannt und ist Mitbegründer der Firma Microsoft.

David S. Cohen, einer der Autoren von *The Simpsons*, veröffentlichte im Jahr 1995 einen Artikel über das *Verbrannte-Pfannkuchen-Problem*, bei dem es darum geht, Pfannkuchen, die auf einer Seite verbrannt sind, so zu wenden, dass sie der Größe nach sortiert sind und die verbrannte Seite jeweils unten ist. Die Anzahl der Wendemanöver bei diesem Problem sei mit V_n bezeichnet. Es gilt:

Satz 5.4.3 (David S. Cohen, 1995)**.** *Es gilt:*

$$\frac{3n}{2} \leq V_n \leq 2n-2$$

5.5 Der Spektralsatz

5.5.1 Eigenräume

Sei $A \in K^{n\times n}$.

Definition 5.5.1. *Der* Eigenraum $E_\lambda(A)$ *zum Eigenwert* λ *ist der Lösungsraum von*

$$A - \lambda \cdot I$$

falls dieser $\neq 0$ *ist.*

Beispiel 5.5.2. *Sei*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Dann ist $E_1(A) = K \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *und* $E_{-1} = K \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Lemma 5.5.3. *Genau dann ist* $\lambda \in K$ *Eigenwert von A, wenn* λ *Nullstelle von*

$$f_A(X) = \det(X \cdot I - A)$$

dem charakteristischen Polynom *ist.*

Beweis. Genau dann ist $A - \lambda \cdot I$ nicht invertierbar, wenn der zugehörige Lösungsraum nicht trivial ist. □

Definition 5.5.4. *Ein* Eigenvektor *von A ist ein nichttriviales Element aus einem Eigenraum von A.*

5.5.2 Basiswechsel

Definition 5.5.5. *Eine Matrix $A \in K^{n\times n}$ heißt* diagonalisierbar*, wenn es eine invertierbare Matrix S gibt, sodass*

$$S^{-1}AS$$

eine Diagonalmatrix ist.

Beispiel 5.5.6. *Sei*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Eine Basis aus Eigenvektoren ist durch folgende Matrix gegeben:

$$S = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Es ist

$$S^{-1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}$$

Nach Basiswechsel ist

$$S^{-1}AS = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Definition 5.5.7. *Die Matrix $S^{-1}AS$ heißt* ähnlich *zu A.*

Bemerkung 5.5.8. *Ähnliche Matrizen haben dieselben Eigenwerte.*

Satz 5.5.9 (Spektralsatz)**.** *Sei $A \in K^{n\times n}$ eine diagonalisierbare Matrix. Dann hat K^n eine Basis S aus Eigenvektoren von A. Die Diagonalmatrix $S^{-1}AS$ hat die Eigenwerte von A als Diagonaleinträge.*

Beweis. Sei

$$S^{-1}AS = \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$$

Dann gilt:

$$AS = S \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$$

also für die i-te Spalte s_i von S:

$$As_i = \lambda_i s_i$$

und die Basis S von K^n besteht aus Eigenvektoren s_i von A zum Eigenwert λ_i. □

Definition 5.5.10. *Die Menge*

$$\operatorname{Spec}(A) = \{\lambda \mid \lambda \textit{ ist Eigenwert von } A\}$$

ist das Spektrum *von A.*

5.5.3 Determinante und Spur

Sei $A = (a_{ij}) \in K^{n\times n}$. Dann ist die *Determinante* definiert als

$$\det(A) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} a_{i,\pi(i)}$$

wobei S_n die Menge aller Permutationen der Zahlen $1,\dots,n$ sei und $\operatorname{sgn}(\pi)$ das *Signum* der Permutation π sei:

$$\operatorname{sgn}(\pi) = \begin{cases} 1, & \pi \text{ ist gerade Permutation} \\ -1, & \pi \text{ ist ungerade Permutation} \end{cases}$$

Beispiel 5.5.11. *Für $n = 1$ ist $A = (a)$ und $\det(A) = a$.*

Für $n = 2$ ist

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

und

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

Da die identische Permutation (1) *eine gerade und und die Vertauschung* (12) *eine ungerade Permutation ist.*

Für $n = 3$ ist

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

und

$$\det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

da die identische Permutation (1) *und die Zyklen* (123) *und* (132) *gerade sind, und die Vertauschungen* (12), (13), (23) *ungerade sind.*

Definition 5.5.12. *Die Spur* $\operatorname{trace}(A)$ *ist die Summe der Diagonaleinträge von A.*

Beispiel 5.5.13. *Sei* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. *Dann ist*

$$\operatorname{trace} A = a + d, \quad \det(A) = ad - bc$$

Betrachten wir das charakteristische Polynom:

$$f_A(X) = \det(X \cdot I - A) = X^2 - (a+d)X + ad - bc = X^2 - \operatorname{trace}(A)X + \det(A)$$

Falls $f_A(X) = (X - \lambda_1)(X - \lambda_2)$ mit $\lambda_1, \lambda_2 \in \mathbb{C}$ ist, so gilt:

$$f_A(X) = X^2 - (\lambda_1 + \lambda_2)X + \lambda_1\lambda_2$$

Koeffizientenvergleich ergibt:

$$\begin{aligned} \operatorname{trace}(A) &= \lambda_1 + \lambda_2 \\ \det(A) &= \lambda_1 \cdot \lambda_2 \end{aligned}$$

Lemma 5.5.14. *Sei $A \in \mathbb{C}^{n\times n}$. Dann gilt:*

$$\operatorname{trace}(A) = \sum_{\lambda\in\operatorname{Spec}(A)} \lambda$$
$$\det(A) = \prod_{\lambda\in\operatorname{Spec}(A)} \lambda$$

Beweis. Es ist

$$\det(X\cdot I - A) = \prod_{\lambda\in\operatorname{Spec}(A)} (X-\lambda)$$

Auf der rechten Seite ist der konstante Term gleich

$$(-1)^n \cdot \prod_{\lambda\in\operatorname{Spec}(A)} \lambda$$

Auf der linke Seite ist der konstante Term gleich

$$\det(-A) = (-1)^n \cdot \det(A)$$

Auf der rechten Seite ist der Koeffizient von X^{n-1} gleich

$$-\sum_{\lambda\in\operatorname{Spec}(A)} \lambda$$

Auf der linken Seite ist der Koeffizient von X^{n-1} gleich

$$-\operatorname{trace}(A)$$

Mit Koeffizientenvergleich folgt die Behauptung. □

5.5.4 Das Futurama-Theorem

Zunächst ein paar Vorbemerkungen zu Permutationen. Jede Permutation σ der Zahlen $1,\dots,n$ lässt sich als Produkt disjunkter Zykeln zerlegen. Dies geht so: Starte mit 1, dann folgt $\sigma(1)$, dann $\sigma^2(1) := \sigma(\sigma(1))$, usw. bis irgendwann erstmalig $\sigma^k(1) = 1$. Dies ist der erste Zyklus. Falls in diesem Zyklus eine Zahl in $\{1,\dots,n\}$ nicht auftaucht, fahre wie eben fort mit einer solchen Zahl. Usw. Irgendwann kommt jede der Zahlen $1,\dots,n$ in genau einem Zyklus vor. Dies ergibt die Zerlegung in disjunkte Zykeln.

In der Episode *The Prisoner of Benda* aus *Futurama* kommt eine von Professor Farnsworth erfundene Bewusstseinstauschmaschine vor. Wird sie auf zwei Personen A und B angewandt, so ist hinterher das Bewusstsein von A im Körper von B und umgekehrt. In einer Gruppe von n Personen[3] haben sich alle mit mehreren verschiedenen Partnern dem Bewusstseinstausch unterzogen und möchten nach einer Weile wieder in ihren ursprünglichen Körper zurück. Das Problem aber ist, dass der Bewusstseinstausch mit demselben Körperpaar nur ein einziges Mal funktioniert.

Ken Keeler, der Hauptautor dieser Episode hatte nun die Aufgabe, herauszufinden wie alle Bewusstseine wieder in ihre ursprünglichen Körper zurückkommen. Nach einiger Mühe bewies er schließlich das *Futurama-Theorem*:

[3] In der Episode sind es 8 Personen

Satz 5.5.15 (Ken Keeler, 2010)**.** *Es genügt, zwei weitere Personen hinzuzufügen, damit jedes Bewusstsein wieder in seinen eigenen Körper kommt.*

Der folgende Beweis wird in der Episode an der Tafel entwickelt:

Beweis. π sei die Permutation von $[n] := \{1, \dots, n\}$, welche durch die Folge der Bewusstseinsvertauschungen gebildet wird.

1. *Fall.* Sei zunächst der Fall angenommen, dass π ein Zyklus der Länge k ist. Ohne Einschränkung kann angenommen werden, dass

$$\pi = \begin{pmatrix} 1 & 2 & \dots & k & k+1 & \dots & n \\ 2 & 3 & \dots & 1 & k+1 & \dots & n \end{pmatrix}$$

Oben stehen die Urbilder (Bewusstsein) und unten die Bilder (Körper) der Permutation. Sei

$$\pi^* := \begin{pmatrix} 1 & 2 & \dots & k & k+1 & \dots & n & x & y \\ 2 & 3 & \dots & 1 & k+1 & \dots & n & x & y \end{pmatrix} \quad \text{mit} \quad x, y \notin [n]$$

und sei mit $(a\ b)$ die Transposition gemeint, die a und b vertauscht.

$$\sigma := (x\ 1) \circ (y\ 2) \cdots (y\ k) \circ (x\ 2) \circ (y\ 1)$$

Dann ist

$$\pi^* \circ \sigma = \begin{pmatrix} 1 & 2 & \dots & n & x & y \\ 1 & 2 & \dots & n & y & x \end{pmatrix} \qquad (*)$$

Also müssen nur noch x und y vertauscht werden, was auch möglich ist, da noch nicht geschehen.

2. Fall. Sei π eine beliebige Permutation von $[n]$. Zerlege π in ein Produkt disjunkter Zykeln und wende den ersten Fall bis $(*)$ auf jeden Zyklus an. Anschließend vertausche x mit y, falls dies nötig sein sollte. □

5.5.5 Positiv definite Matrizen

Definition 5.5.16. *Eine Matrix $A \in \mathbb{C}^{n \times n}$ heißt* hermitesch*, wenn gilt:*

$$A^* := \bar{A}^\top = A$$

wobei A^* die komplex-konjugierte der Transponierten von A ist.

Ein Speziallfall ist bei reellen Matrizen gegeben: Eine reelle Matrix A ist genau dann hermitesch, wenn sie symmetrisch ist:

$$A^\top = A$$

Rechenregeln

Es gilt:

$$(A^*)^* = A$$

und

$$(AB)^* = B^* A^*$$
$$(A^*)^{-1} = \left(A^{-1}\right)^*$$

wann immer die Ausdrücke definiert sind.

Beweis. Sei $A = (\alpha_{ij})$, $B = (\beta_{ij})$, $AB = (\gamma_{ij})$ und $A^* = (\alpha'_{ij})$, $B^* = (\beta'_{ij})$, $(AB)^* = (\gamma'_{ij})$. Weiter sei $(A^*)^* = (\gamma'_{ij})$. Dann gilt:

$$\gamma'_{ij} = \overline{\alpha'}_{ji} = \bar{\bar{\alpha}}_{ij} = \alpha_{ij}$$

also: $(A^*)^* = A$. Weiter gilt:

$$\gamma'_{ij} = \bar{\gamma}_{ji} = \sum_k \bar{\alpha}_{jk}\bar{\beta}_{ki} = \sum_k \bar{\beta}_{ki}\bar{\alpha}_{jk} = \sum_k \beta'_{ik}\alpha'_{kj}$$

also: $(AB)^* = B^*A^*$. Weiter ist

$$\left(A^{-1}\right)^* A^* = \left(AA^{-1}\right)^* = I^* = I$$

also $(A^*)^{-1} = \left(A^{-1}\right)^*$. □

Definition 5.5.17. *Eine hermitesche Matrix $A \in \mathbb{C}^{n\times n}$ heißt* positiv (semi-)definit*, wenn für alle $z \in \mathbb{C}^n$ gilt:*

$$z^*Az > 0 \qquad (z^*Az \geq 0)$$

falls $z \neq 0$ ist.

Bemerkung 5.5.18. *Ist A hermitesch, dann ist z^*Az reell.*

Beweis. Es ist

$$\overline{z^*Az} = (z^*Az)^* = z^*A^*z^{**} = z^*Az$$

□

Bemerkung 5.5.19. *Genau dann ist eine symmetrische reelle Matrix $A \in \mathbb{R}^{n\times n}$ positiv (semi-)definit, wenn für alle $x \in \mathbb{R}^n$ gilt:*

$$x^\top Ax > 0 \qquad (x^\top Ax \geq 0)$$

falls $x \neq 0$ ist.

Beispiel 5.5.20. *Die Einheitsmatrix I ist positiv definit. Denn mit $z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{C}^n$ ist*

$$z^*Iz = z^*z = \sum_{i=1}^n \bar{z}_i z_i = \sum_{i=1}^n |z_i|^2 > 0$$

falls $z \neq 0$ ist.

Beispiel 5.5.21. *Die symmetrische reelle Matrix*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

ist positiv definit. Denn mit $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ ist

$$\begin{aligned} x^\top Ax &= x^\top \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{pmatrix} = 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 > 0 \end{aligned}$$

falls $x \neq 0$ ist.

Beispiel 5.5.22. $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ *ist nicht positiv definit. Denn mit* $z = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ *ist*

$$z^\top Az = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -2 < 0$$

aber $z \neq 0$.

Lemma 5.5.23. *Die Eigenwerte hermitescher Matrizen sind reell.*

Beweis. Sei e Eigenvektor der hermiteschen Matrix A zum Eigenwert $\lambda \in \mathbb{C}$. Dann gilt:

$$\mathbb{R} \ni e^*Ae = e^*(\lambda e) = \lambda \cdot \underbrace{e^*e}_{=\gamma}$$

und $\gamma > 0$ ist reell. Dann ist auch λ reell. □

Satz 5.5.24 (Spektralsatz II)**.** *Eine hermitesche Matrix* $A \in \mathbb{C}^{n\times n}$ *ist diagonalisierbar und hat ausschließlich reelle Eigenwerte. Weiter hat* $\mathbb{C}^n$ *eine Orthonormalbasis aus Eigenvektoren von* A*. Ist* A *zudem reell, so hat* $\mathbb{R}^n$ *eine Orthonormalbasis aus Eigenvektoren von* A*.*

Positiv (semi-)definit kann anhand der Eigenwerte abgelesen werden:

Bemerkung 5.5.25. *Sei* $A \in \mathbb{C}^{n\times n}$ *hermitesch. Dann gilt:* A *ist genau dann positiv (semi-)definit, wenn alle Eigenwerte von* A *positiv (nicht negativ) sind.*

Beweis. $\Rightarrow$. Sei e Eigenvektor von A zum Eigenwert $\lambda \in \mathbb{R}$. Dann ist

$$e^*Ae = e^*\lambda e = \lambda \cdot e^*e = \lambda$$

Der Ausdruck links ist positiv (nicht negativ).

$\Leftarrow$. Sei $z \in \mathbb{C}^n \setminus \{0\}$ und sei $\{e_i\}$ eine Orthonormalbasis von $\mathbb{C}^n$ aus Eigenvektoren von A, und sei λ_i der Eigenwert zu e_i. Dann ist $z = \sum_i \alpha_i e_i$ und

$$z^\top Az = \sum_i \bar{\alpha}_i e_i^\top \sum_j \alpha_j Ae_j = \sum_{i,j} \bar{\alpha}_i \alpha_j \lambda_j \underbrace{e_i^\top e_j}_{=\delta_{ij}} = \sum_i |\alpha_i|^2 \lambda_i$$

Der Ausdruck rechts ist positiv (nicht negativ). □

Aus einer nicht notwendig quadratischen Matrix kann man eine hermitesch Matrix konstruieren:

Lemma 5.5.26. *Sei* $A \in \mathbb{C}^{m\times n}$*. Dann ist* A^*A *hermitesch und positiv semidefinit.*

Proof. Es ist

$$(A^*A)^* = A^*A^{**} = A^*A$$

also A^*A hermitesch. Weiter ist für $x \in \mathbb{C}^n$:

$$x^*A^*Ax = (Ax)^*Ax \geq 0$$

also A^*A positiv semidefinit. □

5.6 Hauptkomponentenanalyse (PCA)

An n „Versuchspersonen“ werden p Merkmale gemessen. Dies ergibt n Punkte in $\mathbb{R}^p$ als Zufallsvektor $X = (X_1, \ldots, X_n)$. Das Ziel der Hauptkomponentenanalyse ist es, diese Datenpunkte so in einen q-dimensionalen Unterraum ($q < p$) zu projizieren, dass dabei möglichst wenig Information verloren geht und Redundanz (d.h. Korrelation) zusammengefasst wird.

Die Idee dabei ist, einen Basiswechsel so vorzunehmen, dass die neuen Variablen dekorreliert sind. Dann ist die Kovarianzmatrix diagonal. Der Nutzen dabei ist, dass, im Falle normalverteilter Daten, die neuen Variablen statistisch unabhängig sind.

Die Kovarianzmatrix

$$\mathrm{Cov}(X) = \mathbb{E}\left((X-\mu)(X-\mu)^\top\right) = (\mathrm{Cov}(X_i, X_j)) \in \mathbb{R}^{n\times n} \qquad (\mu = \mathbb{E}(X))$$

ist symmetrisch, also nach dem Spektralsatz 5.5.24 diagonalisierbar. Sie ist sogar positiv semidefinit. Denn wegen

$$\mathrm{Cov}(S^\top X) = S^\top \mathrm{Cov}(X) S = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$$

(S sei eine Orthonormalbasis des $\mathbb{R}^n$ aus Eigenvektoren von $\mathrm{Cov}(X)$) ist die Diagonalisierung selbst eine Kovarianzmatrix. Deren Diagonaleinträge sind also Varianzen, d.h. nicht negativ. Nach Bemerkung 5.5.25 ist also $\mathrm{Cov}(X)$ positiv semidefinit.

Die Spalten von

$$Y := S^\top X$$

heißen die *Hauptkomponenten* von X. Es gilt:

$$\mathrm{Var}(Y_i) = \lambda_i$$

Die Methode geht nun wie folgt: Ordne S so an, dass die Eigenwerte λ_i der Größe nach absteigend sortiert werden. Wähle q mit $\lambda_1 \geq \ldots \lambda_q$, sodass der Quotient

$$\tau_q := \frac{\sum_{i=1}^{q} \mathrm{Var}(Y_i)}{\sum_{i=1}^{n} \mathrm{Var}(X_i)}$$

groß ist. Dieser Ausdruck ist zwischen 0 und 1. Beachte, dass

$$\sum_{i=1}^{n} \mathrm{Var}(X_i) = \mathrm{trace}(\mathrm{Cov}(X)) = \mathrm{trace}(\mathrm{Cov}(Y)) = \sum_{i=1}^{n} \mathrm{Var}(Y_i)$$

die Gesamtstreuung oder totale Varianz von X ist. Diese ist die Summe aller Eigenwerte. Die Dimension q wird über die größten Eigenwerte bestimmt.

Die Hauptkomponenten liefern auch die beste lineare Approximation an X: Die erste Hauptkomponente ist die Gerade H_1 durch das Zentrum $\mu = \mathbb{E}(X)$ mit dem kleinsten Fehler. Die zweite Hauptkomponente ist die Gerade H_2 durch μ, die orthogonal zu H_1 ist, sodass die Ebene, die von H_1, H_2 aufgespannt wird, den kleinsten Fehler hat, usw. Schließlich bekommen wir Hauptkomponenten $H_1, \ldots, H_q$ zu den Eigenwerten $\lambda_1 \geq \ldots \lambda_q$.

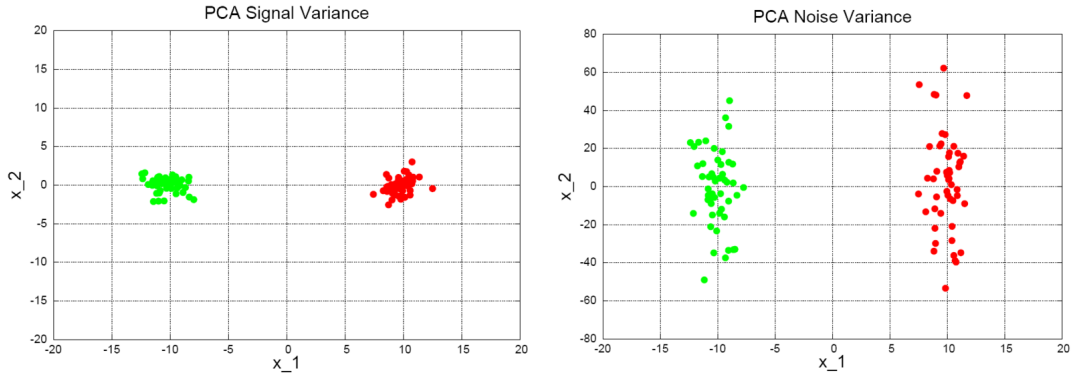


Abbildung 5.6: 2 Cluster mit Signal Variance (links) und Noise Variance (rechts). Quelle: Wikipedia, Author: Rene Andrae.

Grundannahme der PCA

Es wird bei der Hauptkomponentenanalyse angenommen, dass die Richtungen mit der größten Streuung (= Varianz) die meiste Information enthalten.

Diese Annahme ist jedoch nicht immer erfüllt. Dies soll am Beispiel der Clusteranalyse verdeutlicht werden.

In Abbildung 5.6 (links) ist die Streuung innerhalb der beiden Cluster gering im Vergleich zum Abstand der Cluster. Daher ist die erste Hauptkomponente die x_1-Achse. Diese reicht aus, um die Cluster zu trennen. Die Hauptkomponente x_2 kann vernachlässigt werden. Die totale Varianz wird hier vom Signal dominiert, wir haben zwei getrennte Cluster.

In Abbildung 5.6 (rechts) trägt die Streuung innerhalb der Cluster den Hauptanteil an der Gesamtstreuung. Es wird angenommen, dass die Streuung durch Rauschen verursacht wird. Daher heißt dieses Beispiel „noise variance“. Die erste Hauptkomponente ist x_2. Diese hat keine Information über die Trennbarkeit der Cluster.

Als Fazit lässt sich sagen, dass oft, aber nicht immer die dominierenden Hauptkomponenten die meiste *für ein bestimmtes Problem relevante* Information tragen.

5.7 Cholesky-Zerlegung

Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit.

Definition 5.7.1. *Eine* Cholesky-Zerlegung *von A ist eine Faktorisierung*

$$A = G \cdot G^\top$$

wobei G eine untere Dreiecksmatrix mit positiven Einträgen ist.

