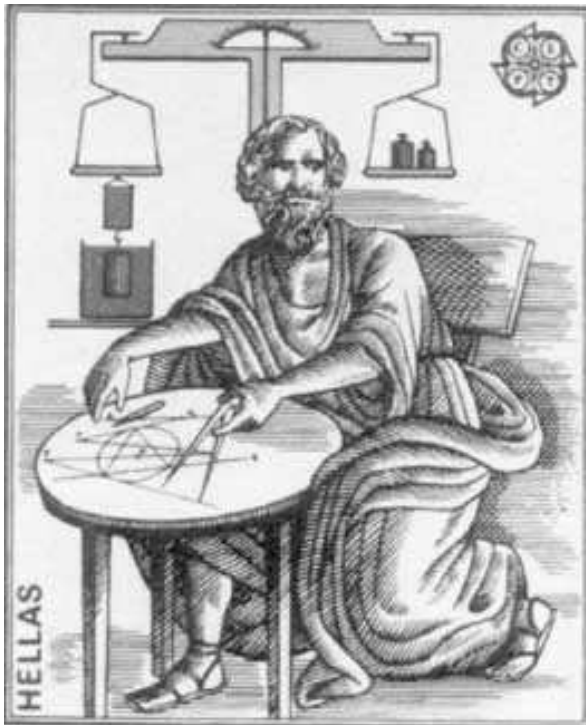


Mathematische Grundlagen für Informatiker

Eine Einführung für Studienanfänger

Stasys Jukna



Logik

Kombinatorik

Zahlentheorie

Algebra

Analysis

Stochastik

Lineare Algebra

Skript zur Vorlesung “Mathematische Grundlagen der Informatik” für Studienanfänger

Johan Wolfgang Goethe Universität

Frankfurt am Main, 2003

Prof. Dr. habil. S. Jukna
Institut für Mathematik und Informatik
Litauische Akademie der Wissenschaften
Vilnius, Litauen

Jetztige Adresse:
Lehrstuhl für Theoretische Informatik
Institut für Informatik
Fachbereich Informatik und Mathematik
J. W. Goethe Universität Frankfurt
Frankfurt am Main

Vorwort

In der Informatik, wie auch in vielen anderen naturwissenschaftlichen Fächern, werden viele Studienanfänger mit mathematischen Methoden und mathematischer Denkweise konfrontiert, auf die sie in der Schule nicht vorbereitet wurden. Dieses Skript bietet Schulabgängern unterschiedlicher Qualifikation einen fairen Einschub aus der Mathematik, der die Einstieg ins Informatik-Studium erleichtern sollte. Insbesondere sollte das Skript denjenigen Studierenden helfen, deren Studiengang – wie zum Beispiel der Studiengang “Bioinformatik” an der Universität Frankfurt – keine weitere Pflichtvorlesungen in der Mathematik vorsieht.

Das Skript ist aus der Vorlesung¹ “Mathematische Grundlagen der Informatik” entstanden. Die Veranstaltung wurde für Anfänger gedacht und darf nur ein Semester (mit 4 SWS) dauern. Es ist deshalb klar, dass nur ein Teil der Mathematik besprochen werden kann. Deshalb habe ich die Themenauswahl sehr gezieht und sehr *pragmatisch* getroffen: Nur die Sachen ausgewählt würden, die (nach der Erfahrung) am häufigsten während des späteren Informatik-Studiums gebraucht werden und die die größten Schwierigkeiten für Studierenden bereiten, wie zu Beispiel Beweisprinzipien (Logik, Induktion und Kombinatorik), asymptotische Analyse (Konvergenz von Folgen und Reihen, klein-*o*/größ-*O* Notation, Rekurrenzen) und insbesondere stochastische Analyse (diskrete Stochastik). Von der ganzen Zahlentheorie werden wir nur die modulare Arithmetik betrachten, da sie am meisten in der Informatik Anwendungen findet und vollkom unbekannt für Schulabgängern ist. Die ganze Algebra ist mit wenigen Ausnahmen (wie Eigenschaften von Gruppen) auf zwei wichtigen Themen – lineare Algebra und Matrizenlgebra – die am häufigsten in der Informatik benutzt werden, reduziert.

Wir werden auf zu detaillierten Formalisierungen wie auch auf zu abstrakten Verallgemeinerungen oft verzichten. Stattdessen, werden wir uns mehr auf die Anwendbarkeit der Konzepte und insbesondere auf die dahinter steckende Intuition konzentrieren. Bevor wir ein Konzept einführen werden wir oft die Frage “Wozu ist das gut?” stellen.

Da hier es um “Mathematik für Anwender” handelt, sollte man auch eine Vorstellung haben, wie diese Gebiete der “reinen” Mathematik zur Anwendungen im “wirklichen Leben” (Informatik inklusiv) kommen. Deshalb werden wir wo möglich die Anwendungen der Mathematik vorstellen: modulare Arithmetik und RSA-Codes, Anwendungen der linearen Algebra in Kodierungstheorie und Kombinatorik (Fisher’s Ungleichung und Expandergraphen), Anwendungen der Matrizenalgebra in Graphentheorie und für die Analyse der Markov-Ketten, Anwendungen der Folgen und Reihen in Finanzmathematik, Anwendungen der Stochastik in der Börse, usw. Ausser dass diese Anwendungen schon selbst nicht trivial sind, sollen sie den Studierenden zeigen, dass auch in Anwendungen eine präzise mathematische Denkweise unverzichtbar ist.

Im Unterschied von vielen Mathematik-Skripten werden wir (mit nur ein paar Ausnahmen) *alle* angegebene Sätze auch *beweisen*. Die Studierende sollen sich gewöhnen, an nichts in der Mathematik zu glauben, bis sie es selbst nicht verifiziert haben. Außerdem, oft sind Beweise viel informative als die Sätze selbst. Beherrscht man halbwegs die mathematische Denkweise und weiss man, was die Konzepte eigentlich bedeuten, kann man (wenn nötig) den Rest in anderen Mathematik-Büchern nachlesen.

Das Skript enthält mehr Stoff, als in der Vorlesung tatsächlich besprochen sein sollte. Insbesondere, habe ich viele motivierende Beispiele, Anwendungen und Aufgaben ausgesucht, die den Stoff ein wenig attraktiver für Studenten machen sollten. Diesen “gutte Nacht” Stoff – der den größten Teil des

¹<http://lovelace.thi.informatik.uni-frankfurt.de/~jukna/Grundlagen/index.html>

Skripts ausmacht – sollten Studenten selber vor der Vorlesung nachlesen.

Zur Darstellung: Die “relative Wichtigkeit” der Sätze ist durch ihre Formatierung gut erkennbar.

Satz 0.1. Hier ist ein “normaler” Satz ...

und

Satz 0.2. Die “wichtigste” Sätze sind zusätzlich im Rahmen gesetzt.

Die mit * markierte Abschnitte sind optional.

Ich danke Georg Schnitger, Gregor Gramlich, Maik Weinard, Markus Schmitz-Bonfigt, Uli Laube und natürlich meine Studenten für zahlreiche Verbesserungsvorschläge.

Stasys Jukna, Frankfurt a. M.

Inhaltsverzeichnis

1	Grundbegriffe und Beweismethoden	1
1.1	Mengen, Relationen und Funktionen	1
1.1.1	Relationen	4
1.1.2	Graphen (binäre Relationen)	8
1.1.3	Abbildungen (Funktionen)	11
1.2	Kardinalität unendlicher Mengen	13
1.3	Aussagenlogik und Beweismethoden	17
1.3.1	Aussageformen (Prädikate)	21
1.3.2	Logische Beweisregeln	23
1.4	Mathematische Induktion: Beweis von Aussagen $\forall x P(x)$	26
1.4.1	Induktion und Entwurf von Algorithmen	33
1.5	Das Taubenschlagprinzip: Beweis von Aussagen $\exists x P(x)$	35
1.6	Kombinatorische Abzählargumente	40
1.6.1	Prinzip der doppelten Abzählung	40
1.6.2	Binomialkoeffizienten	42
1.6.3	Prinzip von Inklusion and Exklusion	47
1.7	Aufgaben	50
2	Algebra und Elementare Zahlentheorie	57
2.1	Division mit Rest	58
2.2	Euklidischer Algorithmus	64
2.3	Primzahlen	65
2.4	Kleiner Satz von Fermat	67
2.4.1	Anwendung in der Kryptographie: RSA-Codes*	68
2.5	Chinesischer Restsatz	72
2.5.1	Anwendung: Schneller Gleichheitstest*	74
2.6	Gruppen	75

2.6.1	Zyklische Gruppen	78
2.7	Ringe und Körper	80
2.7.1	Polynomring	81
2.7.2	Komplexe Zahlen*	84
2.8	Allgemeine Vektorräume	85
2.9	Aufgaben	86
3	Einschub aus der Analysis	89
3.1	Endliche Folgen und Reihen	89
3.2	Unendlicher Folgen	98
3.2.1	Konvergenzkriterien für Folgen	101
3.2.2	Bestimmung des Grenzwertes	104
3.3	Unendliche Reihen	106
3.3.1	Konvergenzkriterien für Reihen	112
3.3.2	Anwendung: Warum Familiennamen aussterben?	117
3.3.3	Umordnungssatz	119
3.4	Grenzwerte bei Funktionen	123
3.5	Differentiation	124
3.6	Mittelwertsätze der Differentialrechnung	126
3.7	Approximation durch Polynome: Taylorentwicklung	129
3.8	Extremalstellen	131
3.9	Die Bachmann-Landau-Notation: klein o und groß O	132
3.10	Rekurrenzen*	140
3.10.1	Das Master Theorem	149
3.11	Aufgaben	151
4	Diskrete Stochastik	157
4.1	Intuition und Grundbegriffe	158
4.2	Drei Modellierungsschritte	161
4.2.1	Das Geburtstagsproblem	162
4.3	Stochastische Unabhängigkeit	163
4.4	Bedingte Wahrscheinlichkeit	165
4.4.1	Multiplikationssatz für Wahrscheinlichkeiten	168
4.4.2	Satz von der totalen Wahrscheinlichkeit	169
4.4.3	Satz von Bayes	171
4.5	Stochastische Entscheidungsprozesse	174