Eine Cholesky-Zerlegung von A lässt sich wie folgt berechnen: Mit $A = (\alpha_{ij})$ und $G = (\gamma_{ij})$ ist

$$\alpha_{ij} = \sum_{k=1}^{j} \gamma_{ik}\gamma_{jk}, \qquad i \geq j$$

Dies ergibt:

$$\gamma_{ij} = \begin{cases} 0, & i < j \\ \sqrt{\alpha_{ii} - \sum_{k=1}^{i-1} \gamma_{ik}^2}, & i = j \\ \frac{1}{\gamma_{jj}} \left(\alpha_{ij} - \sum_{k=1}^{j-1} \gamma_{ik}\gamma_{jk} \right), & i > j \end{cases}$$

Beispiel 5.7.2.

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} \gamma_{11}^2 & \gamma_{11}\gamma_{21} \\ \gamma_{11}\gamma_{21} & \gamma_{21}^2 + \gamma_{22}^2 \end{pmatrix}$$

ergibt:

$$\begin{aligned} \gamma_{11} &= \sqrt{a} \\ \gamma_{22} &= \sqrt{c - \gamma_{21}^2} \\ \gamma_{21} &= \frac{b}{\gamma_{11}} \end{aligned}$$

Beispiel 5.7.3.

$$\begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} = \begin{pmatrix} \gamma_{11}^2 & \gamma_{11}\gamma_{21} & \gamma_{11}\gamma_{31} \\ \gamma_{11}\gamma_{21} & \gamma_{21}^2 + \gamma_{22}^2 & \gamma_{21}\gamma_{31} + \gamma_{22}\gamma_{32} \\ \gamma_{11}\gamma_{31} & \gamma_{21}\gamma_{31} + \gamma_{22}\gamma_{32} & \gamma_{31}^2 + \gamma_{32}^2 + \gamma_{33}^2 \end{pmatrix}$$

ergibt:

$$\begin{aligned} &\gamma_{11} = \sqrt{a}, \qquad \gamma_{22} = \sqrt{d - \gamma_{21}^2}, \qquad \gamma_{33} = \sqrt{f - (\gamma_{31}^2 + \gamma_{32}^2)} \\ &\gamma_{21} = \frac{b}{\gamma_{11}}, \qquad \gamma_{31} = \frac{c}{\gamma_{11}} \\ &\gamma_{32} = \frac{e - \gamma_{21}\gamma_{31}}{\gamma_{22}} \end{aligned}$$

Als Anwendung lösen wir wieder ein lineares Gleichungssystem

$$Ax = b$$

mit A symmetrisch, positiv definit. Mit der Cholesky-Zerlegung $A = G \cdot G^\top$ ist

$$G(\underbrace{G^\top x}_{=:y}) = b \qquad \text{und} \qquad G^\top x = y$$

Die erste Gleichung

$$Gy = b$$

ist durch Vorwärtssubstitution und die zweite

$$G^\top x = y$$

durch Rückwärtssubstitution zu lösen (s. Abschnitt 5.4).

5.8 Gauß-Newton-Verfahren

Seien m Funktionen $r = (r_1, \ldots, r_m)$ in n Variablen $X = (X_1, \ldots, X_n)$ mit $m \geq n$ gegeben. Es soll die Größe

$$S(X) = \sum_{i=1}^{m} r_i(X)^2$$

minimiert werden.

Das Gauß-Newton-Verfahren verläuft iterativ. Startwert ist $X = x_0 \in \mathbb{R}^n$. Die Iteration läuft über ein Inkrement:

$$x_{s+1} = x_s + \epsilon$$

mit $\epsilon^\top \epsilon$ klein. Um das Inkrement zu bestimmen, ziehen wir die Taylor-Entwicklung heran:

$$S(x_s + \epsilon) \approx S(x_s) + \left[\frac{\partial S}{\partial X_i}\right]^\top \epsilon + \frac{1}{2}\epsilon^\top \left[\frac{\partial^2 S}{\partial X_i \partial X_j}\right] \epsilon$$

mit

$$\left[\frac{\partial S}{\partial X_i}\right] = 2J_r(X)^\top r$$

wobei

$$J_r(X) = \left[\frac{\partial r_i}{\partial X_j}\right]$$

die Jacobi-Matrix ist, und die Hesse-Matrix approximiert wird:

$$\left[\frac{\partial^2 S}{\partial X_i \partial X_j}\right] \approx 2J_r(X)^\top J_r(X)$$

für $r^\top r$ klein. Dies ergibt

$$S(x_s + \epsilon) \approx S(x_s) + 2r^\top J_r(X)\epsilon + \epsilon^\top J_r(X)^\top J_r(X)\epsilon$$

Dann ist zu minimieren:

$$\frac{\partial S}{\partial \epsilon}(x_s + \epsilon) \approx 2J_r(X)^\top r + 2J_r(X)^\top J_r(X)\epsilon \stackrel{!}{=} 0$$

was auf die *Normalgleichungen* führt:

$$J_r^\top(X) J_r(X) \cdot \epsilon = -J_r(X)^\top r \tag{5.4}$$

Der Hintergrund ist der Folgende:

- In der Datenmodellierung ist $X = \beta$ ein Parametervektor, für den eine Modellfunktion

  $$y = f(x, \beta)$$

  auf Daten (x_i, y_i) gefittet wird.

- Die Funtkionen

  $$r_i(\beta) = y_i - f(x_i, \beta)$$

  sind die *Residuen*.

- Das Inkrement löst die Normalgleichungen (5.4) mit $X = \beta$.

- In der Regel ist die Cholesky-Zerlegung anwendbar.

Vergleich mit dem Newton-Verfahren

Beim Newton-Verfahren wird die Hesse-Matrix $H(S)$ statt seiner Approximation mit dem doppelten Quadrat der Jacobi-Matrix verwendet. Das Inkrement ist dann:

$$\epsilon = -H(S)^{-1}\nabla S$$

Die Konvergenz des Gauß-Newton-Verfahrens ist also höchstens quadratisch.

5.9 Lisa und Baseball

Als in *The Lisa Series* (2010) Barts Baseballteam *die Isotots* den Trainer verliert, ergreift sie die Chance und wird deren Trainerin. Allerdings hat sie keine Ahnung von Baseball. Aber sie trifft zufällig Professor Frink, der die Meinung vertritt, dass Baseball nur durch tiefgreifende mathematische Analyse verstanden werden kann und ihr einen Stapel Bücher gibt, die sie durcharbeiten soll.

Eines dieser Bücher ist *The Bill James Historical Baseball Abstract*, eine Sammlung der wichtigsten Baseball-Statistiken aus der realen Welt, von Bill James zusammengestellt. Aufgrund ihres Studiums der Bücher holt sie die Isotots aus dem Tabellenkeller bis auf den zweiten Platz. Doch als sie Bart in einem Spiel anweist, keinen Schlag zu machen, missachtet er ihre Anweisung und gewinnt mit einem Homerun das Spiel. Daraufhin wirft sie Bart aus dem Team, weil „er glaubt, er sei besser als die Gesetze der Wahrscheinlichkeit." Die Isotots setzen auch ohne Bart ihre Siegesserie fort. Im Finale der *Little League State Championship* fällt jedoch ein Spieler aus. Daher bittet sie Bart, für ihn einzuspringen. Er zögert, da er weiß, dass er vor einem Dilemma steht: Statistik oder Instinkt. Im letzten Inning widersetzt sich Bart Lisas Anweisung erneut. Doch diesmal geht er out, und die Isotots verlieren das Spiel.

5.10 Innenprodukträume

Sei V ein K-Vektorraum mit $K = \mathbb{R}$ oder $\mathbb{C}$.

Definition 5.10.1. *Eine Abbildung*

$$\langle \cdot, \cdot \rangle \colon V \times V \to K$$

mit

1. $\langle x, y \rangle = \overline{\langle y, x \rangle}$ *(konjugiert-symmetrisch)*
2. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
   $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ *(links linear)*
3. $\langle x, x \rangle \geq 0 \quad mit = 0 \; nur \; für \; x = 0$ *(positiv definit)*

für $x, y, z \in V$, $\alpha \in K$ heißt Innenprodukt. *Das Paar $(V, \langle \cdot, \cdot \rangle)$ heißt* Innenproduktraum.

Bemerkung 5.10.2. *1. $\langle x, x \rangle$ ist stets reell.*

2. Es gilt:

$$\langle x, \alpha y \rangle = \bar{\alpha} \langle x, y \rangle$$
$$\langle x, y + z \rangle = \langle x, z \rangle + \langle y, z \rangle$$

Ein Innenprodukt ist also sesquilinear.

3. Für $K = \mathbb{R}$ ist ein Innenprodukt symmetrisch und bilinear.

Beispiel 5.10.3. *Sei $V = \mathbb{R}^n$. Dann ist*

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i = y^\top x$$

das Standardinnenprodukt.

Beispiel 5.10.4. *Sei $V = \mathbb{C}^n$. Dann ist*

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$$

das Standardinnenprodukt

Beispiel 5.10.5. *Sei $V = C[a, b] = \{f\colon [a, b] \to \mathbb{K} \mid f \text{ stetig}\}$. Dann ist ein Innenprodukt gegeben durch*

$$\langle f, g\rangle = \int_a^b f(t)\overline{g(t)}\, dt$$

Die ersten beiden Axiome des Innenprodukts ergeben sich aus den Rechenregeln für Integrale. Zur Positiv-Definitheit: Ist $f \neq 0$, so ist $f(x_0) \neq 0$ für ein $x_0 \in [a, b]$. Dann aber existiert eine ϵ-Umgebung U von x_0, sodass

$$f(x) \neq 0$$

für alle $x \in U$ gilt. Dann ist

$$\langle f, f\rangle = \int_a^b |f(t)|^2\, dt \geq \int_U |f(t)|^2\, dt > 0$$

Sei $(V, \langle\cdot,\cdot\rangle)$ ein Innenproduktraum.

Definition 5.10.6. *Die Funktion*

$$\|\cdot\|\colon V \to \mathbb{R}, \quad x \mapsto \sqrt{\langle x, x\rangle}$$

heißt Norm *auf V.*

Eigenschaften einer Innenprodukt-Norm

1. **Cauchy-Schwarz-Ungleichung.**

   (5.5) $$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$$

   Beweis. Ist $y = 0$, so gilt die Ungleichung. Ist $y \neq 0$, so setze $\lambda = \frac{\langle x,y\rangle}{\langle y,y\rangle}$. Dann ist:

   $$\begin{aligned} 0 \leq \langle x - \lambda y, x - \lambda y\rangle &= \langle x, x\rangle - \lambda\langle y, x\rangle - \bar{\lambda}\langle y, x\rangle + |\lambda|^2\langle y, y\rangle \\ &= \langle x, x\rangle - \frac{\langle x, y\rangle\langle y, x\rangle}{\langle y, y\rangle} - \frac{\langle y, x\rangle\langle x, y\rangle}{\langle y, y\rangle} + \frac{|\langle x, y\rangle|^2}{\langle y, y\rangle} \\ &= \langle x, x\rangle - \frac{|\langle x, y\rangle|^2}{\langle y, y\rangle} \end{aligned}$$

   Dies ergibt

   $$|\langle x, y\rangle|^2 \leq \langle x, x\rangle\langle y, y\rangle$$

   woraus die Behauptung folgt. □

2. Wegen der Cauchy-Schwarz-Ungleichung (5.5) kann der *Winkel* zwischen zwei Vektoren $x, y \in V \setminus \{0\}$ definiert werden:

$$w(x,y) := \arccos \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}$$

Wir sagen, x und y sind *orthogonal* ($x \perp y$), wenn der Winkel gleich π ist:

$$x \perp y \quad :\Leftrightarrow \quad w(x,y) = \pi \quad \Leftrightarrow \quad \langle x, y\rangle = 0$$

3. **Homogenität.** Für $\alpha \in K$, $x \in V$ gilt:

$$\|\alpha \cdot x\| = |\alpha| \|x\|$$

4. **Dreiecksungleichung.** Für $x, y \in V$ gilt:

$$\|x + y\| \leq \|x\| + \|y\|$$

Beweis. Es ist

$$\begin{aligned}
\|x+y\|^2 = \langle x+y, x+y\rangle &= \langle x, x\rangle + \underbrace{\langle x, y\rangle + \langle y, x\rangle}_{=2\Re(\langle x,y\rangle) \leq 2|\langle x,y\rangle|} + \langle y, y\rangle \\
&\leq \|x\|^2 + 2|\langle x, y\rangle| + \|y\|^2 \\
&\overset{(*)}{\leq} \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 \\
&= (\|x\| + \|y\|)^2
\end{aligned}$$

wobei in $(*)$ die Cauchy-Schwarz-Ungleichung (5.5) verwendet wurde. □

5. **Satz von Pythagoras.** Sind $x_1, \ldots, x_n$ paarweise orthogonal, so gilt:

$$\sum_{i=1}^{n} \|x_i\|^2 = \left\| \sum_{i=1}^{n} x_i \right\|^2 \tag{5.6}$$

6. **Parallelogramm-Identität.** Es gilt für $x, y \in V$:

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

5.11 QR-Zerlegung

Sei $(V, \langle \cdot, \cdot \rangle)$ ein Innenproduktraum. Wir betrachten den Projektionsoperator für $u \in V$.

$$\pi_u \colon V \to V, \quad x \mapsto \frac{\langle x, u\rangle}{\langle u, u\rangle} u$$

Bemerkung 5.11.1. *Es ist*

$$x - \pi_u(x) \perp u$$

Beweis. Es ist

$$\langle x - \pi_u(x), u\rangle = \langle x, u\rangle - \frac{\langle x, u\rangle}{\langle u, u\rangle} \langle u, u\rangle = 0$$

□

Nun können wir linear unabhängige Vektoren $b_1, \ldots, b_n \in V$ nach Gram-Schmidt orthogonalisieren:

$$\begin{aligned} u_1 &= b_1, & e_1 &= \frac{u_1}{\|u_1\|} \\ u_2 &= b_2 - \pi_{u_1}(b_2), & e_2 &= \frac{u_2}{\|u_2\|} \\ &\vdots \\ u_n &= b_n - \sum_{i=1}^{n-1} \pi_{u_i}(b_i), & e_n &= \frac{u_n}{\|u_n\|} \end{aligned}$$

Bemerkung 5.11.2. *Die $u_1, \ldots, u_n$ spannen denselben Untervektorraum auf wie $b_1, \ldots, b_n$ und sind paarweise orthogonal.*

Beweis. Die u_i sind orthogonal. Für $n = 1$ ist nichts zu zeigen. Sei $n > 1$. Seien per Induktionsvoraussetzung $u_1, \ldots, u_{n-1}$ orthogonal. Dann gilt für $j < n$:

$$\begin{aligned} \langle u_n, u_j \rangle &= \left\langle b_n - \sum_{i=1}^{n-1} \frac{\langle b_n, u_i \rangle}{\langle u_i, u_i \rangle} u_i, u_j \right\rangle \\ &= \langle b_n, u_j \rangle - \sum_{i=1}^{n-1} \frac{\langle b_n, u_i \rangle}{\langle u_i, u_i \rangle} \underbrace{\langle u_i, u_j \rangle}_{= \langle u_i, u_i \rangle \delta_{ij}} \\ &= \langle b_n, u_j \rangle - \langle b_n, u_j \rangle = 0 \end{aligned}$$

Der Raum, den die u_i aufspannen. Die $u_1, \ldots, u_n$ sind Linearkombinationen der $b_1, \ldots, b_n$ und orthogonal, also linear unabhängig: Sei

$$x = \sum_j \alpha_j u_j = 0$$

Dann ist

$$\langle x, u_i \rangle = \alpha_i \underbrace{\langle u_i, u_i \rangle}_{\neq 0} = 0$$

Also alle $\alpha_i = 0$. Da es nun gleich viele u_i wie b_i gibt, spannen die u_i denselben Raum auf wie die b_i. □

Es ist

$$\langle e_i, b_j \rangle e_i = \frac{\langle u_i, b_j \rangle}{\|u_i\|} \frac{u_i}{\|u_i\|} = \frac{\langle u_i, b_j \rangle}{\langle u_i, u_i \rangle} u_i = \pi_{u_i}(b_j)$$

Außerdem ist

$$\langle b_k, u_k \rangle = \left\langle u_k + \sum_{i=1}^{k-1} \pi_{u_i}(b_i) \right\rangle = \langle u_k, u_k \rangle + \sum_{i=1}^{k-1} \underbrace{\langle \pi_{u_i}(b_i), u_k \rangle}_{=0} = \langle u_k, u_k \rangle$$

also:

$$\pi_{u_k}(b_k) = \frac{\langle b_k, u_k \rangle}{\langle u_k, u_k \rangle} u_k = \frac{\langle u_k, u_k \rangle}{\langle u_k, u_k \rangle} u_k = u_k$$

Deshalb ist

$$b_k = u_k + \sum_{i=1}^{k-1} \pi_{u_i}(b_i) = \sum_{i=1}^{k} \pi_{u_i}(b_i) = \sum_{i=1}^{k} \langle e_i, b_k \rangle e_i$$

In Matrixform ist dies mit $B = (b_1 \mid \cdots \mid b_n)$:

$$B = Q \cdot R \tag{5.7}$$

mit

$$Q = (e_1 \mid \cdots \mid e_n)$$

und

$$R = \begin{pmatrix} \langle e_1, b_1\rangle & \langle e_1, b_2\rangle & \langle e_1, b_3\rangle & \dots \\ 0 & \langle e_2, b_2\rangle & \langle e_2, b_3\rangle & \dots \\ 0 & 0 & \langle e_3, b_3\rangle & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = Q^* B$$

Definition 5.11.3. *(5.7) ist die QR-Zerlegung von B. Die Spalten von Q sind orthonormal, und R ist obere Dreiecksmatrix.*

Anwendung

Löse ein lineares Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m\times n}$, $m \geq n$, wobei A vollen Rang habe. Mit der QR-Zerlegung von A geht dies folgendermaßen:

$$A = QR$$

mit $Q \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^{n\times n}$. Löse nun

$$Q \cdot \underbrace{Rx}_{=y} = b$$

so:

$$y = Q^\top b$$

und

$$Rx = y$$

durch Rückwärtssubstitution (s. Abschnitt 5.4).

5.12 Eigenwertbestimmung mit der QR-Zerlegung

Sei $A \in \mathbb{C}^{n\times n}$ nicht singulär und alle Eigenwerte seien von unterschiedlichem Betrag. Dann konvergiert die Folge

$$\begin{aligned} A_k &= Q_k R_k && (QR\text{-Zerlegung}) \\ A_{k+1} &:= R_k Q_k = Q_{k+1} R_{k+1} && (QR\text{-Zerlegung}) \end{aligned}$$

gegen eine obere Dreiecksmatrix A_∞.

Es gilt:

1. Wegen
$$A_{k+1} = R_k Q_k = Q_k^* Q_k R_k Q_k = Q_k^* A_k Q_k$$
haben alle A_k dieselben Eigenwerte.

2. Die Eigenwerte von A_k sind die Diagonaleinträge, denn das charakteristische Polynom ist

$$\det \begin{pmatrix} X - a_{11}^{(\infty)} & & * \\ & \ddots & \\ 0 & & X - a_{nn}^{(\infty)} \end{pmatrix} = (X - a_{11}^{(\infty)}) \cdots (X - a_{nn}^{(\infty)})$$

3. Ist A symmetrisch, so sind die Spalten von $Q = Q_1 Q_2 \cdots$ die Eigenvektoren von A, denn

$$A \cdot Q_1 Q_2 \cdots Q_n = Q_1 R_1 Q_1 \cdots Q_n = Q_1 Q_2 R_2 Q_2 \cdots Q_n = \cdots = Q_1 \cdots Q_n A_n$$

und da A symmetrisch ist, ist A_∞ eine Diagolnalmatrix $\mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, und

$$AQ = QA_\infty = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)Q$$

Für die Eigenschaft der Matrix Q in der QR-Zerlegung gibt es auch einen Namen:

Definition 5.12.1. *Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt* orthogonal*, wenn die Spalten von A eine Orthonormalbasis von $\mathbb{R}^n$ bilden.*

Bemerkung 5.12.2. *Nach dem Spektralsatz II (Satz 5.5.24) ist eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ diagonalisierbar mit einer orthogonalen Matrix O:*

$$O^\top AO \quad \textit{ist diagonal}$$

denn es gibt ja eine Orthonormalbasis O des $\mathbb{R}^n$ aus Eigenvektoren von A.

Im Komplexen hat sich ein anderer Name eingebürgert:

Definition 5.12.3. *Eine Matrix $A \in \mathbb{C}^{n \times n}$ heißt* unitär*, wenn die Spalten von A eine Orthonormalbasis von $\mathbb{C}^n$ bilden.*

Bemerkung 5.12.4. *Nach dem Spektralsatz II (Satz 5.5.24) ist eine hermitesche Matrix $A \in \mathbb{C}^{n \times n}$ diagonalisierbar mit einer unitären Matrix U:*

$$U^* AU \quad \textit{ist diagonal}$$

denn es gibt ja eine Orthonormalbasis des $\mathbb{C}^n$ aus Eigenvektoren von A.

5.13 Singulärwertzerlegung

Sei $K = \mathbb{R}$ oder $\mathbb{C}$.

Satz 5.13.1 (Singulärwertzerlegung)**.** *Eine Matrix $M \in K^{m \times n}$ hat eine Zerlegung*

$$M = U\Sigma V^*$$

mit $U \in K^{m \times m}$ unitär, $\Sigma \in K^{m \times n}$ diagonal mit nicht negativen Einträgen und $V^* \in K^{n \times n}$ unitär.

Definition 5.13.2. *Die Diagonaleinträge von Σ heißen die* Singulärwerte *von A.*

Das Ganze hat eine geometrische Interpretation. Sei dazu $T\colon K^n \to K^m$ eine lineare Abbildung. $V^* = (v_1 *, \ldots, v_n^*)$ ist eine Orthonormalbasis von K^n und $U = (u_1, \ldots, un_m)$ eine Orthonormalbasis von K^m, sodass

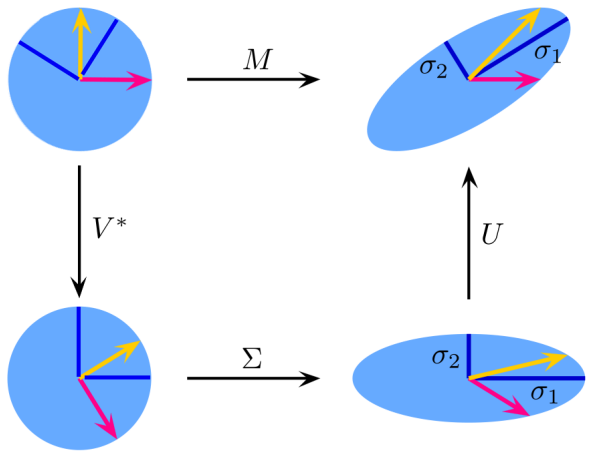
$$T(v_i^*) = \sigma_i u_i$$

für den Singulärwert σ_i ist.