4.5.1	Das „Monty Hall Problem“	177
4.5.2	Stichproben	178
4.5.3	Das “Sekretärinnen-Problem” an der Börse	179
4.6	Zufallsvariablen	184
4.7	Erwartungswert und Varianz	186
4.8	Analytische Berechnung von $E[X]$ und $\text{Var}[X]$	190
4.9	Eigenschaften von $E[X]$ und $\text{Var}[X]$	191
4.10	Verteilungen diskreter Zufallsvariablen	199
4.11	Abweichung vom Erwartungswert	205
4.11.1	Markov-Ungleichung	206
4.11.2	Tschebyschev-Ungleichung	208
4.11.3	Chernoff-Ungleichungen	212
4.12	Das Urnenmodel – Hashing*	218
4.13	Bedingter Erwartungswert*	224
4.14	Summen von zufälliger Länge – Wald’s Theorem	228
4.15	Irrfahrten und Markov-Ketten	232
4.16	Statistisches Schätzen: Die Maximum-Likelihood-Methode*	240
4.17	Die probabilistische Methode*	243
4.18	Aufgaben	247
5	Lineare Algebra	253
5.1	Lineare Vektorräume	254
5.2	Basis und Dimension	256
5.3	Skalarprodukt und Norm	258
5.4	Dimensionsschranke und ihre Anwendungen*	260
5.5	Matrizen	263
5.6	Rang einer Matrix	270
5.7	Lösbarkeit der linearen Gleichungssysteme	272
5.8	Gauß-Verfahren	275
5.9	Inversen von Matrizen	279
5.10	Orthogonalität	282
5.11	Determinanten	285
5.12	Eigenwerte und Eigenvektoren	289
5.13	Einige Anwendungen des Matrizenkalküls*	292
5.13.1	Matrizenkalkül unf komplexe Zahlen	292
5.13.2	Diskrete Fourier-Transformation	293

5.13.3 Fehlerkorrigierende Codes	298
5.13.4 Expandergraphen	301
5.13.5 Expander-Codes	305
5.13.6 Markov-Ketten	307
5.14 Aufgaben	315

Schulstoff: Rechnen mit reellen Zahlen

In diesem Abschnitt stellen wir einige aus der Schule bekannten Fakten zusammen. Da nicht alle diese Fakten in der Schule auch *bewiesen* worden sind, werden wir das im Kapitel 3 “Einschub aus der Analysis” tun.

Einige wichtige Mengen in der Mathematik sind:

Die Menge der natürlichen Zahlen: $\mathbb{N} = \{0, 1, 2, \dots\}$

Die Menge der natürlichen Zahlen ohne Null: $\mathbb{N}_+ = \{1, 2, \dots\}$

Die Menge der ersten n positiven natürlichen Zahlen: $[n] = \{1, 2, \dots, n\}$

Die Menge der ganzen Zahlen: $\mathbb{Z} = \{\dots, -1, 0, 1, 2, \dots\}$

Die Menge der rationalen Zahlen: $\mathbb{Q} = \left\{ \frac{a}{b} : a \in \mathbb{Z} \text{ und } b \in \mathbb{N}_+ \right\}$

Die Menge der reellen Zahlen: \mathbb{R}

Summe, Differenz, Produkt und Quotient von zwei reellen Zahlen ist wieder eine reelle Zahl.



Ausnahme: Division durch 0 ist nicht erlaubt!

Anordnung der reellen Zahlen

Zwei beliebige reelle Zahlen x und y lassen sich vergleichen: d.h. es gilt entweder

$$\underbrace{x < y}_{x < y} \text{ oder } \underbrace{x = y}_{x \leq y} \text{ oder } \underbrace{x > y}_{x > y}$$

Für beliebige reelle Zahlen x, y und z gilt:

$$x < y \text{ und } y < z \implies x < z$$

$$x < y \implies x \pm z < y \pm z$$

$$x < y \text{ und } z > 0 \implies x \cdot z < y \cdot z$$

$$x < y \implies -x > -y$$

$$0 < x < y \implies 0 < \frac{1}{y} < \frac{1}{x}$$

$$x \cdot y > 0 \implies (x > 0, y > 0) \text{ oder } (x < 0, y < 0)$$



Vorsicht: Aus $x \cdot y < z$ folgt $x < \frac{z}{y}$ im Allgemeinen nicht! Dies gilt nur wenn $y > 0$.

Betrag

Ist x eine reelle Zahl, so ist der *Betrag* $|x|$ von x wie folgt definiert:

$$|x| = \begin{cases} x & \text{für } x \geq 0 \\ -x & \text{für } x < 0 \end{cases}$$

Anschauliche Bedeutung:

$|x|$ = Abstand zwischen 0 und x auf der Zahlengeraden.

Es gilt:

$$\begin{aligned} |x| &\geq 0 \\ |x \cdot y| &= |x| \cdot |y| \\ \left| \frac{x}{y} \right| &= \frac{|x|}{|y|} \quad (y \neq 0) \\ |x + y| &\leq |x| + |y| \quad (\text{Dreiecksungleichung}) \end{aligned}$$

Häufige Form:

$|a - b|$ = Abstand zwischen x und y auf der Zahlengeraden.

Summen- und Produktzeichen \sum und \prod

Zur Abkürzung längerer Summen vereinbart man

$$\sum_{i=1}^n a_i = a_1 + a_2 + a_3 + \dots + a_n$$

Zum Beispiel kürzt man $1 + 2 + 3 + \dots + n$ als $\sum_{i=1}^n i$ ab. Dazu müssen die Summanden mit einer Nummer (*Index*) (oben ist das i) versehen sein. Mit den durch ein Summenzeichen ausgedrückten (endlichen) Summen wird genauso gerechnet wie mit "normalen" Summen auch.

Der Name des Indices spielt keine Rolle:

$$\sum_{i=1}^n a_i = \sum_{j=1}^n a_j$$

Ist a_0, a_1, \dots eine Folge von Zahlen und $I \subseteq \{0, 1, \dots\}$ eine endliche Teilmenge der Indices, so ist

$$\sum_{i \in I} a_i$$

die Summe aller Zahlen a_i mit $i \in I$.

Man betrachtet auch Doppelsummen:

$$\sum_{i=1}^n \sum_{j=1}^m a_{ij} := \left(\sum_{j=1}^m a_{1j} \right) + \left(\sum_{j=1}^m a_{2j} \right) + \dots + \left(\sum_{j=1}^m a_{nj} \right)$$

Mit den durch ein Summenzeichen ausgedrückten (endlichen!) Summen wird genauso gerechnet wie mit “normalen” Summen auch. So kann man zum Beispiel die Reihenfolge der Summen vertauschen (eine solche Umformung nennt man auch das *Prinzip der doppelten Abzählung*):

$$\sum_{i=1}^n \sum_{j=1}^m a_{ij} = \sum_{j=1}^m \sum_{i=1}^n a_{ij}$$

Analog vereinbart man

$$\prod_{i=1}^n a_i = a_1 \cdot a_2 \cdot a_3 \cdot \dots \cdot a_n$$

Prominente Summen:

- Arithmetische Summe:

$$\sum_{k=1}^n k = 1 + 2 + 3 \dots + n = \frac{n(n+1)}{2}.$$

- Geometrische Summe:

$$\sum_{k=0}^n x^k = 1 + x + x^2 + \dots + x^n = \frac{1-x^{n+1}}{1-x} \text{ für } x \neq 1.$$

- Harmonische Summe:

$$\sum_{k=1}^n \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} = \ln n + \delta \text{ mit } 0 < \delta \leq 1.$$

Potenzen und Wurzel

Für $a \in \mathbb{R}, n \in \mathbb{N}_+$ wird definiert:

$$\begin{aligned} a^n &= \underbrace{a \cdot a \cdot \dots \cdot a}_{n \text{ mal}} \\ a^0 &= 1 \\ a^{-n} &= \frac{1}{a^n} \quad (a \neq 0) \end{aligned}$$

Für $a \in \mathbb{R}, a \geq 0, n \in \mathbb{N}_+$ gibt es genau eine *nicht-negative* reelle Zahl, die hoch n genommen a ergibt. Diese Zahl wird mit $\sqrt[n]{a}$ bezeichnet, d.h.

$$(\sqrt[n]{a} = x) \stackrel{\text{Def.}}{\iff} (x \geq 0) \text{ und } (x^n = a)$$

Für ungerades n und $a < 0$ ist auch $\sqrt[n]{a} = -\sqrt[n]{-a}$ definiert.

Gebrochene Exponenten: Für $a \in \mathbb{R}, a \geq 0, m, n \in \mathbb{N}_+$ wird definiert:

$$\begin{aligned} a^{m/n} &= \sqrt[n]{a^m} = (\sqrt[n]{a})^m \\ a^{-m/n} &= \frac{1}{a^{m/n}} \end{aligned}$$

Für alle $a, b \in \mathbb{R}$ und $p, q \in \mathbb{Q}$, für die die folgenden Ausdrücke definiert sind, gilt:

$$\begin{aligned} a^p \cdot a^q &= a^{p+q} \\ \frac{a^p}{a^q} &= a^{p-q} \\ a^p \cdot b^p &= (a \cdot b)^p \\ \frac{a^p}{b^p} &= \left(\frac{a}{b}\right)^p \\ (a^p)^q &= a^{p \cdot q} \end{aligned}$$

Die Rechenregeln für Wurzeln ergeben sich aus diesen Regeln durch den Übergang von $\sqrt[n]{a}$ zu $a^{1/n}$. Exponent und entsprechende Wurzel heben sich auf; d.h.

$$\sqrt[n]{x^n} = (\sqrt[n]{x})^n = x \text{ für alle } x \geq 0.$$



Vorsicht mit negativem x : es gilt z.B. $\sqrt{x^2} = |x|$ für alle $x \in \mathbb{R}$; $\sqrt{x^2} = x$ gilt nur, wenn $x \geq 0$.

Die obigen Rechenregeln gelten auch für irrationalen Exponenten a^x mit $x \in \mathbb{R} \setminus \mathbb{Q}$.

Lineare und quadratische Gleichungen

Die Lösung für

$$ax + b = 0 \quad \text{mit } a, b \in \mathbb{R} \text{ und } a \neq 0$$

ist

$$x = -\frac{b}{a}$$

Um die quadratische Gleichung

$$ax^2 + bx + c = 0 \quad \text{mit } a, b, c \in \mathbb{R}, a \neq 0$$

zu lösen, schreibt man zuerst sie um

$$x^2 + \frac{b}{a}x = -\frac{c}{a}$$

und stellt die linke Seite als Quadrat dar

$$\left(x + \frac{b}{2a}\right)^2 = -\frac{c}{a} + \frac{b^2}{4a^2} = \frac{b^2 - 4ac}{4a^2}$$

woraus

$$x = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

folgt.

Logarithmen

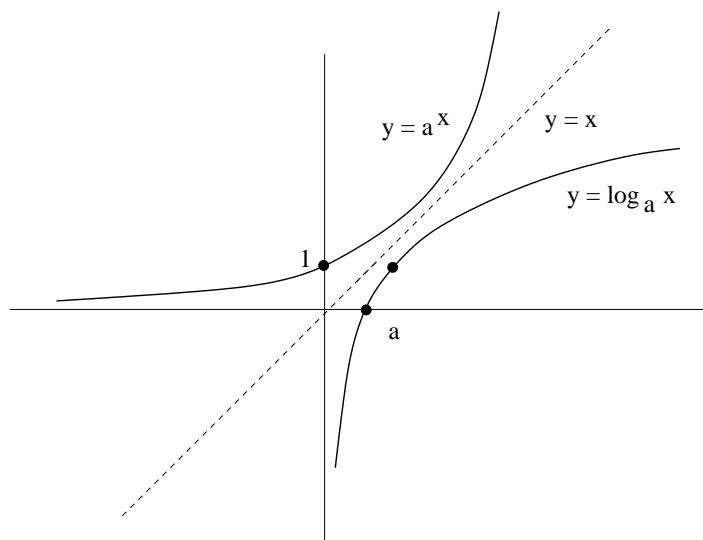
In der Gleichung

$$a^x = r$$

sind a (Basis, $a > 0, a \neq 1$) und r (Numerus, $r > 0$) gegeben. Gesucht ist die Zahl x . Diese Zahl heißt *Logarithmus* von r zur Basis a : schreibweise

$$x = \log_a r$$

Logarithmus $\log_e r$ zur Basis $e = 2,7182818\dots$ (Euler'sche Zahl) bezeichnet man als $\ln r$.



Die Rechenregeln mit Logarithmen sind in folgendem Satz zusammen gefasst.

Satz 0.3. $a, b, n > 1$ seien reelle Zahlen. Dann gilt

- (a) $a^{\log_a n} = n$.
- (b) $\log_n (a \cdot b) = \log_n a + \log_n b$ und $\log_n \left(\frac{a}{b}\right) = \log_n a - \log_n b$
- (c) *Basisvertauschregel:* $\log_a n = \frac{\log_b n}{\log_b a}$
- (d) $\log_b (a^n) = n \cdot \log_b a$
- (e) $(\log_a b) \cdot (\log_b a) = 1$
- (f) $b^{\log_a n} = n^{\log_a b}$

Beweis. (a) folgt aus der Definition.

(b):

$$n^{\log_n a + \log_n b} = n^{\log_n a} \cdot n^{\log_n b} = a \cdot b.$$

(c): Zu zeigen ist: $\log_b n = (\log_b a)(\log_a n)$

$$b^{(\log_b a)(\log_a n)} = a^{\log_a n} \stackrel{(a)}{=} n$$

(d):

$$b^n \cdot \log_b a = \left(b^{\log_b a}\right)^n = a^n \quad (\text{Definition von } \log_b a)$$

(e):

$$(\log_a b) \cdot (\log_b a) \stackrel{(c)}{=} \frac{\log_b b}{\log_b a} \cdot \log_b a = 1.$$

(f):

$$b^{\log_a n} \stackrel{(c)}{=} \left(b^{\log_b n}\right)^{1/\log_b a} = n^{1/\log_b a} \stackrel{(e)}{=} n^{\log_a b}$$

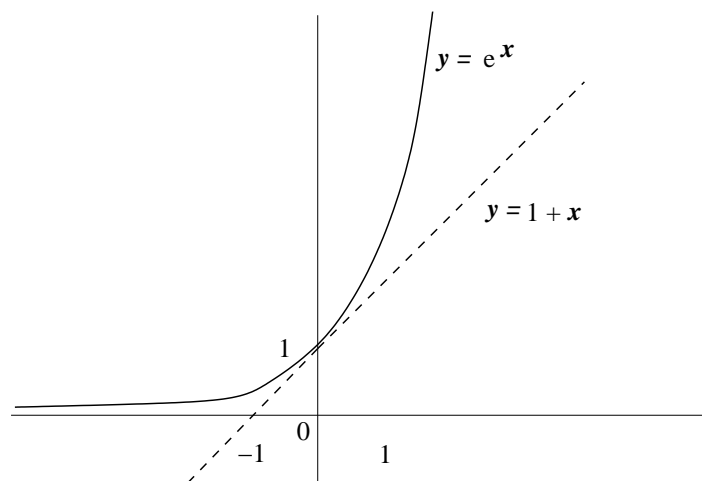
□

Für alle $x \in \mathbb{R}$, $x \neq 0$ gilt die folgende sehr nützliche Ungleichung:

$$1 + x < e^x \quad (1)$$

und für alle $x \in (-1, 1)$ gilt²

$$1 + x \geq e^{x/(1+x)} \quad (2)$$



²Wir werden diese Ungleichungen später beweisen (siehe Lemma 3.62).

Gauß-Klammern $\lfloor x \rfloor$ und $\lceil x \rceil$

Für eine reelle Zahl $x \in \mathbb{R}$ ist

$$\lfloor x \rfloor := \max\{b \in \mathbb{Z} : b \leq x\}$$

$$\lceil x \rceil := \min\{a \in \mathbb{Z} : x \leq a\}$$

Eigenschaften:

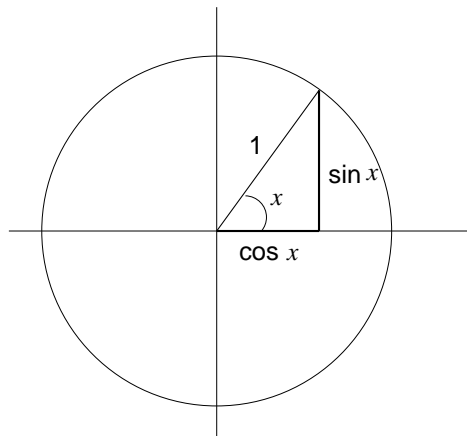
$$x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$$

$$\lfloor -x \rfloor = -\lceil x \rceil; \quad \lceil -x \rceil = -\lfloor x \rfloor$$

Sei $n \in \mathbb{N}$ eine natürliche Zahl, $2^{m-1} \leq n < 2^m$. Dann hat die binäre Darstellung von n genau

$$m = \lfloor \log_2 n \rfloor + 1 = \lceil \log_2(n+1) \rceil$$

Bits.

Sinus und Cosinus

Symmetrie:

$$\cos(-x) = \cos x$$

$$\sin(-x) = -\sin x$$

Pythagoras:

$$\cos^2 x + \sin^2 x = 1$$

Additionstheoreme:

$$\cos(x+y) = \cos x \cdot \cos y - \sin x \cdot \sin y$$

$$\sin(x+y) = \sin x \cdot \cos y + \cos x \cdot \sin y$$

Modulare Arithmetik

- Euklid's Algorithmus zur Berechnung von $\text{ggT}(a, b)$: $\text{ggT}(a, b) = \text{ggT}(b, a \bmod b)$.
- $\mathbb{Z}_m^* = \{a \in \mathbb{Z}_m : a \neq 0 \text{ und } \text{ggT}(a, m) = 1\}$ ist eine multiplikative Gruppe.
- Kleiner Satz von Fermat: p prim $\Rightarrow a^{p-1} \equiv 1 \pmod p$ für alle $a \in \mathbb{N}$.
- Chinesischer Restsatz: Jedes Gleichungssystem

$$x \equiv a_i \pmod{p_i}, \quad i = 1, \dots, n$$

modulo Primzahlen p_i hat genau eine Lösung x mit $0 \leq x \leq p_1 \cdot p_2 \cdot \dots \cdot p_n$. Insbesondere gilt

$$\mathbb{Z}_{p_1 \cdot p_2 \cdot \dots \cdot p_n} \xleftrightarrow{\text{Bijektion}} \mathbb{Z}_{p_1} \times \mathbb{Z}_{p_2} \times \dots \times \mathbb{Z}_{p_n}.$$

- Sind $a, b < \prod_{i=1}^n p_i$ für Primzahlen p_1, \dots, p_n so gilt:

$$a = b \iff a \equiv b \pmod{p_i} \text{ für alle } i = 1, \dots, n.$$

- Jede endliche Gruppe (G, \circ) , deren Ordnung $p = |G|$ eine Primzahl ist, ist zyklisch: Für alle $a \in G$ gilt

$$G = \{a, a^2, a^3, \dots\}.$$

Prominente unendliche Reihen

- Geometrische Reihe: für $|x| < 1$

$$1 + x + x^2 + x^3 + x^4 + \dots + x^n + \dots = \frac{1}{1-x}.$$

- Verallgemeinerte geometrische Reihe: für $|x| < 1$

$$x + 2x^2 + 3x^3 + 4x^4 + \dots + nx^n + \dots = \frac{x}{(1-x)^2}.$$

- Modifizierte harmonische Reihe: für $r > 1$

$$1 + \frac{1}{2^r} + \frac{1}{3^r} + \frac{1}{4^r} + \dots + \frac{1}{n^r} + \dots = a$$

mit

$$1 < a \leq 1 + \frac{1}{2^{r-1} - 1}.$$

Insbesondere gilt

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \dots + \frac{1}{n^2} + \dots = \frac{\pi^2}{6}.$$

Ableitungen

Seien $f, g : I \rightarrow \mathbb{R}$ differenzierbar in $x \in I$.