Im Reellen gibt es folgende geometrische Interpretation: Die lineare Abbildung

$$T\colon \mathbb{R}^n \to \mathbb{R}^m$$

bildet die Einheitssphäre in $\mathbb{R}^n$ auf ein Ellipsoid in $\mathbb{R}^m$ ab. Die positiven Singulärwerte sind dann die Längen der Halbachsen des Ellipsoids (s. Abbildung 5.7).



$$M = U \cdot \Sigma \cdot V^*$$

Abbildung 5.7: Illustration der Singulärwertzerlegung (Quelle: Wikipedia, Autor: Georg-Johann).

5.13.1 Beste Rang-r-Approximation

Sei $A \in \mathbb{R}^{m\times n}$. Dann sieht die Singulärwertzerlegung wie folgt aus:

$$A = U\Sigma V^\top$$

mit $U \in \mathbb{R}^{m\times m}$ orthogonal, $\Sigma \in \mathbb{R}^{m\times n}$ diagonal mit nicht negativen Einträgen und $V \in \mathbb{R}^{n\times n}$ orthogonal. Ausgeschrieben sieht dies so aus:

$$\begin{aligned}
A &= (u_1 \dots u_k \mid u_{k+1} \dots u_m) \left(\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_k & \\ \hline 0 & & & 0 \end{array}\right) \left(\begin{array}{c} v_1^\top \\ \vdots \\ v_k^\top \\ \hline v_{k+1}^\top \\ \vdots \\ v_n^\top \end{array}\right) \\
&= (u_1 \dots u_k) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} v_1^\top \\ \vdots \\ v_k^\top \end{pmatrix} + \underbrace{(u_{k+1} \dots u_m)(0) \begin{pmatrix} v_{k+1}^\top \\ \vdots \\ v_n^\top \end{pmatrix}}_{=0} \\
&= (\sigma_1 u_1 \dots \sigma_k u_k) \begin{pmatrix} v_1^\top \\ \vdots \\ v_k^\top \end{pmatrix} = \sigma_1 u_1 v_1^\top + \dots + \sigma_k u_k v_k^\top
\end{aligned}$$

Bemerkung 5.13.3. *Es ist*

$$\mathrm{Rang}(A) = \mathrm{Rang}(\Sigma) = k$$

und

$$\mathrm{Rang}(u_i v_i^\top) = 1$$

denn jede Spalte von $u_i v_i^\top$ ist ein Vielfaches von u_i.

Für das *Matrix-Innenprodukt*

$$A \odot B := \sum_{i,j} \alpha_{ij}\beta_{ij}$$

mit $A = (\alpha_{ij}), B = (\beta_{ij}) \in \mathbb{R}^{m\times n}$ sind die Matrizen $u_i v_i^\top$ paarweise orthogonal. Denn für $x, u \in \mathbb{R}^m$, $y, v \in \mathbb{R}^m$ ist

$$\begin{aligned} xy^\top \odot uv^\top &= (xy_1 \mid \cdots \mid xy_n) \odot (uv_1 \mid \cdots \mid uv_n) = \sum_{i,j} x_i y_j u_i v_j \\ &= \sum_j xy_j \cdot uv_j = (x \cdot u) \sum_j y_j v_j = (x \cdot u)(y \cdot v) \end{aligned}$$

d.h. falls $x \perp u$ oder $y \perp v$, ist $xy^\top \perp uv^\top$. Wir haben hier das Standard-Innenprodukt mit $\cdot$ bezeichnet.

Wir haben also:

Satz 5.13.4. *Die Singulärwertzerlegung zerlegt $A \in \mathbb{R}^{m\times n}$ in eine Linearkombination*

$$A = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$$

paarweise orthogonaler Matrizen $u_i v_i^\top$ von Rang 1.

Für die *Frobenius-Norm*

$$\|A\|_F := \sqrt{A \odot A}$$

gilt:

$$\left\|u_i v_i^\top\right\|_F^2 = u_i v_i^\top \odot u_i v_i^\top = (u_i \cdot u_i)(v_i \cdot v_i) = 1$$

Also gilt nach dem Satz des Pythagoras (5.6):

$$\|A\|_F^2 = \sum_{i=1}^{k} \left\|\sigma_i u_i v_i^\top\right\|_F^2 = \sum_{i=1}^{k} |\sigma_i|^2 = \sum_{i=1}^{k} \sigma_i^2$$

Sei Σ so angeordnet, dass $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. Dann ist $\sigma_1 u_1 v_1^\top$ die beste Rang-1-Approximation an A. Das Fehlerquadrat beträgt:

$$\left\|A - \sigma_1 u_1 v_1^\top\right\|_F^2 = \sum_{i=2}^{k} \sigma_i^2$$

Allgemein ist $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$ die beste Rang-r-Approximation an A für $r \leq k$. Das Fehlerquadrat beträgt:

$$\left\|A - \sum_{i=1}^{r} \sigma_i u_i v_i^\top\right\|_F^2 = \sum_{i=r+1}^{k} \sigma_i^2$$

5.13.2 Datenkompression als beste Rang-r-Approximation

Ein Grauwertbild aus $m \cdot n$ Pixeln kann als Matrix $A \in \mathbb{R}^{m \times n}$ angesehen werden. Um das volle Bild abzuspeichern, müssen alle $m \cdot n$ Grauwerte gespeichert werden. Bei der besten Rang-1-Approximation

$$\sigma_1 u_1 v_1^\top$$

sind lediglich $m + n + 1$ Werte zu speichern. Bei der besten Rang-r-Approximation

$$\sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

sind es $r(m + n + 1)$ Werte. Für r hinreichend klein, ist diese Zahl kleiner als $m \cdot n$.

Die Methode ist wie folgt: Sei

$$E_r := A - \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

Wähle nun r so, dass

$$\frac{\|E_r\|_F}{\|A\|_F} = \sqrt{\frac{\sum_{i=r+1}^{k} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2}} < \epsilon$$

für einen vorgegebenen Schwellwert $\epsilon > 0$. Dann ist

$$\sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

die *Kompression* von A als beste Rang-r-Approximation.

Grundannahme der Kompression

Es wird bei dieser Methode angenommen, dass die Terme $\sigma_i u_i v_i^\top$ für kleine Singulärwerte σ_i keine relevante Information enthalten, d.h. dass sie aus Rauschen bestehen.

5.13.3 Lineare kleinste Quadrate

Sei $V = \mathbb{R}^m$ und $\{a_1, \ldots, a_n\} \subseteq V$ linear unabhängig sowie $b \in V$. Gesucht sind Koeffizienten $\xi_1, \ldots, \xi_n \in \mathbb{R}$, sodass der Fehler

$$\left\| b - \sum_{i=1}^{n} \xi_i a_i \right\|$$

minimal wird. In Matrixschreibweise bedeutet dies mit $A = (a_1, \ldots, a_n) \in \mathbb{R}^{m \times n}$ die Minimierung von

$$\|b - Ax\|$$

wobei $x = (\xi_1, \ldots, \xi_n)$ ist. Es wird die Situation betrachtet, dass das lineare Gleichungssystem

$$Ax = b$$

überbestimmt ist. Dann wird eine bestmögliche Lösung dieses LGS gesucht. Geometrisch bedeutet dies: suche ein Element des von $a_1, \ldots, a_n$ aufgespannten Unterraums $S \subseteq \mathbb{R}^n$ mit minimalem Abstand zu b.

Die Lösung des Problems ergibt sich mit der Orthogonalprojektion $\pi_S(b)$ von b auf S:

$$\pi_S(b) = Ax$$

Für den Fehlervektor $b - \pi_S(b)$ gilt:

$$b - \pi_S(b) \perp S$$

Dies ist äquivalent zu

$$\begin{aligned} & a_i \perp Ax - b && i = 1, \ldots, n \\ \Leftrightarrow \quad & A^\top (Ax - b) = 0 && \\ \Leftrightarrow \quad & A^\top Ax = A^\top b && \textit{(Normalgleichungen)} \end{aligned}$$

Bemerkung 5.13.5. *$A^\top A$ ist invertierbar, da $a_1, \ldots, a_n$ linear unabhängig sind. Aber die Berechnung von $\left(A^\top A\right)^{-1}$ soll vermieden werden, da die Rechengeschwindigkeit und -Genauigkeit leidet.*

Mit der Singulärwertzerlegung $A = U\Sigma V^\top$ gilt:

$$Ax - b = U\Sigma V^\top x - b = U(\Sigma \underbrace{V^\top x}_{=:y} - \underbrace{U^\top b}_{=:c})$$

Also ist

$$\|Ax - b\| = \|\Sigma y - c\|$$

denn:

$$\|Uz\|^2 = \langle Uz, Uz \rangle = (Uz)^\top Uz = z^\top \underbrace{U^\top U}_{=I} z = z^\top z = \|z\|^2 \tag{5.8}$$

Es ist also y zu finden, dass $\|\Sigma y - c\|$ minimal wird. Da Σ diagonal ist, haben wir:

$$\Sigma y = (\sigma_1 y_1, \ldots, \sigma_k y_k, 0, \ldots, 0)$$

Also

$$\Sigma y - c = (\sigma_1 y_1 - c_1, \ldots, \sigma_k y_k - c_k, -c_{k+1}, \ldots, -c_m)$$

Und dessen Norm wird minimal für

$$y_i = \frac{c_i}{\sigma_i}, \quad i = 1, \ldots, k \tag{5.9}$$

und $y_{k+1}, \ldots, y_n$ sind frei wählbar. Dann ist nämlich

$$\|\Sigma y - c\|^2 = \sum_{i=k+1}^{m} c_i^2$$

Das gesuchte x ist also

$$x = Vy$$

mit y gemäß (5.9).

5.13.4 Konditionszahl von quadratischen Matrizen

Sei $A \in \mathbb{R}^{n\times n}$ und die Singulärwertzerlegung sei

$$A = U\Sigma V^\top$$

mit $\Sigma = \mathrm{Diag}(\sigma_1, \dots, \sigma_n)$ sodass $\sigma_1 \geq \cdots \geq \sigma_n$.

Genau dann ist A invertierbar, wenn $\sigma_n > 0$ ist. Dann ist

$$A^{-1} = V \,\mathrm{Diag}\left(\sigma_1^{-1}, \dots, \sigma_n^{-1}\right) U^\top$$

Definition 5.13.6.

$$\mathrm{cond}(A) := \frac{\sigma_1}{\sigma_n}$$

ist die Konditionszahl *von A.*

Die Konditionszahl gibt an, wie nahe A an einer singulären Matrix ist. Bei $\mathrm{cond}(A) = \infty$ ist A tatsächlich singulär, bei $\mathrm{cond}(A) >> 1$ ist A „fast“ singulär.

Das Problem

$$Ax = b$$

nennt man bei $\mathrm{cond}(A) >> 1$ *schlecht konditioniert*, bei $\mathrm{cond}(A) = \infty$ *schlecht gestellt* und sonst *gut konditioniert.*

5.13.5 Kabsch-Algorithmus

Gesucht ist die optimale Rotationsmatrix zwischen gepaarten Punktmengen in $\mathbb{R}^3$. Die Punktmengen seien

$$P = \{p_1, \dots, p_n\}, \quad Q = \{q_1, \dots, q_n\}$$

Zunächst werden die Zentroide gebildet:

$$C_P = \frac{1}{n}\sum_{i=1}^{n} p_i, \quad C_Q = \frac{1}{n}\sum_{i=1}^{n} q_i$$

Ersetze dann P, Q durch

$$\{p_1 - C_P, \dots, p_n - C_P\}, \quad \{q_1 - C_Q, \dots, q_n - C_Q\}$$

und nenne die Punkte wiederum p_i bzw. q_i.

Es ist die Größe

$$E(U) = \frac{1}{n}\sum_{i=1}^{n} \left\| \underbrace{Up_i}_{=:p_i'} - q_i \right\|^2$$

zu minimieren. Schreibe dabei P, Q als $3 \times n$-Matrizen. Es ist

$$\begin{aligned} nE &= \sum_{i=1}^{n} \left\|p_i' - q_i\right\|^2 = \mathrm{trace}\left((P' - Q)^\top (P' - Q)\right) \\ &= \mathrm{trace}\left(P'^\top P'\right) + \mathrm{trace}\left(Q^\top Q\right) - 2\,\mathrm{trace}\left(Q^\top P'\right) \\ &= \sum_{i=1}^{n} \left(\|p_i\|^2 + \|q_i\|^2\right) - 2\,\mathrm{trace}\left(Q^\top P'\right) \end{aligned}$$

Beachte dabei, dass $\|p_i'\| = \|p_i\|$ wegen (5.8), da U orthogonal ist. Maximiere also

$$\operatorname{trace}\left(Q^\top P'\right) = \operatorname{trace}\left(Q^\top U P\right) = \operatorname{trace}\left(PQ^\top \cdot U\right)$$

wobei $PQ^\top \in \mathbb{R}^{3\times 3}$ ist. Die letzte Gleichung gilt allgemein als Rechenregel für die Spur. Die Singulärwertzerlegung ist

$$PQ^\top = V\Sigma W^\top$$

Dann ist die optimale Rotation

$$U = W \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} V^\top$$

mit

$$d = \operatorname{sgn}\left(\det\left(PQ^\top\right)\right)$$

Eine Anwendung des Kabsch-Algorithmus findet beispielsweise in der Satellitenorientierung statt.

5.14 Hilberträume

Sei $(V, \langle\cdot,\cdot\rangle)$ ein Innenproduktraum. Die zugehörige Norm ist

$$\|\cdot\| \colon V \to \mathbb{R}, \quad x \mapsto \sqrt{\langle x, x\rangle}$$

Definition 5.14.1. *Ist $(V, \|\cdot\|)$ vollständig, so heißt $(V, \langle\cdot,\cdot\rangle)$* Hilbertraum.

Definition 5.14.2. *Sind $b_1, b_2, \dots \in V$ mit $\|b_\nu\| = 1$ und $b_\mu \perp b_\nu$ für $\mu \neq \nu$, und gilt für jedes $x \in V$:*

$$x = \sum_{\nu=1}^{\infty} \alpha_\nu b_\nu$$

für gewisse $\alpha_\nu \in K$, so heißt die Folge $b_1, b_2, \dots$ eine Orthonormalbasis *von V.*

Bemerkung 5.14.3. *Eine derartige Orthonormalbasis ist i.A. keine Basis im Sinne der linearen Algebra!*

Satz 5.14.4. *Jeder Hilbertraum hat eine Orthonormalbasis.*

Bemerkung 5.14.5. *Für $x = \sum_{\nu=1}^{\infty} \alpha_\nu b_\nu$ ist*

$$\langle x, b_\mu\rangle = \sum_{\nu=1}^{\infty} \alpha_\nu \underbrace{\langle b_\nu, b_\mu\rangle}_{=\delta_{\nu\mu}} = \alpha_\mu$$

Definition 5.14.6. *Der Ausdruck*

$$\langle x, b_\mu\rangle$$

heißt Fourier-Koeffizient *von x bezüglich der Orthonormalbasis $b_1, b_2, \dots$.*

Beispiel 5.14.7. *Sei $V = C[-\pi, \pi]$ mit dem Skalarprodukt*

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(t)\overline{g(t)}\, dt$$

Dann ist

$$e_k(t) = \frac{1}{\sqrt{2\pi}} e^{ikt}$$

eine Orthonormalbasis von V. Die Orthogonalitätsrelationen gelten wegen

$$\langle e_k, e_\ell \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(k-\ell)t}\, dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(k-\ell)t\, dt + i\frac{1}{2\pi} \int_{-\pi}^{\pi} \sin(k-\ell)t\, dt = \delta_{k,\ell}$$

Der k-te Fourier-Koeffizient ist

$$\langle f, e_k \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(t) e^{-ikt}\, dt$$

Die Orthogonalbasis-Eigenschaft der e_k besagt zudem, dass jedes Signal eine Überlagerung von reinen Sinus-Schwingungen ist, die als Obertöne vorkommen. Bei einer schwingenden Saite wird der k-te Oberton durch Berühren an der Stelle $\frac{1}{k-1}$ hörbar (s. Abbildung 5.8).

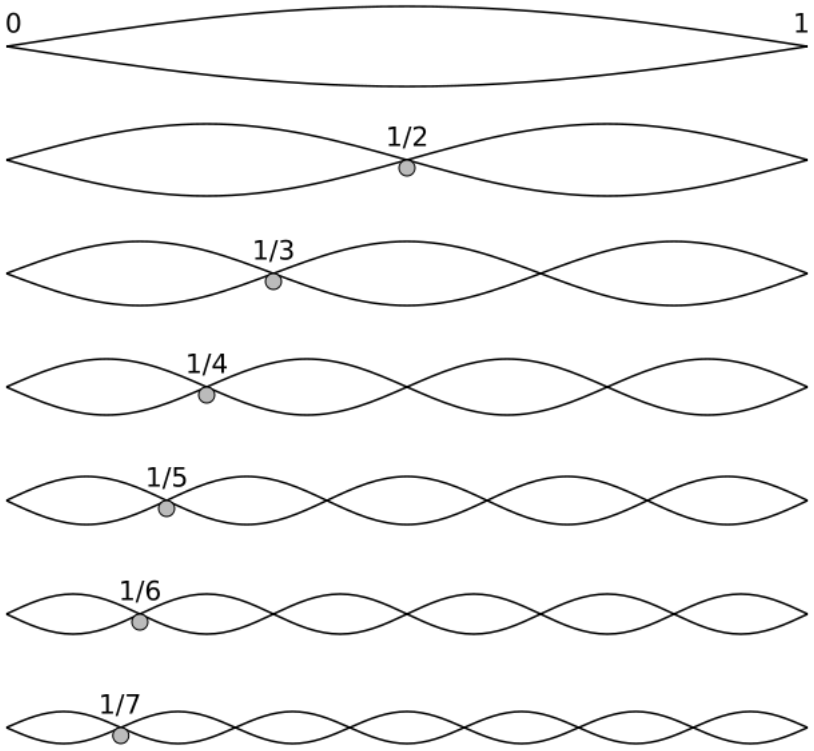


Abbildung 5.8: Obertöne einer schwingenden Saite (Quelle: Wikipedia, Author: Qef).

5.14.1 40000 Stellen von π

In der Episode *Marge wird verhaftet* (1993) steht Marge wegen Ladendiebstahls vor Gericht. Ihr Anwalt will die Glaubwürdigkeit des angeblichen Zeugen Apu Nahasapeemapetilon in Zweifel ziehen, in dem er andeutet, dessen Erinnerungen könnten falsch sein. Apu erwidert, dass er in der Lage sei, die Zahl π bis zur 40 000-sten Stelle hinter dem Komma zu nennen. Die 40 000-ste Ziffer sei eine 1.

Hätte Apu eine Zeitmaschine zur Verfügung gehabt, so hätte er im Jahr 1995 die Bailey-Borwein-Plouffe-Formel nachsehen können, welche eine beliebige Stelle von π ausspuckt, ohne die Kenntnis der Stellen davor. Allerdings verwendet diese Formel das 16er-System.

Satz 5.14.8 (Bailey-Borwein-Plouffe-Formel, 1995)**.** *Es gilt:*

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right)$$

In der Hexadezimaldarstellung von π ergibt sich

$$\pi = \sum_{k=0}^{\infty} \frac{z_k}{16^k}$$

mit

$$z_k = \left\lfloor \left(16^{k-1}\pi \mod 1 \right) \cdot 16 \right\rfloor$$

Dann ist nach Satz 5.14.8

$$16^{n-1}\pi = 4\sigma_1 - 2\sigma_4 - \sigma_5 - \sigma_6$$

mit

$$\sigma_\ell = \sum_{k=0}^{\infty} \frac{16^{n-k-1}}{8k-\ell}$$

Von jedem einzelnen Summanden ist nun der ganzzahlige Teil zu entfernen. Dies geht wie folgt: ändere σ_ℓ ab zu

$$\sigma'_\ell = \sum_{k=0}^{n-1} \frac{\left(16^{n-k-1} \mod (8k+\ell)\right)}{8k+\ell} + \sum_{k=n}^{\infty} \frac{16^{n-k-1}}{8k+\ell}$$

Dann ist:

$$16^{n-1}\pi \equiv 4\sigma'_1 - 2\sigma'_4 - \sigma'_5 - \sigma'_6 \equiv \theta_n \mod 1$$

wobei $\theta_n \in [0,1)$ ist. Dann ist

$$z_n = \lfloor 16 \cdot \theta_n \rfloor$$

die gesuchte Ziffer im Dezimalsystem. Verwendet wurde die Gauß-Klammer

$$\lfloor x \rfloor := n \in \mathbb{Z} \quad \text{mit} \quad x - n \in [0,1)$$

Dann ergibt sich auf einem Rechner $z_{40\,000} = 1$.

Kapitel 6

Trigonometrische Funktionen

6.1 Diskrete Fouriertransformation

Sei $\zeta = e^{2\pi i/N} \in \mathbb{C}$ eine primitive N-te Einheitswurzel. Sie ist eine komplexe Lösung der Gleichung

$$X^N = 1$$

Alle Lösungen dieser Gleichung sind gegeben durch die Potenzen von ζ:

$$\zeta^0, \zeta^1, \ldots, \zeta^{N-1}$$

Daraus formen wir die Vektoren

$$\begin{aligned} z &= \left(1, \zeta, \zeta^2, \ldots, \zeta^{N-1}\right) \in \mathbb{C}^N \\ z^k &= \left(1, \zeta^k, \zeta^{2k}, \ldots, \zeta^{(N-1)k}\right) \in \mathbb{C}^N \end{aligned}$$

Indizieren wir noch Vektoren in $\mathbb{C}^N$ wie folgt:

$$f = (f_0, \ldots, f_N) \in \mathbb{C}^N$$

so schreibt sich das Standard-Innenprodukt auf $\mathbb{C}^N$ wie folgt:

$$\langle a, b \rangle = \sum_{\nu=0}^{N-1} a_\nu \bar{b}_\nu$$

Lemma 6.1.1. *Die Vektoren $z^0, \ldots, z^{N-1}$ bilden eine Orthogonalbasis von $\mathbb{C}^N$.*

Beweis. Dies folgt aus den Orthogonalitätsrelationen:

$$\langle z^k, z^\ell \rangle = \sum_{\nu=0}^{N-1} \zeta^{\nu k} \zeta^{-\nu \ell} = \sum_{\nu=0}^{N-1} e^{\frac{2\pi i}{N}(k-\ell)\nu} = N\delta_{k\ell} \tag{6.1}$$

Letztere Gleichheit gilt, da $\xi = \zeta^{k-\ell}$ für $k \neq \ell$ eine N-te Einheitswurzel ist und

$$0 = \frac{1-\xi^N}{1-\xi} = 1 + \xi + \cdots + \xi^{N-1}$$

Es handelt sich also um N paarweise orthogonale Vektoren in $\mathbb{C}^N$. □

Als Konsequenz ergibt sich, dass ein Vektor $f \in \mathbb{C}^N$ eine Koordinatendarstellung bezüglich $z^0, \ldots, z^{N-1}$ hat:

$$f = \sum_{k=0}^{N-1} \alpha_k z^k$$

Der Koeffizient α_k berechnet sich zu

$$\alpha_k = \frac{1}{N} \sum_{\nu=0}^{N-1} \alpha_\ell N \delta_{k\ell} = \frac{1}{N} \sum_{\ell=0}^{N-1} \alpha_\ell \langle z^\ell, z^k \rangle = \frac{1}{N} \left\langle \sum_{\ell=0}^{N-1} \alpha_\ell z^\ell, z^k \right\rangle = \frac{1}{N} \langle f, z^k \rangle$$

Die Größe $\langle f, z^k \rangle$ heißt *diskreter Fourier-Koeffizient* von f. Dieser hat die Darstellung

$$b_k := \langle f, z^k \rangle = \sum_{\nu=0}^{N-1} f_\nu \zeta^{-k\nu} = \Phi(\zeta^{-k})$$

mit dem Polynom

$$\Phi(X) = \sum_{\nu=0}^{N-1} f_\nu X^\nu \in \mathbb{C}[X]$$

Bemerkung 6.1.2. *Der diskrete Fourier-Koeffizient b_k ist periodisch:*

$$b_k = b_{k+N}$$

Dies ergibt sich aus

$$\Phi\left(\zeta^{-k}\right) = \Phi\left(\zeta^{-k-N}\right)$$

wegen

$$\zeta^{-k-N} = \zeta^{-k}\zeta^{-N} = \frac{\zeta^{-k}}{\zeta^N} = \zeta^{-k}$$

da $\zeta^N = 1$ ist.