- $(f + g)'(x) = f'(x) + g'(x)$ (Summenregel)
- $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$ (Produktregel)
- Ist $f(x) \neq 0$, so

$$\left(\frac{f}{g}\right)'(x) = \frac{g'(x)f(x) - g(x)f'(x)}{(f(x))^2} \quad (\text{Quotientenregel})$$

- Sei $f : I \rightarrow \mathbb{R}$ differenzierbar in $x \in I$, sei $J \supseteq f(I)$ und $g : J \rightarrow \mathbb{R}$ differenzierbar in $f(x)$. Dann ist $g \circ f(x) = g(f(x))$ differenzierbar in $x \in I$ und

$$(g \circ f)'(x) = g'(f(x)) \cdot f'(x) \quad (\text{Kettenregel})$$

Ableitungen einiger Funktionen (für $c \in \mathbb{R}$ und $n \in \mathbb{N}$):

- $f(x) = c \Rightarrow f'(x) = 0$
- $f(x) = x \Rightarrow f'(x) = 1$
- $f(x) = x^n \Rightarrow f'(x) = nx^{n-1}$
- $f(x) = \ln g(x) \Rightarrow f'(x) = \frac{1}{g(x)}g'(x)$ Spezialfall: $(\ln x)' = \frac{1}{x}$
- $f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2}$
- $f(x) = c^{g(x)} \Rightarrow f'(x) = (\ln c)e^{g(x)}g'(x)$ Spezialfall: $(c^x)' = (\ln c)e^x$
- $f(x) = e^{cg(x)} \Rightarrow f'(x) = ce^{cg(x)}g'(x)$ Spezialfall: $(e^x)' = e^x$

Einige Integrale:

$$\int e^x dx = e^x,$$

$$\int \frac{1}{x} dx = \ln x,$$

$$\int x^n dx = \frac{1}{n+1}x^{n+1} \quad \text{für } n \neq -1$$

Extremstellen

Gegeben sei ein offenes Intervall I in \mathbb{R} , eine Funktion $f : I \rightarrow \mathbb{R}$ und eine Stelle $a \in I$. Sei $f'(a) = 0$. Gilt außerdem $f''(a) > 0$ (bzw. $f''(a) < 0$), so ist a lokale Minimalstelle (bzw. Maximalstelle) von f .

Taylorreihen

$$\begin{aligned}
 e^x &= \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots \\
 \ln(1+x) &= \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k} = 1 - x + \frac{x^2}{2} - \frac{x^3}{3} \pm \dots \\
 \cos x &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\
 \sin x &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots
 \end{aligned}$$

Taylorentwicklung von Funktionen ist im Appendix (Abschnitt 3.7) skizziert.

Stochastik

- Satz von der totalen Wahrscheinlichkeit:

$$\Pr\{A\} = \Pr\{B\} \cdot \Pr\{A|B\} + \Pr\{\bar{B}\} \cdot \Pr\{A|\bar{B}\}.$$

- Satz von Bayes:

$$\Pr\{A|B\} = \frac{\Pr\{A\}}{\Pr\{B\}} \cdot \Pr\{B|A\}$$

- Erwartungswert einer Zufallsvariable $X : \Omega \rightarrow I$

$$E[X] := \sum_{a \in I} a \cdot \Pr\{X = a\}.$$

- Linearität des Erwartungswertes: $E[X + Y] = E[X] + E[Y]$.
- Varianz: $\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$.
- Ist X die Indikatorvariable für ein Ereignis, so gilt

$$E[X_A] = \Pr\{A\} \quad \text{und} \quad \text{Var}[X_A] = \Pr\{A\} - \Pr\{A\}^2 = \Pr\{A\} \cdot \Pr\{\bar{A}\}$$

- Geometrische Verteilung = "solange bis" Verteilung: Unabhängige Versuche mit der Erfolgswahrscheinlichkeit $p \neq 0$

$$E[\text{Anzahl der Versuche bis zum ersten Erfolg}] = \frac{1}{p}.$$

- Berechnung von $E[X]$ und $\text{Var}[X]$: Setze

$$f(x) := \sum_{a \in I} x^a \Pr\{X = a\}$$

und berechne $f'(x)$ wie auch $f''(x)$. Dann gilt

$$E[X] = f'(1) \quad \text{und} \quad \text{Var}[X] = f''(1) + E[X] - E[X]^2.$$

- Markov-Ungleichung für $X \geq 0$:

$$\Pr\{X \geq k\} \leq \frac{E[X]}{k}.$$

- Tschebyschev-Ungleichung für X mit $E[X^2] < \infty$:

$$\Pr\{|X - E[X]| \geq k\} \leq \frac{\text{Var}[X]}{k^2}.$$

- Chernoff-Ungleichung für die Summe $X = X_1 + \dots + X_n$ von *unabhängigen* Bernoulli-Variablen mit $\Pr\{X_i = 1\} = p$:

$$\Pr\{X \geq (1 + \delta)np\} \leq e^{-\delta^2 np/3} \quad \text{für jedes } 0 < \delta < 1.$$

Kapitel 1

Grundbegriffe und Beweismethoden

Contents

1.1 Mengen, Relationen und Funktionen	1
1.1.1 Relationen	4
1.1.2 Graphen (binäre Relationen)	8
1.1.3 Abbildungen (Funktionen)	11
1.2 Kardinalität unendlicher Mengen	13
1.3 Aussagenlogik und Beweismethoden	17
1.3.1 Aussageformen (Prädikate)	21
1.3.2 Logische Beweisregeln	23
1.4 Mathematische Induktion: Beweis von Aussagen $\forall x P(x)$	26
1.4.1 Induktion und Entwurf von Algorithmen	33
1.5 Das Taubenschlagprinzip: Beweis von Aussagen $\exists x P(x)$	35
1.6 Kombinatorische Abzählargumente	40
1.6.1 Prinzip der doppelten Abzählung	40
1.6.2 Binomialkoeffizienten	42
1.6.3 Prinzip von Inklusion and Exklusion	47
1.7 Aufgaben	50

1.1 Mengen, Relationen und Funktionen

Der (einzige) Barbier in Sevilla rasiert genau die Männer der Stadt, die *nicht sich selbst rasieren*. Es scheint nichts dagegen zu sprechen, die Menge M aller Männer, die der Barbier rasiert, zu bilden:

$$M = \{x : x \text{ ist ein männlicher Einwohner von Sevilla, den der Barbier rasiert}\}.$$



Ist der Barbier ein Element von M ? Weder noch!

In Anbetracht solcher Paradoxe gibt die folgende, von Cantor 1895 gegebene Erklärung eine zumindest für die praktische Arbeit ausreichend präzise Fassung des Begriffs der Menge:



Erklärung (keine Definition!) Eine *Menge* ist die Zusammenfassung bestimmter, wohlunterschiedener Objekte unserer Anschauung oder unseres Denkens, wobei von *jedem* dieser Objekte *eindeutig* feststeht, ob es zur Menge gehört oder nicht. Die Objekte der Menge heißen *Elemente* der Menge.

Diese Schwierigkeiten (den Begriff einer “Menge” zu definieren) spielen in dieser Vorlesung keine Rolle, man sollte aber davon wissen.

Wir schreiben “ $a \in A$ ”, wenn a ein Element der Menge A ist; “ $a \notin A$ ” ist die Negation von $a \in A$. Wir schreiben “ $A \subseteq B$ ”, wenn A eine Teilmenge von B ist, d.h. wenn jedes Element aus A auch in B ist; wenn $A \subseteq B$ und $A \neq B$, dann schreibt man auch $A \subset B$.

Die *Potenzmenge* $\mathcal{P}(A)$ einer Menge A ist die Menge *aller* Teilmengen von A (anstatt $\mathcal{P}(A)$ schreibt man oft 2^A). Beispiel: $\mathcal{P}(\{1,2\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$.

Man stellt die Mengen dar entweder durch Aufzählung der Elemente, z.B.

$$A = \{1, 3, 4, 7\}$$

oder durch Beschreibung der Eigenschaften der Elemente, z.B.

$$A = \{a : a \text{ ist eine ganze Zahl mit } a^2 = a\} = \{0, 1\}$$

$$A = \{a : a \text{ ist ganze Zahl, } 1 \leq a \leq 5, a \neq 2\} = \{1, 3, 4, 5\}$$

Nach dem Zeichen “:” folgen die Bedingungen, die die Elemente erfüllen müssen; eine Komma “,” bedeutet hier ein “und”. Die Anzahl der Elemente in einer endlichen Menge A , die Mächtigkeit von A , wird mit $|A|$ bezeichnet (manchmal auch $\#A$).

Wichtige Regeln beim Umgang mit Mengen sind:

- Eine Menge enthält jedes Element nur einmal: $a \in A$ oder $a \notin A$. Es gibt also genau eine *leere Menge* $\emptyset = \{\}$, die keine Elemente enthält.
- Eine Menge ist durch ihre Elemente bestimmt, d.h. zwei Mengen A und B sind genau dann gleich, wenn sie die gleichen Elemente haben. Zwei Mengen A und B sind also genau dann gleich, wenn $A \subseteq B$ und $B \subseteq A$.



Diese Regel ist wichtig: So zeigt man Mengengleichheit! Um $A = B$ zu beweisen, muss man also folgendes zeigen:

$$\text{für jedes } x \in A \text{ gilt } x \in B$$

und

$$\text{für jedes } x \in B \text{ gilt } x \in A$$

- Elemente einer Menge können wieder Mengen sein: z.B. $\{\mathbb{N}\}$ ist eine Menge mit genau einem Element, und $\{0, \mathbb{N}\}$ hat genau 2 Elemente.

Man beachte den Unterschied zwischen \in und \subseteq :

$$a \in A, \{a\} \subseteq A \text{ und } \{a\} \in 2^A \text{ sind daher äquivalent.}$$

Verknüpfungen von Mengen:

$$A \cap B = \{x : x \in A \text{ und } x \in B\} \text{ (Schnittmenge)}$$

$$A \cup B = \{x : x \in A \text{ oder } x \in B\} \text{ (Vereinigungsmenge)}$$

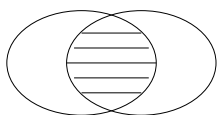
$$A \setminus B = \{x : x \in A \text{ und } x \notin B\} \text{ (Differenz)}$$

$$A \oplus B = \{x : x \in A \setminus B \text{ oder } x \in B \setminus A\} \text{ (symmetrische Differenz)}$$

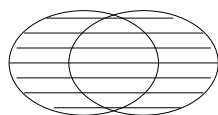
Ist eine Menge U fixiert (ein "Universum") und ist $A \subseteq U$, dann heißt $\bar{A} = U \setminus A$ das Komplement von A (im Universum U).