Definition 6.1.3. *Der Vektor*

$$\mathcal{F}(f) = (b_0, \ldots, b_{N-1})$$

heißt Diskrete Fourier-Transformierte (DFT) *von f.*

Bemerkung 6.1.4. *$\mathcal{F}$ ist die Multiplikation mit F^*, wobei F die Vandermonde-Matrix ist:*

$$F = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & \zeta & \ldots & \zeta^{N-1} \\ \vdots & \vdots & & \vdots \\ 1 & \zeta^{N-1} & \ldots & \zeta^{(N-1)(N-1)} \end{pmatrix}$$

Die Orthogonalitätsrelationen besagen:

$$F^{-1} = \frac{1}{N} F^*$$

Als Konsequenz ergibt sich:

$$f_k = \langle \mathcal{F}(f), z^{-k} \rangle = \frac{1}{N} \sum_{\nu=0}^{N-1} b_\nu e^{\frac{2\pi i}{N} k\nu} = \frac{1}{N} \tilde{\Phi}(\zeta^k)$$

mit dem Polynom

$$\tilde{\Phi}(X) = \sum_{\nu=0}^{N-1} b_\nu X^\nu \in \mathbb{C}[X]$$

für die inverse diskrete Fourier-Transformation (IDFT).

6.1.1 Schnelle Fourier-Transformation

Wir verwenden die Bezeichnungen des vorigen Abschnitts, nur dass jetzt ein Index N zusätzlich verwendet wird. D.h. $\zeta_N = e^{2\pi i/N}$ sei die primitive N-te Einheitswurzel, und es sei

$$z_N^k = \left(1, \zeta_N^k, \zeta_N^{2k}, \ldots, \zeta_N^{(N-1)k}\right) \in \mathbb{C}_N$$

Weiter setzen wir voraus, dass $N = 2^n$ eine Zweierpotenz sei. Den Vektor

$$f = (f_0, \ldots, f_{N-1}) \in \mathbb{C}^N$$

zerlegen wir in einen geraden und einen ungeraden Anteil:

$$\begin{aligned} f &= \tilde{g} + \tilde{u} \\ \tilde{g} &= (f_0, 0, f_2, 0, \ldots, f_{N-2}, 0) \\ \tilde{u} &= (0, f_1, 0, f_3, \ldots, 0, f_{N-1}) \end{aligned}$$

Entfernen der Nullen ergibt:

$$\begin{aligned} g &= (g_\mu) \in \mathbb{C}^{N/2}, \quad g_\mu = \tilde{g}_{2\mu} \\ u &= (u_\mu) \in \mathbb{C}^{N/2}, \quad u_\mu = \tilde{u}_{2\mu+1} \end{aligned}$$

Der diskrete Fourier-Koeffizient $b_{k,N}$ hat auch eine Zerlegung:

$$b_{k,N} := \langle f, z_N^k \rangle = \langle \tilde{g}, z_N^k \rangle + \langle \tilde{u}, z_N^k \rangle$$

Dabei ist

$$\begin{aligned} \langle \tilde{g}, z_N^k \rangle &= \sum_{\mu=0}^{N/2-1} \tilde{g}_{2\mu} e^{-\frac{2\pi i}{N}(2\mu)k} = \sum_{\mu=0}^{N/2-1} \tilde{g}_{2\mu} e^{-\frac{2\pi i}{N/2}\mu k} = \sum_{\mu=0}^{N/2-1} g_\mu \zeta_{N/2}^{-\mu k} = \langle g, z_{N/2}^k \rangle \\ \langle \tilde{u}, z_N^k \rangle &= \sum_{\mu=0}^{N/2-1} \tilde{u}_{2\mu+1} e^{-\frac{2\pi i}{N}(2\mu+1)k} = e^{-\frac{2\pi i}{N}k} \sum_{\mu=0}^{N/2-1} \tilde{u}_{2\mu+1} e^{-\frac{2\pi i}{N/2}k} = \zeta_N^k \langle u, z_{N/2}^k \rangle \end{aligned}$$

Es ist also

$$b_{k,N} = \langle g, z_{N/2}^k \rangle + \zeta_N^k \langle u, z_{N/2}^k \rangle, \quad k = 0, \ldots, N-1$$

d.h. der Fourier-Koeffizient für N zerlegt sich in einen Fourier-Koeffizient für $N/2$ und einen „getwisteten“ Fourier-Koeffizienten für $N/2$. Weiter gilt:

$$z_{N/2}^k = z_{N/2}^{k+N/2}$$

und

$$\zeta_N^{k+N/2} = \zeta_N^{N/2} \zeta_N^k = -\zeta_N^k$$

Es ergibt sich damit:

$$b_{k,N} = \begin{cases} \langle g, z_{N/2}^k \rangle + \zeta_N^k \langle u, z_{N/2}^k \rangle, & k < N/2 \\ \langle g, z_{N/2}^{k-N/2} \rangle + \zeta_N^{k-N/2} \langle u, z_{N/2}^{k-N/2} \rangle, & k \geq N/2 \end{cases}$$

Die DFT bei Länge N reduziert sich somit auf die DFT bei Länge $N/2$, und wir können dies solange fortsetzen, bis wir bei Länge 2 angelangt sind. Dies ergibt einen *Teile-und-Herrsche-Algorithmus* zur Berechnung des diskreten Fourier-Koeffizienten bei Länge $N = 2^n$.

6.1.2 Fourier-Reihen

Bekanntlich lässt sich jede integrierbare Funktion $f\colon [0,L] \to \mathbb{R}$ mit $f(0) = f(L)$[1] in eine Fourier-Reihe entwickeln:

$$f(x) = \sum_{k=-\infty}^{\infty} \beta_k e^{2\pi i k x/L}$$

Für den Fourier-Koeffizient β_k ($k \in \mathbb{Z}$) gilt:

$$\beta_k = \frac{1}{L} \int_0^L f(x) e^{-2\pi i k x/L}\, dx = \langle f, e_k \rangle \tag{6.2}$$

mit

$$e_k\colon [0,L] \to \mathbb{C}, \quad x \mapsto \frac{1}{L} e^{2\pi i k x/L}$$

Diskrete Approximation des Fourier-Koeffizienten

Gehen wir von einer äquidistanten Abtastung der periodischen Funktion f aus, so haben wir zunächst die Zerlegung des Intervalls $[0,L]$:

$$h = \frac{L}{N}, \quad x_\nu = \nu h, \quad \nu = 0, \dots, N-1$$

und die Werte

$$f_\nu = f(x_\nu)$$

Dann ergibt sich mit einer Rechteck-Approximation des Integrals:

$$\beta_k = \frac{1}{L} \int_0^L f(x) e^{-2\pi i k x/L}\, dx = \frac{1}{L} \sum_{\nu=0}^{N-1} \int_{\nu L/N}^{(\nu+1)L/N} f(x) e^{-2\pi i k x/L}\, dx \approx \frac{1}{L} \frac{L}{N} \sum_{\nu=0}^{N-1} f_\nu e^{-\frac{2\pi i}{N}\nu k} = \frac{1}{N} b_k$$

wobei b_k der diskrete Fourier-Koeffizient ist. Bei der Rechteck-Approximation wurde folgendermaßen approximiert:

$$f(x) \approx f_\nu, \quad x \approx \frac{\nu L}{N}$$

wenn $x \in \left[\frac{\nu L}{N}, \frac{(\nu+1)L}{N}\right]$ ist.

Bemerkung 6.1.5. *Die Approximation*

$$\beta_k \approx \frac{b_k}{N}$$

ist nur gut für $|k|$ klein, denn b_k ist periodisch, während oft $\beta_k \to 0$ für $|k| \to \infty$.

Approximierte Fourier-Reihe

Sei N aus Bemerkung 6.1.5 gerade und $\alpha_k = \frac{b_k}{N}$. Dann gilt:

$$\begin{aligned} \alpha_k &\approx \beta_k, & k &= 0, \dots, N/2-1 \\ \alpha_k = \alpha_{k+N} &\approx \beta_k, & k &= -N/2, \dots, -1 \end{aligned}$$

und wir approximieren:

$$f(x) \approx \sum_{k=-N/2}^{N/2} \alpha_k e^{2\pi i k x/L}$$

[1] Die Bedingung bedeutet, dass f zu einer periodischen Funtion auf $\mathbb{R}$ fortgesetzt werden kann.

Digitale Signalübertragung

Zur digitalen Übertragung eines analogen Signals $f\colon [0, L] \to \mathbb{R}$ taste N Werte äquidistant ab und erhalte (normierte, periodische) diskrete Fourier-Koeffizienten α_k ($k \in \mathbb{Z}$). Übermittle diese für $k = -N/2, \ldots, N/2$ an den gewünschten Zielort und rekonstruiere dort das analoge Signal

$$\sum_{k=-N/2}^{N/2} \alpha_k e^{2\pi i k x/L}$$

Das Ergebnis ist eine Glättung bzw. Datenkompression, bei der die hochfrequenten Anteile vernachlässigt werden.

6.2 Trigonometrische Interpolation

Wir stellen uns nun die Aufgabe, eine 2π-periodische Funktion f an äquidistanten Stützstellen mit trigonometrischen Summen zu interpolieren. D.h. für unsere Funktion f gilt:

$$f(x + 2\pi) = f(x)$$

und die trigonometrischen Summen haben die Form:

$$T_n(x) = \sum_{k=0}^{n} \gamma_k e^{ikx}$$

Das Interpolationsintervall ist $[0, 2\pi]$ mit äquidistanten Stützstellen

$$x_k = \frac{2\pi}{n+1} k, \quad k = 0, \ldots, n$$

Satz 6.2.1. *Das Interpolationsproblem ist eindeutig lösbar. D.h. zu $y_0, \ldots, y_n \in \mathbb{C}$ gibt es genau eine Funktion*

$$T_n(x) = \sum_{k=0}^{n} \gamma_k e^{ikx}$$

mit $T_n(x_\nu) = y_\nu$ für $\nu = 0, \ldots, n$.

Beweis. Setze $\omega = e^{ix}$, $\omega_k = e^{ix_k} = e^{\frac{2\pi i}{n+1} k}$ und

$$P_n(X) = \sum_{k=0}^{n} \gamma_k X^k$$

Dann gilt:

$$\begin{aligned} T_n(x) &= P_n(\omega) \\ y_\nu = T_n(x_\nu) &= P_n(\omega_\nu) \end{aligned}$$

Weil das Interpolationspolynom $P_n(X)$ nach Satz 4.1.3 eindeutig bestimmt ist, ist es auch $T_n(x)$. □

Berechnung der Koeffizienten

Die Koeffizienten γ_k des trigonometrischen Interpolationspolynoms $T_n(x)$ berechnen sich wie folgt:

$$\gamma_k = \frac{1}{n+1}\sum_{\nu=0}^{n} y_\nu e^{-i\nu x_k} = \frac{1}{n+1}\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu}$$

Beweis.

$$\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu} = \sum_{\nu=0} P_n(\omega_\nu)\omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^\ell \omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^{\ell-k}$$

wobei die letzte Gleichheit gilt wegen

$$\omega_k^\nu = e^{\frac{2\pi i}{n+1}k\nu} = \omega_\nu^k$$

Somit ergibt sich

$$\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^{\ell-k} = \sum_{\ell=0}^{n} \gamma_\ell \sum_{\nu=0}^{n} \omega_\nu^{\ell-k} \overset{(*)}{=} \sum_{\ell=0}^{n} \gamma_\ell \cdot (n+1)\delta_{k,\ell} = \gamma_k(n+1)$$

woraus die Behauptung folgt. Dabei gilt $(*)$ wegen der Orthogonalitätsrelationen (6.1). □

6.3 Multiplikation großer Zahlen

Eine Anwendung der schnellen Multiplikation großer Zahlen tritt bei der Verschlüsselung von Daten im Internet auf. Diese wird in Abschnitt 7.1 behandelt.

6.3.1 Multiplikation über komplexe DFT

Die Multiplikation zweier m-stelliger natürlicher Zahlen nach der Schulmethode lässt sich zurückführen auf m^2 Multiplikationen einstelliger Zahlen. Dies ist für große m schon sehr aufwändig.

Eine erste Idee zur Abhilfe ist es, Zahlen als Polynome aufzufassen. So ist etwa die Zahl $q = 5821$ in ihrer Dezimaldarstellung:

$$q = 1 + 2 \cdot 10 + 8 \cdot 10^2 + 5 \cdot 10^3 = Q(10)$$

für das Polynom

$$Q(X) = 1 + 2X + 8X^2 + 5X^3 \in \mathbb{Z}[X]$$

Das Produkt zweier Zahlen ergibt sich dann über das Produkt $R(X) = P(X) \cdot Q(X)$ zweier Polynome mit anschließender Auswertung

$$pq = R(10)$$

Dies funktioniert übrigens nicht nur für Dezimalzahlen sondern auch für eine beliebige Basis g einer g-adischen Darstellung von Zahlen:

$$a = \sum_{i=0}^{n} a_i g^i$$

mit $a_i \in \{0, \ldots, g-1\}$.

Das Problem ist dabei, dass die direkte Multiplikation zweier Polynome von Grad $n-1$ ebenfalls n^2 Multiplikationen benötigt. Ziel ist es nun, diese Multiplikation zu beschleunigen. Nehmen wir an, dass $\deg R = m-1$ ist. Dann betrachten wir folgende Methode:

1. Evaluiere P und Q an m Stellen $x_0, \ldots, x_{m-1}$.
2. $R(x_s) = P(x_s) \cdot Q(x_s)$ evaluiert R an m Stellen.
3. Bestimme daraus die Koeffizienten von R (d.h. interpoliere).

Schritt 2 benötigt nur m Multiplikationen. Es werden also effiziente Wege zur Realisierung der Schritte 1 und 3 gesucht.

Hilfreich ist es, für x_s m-te Einheitswurzeln, also Lösungen der Gleichung $X^m = 1$, zu verwenden. Im Komplexen ist dann

$$x_s = e^{-2\pi is/m} = \omega^{-s}$$

mit der primitiven m-ten Einheitswurzel $\omega = e^{2\pi i/m}$. Die Auswertung eines Polynoms $A(X) = \sum_{t=0}^{m-1} a_t X^t$ an der Stelle x_s ist dann

$$\tilde{a}_s = A(\omega^{-s}) = \sum_{t=0}^{m-1} a_t e^{-2\pi ist/m}$$

was nichts anderes ist als der s-te Koeffizient der Diskreten Fourier-Transformation (DFT). Diese Fourier-Koeffizienten kann man effizient mit der schnellen Fourier-Transformation (FFT), wie in Abschnitt 6.1.1 behandelt, berechnen. Nun gilt:

Satz 6.3.1 (Faltungssatz)**.** *Die Fourier-Koeffizienten eines Produkts $R(X) = P(X) \cdot Q(X)$ sind die Produkte der Fourier-Koeffizienten der Polynome $P(X)$ und $Q(X)$.*

Dies hat zur Folge, dass die Koeffizienten des Produkts $R(X)$ über die inverse DFT (IDFT) gewonnen werden können. Genauer ist

$$a_t = \frac{1}{m}\tilde{A}(\bar{\omega}^{-t}) = \frac{1}{m}\sum_{s=0}^{m-1} a_s e^{2\pi ist/m} \tag{6.3}$$

falls $\tilde{A}(X) = \sum_{s=0}^{m-1} \tilde{a}_s X^s$ das Fourier-transformierte Polyonom zu $A(X)$ ist. (6.3) besagt, dass die IDFT ebenfalls mit FFT berechnet werden kann, wobei die Einheitswurzel ω durch ihre Komplex-Konjugierte $\bar{\omega}$ zu ersetzen ist.

Diese Methode ist für große Zahlen wesentlich effizienter als die naive Multiplikation von Zahlen. Allerdings hat man mit Rundungsfehlern bei der DFT im Komplexen zu tun.

6.3.2 Multiplikation über modulare DFT

Um die FFT-Methode für das Multiplizieren von Polynomen für die Multiplikation großer Zahlen zu verwenden, ohne dass Rundungsfehler auftreten, wird in Kongruenzen modulo einer großen Zahl der Form $N = 2^{2^w} + 1$ gerechnet. Für große N ist ein Produkt ganzer Zahlen $p \cdot q$ dasselbe wie $p \cdot q \mod N$. Der Vorteil bei dieser Wahl von N ist, dass wegen

$$2^{2^w} \equiv -1 \mod N \quad \text{und} \quad 4^{2^w} \equiv 2^{2\cdot 2^w} \equiv \left(2^{2^w}\right)^2 \equiv 1 \mod N$$

4 eine primitive 2^w-te Einheitswurzel ist. Dies bedeutet, dass 2^k-te Einheitswurzeln für $k \leq w$ Zweierpotenzen sind. Die Fouriertransformation mit 2^k-ten Einheitswurzeln modulo N lässt sich

unter Verwendung von Shift-Operationen auf Binärzahlen effizient berechnen. Dies lässt sich ausnutzen, um die Multiplikation großer ganzer Zahlen mit n Bits via modularer Arithmetik mit einer Komplexität von

$$O(n \log(n) \log(\log(n)))$$

zu berechnen, was wesentlich effizienter ist als die Schulmethode mit einer Komplexität von $O(n^2)$. Es wird vermutet, dass $O(n \log(n))$ eine untere Komplexitätsschranke für die Multiplikation zweier ganzer Zahlen ist.

Sowohl die komplexe als auch die modulare Variante der Multiplikation großer Zahlen sind als *Schönhage-Strassen-Algorithmus* bekannt.

6.4 Eulers Formel und die Existenz Gottes

Die Gleichung

$$e^{i\pi} + 1 = 0$$

von Leonhard Euler kommt als Buchtitel vor in Lisas Büchersammlung, mit der sie sich auf ihre Karriere als Baseball-Trainerin vorbereitet.

Ein weiterer Auftritt dieser Gleichung ist in *Homer*^{3}, wo sie Homer Simpson in der dritten Dimension erscheint.

Für manche ist diese Formel ein Beweis für die Existenz Gottes, da in ihr die verschiedenen mathematischen Disziplinen: Arithmetik (0 und 1), Algebra (i), Geometrie (π) und Analysis (e) vereint sind, was kein Zufall sein kann.

Allerdings kann das Wesen Gottes mit unserem begrenzeten menschlichen Verstand nicht erfasst werden!

Obwohl rationale Argumente für die Existsez Gottes bemerkenswert sind und auch von den Kirchenvätern gelegentlich ins Feld geführt wurden, ist die Erkenntnis Gottes, die aus *persönlicher geistlicher Erfahrung* stammt, von viel größerer Bedeutung:

> Selig die Reinen im Herzen, denn sie werden Gott schauen. (Mt 5, 8)

Die Heiligen Kirchenväter bekennen die Wahrheit dieser Aussage, da ihnen die Vision Gottes gewährt wurde nach einem Läuterungsprozess, in welchem sie die *Geistkraft*[2] von aller durch Sünden erzeugten Unreinheit befreiten und so mit diesem das Unerschaffene Licht sahen. Sie sagen, dass jeder Mensch diesen Zustand der Seele erreichen kann durch die Heilige Taufe in der Orthodoxen Kirche und danach (da wir ständig in die Sünde fallen) durch Gebet, durch den geistigen Kampf gegen die Leidenschaften, durch sorgfältige Gewissensprüfung und indem wir während der Heiligen Beichte einem Priester alle Unreinheiten unserer Seele offenbaren.

Es geht der Orthodoxen Tradition gerade nicht um eine spekulative Aussage *über* Gott, sondern deren Ziel ist die Teilhabe am Göttlichen Leben. Die Propheten, die Apostel und die Heiligen erfuhren diese Teilhabe und zeigten danach den Weg dazu auf. Dabei sprechen sie von Erfahrungen, die in menschlichen Worten nicht fassbar sind. Der Weg zu diesem Ziel des geistlichen Lebens ist zuerst die Erkenntnis, dass ich in meinem gegenwärtigen gefallenen

[2]Griechisch: νους (Nous). Dieses Wort hat in den westlichen Sprachen keine adäquate Entsprechung. Es beschreibt das Organ der Seele, mit welchem der Mensch mit Gott kommunizieren kann. Durch den Sündenfall ist dieses Organ erkrankt und bedarf der Heilung. In einem geistlich gesunden Menschen funktioniert es korrekt und befähigt ihn zur Teilhabe am Göttlichen Leben.

Zustand nicht in der Lage bin, die Gebote Gottes zu erfüllen. Daraufhin beschreite ich den Weg der Umkehr[3] und bin bereit, mit Gott zusammen an meiner eigenen Heilung des *Nous* zu arbeiten. Dies wird *Synergie* genannt. Die Kirche sieht sich dabei als Krankenhaus und wendet die vom Heiligen Geist inspirierten Heilmittel (die *Mysterien*) an. Christus selbst ist der Arzt. Der in der Kirche geheilte Mensch ist fähig, am Göttlichen Leben teilzuhaben und ist auch fähig zur selbstlosen Liebe gemäß göttlichem Gebot.

Die Orthodoxie ist also der Weg

$$\text{Läuterung} \preceq \text{Erleuchtung} \preceq \text{Verherrlichung}$$

mit $\preceq$ wie in Beispiel 5.1.22 definiert, der für jeden Menschen bereit steht und mit der Läuterung des Nous beginnt. Wenn dabei der Eindruck entsteht, dass man schon auf der Stufe der Erleuchtung ist, ist dies ein sicheres Zeichen der eigenen Verblendung und des tief Gefallenseins. Denn dann hat man nicht die wahre Demut erreicht. In jedem Fall ist es sehr empfehlenswert, mit einem geistlichen Vater über den eigenen geistlichen Zustand zu sprechen. Der geistliche Vater soll dabei danach ausgewählt werden, in wieweit er fortgeschritten ist auf dem Weg des geistlichen Kampfes, am besten selbst schon auf der Stufe der Erleuchtung stehend.