A und B heißen *disjunkt*, falls $A \cap B = \emptyset$ gilt.

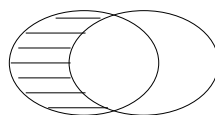
Die eingeführten Mengenoperationen lassen sich mit Hilfe der sogenannten *Venn Diagramme* graphisch gut veranschaulichen.



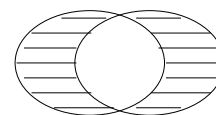
Durchschnitt



Vereinigung



Differenz



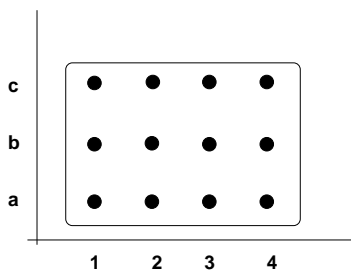
Symmetrische
Differenz

Das *kartesische Produkt*¹ ist definiert durch

$$A \times B = \{(a, b) : a \in A \text{ und } b \in B\}$$



Sind $a \neq b$, so gilt $\{a, b\} = \{b, a\}$ aber $(a, b) \neq (b, a)$.



Die n -te kartesische Potenz von A ist $A^n := \overbrace{A \times A \times \dots \times A}^{n \text{ mal}}$. Die Elemente von A^n sind n -Tupel (oder Vektoren) (a_1, \dots, a_n) mit $a_i \in A$.

¹Der Name "kartesisch" hat eine Geschichte: René Descartes, Philosoph und Mathematiker des 17. Jahrhunderts, hat den systematischen Gebrauch von Koordinaten in der Geometrie eingeführt. Koordinaten im k -dimensionalen Raum sind natürlich k -Tupel (in der Ebene: Paare, im dreidimensionalen: Tripel) von Zahlen. Und weil es damals üblich war, seinen Namen zu latinisieren, gab sich Descartes den Namen Cartesius. Ihm zu Ehren spricht man von kartesischen Koordinaten und vom kartesischen Produkt.



Ist zum Beispiel $A = \{0, 1\}$, so ist $A^n = \{0, 1\}^n$ der n -dimensionale binäre Würfel. Dieses Objekt stellt eine wichtige Struktur mit Anwendungen in der Informatik dar. Ist die Reihenfolge der Elemente in $A = \{a_1, a_2, \dots, a_n\}$ fixiert, so kann man die Teilmengen $S \subseteq A$ mit 0-1 Vektoren $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \{0, 1\}^n$ mit

$$x_i = \begin{cases} 1 & \text{falls } a_i \in S \\ 0 & \text{falls } a_i \notin S \end{cases}$$

kodieren. Ein solcher Vektor \mathbf{x} heißt dann *charakteristischer Vektor* (oder *Inzidenzvektor*) von S . Sind zum Beispiel $A = \{1, 2, 3, 4, 5\}$ und $S = \{1, 3, 4\}$, so kann man den Vektor $\mathbf{x} = (1, 0, 1, 1, 0)$ als Code von S betrachten. Diese Kodierung ist in der Informatik sehr wichtig: So stellt man Mengen im Computer dar!

1.1.1 Relationen

Eine (binäre) Relation ist eine Beziehung zwischen Dingen. Zum Beispiel: zwei Menschen können verwandt sein oder nicht. Ein Auto kann länger sein als ein anderes oder nicht. Zwei Mengen können identisch sein oder nicht.

Will man verschiedene Mengen miteinander vergleichen, braucht man eine Beziehung zwischen diesen. Und auch eine einzelne Menge ist strukturlos, solange die einzelnen Elemente völlig beziehungslos zueinander sind. Hat man aber eine Beziehung (Relation), so entsteht aus dem Chaos eine Struktur, und die Untersuchung der Eigenschaften dieser Beziehungen ist eine der Hauptaufgaben der Mathematik.

Definition: Eine (binäre) *Relation* zwischen zwei Mengen A und B ist eine Teilmenge

$$R \subseteq A \times B.$$

Im Falle $A = B$ sprechen wir von einer *Relation in A* .

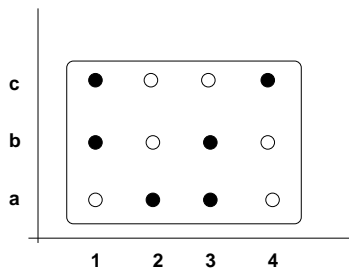


Abbildung 1.1: $A = \{a, b, c\}$, $B = \{1, 2, 3, 4\}$ und $R = \{(a, 2), (a, 3), (b, 1), (b, 3), (c, 1), (c, 4)\}$

Eine Relation R ist also eine Menge geordneter Paare. Schreibweise: anstatt $(a, b) \in R$ schreibt man oft $a \sim_R b$ oder aRb .

Wenn wir irgendeine Beziehung als eine Relation abstrakt modelliert haben, können wir über einige Eigenschaften dieser Relation sprechen.

Seien $a, b, c \in A$ beliebig. Eine binäre Relation $R \subseteq A \times A$ ist:

- *reflexiv*, wenn $a \sim_R a$

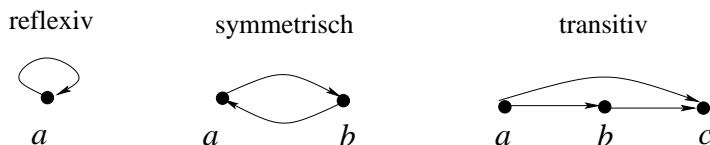
- *symmetrisch*, wenn aus $a \sim_R b$ stets $b \sim_R a$ folgt
- *antisymmetrisch*, wenn aus $a \sim_R b$ und $b \sim_R a$ stets $a = b$ folgt
- *asymmetrisch*, wenn aus $a \sim_R b$ stets $\neg(b \sim_R a)$ folgt
- *transitiv*, wenn aus $a \sim_R b$ und $b \sim_R c$ stets $a \sim_R c$ folgt

Der Unterschied zwischen antisymmetrischen und asymmetrischen Relationen ist, dass in antisymmetrischen Relationen auch $a \sim_R a$ gelten kann, während diese Eigenschaft in asymmetrischen Relationen verboten ist.

Äquivalenzrelationen sind deshalb so wichtig, weil man oft nur an bestimmten Eigenschaften der untersuchten Objekte interessiert ist. Unterscheiden sich zwei Objekte nicht (bezüglich der Eigenschaften, die man gerade untersucht), so sagt man, sie sind *äquivalent*.

Eine Relation $\sim \subseteq A \times A$ heißt *Äquivalenzrelation*, wenn sie die folgenden drei Eigenschaften hat:

- (i) $a \sim a$ für alle $a \in A$ (reflexiv)
- (ii) Wenn $a \sim b$, dann auch $b \sim a$ (symmetrisch)
- (iii) Wenn $a \sim b$ und $b \sim c$, dann gilt auch $a \sim c$ (transitiv)

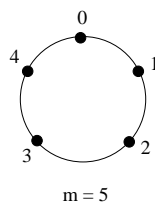


Ist eine Äquivalenzrelation $\sim \subseteq A \times A$ auf einer Menge A gegeben, so heißt eine Teilmenge $S \subseteq A$ *Äquivalenzklasse* (bezüglich \sim), falls gilt:

1. $S \neq \emptyset$,
2. $x, y \in S \Rightarrow x \sim y$,
3. $x \in S, y \in A$ und $x \sim y \Rightarrow y \in S$.

► *Beispiel 1.1*: 1. Eine Äquivalenzrelation in \mathbb{Z} : $a \sim b \iff a$ und b den selben Rest r modulo 5 haben. Es gibt fünf Äquivalenzklassen: $[r] = \{x \in \mathbb{Z} : x = 5y + r\}$, $r = 0, 1, 2, 3, 4$ (Restklassen modulo 5):

...	-10	-5	0	5	10	15	20	...
...	-9	-4	1	6	11	16	21	...
...	-8	-3	2	7	12	17	22	...
...	-7	-2	3	8	13	18	23	...
...	-6	-1	4	9	14	19	24	...



2. Viele praktische Beispiele ..., z.B. Post: Briefe mit gleicher Postleitzahl gelten für die Sortierung in Postsäcken (Äquivalenzklassen) als äquivalent.

Eine *disjunkte Zerlegung* einer Menge A besteht aus paarweise disjunkten Teilmengen A_1, \dots, A_n von A , deren Vereinigung die ganze Menge A ergibt, d.h. es muß $A = A_1 \cup A_2 \cup \dots \cup A_n$ und $A_i \cap A_j = \emptyset$ für alle $1 \leq i \neq j \leq n$ gelten.

Satz 1.2. Äquivalenzklassen bilden eine disjunkte Zerlegung.

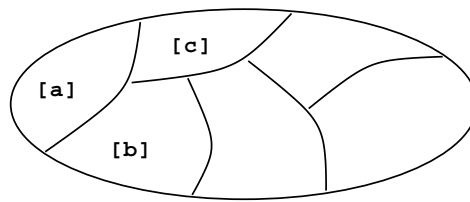
Beweis. Sei \sim eine Äquivalenzrelation auf einer Menge A . Zu zeigen: (i) jedes Element $a \in A$ gehört zu *genau einer* Äquivalenzklasse, und (ii) verschiedene Äquivalenzklassen sind disjunkt. Dazu definieren wir für ein festes gegebenes $a \in A$ die Menge ²

$$S_a := \{x \in A : x \sim a\}$$

aller Elemente, die äquivalent zu a sind. Wegen $a \sim a$ gilt $a \in S_a$, und es folgt $S_a \neq \emptyset$. Wir zeigen, dass S_a eine Äquivalenzklasse ist. Sind $x, y \in S_a$, so gilt $x \sim a$ und $y \sim a$, also $x \sim y$, da \sim symmetrisch und transitiv ist. Gilt nun $x \in S_a$, $y \in A$ und $x \sim y$, dann haben wir $x \sim a$, also auch $y \sim a$ (Transitivität) und daher $y \in S_a$. Damit ist gezeigt, dass a in mindestens einer Äquivalenzklasse enthalten ist.

Es bleibt zu zeigen, dass zwei Äquivalenzklassen S und S' entweder gleich oder disjunkt sind. Angenommen, S und S' sind nicht disjunkt und $a \in S \cap S'$. Gilt nun $x \in S$, so gilt $x \sim a$, und wegen $a \in S'$ folgt auch $x \in S'$. Also ist $S \subseteq S'$. Ebenso beweist man $S' \subseteq S$, woraus $S = S'$ folgt. \square

Jede Äquivalenzrelation \sim auf einer Menge A liefert also eine *Zerlegung* von A in disjunkte Äquivalenzklassen.



Diese Äquivalenzklassen betrachtet man als Elemente einer neuen Menge, die man mit A/\sim bezeichnet. Man nennt sie die *Quotientenmenge* von A nach der Äquivalenzrelation \sim . Die *Elemente* von A/\sim sind also spezielle *Teilmengen* von A . Indem man jedem Element $a \in A$ die Äquivalenzklasse S_a zuordnet, in der es enthalten ist, erhält man eine *kanonische* (d.h. in der gegebenen Situation eindeutig festgelegte) Abbildung

$$A \rightarrow A/\sim \quad \text{mit} \quad a \mapsto S_a$$

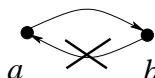
Eine Relation $< \subseteq A \times A$ heißt *Ordnung*, wenn sie die folgenden drei Eigenschaften hat:

- (i) $\neg(a < a)$ für alle $a \in A$ (antireflexiv)
- (ii) Wenn $a < b$ und $a \neq b$, dann $\neg(b < a)$ (antisymmetrisch)
- (iii) Wenn $a < b$ und $b < c$, dann ist auch $a < c$ (transitiv)

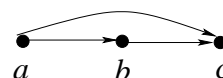
antireflexiv



antisymmetrisch



transitiv



²So definierte Mengen S_a bezeichnet man oft auch mit $[a]$; a ist dann ein Repräsentant der Äquivalenzklasse $[a]$. Beachte, dass aus $a \sim b$ immer $[a] = [b]$ folgt. D.h. *jedes* Element aus $[a]$ kann man als den Repräsentanten der Klasse $[a]$ nehmen; bis auf die Äquivalenzrelation \sim sind sie alle "gleich".

Einige Beispiele:

- Die Potenzmenge $\mathcal{P}(S)$ mit der Ordnung: $A < B$ genau dann, wenn $A \subset B$ (siehe Abb. 1.2(a)).
- Die Menge aller positiven natürlichen Zahlen mit der Ordnung: $a < b$ genau dann, wenn a kleiner als b ist.
- Die Menge aller positiven natürlichen Zahlen mit der Ordnung: $a < b$ genau dann, wenn b ohne Rest durch a teilbar ist (siehe Abb. 1.2(b)).
- Die Menge aller Vektoren in \mathbb{R}^n mit der Ordnung $(a_1, \dots, a_n) < (b_1, \dots, b_n)$ genau dann, wenn $a_i \leq b_i$ für alle i , und $a_i < b_i$ für mindestens ein i gilt.

Kleine Ordnungen kann man mittels sogenannter *Hasse-Diagramme* darstellen, wobei die Elemente als Punkte und die Relationen als Verbindungen vom kleineren unteren Elementen zum größeren oberen Elementen dargestellt werden (siehe Abb. 1.2).

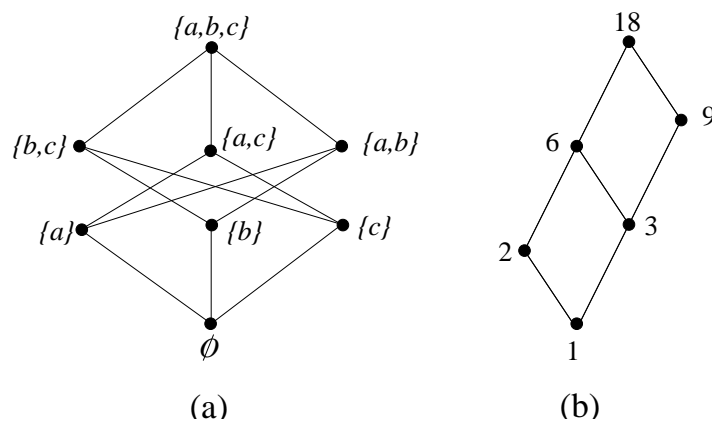


Abbildung 1.2:

Die Relation “kleiner” (“<”) auf \mathbb{R} hat eine weitere interessante Eigenschaft: sie ist *vollständig*, d.h. sie erfüllt die Forderung, dass je zwei Elemente vergleichbar sind. Eine Ordnung mit dieser Eigenschaft heißt *vollständige Ordnung* (oder “totale Ordnung” oder “lineare Ordnung”). Ist eine Ordnungsrelation nicht vollständig, so nennt man sie manchmal *partielle Ordnung*. Welche von der oben aufgelisteten Ordnungen sind vollständig und welche nur partiell? (Übungsaufgabe!)

Da Relationen die Teilmengen $R \subseteq A \times B$ des kartesischen Produkts $A \times B$ sind (bei festen Mengen A und B), sind Durchschnitt, Vereinigung, Differenz und Komplement von Relationen erklärt und ergeben wieder eine Relation zwischen A und B ; ebenso ist die Inklusion für Relationen definiert und es gelten alle für Mengenoperationen üblichen Regeln. Es gibt aber auch eine spezielle Operation zwischen Relationen – ihre “Verknüpfung” oder “Komposition.”

Sei $R \subseteq A \times B$ und $S \subseteq B \times C$. Dann wird die *Komposition* $R \circ S$ als Relation zwischen A und C definiert durch

$$R \circ S := \{(a, c) : \text{es gibt ein } b \in B \text{ mit } (a, b) \in R \text{ und } (b, c) \in S\}.$$

► *Beispiel 1.3* : Sei M eine Menge von Menschen,

$$R = \{(x, y) : x \text{ ist Mutter von } y\}$$

$$S = \{(y, z) : y \text{ ist verheiratet mit } z\}$$

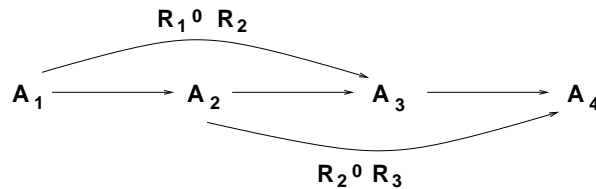
Dann ist

$$R \circ S = \{(x, z) : x \text{ ist Schwiegermutter von } z\}$$

Eine wichtige Eigenschaft dieser Komposition von Relationen (und somit auch von Abbildungen) ist ihre Assoziativität: Sind $R_i \subseteq A_i \times A_{i+1}$, $i = 1, 2, 3$ Relationen, dann gilt:

$$(R_1 \circ R_2) \circ R_3 = R_1 \circ (R_2 \circ R_3).$$

Man mache sich diese Aussage an Hand eines Diagramms klar:



Eine Sonderrolle, die dann besonders bei Abbildungen Bedeutung erlangt, spielt bezüglich der Komposition die *identische Relation* I_A : Sie wird in einer beliebigen Menge A durch die Gleichheit definiert, also durch:

$$xI_Ay \iff x = y$$

Vor allem dann, wenn man sie als Teilmenge von $A \times A$ versteht (so haben wir Relationen ja definiert), wird sie auch als *Diagonale* bezeichnet. Mit dieser identischen Relation gilt dann:

$$R \circ I_A = I_A \circ R = R.$$

Dreht man alle Paare (a, b) in R um, so bekommt man die *inverse Relation* R^{-1} :

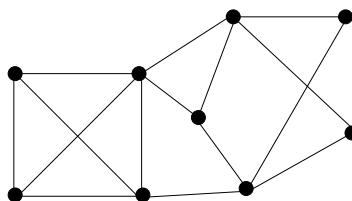
$$R^{-1} := \{(b, a) : (a, b) \in R\}$$



Ist $R \subseteq A \times B$ eine Abbildung, so ist dennoch R^{-1} im Allgemeinen keine Abbildung. Im Allgemeinen gilt auch weder $R \circ R^{-1} = I_A$ noch $R^{-1} \circ R = I_B$. Sei zum Beispiel $A = \{1, 2\}$ und $B = \{a, b\}$ sowie $R = \{(1, a), (1, b)\}$. Dann ist $R^{-1} = \{(a, 1), (b, 1)\}$ und man erhält $R \circ R^{-1} = \{(1, 1)\}$ und $R^{-1} \circ R = B \times B$.

1.1.2 Graphen (binäre Relationen)

Binäre Relationen auf endlichen Mengen nennt man auch "Graphen." Solche Relationen sind anschaulich und vielseitig anwendbar. Wir reden hier nicht von Graphen einer Funktion, für uns ist ein Graph so etwas:



Es gibt Punkte (*Knoten*) und Linien (*Kanten*) zwischen einigen dieser Punkte. Graphen sind so vielseitig, weil die Idee so einfach ist: es gibt eine Menge von Objekten (die Knoten), und zwei Knoten sind entweder verbunden oder nicht.

Ein *gerichteter Graph* ist also ein Paar $G = (V, E)$ mit $E \subseteq V \times V$. Ist die Relation E symmetrisch und anti-reflexiv, so ist $G = (V, E)$ ein *ungerichteter Graph* (oder einfach ein *Graph*). Wir werden fast ausschließlich nur solche (ungerichtete) Graphen betrachten. Die Elemente v der Menge V werden Knoten genannt, die Elemente $e = \{v, u\}$ (oft schreibt man $e = uv$, um Klammern zu sparen) der Menge E sind die Kanten des Graphen. Man sagt, dass die Kante $e = \{v, u\}$ die Knoten v und u verbindet; die Knoten u, v selbst sind *Endknoten* von e . Zwei Knoten, die in einem Graphen durch eine Kante verbunden sind, heißen *adjazent* oder *benachbart*. Die Anzahl aller Nachbarn von u nennt man als *Grad* von u und bezeichnet ihm mit $d(u)$ (engl. Grad = degree).