[3] gr. Metanoia (Reue)

Kapitel 7

Kryptographie

7.1 RSA-Kryptographie

RSA[1] ist ein asymmetrisches kryptographisches Verfahren zum Verschlüsseln oder zum digitalen Signieren.

7.1.1 Die eulersche Phi-Funktion

Die *eulersche Phi-Funktion phi*(n) gibt für jede natürliche Zahl n an, wieviele zu n teilerfremde Zahlen es zwischen 1 und n gibt:

$$\phi(n) := |\{a \in \mathbb{N} \mid 1 \leq a \leq n \text{ und } \mathrm{ggT}(a,n) = 1\}|$$

Die Phi-Funktion ist *schwach multiplikativ*, d.h. für teilerfremde m, n gilt:

$$\phi(m \cdot n) = \phi(m) \cdot \phi(n)$$

Beispielsweise ist

$$\phi(18) = \phi(2) \cdot \phi(9) = 1 \cdot 6 = 6$$

Bemerkung 7.1.1. *$\phi(n)$ ist die Anzahl der invertierbaren Elemente modulo n.*

Zur Berechnung von $\phi(n)$ lässt sich folgendes sagen:

Lemma 7.1.2. *Ist p eine Primzahl, so gilt:*

1. $\phi(p) = p - 1$
2. $\phi\left(p^k\right) = p^k \cdot \left(1 - \frac{1}{p}\right)$ *für* $k \geq 1$.

Beweis. 1. p ist zu allen Zahlen zwischen 1 und $p-1$ teilerfremd, aber nicht zu p. Dies sind genau $p - 1$ Zahlen.

2. p^k ist genau zu den Zahlen $p \cdot 1, p \cdot 2, \ldots, p \cdot p^{k-1}$ zwischen 1 und p^k nicht teilerfremd. Dies sind genau

$$p^k - p^{k-1} = p^k \cdot \left(1 - \frac{1}{p}\right)$$

Stück. □

[1] benannt nach R.L. Rivest, A. Shamir und L. Adleman

Es ergibt sich aus der Primfaktorzerlegung von n:

$$n = \prod_{p|n} p^{\alpha_p}$$

und der schwachen Multiplikativität die Berechnungsformel

$$\phi(n) = \prod_{p|n} p^{\alpha_p} \left(1 - \frac{1}{p}\right) = n \prod_{p|n} \left(1 - \frac{1}{p}\right)$$

Wichtig für die Kryptographie ist der Satz von Fermat-Euler:

Satz 7.1.3 (Fermat-Euler)**.** *Falls* $\mathrm{ggT}(a, n) = 1$*, so gilt:*

$$a^{\phi(n)} \equiv 1 \mod n$$

7.1.2 RSA-Kryptosystem

RSA ist ein asymetrisches kryptographisches Verfahren, das Paare von Schlüsseln verwendet. Der private Schlüssel ist zum Entschlüsseln oder Signieren von Daten, und der öffentliche Schlüssel ist zum Verschlüsseln oder Überprüfen von Signaturen. Der private Schlüssel wird geheim gehalten und ist nur sehr schwierig aus dem öffentlichen Schlüssel zu berechnen.

Der öffentliche Schlüssel (public key) ist ein Paar (e, N), und der private Schlüssel (private key) ist ein Paar (d, N). N heißt der *RSA-Modul*, e der *private Exponent* und d der *öffentliche Exponent*. Die Schlüssel werden folgendermaßen erzeugt:

1. Wähle zufällig und stochastisch unabhängig zwei Primzahlen $p \neq q$, für die in etwa gilt:

   $$0.1 < |\log_2 p - \log_2 q| < 30$$

   In der Praxis werden Zahlen der entsprechenden Länge erzeugt und mit einem Primzahltest geprüft.

2. Berechne den RSA-Modul $N = p \cdot q$ und die Eulersche Phi-Funktion

   $$\phi(N) = (p - 1) \cdot (q - 1)$$

3. Wähle eine zu $\phi(N)$ teilerfremde Zahl e mit $1 < e < \phi(N)$ als öffentlichen Exponenten.

4. Berechne den privaten Exponenten d als Lösung von

   $$e \cdot d \equiv 1 \mod \phi(N) \tag{7.1}$$

Bemerkung 7.1.4. *Die Kongruenz (7.1) wird mit dem erweiterten euklidischen Algorithmus (Satz 5.2.3) gelöst. Aus Effizienzgründen wird e nicht zu groß gewählt. Üblich ist die vierte Fermat-Zahl:*

$$e = 2^{16} + 1 = 65537$$

Kleiner sollte e nicht sein, um nicht weitere Angriffsmöglichkeiten zu bieten.

Das Verschlüsseln einer Nachricht m geht so:

$$c \equiv m^e \mod N$$

Der Geheimtext c wird dann an den Empfänger mit öffentlichen Schlüssel (e, N) verschickt. Dabei muss $1 < m < N$ gelten.

Der Geheimtext wird mit dem privaten Schlüssel (d, N) folgendermaßen entschlüsselt:

$$m \equiv c^d \mod N$$

Dies klappt aus folgendem Grund: es ist

$$1 = \text{ggT}(e, \phi(N)) = d \cdot e + k \cdot \phi(N)$$

Dies ergibt:

$$c^d \equiv m^{d \cdot e} \overset{(*)}{\equiv} m^{d \cdot e + k \cdot \phi(N)} \equiv m^1 \equiv m \mod N$$

wobei $(*)$ gilt wegen $m^{\phi(N)} \equiv 1 \mod N$ nach dem Satz von Fermat-Euler (Satz 7.1.3).

Beispiel 7.1.5. *Schlüsselerzeugung für Person B.*

1. *Wähle $p = 11$ und $q = 13$ als Primzahlen.*
2. *Der RSA-Modul ist $N = p \cdot q = 143$. $\phi(N) = 10 \cdot 12 = 120$.*
3. *Wähle $e = 23$: e ist teilerfremd zu N und kleiner als N.*
4. *Der erweiterte euklidische Algorithmus (Satz 5.2.3) liefert:*

$$1 = \text{ggT}(23, 120) = 23 \cdot d + k \cdot 120$$

mit $d = 47$ und $k = -9$. Also ist $d = 47$ ist private Exponent.

Der Absender A möchte an B eine Nachricht $m = 7$ verschlüsselt senden. Dazu berechnet A:

$$7^{23} \equiv 2 \mod 143$$

B entschlüsselt den Geheimtext $c = 2$:

$$2^{47} \equiv 7 \mod 143$$

Der Klartext ist also $m = 7$.

7.1.3 Binäre Exponentiation

Das Ver- und Entschlüsseln einer Nachricht m geschieht durch Exponenentiation. Ganzzahlige Potenzen können durch „fortgesetztes Quadrieren und gelegentliches Multiplizieren“ effizient berechnet werden. Dies funktioniert für reelle Zahlen, Matrizen, elliptische Kurven bzw. für beliebige *Halbgruppen*, d.h. wenn eine Verknüpfung das Assiziativgesetz befolgt.

Algorithmus 7.1.6. *1. Der Exponent k wird in seine Binärdarstellung umgewandelt.*

2. *Ersetze jede 0 durch Q und jede 1 durch QM.*
3. *Q bedeutet „Quadrieren“ und M bedeutet „Multiplizieren mit x“*

4. Wende die resultierende Zeichenkette von links nach rechts auf 1 *an.*

Für $k > 0$ beginnt die Binärdarstellung immer mit der Ziffer 1. Also ergibt sich stets am Anfang die Anweisung QM bzw. $1^2 \cdot x = x$. Deshalb kann die erste Anweisung QM durch x ersetzt werden.

Beispiel 7.1.7. *Sei $k = 23$. In Binärdarstellung ist $k = 10111$. Dies ergibt $QM\,Q\,QM\,QM\,QM$. Mit der Vereinfachung ergibt sich: $Q\,QM\,QM\,QM$ angewandt auf x. D.h.*

$$x^{23} = \left(\left(\left(x^2\right)^2 \cdot x\right)^2 \cdot x\right)^2 \cdot x$$

Bemerkung 7.1.8. *Beim Rechnen modulo N wird nach jedem Rechenschritt Q oder M jeweils modulo N reduziert.*

7.1.4 Padding

In der Praxis wird das oben beschriebene RSA-Verfahren nicht eingesetzt, da es mehrere Schwächen hat.

Zunächst ist das Verfahren deterministisch. Ein Angreifer kann also einen Klartext raten, ihn mit dem öffentlichen Schlüssel verschlüsseln und dann mit einem Chiffrat vergleichen. Legt der Angreifer eine große Tabelle von Klartext-Chiffrat-Paaren an, so hat er ein „Wörterbuch“ zur Hand, welches ihm beim Analysieren von verschlüsselten Nachrichten hilft.

Gilt $c = m^e < N$, so kann ein Angreifer die ganzzahlige e-te Wurzel von c ziehen und erhält den Klartext m.

Da das Produkt zweier Chiffrate selbst ein Chiffrat ist:

$$m_1^e \cdot m_2^e \equiv (m_1 \cdot m_2)^e \mod N$$

kann ein Angreifer ein Chiffrat $c \equiv m^e \mod N$ modifizieren zu $c' \equiv c \cdot r^e \mod N$ und den Empfänger bitten, den unverfänglichen Text c' zu entschlüsseln, was $m' \equiv m \cdot r \mod N$ ergibt. Mit dem erweiterten euklidischen Algorithmus (Satz 5.2.3) hat der Angreifer dann den Klartext $m \equiv m' \cdot r^{-1} \mod N$.

Angenommen, es wird dieselbe Nachricht m an e verschiedene Empfänger mit demselben öffentlichen Exponenten e, aber paarweise teilerfremden Moduli N_i versandt. Dann muss ein Angreifer nur die Kongruenzen

$$x \equiv m^e \mod N_i, \quad i = 1, \ldots, e$$

simultan lösen, was mit dem Chinesischen Restsatz (Satz 7.1.9) ein $x \equiv m^e \mod \prod N_i$ ergibt. Wegen $x < \prod N_i$, kann nun die ganzzahlige e-te Wurzel aus m^e gezogen werden, um den Klartext m zu berechnen.

Satz 7.1.9 (Chinesischer Restsatz)**.** *Seien $m_1, \ldots, m_e$ paarweise teilerfremde Zahlen. Dann existiert für jedes Tupel $a_1, \ldots, a_e$ ganzer Zahlen eine ganze Zahl x, welche die simultane Kongruenz*

$$x \equiv a_i \mod m_i, \quad i = 1, \ldots, e$$

erfüllt. Alle Lösungen dieser Kongruenz sind kongruent modulo $M := m_1 \cdots m_e$.

Beweis. Für jedes i sind m_i und $M_i := M/m_i$ teilerfremd. Daher existieren nach Satz 5.2.3 zwei Zahlen r_i und s_i mit

$$1 = r_i m_i + s_i M i$$

Setze $e_i := s_i M_i$. Dann gilt:

$$\begin{aligned} e_i &\equiv 1 \mod m_i \\ e_i &\equiv 0 \mod m_j, \quad j \neq i \end{aligned}$$

Die Zahl $x := \sum_{i=1}^{e} a_i e_i$ ist dann eine Lösung der simultanen Kongruenz. □

Um derartige Angriffe zu verhindern, wird der Klartext durch eine Zeichenfolge R mit vorgegebener Struktur ergänzt, die eine Randomisierung beinhaltet *(Padding)*. Es wird also nicht die Nachricht M, sonder der Klartext mit angehängtem R verschlüsselt, was bei geeigneter Wahl einer Padding-Methode Angriffe erschwert. Zur Berechnung von R kommen oft auch Zufallszahlen zum Einsatz.

7.1.5 Sicherheit von RSA

Die Sicherheit des RSA-Kryptosystems beruht auf zwei mathematischen Problemen:

1. Faktorisierung großer Zahlen
2. RSA-Problem

Das RSA-Problem besagt: Für gegebenes $m^e \mod N$ und ein Paar (e, N) bestimme m. Ziehe also die e-te Wurzel modulo zusammengesetztem N. Am vielversprechendsten scheint der Ansatz zu sein, N zu faktorisieren: Hat ein Angreifer die Zerlegung $N = p \cdot q$, so berechnet er $\phi(N) = (p-1)(q-1)$ und kann aus e mit dem erweiterten euklidischen Algorithmus den privaten Exponenten d effizient berechnen. Bisher wurde allerdings kein Algorithmus zur Faktorisierung einer ganzen Zahl auf herkömmlichen Computern gefunden, der in polynomieller Zeit abläuft. Zurzeit wird empfohlen, dass N mindestens 2048 Bit lang sein soll, um die Dauer einer Faktorisierung hinreichend lang zu halten.

Auf einem Quantencomputer sieht die Sache anders aus: 1994 entwickelte Peter Shor einen Quanten-Algorithmus, der natürliche Zahlen in polynomieller Zeit faktorisieren kann. Diese Methode, falls sie eines Tages implementiert werden kann, macht RSA also unsicher.

Aus Effizienzgründen wird RSA oft als Teil eines hybriden Kryptosystems genutzt. Die eigentliche Nachricht wird dabei mit einem *symmetrischen Verschlüsselungsverfahren* verschlüsselt, wobei zum Ver- und Entschlüsseln derselbe Schlüssel benutzt wird. RSA wird dabei genutzt, um den Schlüssel auszutauschen. Dieses System kommt z.B. beim TLS-Protokoll im Internet zum Einsatz.

Da im Laufe der Zeit Schlüssellängen immer größer werden bei gleich bleibender Sicherheit, wird allmählich RSA durch Kryptographie mit *elliptischen Kurven* ersetzt.

7.1.6 Ein Eine-Million-Dollar-Problem

Eine weitere Gleichung, die Homer Simpson in der dritten Dimension in *Homer*^{3} erscheint, ist

$$P = NP$$

Es handelt sich hierbei um eine Antwort auf das *P-versus-NP-Problem* der Informatik. P ist die Klasse der Probleme, für die es einen Algorithmus gibt, der es in polynomieller Zeit löst. NP ist die Klasse aller Probleme, für die eine Antwort in polynomieller Zeit überprüft werden kann. NP steht für *non-deterministic polynomial time*.

Ein Algorithmus läuft in *polynomieller Zeit* ab, wenn seine Laufzeit $T(n)$ nach oben durch ein Polynom in der Größe n der Eingabedaten beschränkt ist, z.B. heißt

$$T(n) = O(n^2)$$

dass die Laufzeit schlimmstenfalls quadratisch in n ist.

Ein Beispiel für ein Problem, das in NP liegt, ist die Faktorisierung natürlicher Zahlen. Es kann in polynomieller Zeit überprüft werden, ob eine gegebene Faktorisierung einer Zahl stimmt. Allerdings ist kein polynomieller Faktorisierungsalgorithmus auf einem herkömlichen Rechner bekannt. Falls $P = NP$ gilt, dann lässt sich jedes Problem, dass in polynomieller Zeit verifizierbar ist, auch in polynomieller Zeit lösen. So muss es dann auch einen Algorithmus zur Faktorisierung natürlicher Zahlen geben, der in polynomieller Zeit abläuft. Bei $P \neq NP$ ist dies nicht zwingend der Fall.

Das P-versus-NP-Problem ist eines der sieben Millenium-Probleme, die das Clay Mathematics Institute im Jahr 2000 ausgeschrieben hat. Wer eines davon zuerst löst, bekommt eine Million Dollar Preisgeld. Bisher wurde ein Millenium-Problem gelöst, und zwar im Jahr 2002 von G.J. Perelman die *Poincaré-Vermutung*, der jedoch den Preis ablehnte.

7.2 Elliptische-Kurven-Kryptographie

RSA-Kryptographie wird derzeit almählich durch Kryptographie mit *elliptischen Kurven* ersetzt, da diese bei gleicher Schlüssellänge eine höhere Sicherheit bieten. Diese Art von Kryptographie beruht auf dem Diffie-Hellman-Schlüsselaustausch, der im folgenden Unterabschnitt vorgestellt wird.

7.2.1 Diffie-Hellman-Schlüsselaustausch

Für den Schlüsselaustausch von Diffie und Hellman wird eine *abelsche Gruppe* $(G, +)$ benötigt, bei welcher das *diskrete Logarithmus-Problem* (DLP) schwierig zu lösen ist. Die Bedeutung der beiden Begriffe wird sogleich erklärt:

Das diskrete Logarithmus-Problem (DLP). *Gegeben seien in einer abelschen Gruppe die Elemente P und $n \cdot P$, wobei n eine natürliche Zahl sei. Bestimme nun n.*

Definition 7.2.1. *Eine* abelsche Gruppe *ist ein Paar $(G, +)$, wobei G eine Menge und*

$$+ \colon G \times G \to G$$

eine Abbildung (die Gruppenaddition*) sei, für die folgende Bedingungen gelten:*

1. $(a + b) + c = a + (b + c)$ *(Assoziativgesetz)*
2. *Es gibt ein Element $0 \in G$, sodass stets $a + 0 = 0 + a = a$ gilt.* *(Nullelement)*
3. *Zu jedem $a \in G$ existiert ein $-a \in G$ mit $a + (-a) = (-a) + a = 0$.* *(Inverse)*
4. $a + b = b + a$ *(Kommutativgesetz)*

Hierbei seien a, b, c beliebige Elemente aus G.

Beispiel 7.2.2. *Beispiele für abelsche Gruppen sind $(\mathbb{Z}, +)$ und $(K, +)$ sowie $(K \setminus \{0\}, \cdot)$ mit $K \in \{\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{F}_p\}$.*

Zur Schreibweise. Die Gruppenaddition wird meist mit dem Symbol + bezeichnet. Dann wird oft vereinfacht geschrieben:

$$nP := \underbrace{P + \cdots + P}_{n \text{ mal}}$$

wobei $n \in \mathbb{N}$ ist.

Der Schlüsselaustausch. Öffentlich ist ein Element $P \in G$.

Alice denkt sich eine geheime Zahl $n \in \mathbb{N}$ aus und veröffentlicht nP.

Bob denkt siche eine geheime Zahl $m \in \mathbb{N}$ aus und veröffentlicht mP.

Alice berechnet $n(mP) = nmP = Q$.

Bob berechnet $m(nP) = mnP = Q$.

Q ist der geheime Schlüssel.

Sicherheit. Die Sicherheit des Diffie-Hellmann-Schlüsselaustauschs beruht auf der Schwierigkeit des diskreten Logarithmus-Problems. Denn findet ein Angreifer die Zahlen n und m heraus, so hat er auch den Schlüssel

$$Q = nmP$$

Beispiel 7.2.3. *Die abelsche Gruppe $(\mathbb{F}_p^\times, \cdot)$ mit*

$$\mathbb{F}_p^\times := \{1, \ldots, p-1\}$$

für p prim hat die besondere Eigenschaft, dass sie zyklisch *ist. Dies bedeutet, dass jedes $a \in \mathbb{F}_p^\times$ eine Potenz eines festen Elements $t \in \mathbb{F}_p^\times$ ist:*

$$a = t^n$$

Das Element t heißt ein Erzeuger *der zyklischen Gruppe. Das diskrete Logarithmus-Problem bedeutet hier: gegeben sei ein Erzeuger t und $a = t^n$. Bestimme n.*

Endliche Körper. Außer den endlichen Körpern $\mathbb{F}_p$ gibt es noch weitere endliche Körper. Diese werden folgendermaßen konstruiert: Es sei

$$\mathbb{F}_p[t] := \{\text{Polynome in } t \text{ mit Koeffizienten aus } \mathbb{F}_p\}$$

Ein Polynom $\pi \in \mathbb{F}_p[t]$ heißt *irreduzibel*, wenn $\deg(\pi) > 0$ und es nur triviale Zerlegungen hat:

$$\pi = f \cdot g \quad \Rightarrow \quad f \in \mathbb{F}_p^\times \quad \text{oder} \quad g \in \mathbb{F}_p^\times$$

Gegeben sei ein irreduzibles Polynom $\pi \in \mathbb{F}_p[t]$, und es sei $n := \deg(\pi)$. Dann ist

$$\mathbb{F}_{p^n} := \mathbb{F}_p[t]/\pi\mathbb{F}_p[t] := \{\text{Reste von } f \in \mathbb{F}_p[t] \text{ modulo } \pi\}$$

mit der Addition und Multiplikation von Polynomen modulo π ein Körper.

Die erste Beobachtung ist, dass jedes Element $a \in \mathbb{F}_{p^n}$ eine Linearkombination der Elemente $1, t, \ldots, t^{n-1}$ ist. Dies bedeutet, dass $\mathbb{F}_{p^n}$ ein Vektorraum über $\mathbb{F}_p$ der Dimension n ist. Folglich hat der Körper $\mathbb{F}_{p^n}$ genau p^n Elemente.

Die Konstruktion liefert auch alle möglichen endlichen Körper. Denn es gilt:

Satz 7.2.4. *Jeder endliche Körper ist isomorph zu einem der Körper $\mathbb{F}_{p^n}$.*

Beispiel 7.2.5. *Das Polynom $\pi = t^2 + t + 1 \in \mathbb{F}_2[t]$ ist irreduzibel, da es keine Nullstellen in $\mathbb{F}_2$ hat (Nullstelle ist gleichbedeutend mit Linearfaktor). Also ist*

$$\mathbb{F}_{2^2} = \mathbb{F}_2 \cdot 1 + \mathbb{F}_2 t = \{0, 1, t, t+1\}$$

Die Multiplikationstabelle ist

$\cdot$	1	t	$t+1$
1	1	t	$t+1$
t	t	$t+1$	1
$t+1$	$t+1$	1	t

Charakteristik eines Körpers. Die *Charakteristik* eines Körpers K ist die kleinste positive natürliche Zahl n, für die gilt

$$n \cdot 1_K = 0_K$$

wobei 1_K und 0_K das Einselement bzw. das Nullelement von K seien. Falls keine derartige natürliche Zahl existiert, so definiert man dass die Charakteristik von K Null sei. Die Charakteristik von K wird als $\mathrm{char}(K)$ geschrieben.

Satz. *Die Charakteristik eines Körpers ist entweder Null oder eine Primzahl.*

Beweis. Sei $n = \mathrm{char}\, K > 0$, und sei $n = a \cdot b$ eine Zerlegung. Beachte, dass $a \neq 0$ und $b \neq 0$ ist. Dann gilt

$$a \cdot 1_K \cdot b \cdot 1_K = 0_K$$

wobei $a \cdot 1_K \neq 0_K$ und $b \cdot 1_K \neq 0_K$, was in einem Körper nicht sein kann. Denn sonst würde in K gelten:

$$b \cdot 1_K = baa^{-1} \cdot 1_K = b \cdot 1_K \cdot a \cdot 1_K \cdot a^{-1} \cdot 1_K = 0_K \cdot a^{-1} \cdot 1_K = 0_K$$

ein Widerspruch. □

Beispiel. *Es gilt:*

$$\mathrm{char}\,(\mathbb{F}_{p^n}) = p$$

Bemerkung 7.2.6. *Die multiplikative Gruppe $(\mathbb{F}_{p^n}^{\times}, \cdot)$ ist zyklisch. Dies bedeutet, dass sie ebenfalls für den Diffie-Hellman-Schlüsselaustausch interessant ist.*

7.2.2 Elliptische Kurven

Historisch sind elliptische Kurven aus dem Versuch der Berechnung der Bogenlänge einer Ellipse hervorgegangen. Wir werden im Folgenden diesen Weg nachzeichnen.