► *Beispiel 1.4*: Der n -dimensionale binäre Würfel ist ein Graph $Q_n = (V, E)$ dessen Knoten den 2^n binären Strings der Länge n entsprechen. Zwei Knoten (=Strings) sind genau dann adjazent, wenn sie sich in genau einem Bit unterscheiden (siehe Abb. 1.3).

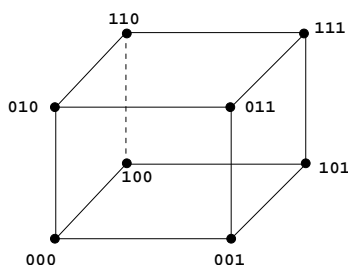
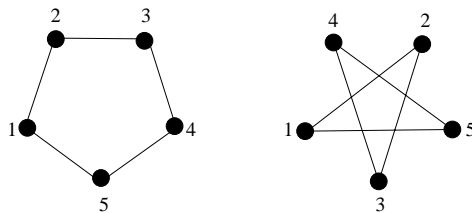


Abbildung 1.3: Der 3-dimensionale binäre Würfel Q_3 . Er hat 8 Knoten, je vom Grad 3, und 12 Kanten.

Graphen werden gewöhnlich mit Hilfe geometrischer Diagramme dargestellt. Oft aber wird zwischen einem Graphen und einem diesen Graphen darstellenden Diagramm nicht deutlich unterschieden. Es muss aber ausdrücklich davor gewarnt werden, Graphen und Diagramme gleichzusetzen. Spezielle geometrische Darstellungen können das Vorhandensein von Eigenschaften suggerieren, die der dargestellte Graph als eine Struktur, die lediglich aus einer Knotenmenge und einer Relation über dieser Menge besteht, gar nicht besitzen kann. Zum Beispiel kann ein Kreis der Länge 5 sowohl als 5-zackiger Stern als auch als 5-Eck dargestellt werden:



Um verschiedene Diagramme, die demselben Graph entsprechen, als ein Objekt zu betrachten, benutzt man den Begriff der "Isomorphie". Nämlich, man sagt, dass zwei Graphen $G = (V, E)$ und $G' = (V', E')$ *isomorph* sind, falls es eine bijektive Abbildung $f : V \rightarrow V'$ mit der folgendem Eigenschaft gibt: Für jede zwei Knoten $u \neq v$ in V

$$f(u) \text{ und } f(v) \text{ sind Nachbarn in } G' \iff u \text{ und } v \text{ sind Nachbarn in } G.$$

D.h. zwei Graphen sind isomorph, falls sie sich nur bis auf der Numerierung ihrer Knoten unterscheiden. Die folgende Abbildung gibt stellt einen sogenannten *Petersen's Graphen* (links) dar und alle diese Graphen sind isomorph:

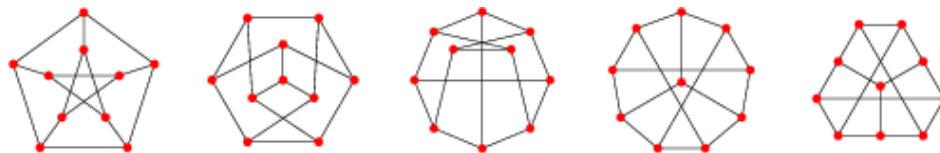
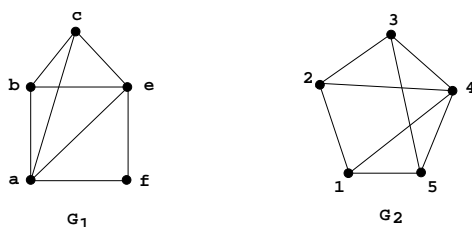
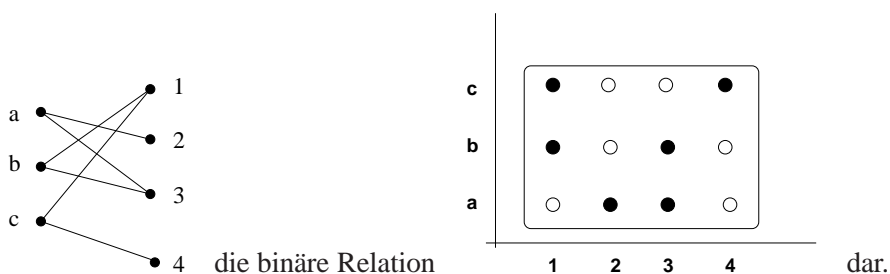


Abbildung 1.4: Petersen's Graph und seine 4 isomorphe Kopien.

Die zwei im nächsten Abbildung dargestellte Graphen sind aber bereits verschieden (nicht isomorph), da z.B. G_1 zwei Knoten (a und e) von Grad 4 hat, während G_2 nur einen solchen Knoten (Knoten 4) hat.



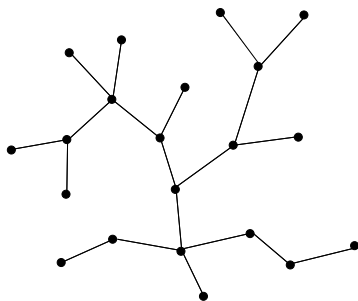
Eine interessante Klasse von Graphen sind die *bipartiten Graphen*. Diese Graphen haben die Eigenschaft, dass es eine Zerlegung der Knotenmenge V in zwei disjunkte Teilmengen A und B gibt, so dass von sämtlichen Kanten der eine Endknoten zu A gehört und der andere zu B . Bipartite Graphen haben eine große Bedeutung, liefern sie doch unmittelbar eine Veranschaulichung der binären Relationen. Tatsächlich können nämlich die Elemente einer *beliebigen* Relation $R \subseteq A \times B$ als Kanten von Knoten aus A nach Knoten aus B aufgefasst werden. Zum Beispiel stellt der bipartite Graph



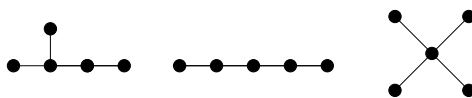
Wichtige Objekte in einem Graphen sind "Wege" und "Kreise." Ein *Weg* (engl. walk) von u nach v ist eine Folge u_0, u_1, \dots, u_r von (nicht notwendig verschiedenen) jeweils benachbarten Knoten mit $u = u_0$ und $v = u_r$. Die *Länge* dieses Weges ist die Anzahl l der Kanten, die er erhält; u und v sind seine *Endknoten*. Beachte, dass i.A. ein Weg sowohl einen Knoten wie auch eine Kante *mehrmals* durchlaufen kann! Ein Weg ist *einfach*, falls er keinen Knoten mehr als einmal durchläuft. Ein einfacher Weg mit $u_0 = u_r$ heißt *Kreis* oder *Zyklus*.

Definition: Ein Graph $G = (V, E)$ heißt *zusammenhängend*, wenn es für zwei beliebige Knoten $u, v \in V$ mindestens einen Weg von u nach v gibt.

Ein Graph heißt *zyklenfrei*, wenn er keine Kreise (= geschlossene Wege) besitzt. Ein ungerichteter Graph heißt *Wald*, wenn er zyklenfrei ist. Ein ungerichteter Graph heißt *Baum*, wenn er zyklenfrei und zusammenhängend ist.



Alle Bäume (bis auf Isomorphie) mit 5 Knoten:



Ein wesentliches Charakteristikum von Bäumen ist die Tatsache, dass jedes Paar von Knoten in einem Baum durch genau einen Weg verbunden ist.

Wir beobachten weiter, dass das Streichen von Kanten oder Knoten schlimmstenfalls aus Bäumen Wälder machen kann, während das Hinzufügen nur einer Kante zu einem Baum, die Baumstruktur sofort vollkommen zerstört.

Behauptung 1.5. Werden in einem Baum Kanten gestrichen, dann entsteht ein Wald. Wird in einem Baum eine Kante zwischen zwei seiner Knoten hinzugefügt, dann verliert man die Baumstruktur.

Beweis. Die Beobachtung, dass durch das Streichen von Kanten in einem Baum keine Zyklen entstehen können, beweist erste Aussage. Den Nachweis von der zweiten Aussage liefert folgende Überlegung: Sei T ein Baum, also ein zusammenhängender, zyklenfreier Graph. Aufgrund des Zusammenhangs von T sind zwei beliebige Knoten u, v durch einen Weg $p_{u,v}$ in T verbunden. Wird nun dem Graph eine zusätzliche Kante $e = \{u, v\}$ hinzugefügt, dann entsteht aus $p_{u,v}$ und e ein Kreis, der die Baumstruktur von T zerstört. \square

1.1.3 Abbildungen (Funktionen)

Eine Relation $f \subseteq A \times B$ derart, dass für jedes $a \in A$ genau ein Element $b \in B$ mit $(a, b) \in f$ gibt, heißt *Abbildung* (oder *Funktion*) von A nach B ; A ist der *Definitionsbereich* der Abbildung und B ihr *Bildbereich* (oder *Wertebereich*). Bei einer Abbildung wird also jedem Element a aus A in eindeutiger Weise das Element $b = f(a)$ in B zugeordnet.

Eine Abbildung können wir uns als “black box” $x \mapsto f(x)$ vorstellen, in die wir etwas hineinstecken und dafür etwas neues herausbekommen. Beispiel: $x \mapsto x^2$ ergibt das Quadrat einer Zahl, $A \mapsto |A|$ die Mächtigkeit einer Menge und $x \mapsto |x|$ den Betrag einer Zahl. Bei einer Funktion ist es wichtig zu wissen, was man hineinstecken darf und aus welchem Bereich das Ergebnis ist. Wir schreiben

$$f : A \rightarrow B$$

um anzuzeigen, dass die Funktion f Eingaben aus der Menge A akzeptiert und die Ausgabe $f(x)$ zu der Menge B gehört; A heißt *Definitionsbereich* und B *Bildbereich* von f . Also beschreibt $f : A \rightarrow B$ den Typ der Funktion und $x \mapsto f(x)$ ihr Ergebnis. Um beides in einem zu beschreiben, benutzt man manchmal die Notation $A \ni x \mapsto f(x) \in B$.