Elliptische Integrale. Die Bogenlänge L einer Ellipse ist als folgendes Integral darstellbar:

$$L = 4a \int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 t}\, dt$$

mit $k = \frac{\sqrt{a^2-b^2}}{a}$ und den Halbachsen a und b. Das Integral ist ein Beispiel eines *elliptischen Integrals 1. Art*:

$$E(\phi) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2 t}\, dt$$

Ein solches Integral kommt auch in den Vorlesungen *Kartenprojektionslehre* (Master) sowie *Grundlagen kinematischer und dynamischer Modelle der Geodäsie* (Bachelor) vor. Mit der Substitution $x = \sin t$ wird daraus

$$E(u) = \int_0^{u} \frac{1 - k^2 x^2}{\sqrt{(1-x^2)(1-k^2x^2)}}\, dx$$

Ein allgemeines *elliptisches Integral* ist gegeben als

$$f(x) = \int \frac{A(x) + B(x)}{C(x) + D(x)\sqrt{S(x)}}\, dx$$

wobei A, B, C, D Polynome und S ein Polynom von Grad 3 oder 4 sind.

Beispiel.

$$u = f(x) = \int_0^{x} \frac{dt}{\sqrt{1-t^2}} = \arcsin(x)$$

ist ein elliptisches Integral. Abel erkannte, dass es besser ist die Umkehrfunktion zu betrachten. In diesem Beispiel ist diese $\sin(x)$*, eine periodische Funktion.*

Elliptische Funktion. Eine elliptische Funktion p ist die Umkehrfunktion eines elliptischen Integrals 2. Art. Für $k \neq 0$ sind elliptische Funktionen *doppelt periodisch*:

$$p(u + m\alpha) = p(u + n\beta) = p(u)$$

für gewisse $\alpha, \beta \in \mathbb{C}$ mit $\frac{\alpha}{\beta} \notin \mathbb{R}$. Eisenstein erkannte umgekehrt, dass doppelt periodische Funktionen elliptisch sind.

Die allgemeine Form einer elliptischen Funktion ist

$$f(z) = \sum_{m,n \in \mathbb{Z}} (z + m\omega_1 + n\omega_2)^{-2}$$

mit den Perioden $\omega_1, \omega_2 \in \mathbb{C}$, $\frac{\omega_1}{\omega_2} \notin \mathbb{R}$. Die Funktion

$$y(z) = \sum_{m,n \in \mathbb{Z}} (z + m\omega_1 + n\omega_2)^{-2} - \sum_{m,n \in \mathbb{Z}\setminus\{0\}} (m\omega_1 + n\omega_2)^{-2}$$

erfüllt eine Differenzialgleichung der Form

$$y'(z)^2 = p(y(z))$$

wobei $p(X)$ ein Polynom von Grad 3 mit nur einfachen Nullstellen ist.

Weierstraß-$\wp$-Funktion. Die Weierstraß-$\wp$-Funktion ist

$$\wp(z) = z^{-2} + \sum_{m,n\in\mathbb{Z}\setminus\{0\}} (z + m\omega_1 + n\omega_2)^{-2} - (m\omega_1 + n\omega_2)^{-2}$$

Diese erfüllt die Differenzialgleichung

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$$

mit Koeffizienten g_2, g_3. Setzt man $x := \wp(z)$ und $y := \wp'(z)$, so erhält man die Gleichung

$$E\colon y^2 = 4x^3 - g_2x - g_3$$

einer *elliptischen Kurve*. Auf Grund der Konstruktion kann man einsehen, dass die elliptische Kurve E isomorph zu $\mathbb{C}/\Lambda$ mit dem Gitter

$$\Lambda = \{m\omega_1 + n\omega_2 \mid m, n \in \mathbb{Z}\}$$

eine abelsche Gruppe ist.

Elliptische Kurven über einem Körper K. Formal ist eine elliptische Kurve über dem Körper K eine nicht-singuläre projektive algebraische Kurve von Geschlecht 1. Die Gleichung

$$E\colon y^2 = x^3 + ax + b$$

mit $a, b \in K$ ist die *Weierstraß-Normalform* und gilt falls $\mathrm{char}(K) \neq 2, 3$ ist. Das Polynom $f(x) = x^3 + ax + b$ hat keine doppelte Nullstelle. Auf einer elliptischen Kurve kann wie folgt geometrisch definiert werden: Zunächst sind die *K-rationalen* Punkte gegeben als

$$E(K) = \left\{(x, y) \in K^2 \mid y^2 = x^3 + ax + b\right\} \cup \{O\}$$

mit dem Punkt O „im Unendlichen“. Dieser Punkt kann in der projektiven Ebene $\mathbb{P}^2$ über die Homogenisierung der Gleichung für E gefunden werden:

$$y^2z = x^3 + axy^2 + bz^3$$

und für $z = 0$ ergibt sich $x = 0$ und $y = 1$ (projektive Koordinaten!). Also ist $O = (0 : 1 : 0)$.

Als nächstes wird beobachtet, dass E symmetrisch zur x-Achse ist:

$$P = (x, y) \in E(K) \Rightarrow -P := (x, -y) \in E(K)$$

Definiere noch $-O := O$. Weiter seien $P, Q \in E(K)$. Dann schneidet die Gerade L durch P und Q die Kurve E in einem dritten Punkt $R \in E(K)$. Also kann definiert werden:

$$\begin{aligned} P + Q &:= Q + P := -R \\ P + O &:= O + P := P \\ P + (-P) &:= O \end{aligned}$$

Falls $P = Q$, so sei L die Tangente an P. Falls L einen zweiten Punkt $R \in E(K)$ trifft, so sei dann

$$2P := -R$$

andernfalls sei

$$2P = -P$$

Es gilt:

Satz. *$(E(K), +)$ ist eine abelsche Gruppe mit dem Nullelement O.*

Elliptische Kurven über $\mathbb{F}_{p^n}$. Sei nun $K = \mathbb{F}_{p^n}$. Es gilt mit $q = p^n$:

Satz (Hasse-Schranke)**.** *Für die Anzahl $|E(K)|$ der K-rationalen Punkte auf einer elliptischen Kurve E gilt:*

$$||E(K)| - (q+1)| \leq 2\sqrt{q}$$

Folglich gibt es für große n in etwa $q = p^n$ K-rationale Punkte auf der elliptischen Kurve E. Dies ist für die Erzeugung von Schlüsseln interessant. Weiter gilt: $E(K)$ ist zyklisch oder Produkt zweier zyklischer Gruppen. Dies ist für den Schlüsselaustausch von Bedeutung, da der öffentliche Punkt P ein Element von möglichst hoher Ordnung sein soll:

$$nP = O \quad \text{mit minimalem } n > 0 \text{ möglichst groß}$$

ECDLP.

- Es gibt zahlreiche geeignete elliptische Kurven über endlichen Körpern.
- Das elliptische-Kurven-diskrete-Logarithmus-Problem (ECDLP) ist schwerer als die Faktorisierung natürlicher Zahlen oder als das DLP in $\mathbb{F}_q^\times$ mit $q = p^n$.
- Bestmögliche Körper sind $K = \mathbb{F}_p$ (p prim) oder $K = \mathbb{F}_{2^n}$.

7.3 Quantenkryptographie

Da es in naher Zukunft leistungsfähige Quantenrechner geben wird, wird die RSA- oder elliptische-Kurven-Verschlüsselung unsicher werden. Ein Ausweg bietet die *Quantenkryptographie*. Bei der Quantenkryptographie werden polarisierte Photonen verwendet, und die Gesetze der Quantenmechanik garantieren, dass ein sicherer Schlüsselaustausch stattfinden kann.

Im BB84-Protokoll werden zwei Paare orthogonaler Polarisationszustände verwendet:

- die *rektilineare Basis*: 0° und 90°
- die *diagonale Basis*: 45° und 135°

Aus den Gesetzen der Quantenmechanik folgt, dass keine Messung die 4 verschiedenen Zustände unterscheiden kann, da sie nicht alle orthogonal zueinander sind. Denn eine Messung wählt eine Orthonormalbasis aus und das Messergebnis ist dann einer dieser orthogonalen Zustände. Beispielsweise können in der rektilinearen Basis nur die Zustände „horizontal“ oder „vertikal“ gemessen werden. Nach einer Messung ist das Photon im gemessenen Zustand, d.h. es findet eine Veränderung durch die Messung statt.

Das BB84-Protokoll.

1. Alice legt eine Kodierungstabelle an, z.B.

	0	1
+	↑	→
×	↗	↘

2. Alice erzeugt ein zufälliges Bit (0 oder 1) und wählt zufällig eine ONB aus (rektilinear oder diagonal) und sendet Bob ein Photon im entsprechenden Zustand. Diesen Prozess wiederholt sie mehrfach.

3. Bob wählt jeweils zufällig eine Basis und misst den Zustand des Photons.

4. Alice und Bob vergleichen ihre Folgen von Basen. Bei Übereinstimmung behalten sie jeweils das entsprechende Bit, andernfalls verwerfen sie es.

In etwa 50% der Fälle haben Alice und Bob ein gemeinsames Bit. Die Folge dieser beibehaltenen Bits ist der gemeinsame Schlüssel. Um festzustellen, ob sie belauscht wurden, vergleichen Alice und Bob eine ausgewählte Teilfolge ihrer Versionen des Schlüssels. Falls Eva Informationen über die Polarisierungen erlangt hat, muss es zu Übertragungsfehlern gekommen sein. Falls zuviele Bits unterschiedlich sind, verwerfen sie ihren Schlüssel und wiederholen die Prozedur auf einem anderen Quantenkanal.

Beispiel 7.3.1. *Nehmen wir an, dass Alice die Basis + wählt, und dass ihr Photon die Polarierung $\rightarrow$ hat. Wählt Bob auch +, dann wird er $\rightarrow$ messen, und ihrer beiden Bits stimmen überein. Wählt er jedoch $\times$, so wird er entweder $\nearrow$ oder $\searrow$ mit jeweils Wahrscheinlichkeit $\frac{1}{2}$ messen. Also gibt es eine Wahrscheinlichkeit von 50%, dass ihre Bits nicht übereinstimmen.*

Bemerkung 7.3.2. *Nehmen wir an, dass Eva die Verbindung belauscht und ein Photon abfängt. Sie weiss nicht, welche ONB für dessen Polarisierung verwendet wurde. Also wählt sie zufällig eine und führt eine Messung durch. Falls die beiden ONB nicht übereinstimmen, wird ihre Messung den Zustand des Photons ändern. Dadurch kann sich ein Fehler bei der Übertragung ergeben, der von Alice und Bob bemerkt wird.*

Kapitel 8

Approximation

Bei der Approximation geht es darum, eine (auch unbekannte) Funktion f durch eine einfachere Funktion (z.B. Polynom) zu approximieren. Wir werden uns mit der *linearen Approximation* beschäftigen, bei der f durch eine Linearkombination vorgegebener linear unabhängiger Funktionen $f_1, \ldots, f_n$ approximiert werden soll:

$$f \approx \sum_{i=1}^{n} \gamma_i f_i$$

Hierbei spannen die Funktionen $f_1, \ldots, f_n$ einen Untervektorraum U des Raumes $V = C[a, b]$ der stetigen Funktionen auf dem Intervall $[a, b]$ auf. Die Approximationsaufgabe besteht also darin, f so gut als möglich durch ein Element eines vorgegebenen Untervektorraums U von V zu approximieren.

Beispiel. *Beispiele für vorgegebene Unterräume von $V = C[a, b]$ sind solche, die von Polynomen, trigonometrischen Funktionen, Exponentialfunktionen oder rationalen Funktionen aufgespannt werden. Z.B. von*

1. $f_1 = 1,\ f_2 = x,\ f_3 = x^2, \ldots,\ f_n = x^{n-1}$.
2. $f_1 = 1,\ f_2 = \cos x,\ f_3 = \sin x,\ f_4 = \cos 2x,\ f_5 = \sin 2x,\ \ldots$
3. $f_1 = 1,\ f_2 = e^{\alpha_1 x},\ f_2 = e^{\alpha_2 x}, \ldots$
4. $f_1 = 1,\ f_2 = \frac{1}{(x-a_1)^{p_1}},\ f_3 = \frac{1}{(x-a_2)^{p_2}}, \ldots$

8.1 Beste Approximation

Um eine Approximation bewerten zu können oder einen Approximationsfehler angeben zu können, wird eine *Norm* auf dem Vektorraum V verwendet.

Sei $K = \mathbb{R}$ oder $K = \mathbb{C}$ und V ein K-Vektorraum.

Definition 8.1.1. *Eine* Norm *auf V ist eine Funktion $\|\cdot\| \colon V \to \mathbb{R}_{\geq 0}$ mit folgenden Eigenschaften:*

1. $\|f\| = 0$ *genau dann, wenn* $f = 0$.
2. $\|\alpha f\| = |\alpha| \|f\|$ *für* $\alpha \in K$
3. $\|f + g\| \leq \|f\| + \|g\|$

Eine Norm definiert stets eine Metrik auf V durch:

$$d(f,g) := \|f-g\|$$

Überprüfen wir noch, dass d tatsächlich eine Metrik ist:

Beweis. 1. $d(f,g) = 0$ gilt genau dann, wenn $\|f-g\| = 0$. Dies gilt genau dann, wenn $f-g = 0$. Genau dann ist $f = g$. Dies zeigt die Positivität.

2. Die Symmetrie ergibt sich aus:

$$d(f,g) = \|f-g\| = \|(-1)(g-f)\| = |-1|\|g-f\| = \|g-f\| = d(g,f)$$

3. Die Dreiecksungleichung gilt wegen:

$$d(f,h) = \|f-h\| = \|f-g+g-h\| \leq \|f-g\| + \|g-h\| = d(f,g) + d(g,h)$$

□

Beispiel 8.1.2. *Beispiele für Normen auf $V = C[a,b]$ sind:*

1. *L^1-Norm:*
$$\|f\|_1 := \int_a^b |f(t)|\,dt$$

2. *L^2-Norm:*
$$\|f\|_2 := \left(\int_a^b |f(t)|^2\,dt\right)^{\frac{1}{2}}$$

3. *L^∞-Norm:*
$$\|f\|_\infty := \max_{t\in[a,b]} |f(t)|$$

Nun können wir eine *beste Approximation* aus dem Untervektorraum U an f definieren. Es handelt sich dabei um ein $\hat{\phi} \in U$ mit

$$d(f,\hat{\phi}) = \min_{\phi\in U} d(f,\phi)$$

Satz 8.1.3 (Existenzsatz)**.** *Zu jeder Funktion $f \in V = C[a,b]$ und jedem endlich-dimensionalen Untervektorraum U von V und jeder Norm $\|\cdot\|$ auf V existiert mindestens eine beste Approximation $\hat{\phi} \in U$ an f.*

8.2 Gauß-Approximation

Wir versehen nun $V = C[a,b]$ mit der L^2-Norm. Eine beste Approximation bezüglich der L^2-Norm heißt *beste L^2-Approximation.*

Im Folgenden nutzen wir aus, dass die L^2-Norm von einem Innenprodukt auf V herkommt. Es ist nämlich

$$\|f\|_2 = \sqrt{\langle f,f\rangle}$$

wobei $\langle \cdot, \cdot \rangle$ das Standardinnenprodukt auf V ist:

$$\langle f, g \rangle = \int_a^b f(t)\overline{g(t)}\, dt$$

Sei U ein endlich-dimensionaler Untervektorraum von V und $\hat{\phi} \in U$. Wir betrachten nun den Approximationsfehler für die Approximation von $f \in V$ durch $\hat{\phi}$. Es gilt:

Lemma 8.2.1. *Genau dann ist $\hat{\phi} \in U$ beste L^2-Approximation an f, wenn $f - \hat{\phi}$ orthogonal zu U ist.*

Beweis. $\Leftarrow$. Sei $f - \hat{\phi}$ orthogonal zu U. Dann gilt mit $\phi \in U$ beliebig:

$$\left\| f - \hat{\phi} \right\|^2 = \langle f - \hat{\phi}, f - \hat{\phi} \rangle = \langle f - \hat{\phi}, f - \phi + \phi - \hat{\phi} \rangle = \langle f - \hat{\phi}, f - \phi \rangle + \underbrace{\langle f - \hat{\phi}, \phi - \hat{\phi} \rangle}_{=0}$$

$$= \langle f - \hat{\phi}, f - \phi \rangle \leq \left\| f - \hat{\phi} \right\| \|f - \phi\|$$

Dabei gilt $\langle f - \hat{\phi}, \phi - \hat{\phi} \rangle = 0$, da $\phi - \hat{\phi} \in U$ ist. Die letzte Ungleichung ist die Cauchy-Schwarz-Ungleichung. Es folgt, dass

$$\left\| f - \hat{\phi} \right\| \leq \min_{\phi \in U} \|f - \phi\|$$

$\Rightarrow$. Sei $K = \mathbb{R}$. Ist $\hat{\phi} \in U$ beste L^2-Approximation und $\phi \in U$, so hat

$$F_\phi(t) := \left\| f - \hat{\phi} - t\phi \right\|^2$$

in $t = 0$ ein Minimum. Dann ist

$$0 = \frac{d}{dt} F_\phi(t)\Big|_{t=0} = \frac{d}{dt} \left\| f - \hat{\phi} - t\phi \right\|^2 \Big|_{t=0} = 2 \left\langle f - \hat{\phi} - t\phi, \phi \right\rangle \Big|_{t=0}$$

wegen

$$\frac{d}{dt} \langle a + bt, a + bt \rangle = \frac{d}{dt} \left(\langle a, a \rangle + 2t\langle a, b \rangle + t^2 \langle b, b \rangle \right) = 2\langle a, b \rangle + 2t\langle b, b \rangle = 2\langle a + bt, b \rangle$$

Es folgt, dass für jedes $\phi \in U$ gilt:

$$\langle f - \hat{\phi}, \phi \rangle = 0$$

d.h. $f - \hat{\phi}$ orthogonal zu U ist.

Sei nun $K = \mathbb{C}$. Falls $f - \hat{\phi}$ nicht orthogonal zu U ist, so existiert ein $\psi \in U$ mit

$$\langle f - \hat{\phi}, \psi \rangle \neq 0$$

Ohne Einschränkung sei $\langle f - \hat{\phi}, \psi \rangle < 0$. Ansonsten ersetze ψ durch $e^{i\alpha}\psi$ für geeignetes Argument α. Für $0 < t << 1$ gilt nun:

$$\left\| f - \hat{\phi} + t\psi \right\|^2 = \langle f - \hat{\phi} + t\psi, f - \hat{\phi} + t\psi \rangle$$

$$= \langle f - \hat{\phi}, f - \hat{\phi} \rangle + \underbrace{t\langle f - \hat{\phi}, \psi \rangle + t\langle \psi, f - \hat{\phi} \rangle + t^2 \langle \psi, \psi \rangle}_{<0}$$

$$< \langle f - \hat{\phi}, f - \hat{\phi} \rangle = \left\| f - \hat{\phi} \right\|^2$$

Dies bedeutet, dass $\hat{\phi} - t\psi \in U$ eine bessere L^2-Approximation als $\hat{\phi}$ ist, also $\hat{\phi}$ nicht beste L^2-Approximation ist. $\square$

Hieraus ergibt sich die Eindeutigkeit:

Satz 8.2.2 (Gauß-Approximation)**.** *Zu $f \in V = C[a,b]$ und endlich-dimensionalem Untervektorraum U von V gibt es genau eine beste L^2-Approximation $\hat{\phi} \in U$.*

Beweis. Seien $\hat{\phi}_1, \hat{\phi}_2 \in U$ beste L^2-Approximationen an f. Dann gilt für alle $\phi \in U$:

$$\langle f - \hat{\phi}_1, \phi \rangle = \langle f - \hat{\phi}_2, \phi \rangle = 0$$

Dann gilt auch:

$$0 = \langle f - \hat{\phi}_2 - (f - \hat{\phi}_1), \phi \rangle = \langle \hat{\phi}_1 - \hat{\phi}_2, \phi \rangle$$

Also auch wegen $\hat{\phi}_1 - \hat{\phi}_2 \in U$:

$$0 = \langle \hat{\phi}_1 - \hat{\phi}_2, \hat{\phi}_1 - \hat{\phi}_2 \rangle = \left\| \hat{\phi}_1 - \hat{\phi}_2 \right\|^2$$

Also $\phi_1 = \phi_2$. Dies beweist die Eindeutigkeit.

Die Existenz besagt schon Satz 8.1.3. □

Was noch fehlt, ist eine Methode, mit der die beste L^2-Approximation berechnet werden kann. Dazu gehen wir von einer Basis $f_1, \ldots, f_n$ von U aus. Die beste L^2-Approximation $\hat{\phi}$ hat also die Form

$$\hat{\phi} = \sum_{k=1}^{n} \gamma_k f_k$$

Für den Approximationsfehler $f - \hat{\phi}$ ergibt sich mit $\phi \in U$:

$$0 = \langle f - \hat{\phi}, \phi \rangle = \left\langle f - \sum_{k=1}^{n} \gamma_k f_k, \phi \right\rangle = \langle f, \phi \rangle - \sum_{k=1}^{n} \gamma_k \langle f_k, \phi \rangle$$

Dies bedeutet, dass die Koeffizienten $\gamma_1, \ldots, \gamma_n$ das lineare Gleichungssystem

$$\sum_{k=1}^{n} \langle f_k, f_\ell \rangle \gamma_k = \langle f, f_\ell \rangle, \quad \ell = 1, \ldots, n \tag{8.1}$$

lösen. Hierbei ist die Koeffizientenmatrix

$$A = (\langle f_k, f_\ell \rangle)$$

hermitesch bzw. symmetrisch. Sie ist sogar positiv definit. Denn für $g = (\gamma_k)$ gilt:

$$g^* A g = \sum_{k,\ell=1}^{n} \bar{\gamma}_\ell \gamma_k \langle f_k, f_\ell \rangle = \langle \hat{\phi}, \hat{\phi} \rangle = \|\phi\|^2$$

Die Gleichungen (8.1) heißen *Normalgleichungen*. Deren eindeutig bestimmte Lösung ergibt die beste L^2-Approximation. Idealerweise wird für U eine Orthonormalbasis $f_1, \ldots, f_n$ herangezogen. Dann folgt aus den Normalgleichungen

$$\gamma_k = \langle f, f_k \rangle$$

d.h. die beste Approximation ist in diesem Fall durch

$$\hat{\phi} = \sum_{k=1}^{n} \langle f, f_k \rangle f_k$$

gegeben.

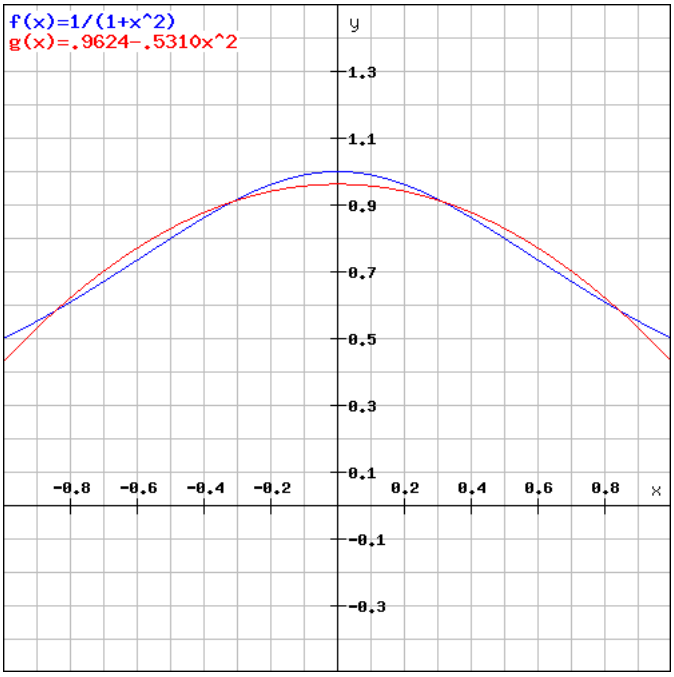


Abbildung 8.1: Eine Funktion und ihre beste L^2-Approximation durch quadratische Polynome.

Beispiel 8.2.3. *Gesucht ist die beste L^2-Approximation an*

$$f\colon [-1,1] \to \mathbb{R}, \quad x \mapsto \frac{1}{1+x^2}$$

durch quadratische Polynome.

Eine Basis für den Untervektorraum U von $C[-1,1]$ aller quadratischen Polynome ist durch

$$f_1 = 1,\ f_2 = x,\ f_3 = x^2$$

gegeben. Die Normalgleichungen sind durch folgendes lineare Gleichungssystem gegeben:

$$\begin{pmatrix} \int_{-1}^{1} dx & \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx \\ \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx \\ \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx & \int_{-1}^{1} x^4\,dx \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} \frac{dx}{1+x^2} \\ \int_{-1}^{1} \frac{x\,dx}{1+x^2} \\ \int_{-1}^{1} \frac{x^2\,dx}{1+x^2} \end{pmatrix}$$

oder, äquivalent:

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} 2\arctan 1 \\ 0 \\ 2 - 2\arctan 1 \end{pmatrix}$$

Dessen Lösung ist

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} 0.9624 \\ 0 \\ -0.5310 \end{pmatrix}$$

Somit ist $\hat{\phi} = 0.9624 - 0.5310x^2$ die beste L^2-Approximation aus U an f. Abbildung 8.1 zeigt f und das beste L^2-approximierende quadratische Polynom an f. Effizienter ist die Verwendung orthogonaler Polynome *aus dem nachfolgenden Abschnitt.*

Trigonometrische Funktionen

Sei

$$f_k = e^{ikx}, \quad k = -n, \ldots, n$$

Diese Funktionen sind ein Orthonormalsystem für den Untervektorraum von $C[0, 2\pi]$, den sie aufspannen. Ist

$$f = \sum_{\nu=-\infty}^{\infty} \gamma_\nu e^{i\nu x}$$

eine Fourier-Reihe, so gilt

$$\gamma_k = \frac{1}{2\pi}\langle f, f_k\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx$$

Demnach ist

$$\hat{\phi} = \sum_{k=-n}^{n} \gamma_k e^{ikx} \in U$$

die beste L^2-Approximation an f. Glättung durch Herausfiltern der hochfrequenten Anteile ist hier also die Verwendung einer besten L^2-Approximation.

8.3 Orthogonale Polynome

Wir verwenden hier eine stetige Gewichtsfunktion

$$w\colon [a, b] \to \mathbb{R}_{>0}$$

und definieren damit ein gewichtetes Innenprodukt

$$\langle f, g\rangle_w = \int_a^b f(t)\overline{g(t)}w(t)\,dt$$

auf $V = C[a, b]$. Es handelt sich dabei in der Tat um ein Innenprodukt:

Beweis. 1. Ist $f \neq 0$, so existiert ein Intervall I_ϵ mit $f(x) \neq 0$ für alle $x \in I_\epsilon$. Dann ist

$$\langle f, f\rangle_w = \int_a^b |f(t)|^2 w(t)\,dt \geq \int_{I_\epsilon} |f(t)|^2 w(t)\,dt > 0$$

D.h. $\langle f, f\rangle_w = 0$ impliziert $f = 0$.

2. $\langle f, g\rangle_w = \overline{\langle g, f\rangle_w}$ ergibt sich durch Einsetzen in das Integral.

3. Ebenso ergibt sich

$$\langle \alpha f + g\rangle_w = \alpha\langle f, g\rangle_w$$
$$\langle f + g, h\rangle_w = \langle f, h\rangle_w + \langle g, h\rangle_w$$

□

Das Innenprodukt $\langle \cdot, \cdot \rangle_w$ induziert auf V eine Norm $\|\cdot\|_w$ via

$$\|f\|_w := \sqrt{\langle f, f \rangle_w} = \left(\int_a^b |f(t)|^2 w(t)\, dt \right)^{\frac{1}{2}}$$

Orthogonale Polynome entstehen durch Orthogonalisierung von $1, X, X^2, \ldots$:

$$\begin{aligned} p_0 &= 1 \\ p_n &= X^n - \sum_{\mu=0}^{n-1} \frac{\langle X^n, p_\mu \rangle_w}{\langle p_\mu, p_\mu \rangle_w} p_\mu, \qquad\qquad n = 1, 2, 3, \ldots \end{aligned}$$

nach der Methode von Gram-Schmidt (s. Abschnitt 5.11). Es gilt, dass p_n ein normiertes Polynom ist und orthogonal zu $K[X]_{n-1}$, den Polynomen von Grad $\leq n-1$, ist.

Orthogonale Polynome erfüllen eine 3-Terme-Rekursion:

Satz 8.3.1 (3-Terme-Rekursion)**.** *Für orthogonale Polynome $p_0, p_1, \ldots$ gilt:*

$$p_0 = 1, \quad p_1 = X - \beta_0, \quad p_{n+1} = (X - \beta_n) p_n - \gamma_n^2 p_{n-1}$$

mit $n = 1, 2, \ldots$ Dabei ist

$$\beta_n = \frac{\langle X p_n, p_n \rangle}{\langle p_n, p_n \rangle}, \quad \gamma_n^2 = \frac{\langle p_n, p_n \rangle}{\langle p_{n-1}, p_{n-1} \rangle}$$

Beweis. $p_0 = 1$ nach Konstruktion (Gram-Schmidt). Ebenso:

$$p_1 = X - \frac{\langle X, p_0 \rangle}{\langle p_0, p_0 \rangle} p_0 = X - \beta_0$$

Sei $n \geq 1$ und sei

$$q_{n+1} := (X - \beta_n) p_n - \gamma_n^2 p_{n-1}$$

Wir haben zu zeigen, dass $q_{n+1} = p_{n+1}$. Zunächst sind q_{n+1} und p_{n+1} normierte Polynome von Grad $n+1$. Daher ist

$$r := p_{n+1} - q_{n+1} \in K[X]_n$$

Nun zeigen wir, dass q_{n+1} orthogonal zu $K[X]_n$ ist. Dann ist auch $r = p_{n+1} - q_{n+1}$ orthogonal zu $K[X]_n$, insbesondere

$$\langle r, r \rangle = 0$$

also: $q_{n+1} = p_{n+1}$. Um unsere Orthogonalitätsbehauptung zu zeigen, zeigen wir nacheinander, dass q_{n+1} orthogonal zu p_n, zu p_{n-1} und zu $K[X]_{n-2}$ ist. Es gilt:

$$\langle q_{n+1}, p_n \rangle = \langle X p_n, p_n \rangle - \beta_n \langle p_n, p_n \rangle - \gamma_n^2 \underbrace{\langle p_{n-1}, p_n \rangle}_{=0} = 0$$

nach Definition von β_n. Also ist q_{n+1} orthogonal zu p_n. Weiter gilt:

$$\begin{aligned} \langle q_{n+1}, p_{n-1} \rangle &= \langle X p_n, p_{n-1} \rangle - \beta_n \underbrace{\langle p_n, p_{n-1} \rangle}_{=0} - \underbrace{\gamma_n^2 \langle p_{n-1}, p_{n-1} \rangle}_{= \langle p_n, p_n \rangle} \\ &= \underbrace{\langle X p_n, p_{n-1} \rangle}_{\overset{(*)}{=} \langle p_n, X p_{n-1} \rangle} - \langle p_n, p_n \rangle = \langle p_n, \underbrace{X p_{n-1} - p_n}_{\in K[X]_{n-1}} \rangle = 0 \end{aligned}$$

Intervall	$w(x)$	orthogonale Polynome
$[-1,1]$	1	Legendre-Polynome
$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	Tschebyschoff-Polynome
$[-1,1]$	$(1-x)^\alpha(1+x)^\beta,\ \alpha,\beta>-1$	Jacobi-Polynome
$(-\infty,\infty)$	e^{-x^2}	Hermite-Polynome
$(0,\infty)$	$e^{-x}x^\alpha,\ \alpha>-1$	Laguerre-Polynome

Tabelle 8.1: Einige Klassen orthogonaler Polynome.

wobei $(*)$ gilt, da X nur reelle Werte annimmt:

$$\langle Xp_n, p_{n-1}\rangle = \int_a^b tp_n(t)\overline{p_{n-1}(t)}w(t)\,dt = \int_a^b p_n(t)\overline{tp_{n-1}(t)}w(t)\,dt$$

Also ist q_{n+1} orthogonal zu p_{n-1}. Sei nun $q \in K[X]_{n-2}$. Dann ist

$$\langle q_{n+1}, q\rangle = \underbrace{\langle Xp_n, q\rangle}_{=\langle p_n, Xq\rangle=0} - \beta_n \underbrace{\langle p_n, q\rangle}_{=0} - \gamma_n^2 \underbrace{\langle p_{n-1}, q\rangle}_{=0} = 0$$

Also ist q_{n+1} orthogonal zu $K[X]_{n-2}$. Da p_n, p_{n-1} und $K[X]_{n-2}$ den Raum $K[X]_n$ aufspannen, folgt die Behauptung. □

Beispiel 8.3.2. *Spezielle Gewichtsfunktionen auf speziellen Intervallen liefern orthogonale Polynome mit gewissen Namen. Einige davon sind in Tabelle 8.1 aufgelistet.*

Beispiel 8.3.3. *Gesucht ist wieder die beste L^2-Approximation an $f = \frac{1}{1+x^2}$ durch quadratische Polynome. Diesmal sollen orthogonale Polynome verwendet werden.*

Zunächst konstruieren wir die orthogonalen Polynome. Die Gewichtsfunktion ist $w = 1$, das Intervall ist $[-1,1]$. Nach Tabelle 8.1 handelt es sich um Legendre-Polynome. In der 3-Terme-Rekursion (Satz 8.3.1) ergeben sich die Parameter:

$$\beta_0 = \beta_1 = 0, \quad \gamma_1^2 = \frac{1}{3}$$

Dies bedeutet, dass die ersten drei Legendre-Polynome

$$p_0 = 1, \quad p_1 = X, \quad , p_2 = X^2 - \frac{1}{3}$$

sind. Um eine Orthonormalbasis von $K[X]_2$ zu bekommen, finden wir noch die Normierungsfaktoren:

$$\langle p_0, p_0\rangle = 2$$
$$\langle p_1, p_1\rangle = \int_{-1}^1 t^2\,dt = \frac{2}{3}$$
$$\langle p_2, p_2\rangle = \int_{-1}^1 \left(t^4 - \frac{2}{3}t^2 + \frac{1}{9}\right)\,dx = \frac{8}{45}$$

Damit erhalten wir die Polynome:

$$P_0 = \frac{1}{\sqrt{2}}, \quad P_1 = \sqrt{\frac{3}{2}}X, \quad P_2 = \frac{3}{2}\sqrt{\frac{5}{2}}\left(X^2 - \frac{1}{3}\right)$$

Die Koeffizienten sind dann:

$$\begin{aligned}\gamma_0 &= \langle f, P_0\rangle = \sqrt{2}\arctan 1 \approx 1.111\\ \gamma_1 &= \langle f, P_1\rangle = 0\\ \gamma_2 &= \langle f, P_2\rangle = \frac{2}{3}\sqrt{\frac{5}{2}}\left(2 - \frac{8}{3}\arctan 1\right) \approx -0.2239\end{aligned}$$

Somit ist

$$\hat{\phi} = \gamma_0 P_0 + \gamma_1 P_1 + \gamma_2 P_2 \approx 0.9624 - 0.5310x^2$$

die beste L^2-Approximation an f durch quadratische Polynome.

8.4 Tschebyschoff-Approximation

Hier sei $K = \mathbb{R}$ und es bezeichnet

$$C[a,b] = \{f\colon [a,b] \to \mathbb{R} \mid f \text{ stetig}\}$$

diesmal versehen mit der L^∞-norm $\|\cdot\|_\infty$:

$$\|f\|_\infty = \max_{t\in[a,b]} |f(t)|$$

Sei U ein endlich-dimensionaler Untervektorraum von $V = C[a,b]$ und $f \in V$. Hier ist im Allgemeinen die beste Approximation $\hat{\phi} \in U$ mit

$$\left\|f - \hat{\phi}\right\|_\infty = \min_{\phi\in U} \|f - \phi\|_\infty$$

nicht eindeutig bestimmt.

Beispiel 8.4.1. *Sei $[a,b] = [0,1]$, $f = 1$ und $U = \mathbb{R}x$. Nun gilt für $\phi \in U$:*

$$\|f - \phi\|_\infty \geq 1$$

und für alle ϕ der Form $\phi = \alpha x$ mit $0 \leq \alpha \leq 2$:

$$\|f - \phi\|_\infty = 1$$

Hier liegt also eine Mehrdeutigkeit der besten L^∞-Approximation vor.

Die Eindeutigkeit ist gegeben durch die

Haarsche Bedingung. *Es sei $\dim U = n$ und die Interpolationsaufgabe*

$$\phi(x_i) = y_i, \quad i = 1, \ldots, n$$

mit beliebigen Stützstellen $a \leq x_1 < \cdots < x_n \leq b$ und Werten $y_1, \ldots, y_n$ stets durch ein $\phi \in U$ lösbar.

Beweis. Sei $f_1, \ldots, f_n$ eine Basis von U. Ein interpolierendes $\phi = \sum_{i=1}^{n} \gamma_i f_i$ existiert genau dann, wenn das lineare Gleichungssystem

$$\sum_{i=1}^{n} \gamma_i f_i(x_j) = y_j, \quad j = 1, \ldots, n \tag{8.2}$$

für $g = (x_i) \in \mathbb{R}^n$ lösbar ist. Die haarsche Bedingung besagt, dass das lineare Gleichungssystem eindeutig lösbar ist. Schreibt man (8.2) um als

$$Ag = y \tag{8.3}$$

mit $A = (f_i(x_j)) \in \mathbb{R}^{n \times n}$ und $y = (y_j) \in \mathbb{R}^n$, so ist nach der haarschen Bedingung die Gleichung (8.3) für jede rechte Seite y lösbar. Wählen wir als rechte Seite jede Spalte der Einheitsmatrix I, so folgt, dass die Matrix-Gleichung

$$AX = I$$

lösbar ist. Also ist A invertierbar, und somit (8.3) eindeutig lösbar. □

8.5 Tschebyschoff-Polynome 1. Art

Die Tschebyschoff-Polynome 1. Art können direkt definiert werden als

$$T_n(x) = \cos(n \arccos(x)), \quad x \in [-1, 1],\ n = 0, 1, 2, \ldots$$

Es gilt für $\theta \in [0, \pi]$

$$T_n(\cos\theta) = \cos(n\theta)$$

Diese Polynome erfüllen die Rekursion:

$$\begin{aligned} T_0(X) &= 1 \\ T_1(X) &= X \\ T_{n+1}(X) &= 2XT_n(X) - T_{n-1}(X), \quad n = 1, 2, 3, \ldots \end{aligned}$$

Beweis. Aus dem Additionstheorem für den Cosinus:

$$\cos x + \cos y = 2 \cos\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right)$$

ergibt sich:

$$2\cos\theta\cos(n\theta) = \cos((n+1)\theta) + \cos((n-1)\theta)$$

Folglich gilt mit $t = \cos\theta$:

$$2tT_n(t) - T_{n-1}(t) = 2\cos\theta\cos(n\theta) - \cos((n-1)\theta) = \cos((n+1)\theta) = T_{n+1}(t)$$

□

Es folgt, dass $T_n(X) \in \mathbb{Z}[X]_n$, also ein Polynom mit ganzzahligen Koeffizienten ist. Der führende Koeffizient ist 2^{n-1} und $\deg(T_n) = n$.

Eigenschaften der Tschebyschoff-Polynome 1. Art

1. Es gilt:
$$\max_{t\in[-1,1]} |T_n(t)| = 1$$

2. T_n hat in $[-1,1]$ insgesamt $n+1$ Extrema:
$$s_k^{(n)} = \cos\left(\frac{k\pi}{n}\right), \quad T_n\left(s_k^{(n)}\right) = (-1)^k, \quad k = 0,1,\ldots,n$$

3. T_n hat in $[-1,1]$ insgesamt n einfache Nullstellen
$$t_k^{(n)} = \cos\left(\frac{(2k-1)\pi}{2n}\right), \quad T_n\left(t_k^{(n)}\right) = 0, \quad k = 1,\ldots,n$$

4. Es gilt:
$$\max_{t\in[-1,1]} \prod_{k=1}^{n+1} \left|t - t_k^{(n+1)}\right| = 2^{-n}$$

5. Orthogonalitätsrelationen:
$$\int_{-1}^{1} T_n(t)T_m(t)\,\frac{dt}{\sqrt{1-t^2}} = \begin{cases} 0, & n \neq m \\ \pi, & n = m = 0 \\ \frac{\pi}{2}, & n = m \neq 0 \end{cases}$$

   Tschebyschoff-Polynome sind also orthogonale Polynome.

Beweis von 4. Dies folgt aus:
$$\frac{1}{2^n}T_{n+1}(X) = \prod_{k=0}^{n+1}\left(X - t_k^{(n+1)}\right)$$
und Eigenschaft 1. □

Beispiel 8.5.1. *In Abbildung 8.2 sind die Tschebyschoff-Polynome 1. Art T_2 bis T_7 abgebildet.*

8.6 Optimale Lagrange-Interpolation

Sei $f \in C[a,b]^{n+1}$. Wir wollen f bestmöglich durch Polynom-Interpolation an $n+1$ Stützstellen approximieren. Ist $P_n(X) \in \mathbb{R}[X]_n$ das Lagrange-Interpolationspolynom, so ist der Interpolationsfehler gegeben durch:
$$f(t) - P_n(t) = \frac{f^{(n+1)}(\xi)}{(n+1)!}N_{n+1}(t)$$
(4.1.8), wobei $\xi \in I_t$ und
$$N_{n+1}(X) = \prod_{\nu=0}^{n}(X - x_\nu)$$
sei und I_t das kleinste Intervall sei, das die Stützstellen $x_0 < \cdots < x_n$ (im Intervall $[a,b]$) und t enthält.

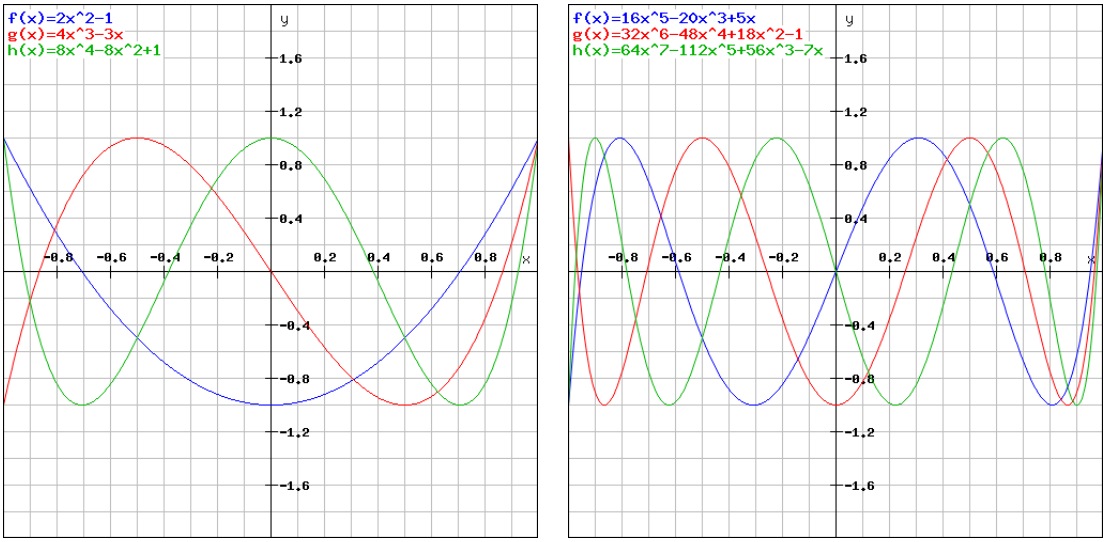


Abbildung 8.2: Die Tschebyschoff-Polynome 1. Art T_2 bis T_7.

Die Aufgabe lautet nun, die Stützstellen $x_0, \ldots, x_n$ so zu wählen, dass $\|N_{n+1}\|_\infty$ minimal wird.

Das normierte Polynom $N_{n+1}(X)$ hat die Gestalt

$$N_{n+1} = X^{n+1} - \phi$$

mit $\phi \in \mathbb{R}[X]_n$. Gesucht ist also die beste L^∞-Approximation $\hat{\phi} \in U = \mathbb{R}[X]_n$ an $f = X^{n+1}$. Da U der haarschen Bedingung (Abschnitt 8.4) genügt, ist die beste L^∞-Approximation eindeutig bestimmt. Es gilt:

Satz 8.6.1. *Auf $[a, b] = [-1, 1]$ ist die beste L^∞-Approximation $\hat{\phi} \in \mathbb{R}[X]_n$ an $f = X^{n+1}$ durch*

$$\hat{\phi} = X^{n+1} - 2^{-n} T_{n+1}(X)$$

gegeben, wobei T_{n+1} das $n+1$-te Tschebyschoff-Polynom 1. Art sei. Die Nullstellen von T_{n+1} sind die optimalen Stützstellen der Lagrange-Interpolation auf $[-1, 1]$.

Kapitel 9

Numerische Integration

Bei der numerischen Integration geht es um die näherungsweise Bestimmung bestimmter Integrale:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} \alpha_i f(x_i)$$

mit Stützstellen $a \leq x_0 < \cdots < x_n \leq b$ und Gewichten $\alpha_i \in \mathbb{R}$.

Beispiel. *Die* Rechteckregel

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$$

ist eine Form der numerischen Integration. Diese ist in Abbildung 9.1 dargestellt.

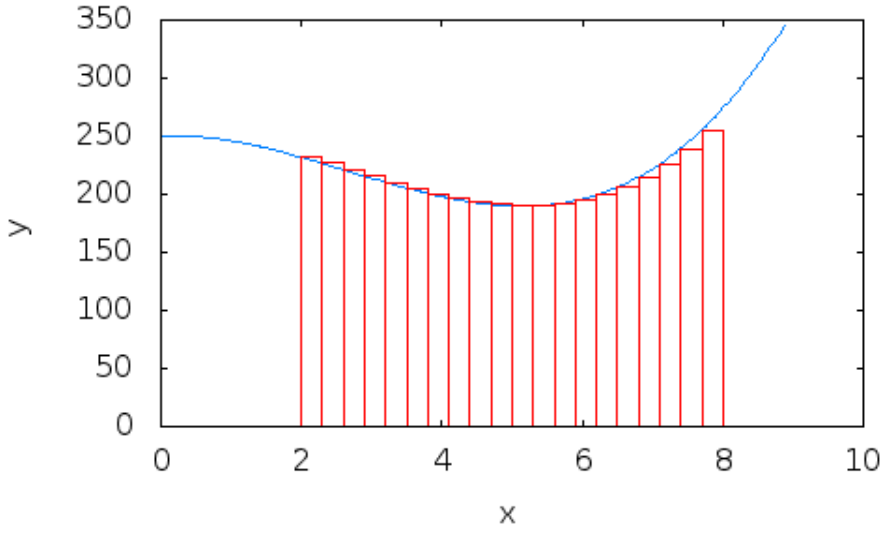


Abbildung 9.1: Rechteckregel (Quelle: Wikipedia, Autor: Mkwadee).

9.1 Interpolatorische Quadratur

9.1.1 Trapezregel

Bei der *Trapezregel* wird die Fläche unter $y = f(x)$ von $x = 0$ bis $x = h$ durch ein Trapez $ABCD$ approximiert (s. Abbildung 9.2). Dann ist

$$\int_a^b f(x)\,dx \approx h\frac{f(0)+f(h)}{2}$$

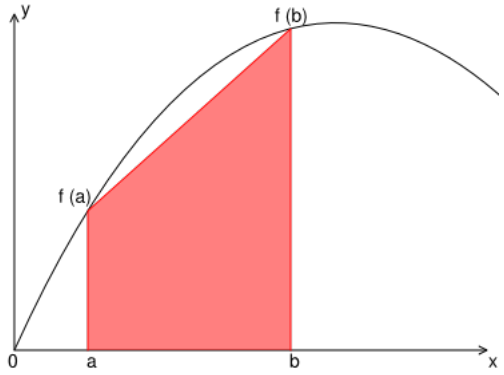


Abbildung 9.2: Die Trapezregel (Quelle: Wikipedia, Author: Boris23).

Die Idee, die hinter der Trapezregel steckt, ist es, f in 0 und h durch ein lineares Polynom $\ell(x)$ zu interpolieren:

$$\int_a^b f(x)\,dx \approx \int_a^b \ell(x)\,dx$$

Der Interpolant ist dabei

$$\ell(x) = f(0) + \frac{f(h)-f(0)}{h}x$$

Es ergibt sich:

$$\int_a^b \ell(x)\,dx = f(0)x + \frac{1}{2}\frac{f(h)-f(0)}{h}x^2\Big|_0^h = h\frac{f(0)+f(h)}{2}$$

also die Trapezregel.

Der Quadraturfehler ergibt sich aus dem Interpolationsfehler

$$f(x) - \ell(x) = \frac{f''(\xi_x)}{2}x(x-h), \quad \xi_x \in [0,h]$$

(4.1.8). Somit ist

$$\int_0^h f(x)\,dx - \int_0^h \ell(x)\,dx = \int_0^h (f(x)-\ell(x))\,dx = \frac{1}{2}\int_0^h f''(\xi_x)x(x-h)\,dx$$

$$\overset{\text{MWSI}}{=} \frac{f''(\eta)}{2}\int_0^h x(x-h)\,dx = -\frac{f''(\eta)}{12}h^3, \quad \eta \in [0,h]$$

Mit MWSI ist der *Mittelwertsatz der Integralrechnung* gemeint:

Satz 9.1.1 (Mittelwertsatz der Integralrechnung)**.** *Sei* $f\colon [a,b] \to \mathbb{R}$ *eine stetige Funktion und* $g\colon [a,b] \to \mathbb{R}$ *integrierbar mit entweder* $g \geq 0$ *oder* $g \leq 0$*. Dann existiert ein* $\eta \in [a,b]$*, sodass*

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

Beachte dabei, dass $x(x-h) \leq 0$ für $x \in [0,h]$ gilt.

9.1.2 Zusammengesetzte Trapezregel

Ist das Intervall groß, so wird die einfache Trapezregel ungenau gemäß Quadraturfehlerbetrachtung im vorigen Abschnitt. Abhilfe kann durch äquidistante Unterteilung des Intervalls $[a,b]$ geschaffen werden:

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$
$$x_i = a + ih, \quad h = \frac{b-a}{n}, \quad i = 0, \ldots, n$$

Wende nun die Trapezregel auf jedes der Teilintervalle $[x_{i-1}, x_i]$ an:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx h\frac{f(x_{i-1}) + f(x_i)}{2}$$

Dies ergibt:

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \sum_{i=1}^{n} h\frac{f(x_{i-1}) + f(x_i)}{2} \\ &= h\left(\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right) =: \mathrm{ZT}_h(f)\end{aligned}$$

Der Quadraturfehler ist nun:

$$\sum_a^b f(x)\,dx - \mathrm{ZT}_h(f) = \sum_{i=1}^{n} -\frac{h^3}{12}f''(\eta_i) = -\frac{h^2}{12}\frac{b-a}{n}\sum_{i=1}^{n} f''(\eta_i)$$

mit $\eta_i \in [x_{i-1}, x_i]$. Die Größe $\frac{1}{n}\sum\limits_{i=1}^{n} f''(\eta_i)$ ist das arithmetische Mittel der Werte $f''(\eta_i)$. Sie liegt also zwischen dem größten und dem kleinsten Wert. Nach dem Zwischenwertsatz ergibt sich, dass ein $\eta \in [a,b]$ existiert mit

$$f''(\eta) = \sum_{i=1}^{n} f''(\eta_i)$$

Somit ist

$$\int_a^b f(x)\,dx - \mathrm{ZT}_h(f) = -\frac{(b-a)f''(\eta)}{12}h^2$$

Konsequenz. *Die Approximation durch die zusammengesetzte Trapezregel wird durch Hinzunahme von Stützstellen (d.h. Verkleinerung von* h*) beliebig genau. Der Quadraturfehler ist sogar quadratisch in* h*.*

9.1.3 Newton-Cotes-Formeln

Bei der Trapezregel wurde mit linearen Polynomen interpoliert. Nun interpolieren wir mit einem Polynom von Grad bis zu n an Stützstellen $x_0, \dots, x_n \in [a, b]$. Finde dazu Gewichte $\alpha_0, \dots, \alpha_n \in \mathbb{R}$, sodass Polynome $f \in \mathbb{R}[X]_n$ exakt integriert werden, d.h.

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} f(x_i)\alpha_i, \quad \text{falls } f \in \mathbb{R}[X]_n$$

Die Lösung ist gegeben durch die Lagrange-Basis-Polynome

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

gegeben. Nämlich

$$\alpha_i := \int_a^b \ell_i(x)\,dx$$

löst diese Aufgabe.

Beweis. Die Quadratur ist exakt für $f \in \mathbb{R}[X]_n$: Zunächst ist

$$f(X) = \sum_{i=0}^{n} f(x_i)\ell_i(X)$$

Somit ist

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

□

Abgeschlossene Newton-Cotes-Formeln

Hier werden die Stützstellen äquidistant gewählt:

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

Folglich ist jedes $x \in [a, b]$ von der Form

$$x = a + th, \quad t \in [0, n]$$

Die Lagrange-Basispolynome ℓ_i schreiben sich dann als

$$\ell_i(X) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{X - x_j}{x_i - x_j} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{a + th - a - jh}{a + ih - a - jh} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - j}{i - j}$$

Also gilt, wegen $dx = h\,dt$:

$$\alpha_i = h \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - j}{i - j}\,dt, \quad i = 0, \dots, n$$

Beispiel 9.1.2. *Sei $n = 2$. Dann ist $h = \frac{b-a}{2}$. Es gilt:*

$$\alpha_0 = h \int_0^2 \frac{t-1}{0-1}\frac{t-2}{0-2}\,dt = \frac{h}{3}$$

$$\alpha_1 = h \int_0^2 \frac{t-0}{1-0}\frac{t-2}{1-2}\,dt = \frac{4}{3}h$$

$$\alpha_2 = h \int_0^2 \frac{t-0}{2-0}\frac{t-1}{2-1}\,dt = \frac{h}{3}$$

Dies ergibt die Simpson-Regel[1]

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

Einige Newton-Cotes-Formeln explizit

Es sei $n = 1, 2, 3, 4$ und $h = \frac{b-a}{n}$. Dann gilt mit $I = \int_a^b f(x)\,dx$:

$n = 1:$ $\quad I \approx \dfrac{b-a}{2}(f(a) + f(b))$ *Trapezregel*

$n = 2:$ $\quad I \approx \dfrac{b-a}{6}\left(f(a) + 4f\left(\dfrac{a+b}{2}\right) + f(b)\right)$ *Simpson-Regel*

$n = 3:$ $\quad I \approx \dfrac{b-a}{8}\left(f(a) + 3f(a+b) + 3f(b-h) + f(b)\right)$ *3/8-Regel*

$n = 4:$ $\quad I \approx \dfrac{b-a}{90}\left(7f(a) + 32f(a+h) + 12f\left(\dfrac{a+b}{2}\right) + 32f(b-h) + 7f(b)\right)$ *Boole-Regel*

Beispiel 9.1.3. *Wir approximieren $I = \int_0^1 \frac{dx}{1+x^2}$ mit den ersten 4 Newton-Cotes-Formeln.*

1. *Trapezregel: $I \approx 0.75000$.*
2. *Simpson-Regel: $I \approx 0.78333$.*
3. *3/8-Regel: $I \approx 0.78462$.*
4. *Boole-Regel: $I \approx 0.78553$.*

Fehler der Simpson-Regel

Der Quadraturfehler bei der Simpson-Regel beträgt

$$\int_a^b f(x)\,dx - \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = \frac{(b-a)^5}{2880}f^{(4)}(\eta), \quad \eta \in [a, b]$$

[1] diesmal benannt nach Thomas Simpson (1710–1761)

9.2 Gauß-Quadratur

Wir definieren zunächst

$$\int f := \int_a^b f(x)w(x)\,dx$$

mit fester positiver, stetiger Gewichtsfunktion $w\colon [a,b] \to \mathbb{R}$ und bemerken, dass $\int$ linear ist:

$$\int \alpha f = \alpha \int f, \quad \alpha \in \mathbb{R}$$
$$\int (f+g) = \int f + \int g, \quad f, g \in C[a,b]$$

Wir erinnern, dass $\int fg$ ein Innenprodukt auf $C[a,b]$ definiert (s. Abschnitt 8.3).

Nullstellen orthogonaler Polynome

Für die Gauß-Quadratur werden wir die Nullstellen orthogonaler Polynome verwenden. Es gilt:

Satz 9.2.1. *Sei $p_0, p_1, p_2, \ldots$ eine Folge orthogonaler Polynome in $C[a,b]$ mit $\deg p_i = i$. Dann sind deren Nullstellen einfach, reell und liegen im Intervall $[a,b]$.*

Beweis. Seien $x_0, \ldots, x_k$ die verschiedenen Nullstellen von p_{n+1}, die im Intervall $[a,b]$ liegen. Falls $k = n$, so ist alles gezeigt. Falls jedoch $k < n$, sei

$$q(X) := (X - x_0) \cdot (X - x_k)$$

Es ist $\deg q = k + 1 < n + 1$. Deshalb ist

$$\int p_{n+1} q = 0 \tag{9.1}$$

Aber $p_{n+1}q$ hat in $[a,b]$ keinen Vorzeichenwechsel, da jede Nullstelle mit gerader Multiplizität auftritt. Folglich ist

$$\int p_{n+1} q \neq 0$$

im Widerspruch zu (9.1). □

In der Tat eignen sich die Nullstellen orthogonaler Polynome gut als Stützstellen für die Integration. Seien $x_0, \ldots, x_n \in [a,b]$ die Nullstellen von p_{n+1}, wobei $p_0, p_1, \ldots$ eine Folge orthogonaler Polynome in $C[a,b]$ mit $\deg p_i = i$ sei. Wir setzen

$$\alpha_i := \int \ell_i, \quad i = 1, \ldots, n$$

wobei

$$\ell_i(X) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{X - x_j}{x_j - x_j}$$

das i-te Lagrange-Basispolynom sei. Die *Gauß-Quadraturformel* ist

$$G_n f = \sum_{i=0}^{n} \alpha_i f(x_i)$$

Es gilt:

Satz 9.2.2. *Ist $f \in \mathbb{R}[X]_{2n+1}$, so integriert $G_n f$ exakt:*

$$G_n f = \int f$$

Beweis. $G_n f$ ist exakt für Polynome von Grad $\leq n$ nach Abschnitt 9.1.3. Sei f ein Polynom mit $\deg f \leq 2n+1$. Es gilt: $\deg p_{n+1} = n+1$. Somit ist bei der Division mit Rest:

$$f = p_{n+1}q + r, \quad \deg q, \deg r \leq n$$

Also ist

$$G_n f = \sum_{i=1}^{n} \alpha_i f(x_i) = \sum_{i=1}^{n} \alpha_i (\underbrace{p_{n+1}(x_i)}_{=0} q(x_i) + r(x_i)) = \sum_{i=1}^{n} \alpha_i r(x_i) = G_n r$$
$$\overset{\deg r \leq n}{=} \int r \overset{(*)}{=} \int (p_{n+1}q + r) = \int f$$

wobei $(*)$ gilt, da p_{n+1} orthogonal zu q ist wegen $\deg q \leq n$. □

Konsequenz. *Die Gewichte α_i sind alle positiv und $\leq \int 1$.*

Proof. 1. Es gilt:

$$0 < \int \ell_i^2 \overset{(*)}{=} G_n \ell_i^2 = \sum_j \alpha_j \underbrace{\ell_i^2(x_j)}_{=\delta_{ij}} = \alpha_i$$

wobei $(*)$ gilt wegen $\deg \ell_i^2 \leq 2n+1$.

2. Es gilt:

$$\sum \alpha_i 1 = G_n 1 = \int 1$$

Ist eine Summe positiver reeller Zahlen kleiner gleich $\int 1$, so sind alle Summanden kleiner gleich $\int 1$. □

Gauß-Quadraturfehler

Es gilt:

$$\int f - G_n f = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int p_{n+1}^2, \quad \xi \in [a,b]$$

Konvergenz

Für $f \in C[a,b]$ gilt:

$$\lim_{n\to\infty} G_n f = \int f$$

Beispiel 9.2.3.

1. *Für $[a,b] = [-1,1]$ und $w = 1$ spricht man von der* Gauß-Legendre-Quadratur.
2. *Für das Intervall $[0,\infty)$ und $w(x) = e^{-x}$ ist es die* Gauß-Laguerre-Quadratur.
3. *Für $(-\infty,\infty)$ und $w(x) = e^{-x^2}$ ist es die* Gauß-Hermite-Quadratur.

Gauß-Legendre-Formeln

Bei der Gauß-Legendre-Quadratur werden die Nullstellen der *Legendre-Polynome* $P_0, P_1, \ldots$ verwendet. Sie erfüllen die Orthogonalitätsrelationen

$$\int_{-1}^{1} P_n(x)P_m(x)\,dx = \frac{2}{2n+1}\delta_{m,n}$$

und die Rekursionsformel

$$(n+1)P_{n+1}(X) + nP_{n-1}(X) = (2n+1)XP_n(X)$$

Die ersten drei Legendre-Polynome lauten:

$$P_0 = 1,\ P_1 = X,\ P_2 = \frac{3}{2}X^2 - \frac{1}{2}$$

Tabelle 9.1 gibt die Knoten und Gewichte der Gauß-Legendre-Quadratur für $n \leq 5$ wieder.

n	Knoten x_i	Gewicht α_i
1	0	2
2	$\pm 1/\sqrt{3}$	1
3	0	$8/9$
	$\pm\sqrt{3/5}$	$5/9$
4	$\pm\sqrt{(3-2\sqrt{6/5})/7}$	$\frac{18+\sqrt{30}}{36}$
	$\pm\sqrt{(3+2\sqrt{6/5})/7}$	$\frac{18-\sqrt{30}}{36}$
5	0	$128/225$
	$\pm\frac{1}{3}\sqrt{5-2\sqrt{10/7}}$	$\frac{322+13\sqrt{70}}{900}$
	$\pm\frac{1}{3}\sqrt{5+2\sqrt{10/7}}$	$\frac{322-13\sqrt{70}}{900}$

Tabelle 9.1: Knoten und Gewichte der Gauß-Legendre-Quadratur.

9.3 Intervalltransformation

Gegeben seien Gewichte und Knoten einer Quadraturformel auf dem Intervall $[-1, 1]$:

$$\int_{-1}^{1} g(x)\,dx \approx \sum_{i=0}^{n} \alpha_i g(x_i)$$

Angenommen, diese sollen verwendet werden, um das Integral

$$\int_{a}^{b} f(t)\,dt$$

näherungsweise zu berechnen. Dann können wir folgende Transformation verwenden:

$$t\colon [-1,1] \to [a,b], \quad x \mapsto \frac{b-a}{2}(x+1) + a$$

Diese ist affin-linear und bildet $[-1,1]$ bijektiv auf $[a,b]$ ab. Mit dieser Substitution wird

$$\int_a^b f(t)\,dt = \int_{-1}^{1} f(t(x)) \underbrace{\frac{b-a}{2}\,dx}_{=dt} = \frac{b-a}{2}\int_{-1}^{1} g(x)\,dx$$

mit $g(x) = f(t(x))$. Dies ergibt folgende Quadraturformel für f:

$$\int_a^b f(t)\,dt \approx \frac{b-a}{2}\sum_{i=1}^{n} \alpha_i f(t_i), \quad t_i = t(x_i)$$

Die neuen Knoten sind also $t(x_i)$ und die neuen Gewichte

$$\frac{b-a}{2}\alpha_i$$

vermöge der Transformation $t\colon [-1,1] \to [a,b]$.

Beispiel 9.3.1. *Wir approximieren*

$$I = \int_0^1 \frac{dx}{1+x^2}$$

mithilfe der Gauss-Legendre-Quadraturformeln. Dazu verwenden wir die Transformation

$$t\colon [-1,1] \to [0,1], \quad x \mapsto \frac{1}{2}(x+1)$$

Es ist $\frac{b-a}{2} = \frac{1}{2}$*. Unter Verwendung von Tabelle 9.1 erhalten wir demnach:*

$n = 1$. $t_1 = t(0) = \frac{1}{2}$, $\alpha_1 = \frac{1}{2}\cdot 2 = 1$.

$$I \approx f\left(\frac{1}{2}\right) = 0.8000$$

$n = 2$.

$$\begin{aligned} t_1 &= t\left(1/\sqrt{3}\right) \approx 0.7887, \quad \alpha_1 = \frac{1}{2}\cdot 1 = \frac{1}{2} \\ t_2 &= t\left(-1/\sqrt{3}\right) \approx 0.2113, \quad \alpha_2 = \frac{1}{2} \\ I &\approx \frac{1}{2}f(0.7887) + \frac{1}{2}f(0.2113) \approx 0.7869 \end{aligned}$$

$n = 3$.

$$\begin{aligned} t_1 &= t(0) = \frac{1}{2}, \quad \alpha_1 = \frac{1}{2}\cdot\frac{8}{9} = \frac{4}{9} \\ t_2 &= t\left(\sqrt{3/5}\right) \approx 0.8873, \quad \alpha_2 = \frac{1}{2}\cdot\frac{5}{9} = \frac{5}{18} \\ t_3 &= t\left(-\sqrt{3/5}\right) \approx 0.1127, \quad \alpha_3 = \frac{5}{18} \\ I &\approx \frac{4}{9}f\left(\frac{1}{2}\right) + \frac{5}{18}\left(f(0.8873) + f(0.1127)\right) \approx 0.7853 \end{aligned}$$

9.4 Romberg-Integration

Sei zunächst T_k^0 die zusammengesetzte Trapezregel für $n = 2^k$ äquidistante Teilintervalle von $[a, b]$:

$$T_k^0 = \mathrm{ZT}_{2^k}(f), \quad k = 0, 1, 2, \ldots$$

mit ZT_n wie in Abschnitt 9.1.2. Nun bildet man höhere Differenzenquotienten:

$$T_k^i = \frac{4^i T_k^{i-1} - T_{k-1}^{i-1}}{4^i - 1}, \quad i = 1, 2, \ldots, k$$

Im Tableau

$$\begin{array}{ccccccc}
T_0^0 & & & & & & \\
 & \searrow & & & & & \\
T_1^0 & \to & T_1^1 & & & & \\
 & \searrow & & \searrow & & & \\
T_2^0 & \to & T_2^1 & \to & T_2^2 & & \\
 & \searrow & & \searrow & & \searrow & \\
T_3^0 & \to & T_3^1 & \to & T_3^2 & \to & T_3^3 \\
\ldots & & & & & &
\end{array}$$

hängt jedes Element nur von seinem linken und seinem linken oberen Nachbarn ab.

Bemerkung 9.4.1. *Der Fehler für T_k^n ist $O(h_k^{2i+2})$, wobei $h_k = \frac{b-a}{2^k}$ ist.*

Beispiel 9.4.2. *Wir berechnen*

$$\ln 2 = \int_1^2 \frac{dx}{x}$$

mit dem Romberg-Verfahren. Zunächst:

$$\begin{aligned}
T_0^0 &= \frac{1}{2}\left(1 + \frac{1}{2}\right) = 0.75 \\
T_1^0 &= \frac{1}{2}\left(1 + \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\right) = 0.708333333 \\
T_2^0 &= \frac{1}{4}\left(1 + \frac{1}{2}\cdot\frac{4}{5} + \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\cdot\frac{4}{7} + \frac{1}{2}\right) = 0.69702380952
\end{aligned}$$

Dann:

$$\begin{aligned}
T_1^1 &= \frac{4T_1^0 - T_0^0}{3} = 0.694444 \\
T_2^1 &= \frac{4T_3^0 - T_2^0}{3} = 0.693253, \quad T_2^2 = \frac{16T_2^1 - T_1^1}{15} = 0.69317460
\end{aligned}$$

Vergleich mit $\ln 2 \approx 0.69214718$ ergibt, dass T_2^2 bereits 4 korrekte Nachkommastellen hat.