Die Menge aller Abbildungen von A nach B bezeichnet man mit B^A , d.h.

$$B^A := \{f : f \text{ ist eine Abbildung von } A \text{ nach } B\}.$$

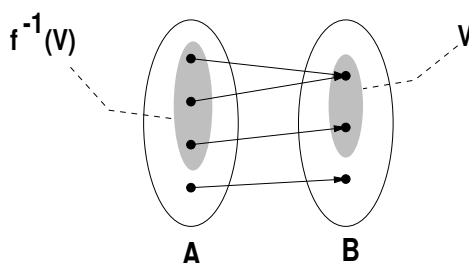
Für $U \subseteq A$ heißt

$$f(U) = \{f(a) : a \in U\}$$

Bild von U unter f . Für $V \subseteq B$ heißt

$$f^{-1}(V) = \{a \in A : f(a) \in V\}$$

Urbild von V unter f .



Für $b \in B$ setzt man

$$f^{-1}(b) = \{a \in A : f(a) = b\}.$$

Eine Funktion $f : A \rightarrow B$ heißt

- *surjektiv*, falls $f(A) = B$, d.h. für jedes $b \in B$ ein $a \in A$ mit $f(a) = b$ gibt.
- *injektiv*, falls aus $a_1 \neq a_2 \in A$ stets $f(a_1) \neq f(a_2)$ folgt,
- *bijektiv*, falls f surjektiv und injektiv ist.

► *Beispiel 1.6* : Die Funktion $f : \mathbb{N} \rightarrow \mathbb{N}$ mit $f(x) = x^2$ ist injektiv, aber die Funktion $g : \mathbb{Z} \rightarrow \mathbb{Z}$ mit $g(x) = x^2$ ist nicht injektiv: $-1 \neq 1$ aber $(-1)^2 = 1^2$. Die Funktion $g(x)$ ist auch nicht surjektiv: $g^{-1}(2) = \sqrt{2}$ gehört nicht zu \mathbb{Z} .

Die Abbildung, die jedem lebenden Menschen sein momentanes Lebensalter zuordnet, ist eine Funktion in die Menge der natürlichen Zahlen \mathbb{N} . Sie ist nicht injektiv, da viele Menschen das gleiche Lebensalter haben. Diese Funktion ist nicht surjektiv, da es keine Menschen mit 4-stelligem Lebensalter gibt.

Injektive Funktionen $f : A \rightarrow B$ kann man auf ihrem Bild $B' = f(A)$ umkehren und erhält dann die *Umkehrfunktion* $f^{-1} : B' \rightarrow A$ mit $f^{-1}(y) = x \iff y = f(x)$. Man verwechsle die Umkehrfunktion nicht mit der Funktion $x \mapsto 1/f(x)$.

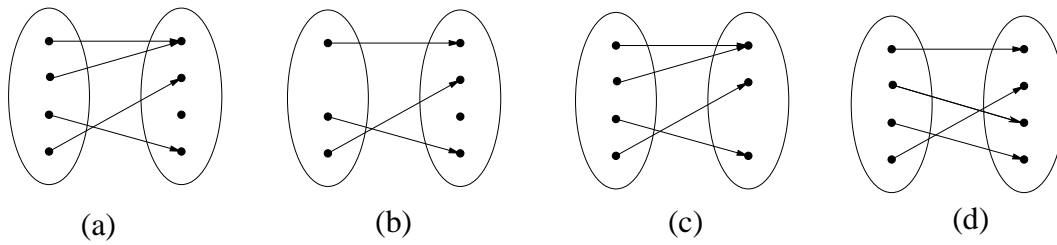


Abbildung 1.5: (a) weder injektiv noch surjektiv; (b) injektiv aber nicht surjektiv; (c) surjektiv aber nicht injektiv; (d) sowohl injektiv wie auch surjektiv, also bijektiv.

Wenn $f : A \rightarrow A$ bijektiv und A endlich sind, dann heißt f eine *Permutation* von A , ein sehr wichtiger mathematischer Begriff. Ist $A = \{1, 2, \dots, n\}$ so bezeichnet man eine Permutation $f : A \rightarrow A$ oft als

$$\begin{pmatrix} 1 & 2 & \dots & n \\ f(1) & f(2) & \dots & f(n) \end{pmatrix}.$$

Sind $f : A \rightarrow B$ und $g : B \rightarrow C$ Funktionen, so definiert man die *Komposition* oder *Hintereinanderausführung* $h := f \circ g$ der Funktionen durch $h(x) = f(g(x))$.

Im Laufe der Vorlesung werden wir verschiedene mathematische Strukturen kennenlernen: Graphen, Gruppen, Ringe, Körper, Vektorräume, usw. Jede Struktur besteht aus einer Menge A (dem “Universum”) und einer (oder mehreren) Relationen (oder Abbildungen) $\sim_R \subseteq A \times A$. Hat man zwei solche Strukturen (A, \sim_R) und (B, \sim_S) , so heißen diese Strukturen *isomorph*, wenn es eine bijektive Abbildung $f : A \rightarrow B$ mit der Eigenschaft

$$\forall x, y \in A : \quad f(x) \sim_S f(y) \iff x \sim_R y$$

gibt. Diese Eigenschaft bedeutet, dass die beide Strukturen im wesentlichen eine und dieselbe Struktur darstellen.

1.2 Kardinalität unendlicher Mengen

Wir wollen die Mengen gemäß ihrer “Größe” vergleichen. Mit endlichen Mengen haben wir kein Problem: Die Größe $|A|$ einer solchen Menge A ist einfach die Anzahl ihrer Elemente

$$|A| = \text{Anzahl der Elemente in } A.$$

Was aber wenn wir unendliche Mengen haben?

Die Intuition hat in der Unendlichkeit keinen festen Platz, wie man an Hilberts Hotel feststellt: Hier gibt es unendlich viele durchnummerierte Zimmer, die alle belegt sind. Ein Neuankömmling erhält dennoch ein freies Zimmer, ohne dass die anderen Gäste sich einen Raum teilen oder aus dem Hotel verschwinden müssen: Der neue Gast bekommt einfach das Zimmer 1, dessen ursprünglicher Bewohner zieht nach Zimmer 2 um, während der Bewohner aus Zimmer 2 in das Zimmer 3 einzieht ... Es können sogar unendlich viele neue Gäste unterkommen - die alten Gäste verlegt man von Zimmer n nach Zimmer $2n$ und die neuen Gäste erhalten die Zimmer mit den ungeraden Nummern.

Deshalb vergleicht man die “Mächtigkeiten” unendlichen Mengen nicht durch einen Zählvorgang, sondern nach dem “Omnibus-Prinzip”:

Das Omnibus-Prinzip: In einem Bus gibt es ebenso viele Sitzplätze wie Fahrgäste, wenn kein Fahrgast stehen muss und kein Sitz frei bleibt. Die Passagiere sitzen injektiv, also nicht gestapelt und eindeutig (keine Person nimmt mehr als einen Platz ein).

Genau dann existiert eine bijektive (injektive und surjektive) Abbildung von der Menge aller Fahrgäste auf die Menge aller Sitzplätze. “Surjektiv” bedeutet hier, dass alle Sitze besetzt sind.

Nach diesem Prinzip hat Cantor 1874 die folgende Definition eingeführt. Man sagt, dass eine Menge A nicht größer als eine andere Menge B ist, falls es eine *injektive* Abbildung $f : A \rightarrow B$ gibt. Ist f bijektiv, so sagt man, dass A und B gleich groß sind (oder die gleiche *Kardinalität* haben).



Sind die Mengen A und B endlich, so gibt es eine Injektion $f : A \rightarrow B$ genau dann, wenn $|A| \leq |B|$ gilt. Deshalb ist für endlichen Mengen die Größe genau die Anzahl der Elemente. Für unendlichen Mengen gilt das nicht mehr, da sie alle die *gleiche* Anzahl der Elemente haben – nämlich ∞ (unendlich viele).

▷ *Beispiel 1.7:* Sei $\mathbb{N} = \{0, 1, 2, 3, 4, 5, \dots\}$ und sei $\mathbb{E} = \{0, 2, 4, 6, \dots\}$ die Menge aller geraden Zahlen. Es ist klar, dass \mathbb{E} nicht größer als \mathbb{N} ist und, wenn man die Zahlen auf einer Linie dargestellt vorstellt, hat \mathbb{E} sehr viele “Lücken”: jede zweite Zahl fehlt! Also sollte \mathbb{E} *echt* kleiner als \mathbb{N} sein? Die Antwort ist: nein, die Mengen \mathbb{E} und \mathbb{N} sind gleich gross!

Beweis: $f(n) = 2n$ ist eine bijektive Abbildung von \mathbb{N} nach \mathbb{E} (da $x \neq y \iff 2x \neq 2y$ gilt).

Vom besonderen Interesse (insbesondere in der Informatik) sind Mengen, die nicht größer als die Menge $\mathbb{N} = \{0, 1, 2, 3, 4, 5, \dots\}$ aller natürlichen Zahlen sind. Solche Mengen nennt man “abzählbar”.

Eine Menge A heißt *abzählbar*, wenn es eine Injektion $f : A \rightarrow \mathbb{N}$ gibt.

D.h. in diesem Fall kann man jedem Element $a \in A$ (Fahrgast) einen eindeutigen Nummer $f(a) \in \mathbb{N}$ (Sitzplatz) zuweisen. Die Menge A ist also abzählbar, wenn man ihre Elemente durchnummerieren kann, $A = \{a_0, a_1, a_2, \dots\}$. Ist eine Menge nicht abzählbar, so heißt sie *überzählbar*.

Behauptung 1.8. Ist eine Menge A abzählbar, so ist jede Teilmenge $S \subseteq A$ abzählbar.

Beweis. Wenn S nicht onehin endlich ist, kann man dies folgendermaßen einsehen. Sei $f : \mathbb{N} \rightarrow A$ eine Bijektion. Wir definieren $g : \mathbb{N} \rightarrow S$, indem wir $g(0), g(1), g(2), \dots$ jeweils auf das nächste Element von $f(0), f(1), f(2), \dots$ setzen, das in S liegt. Zum Beispiel für $A = \{2n : n \in \mathbb{N}\}$ und $S = \{4n : n \in \mathbb{N}\}$:

$$\begin{aligned} f(0) = 0 \in S &\Rightarrow g(0) = f(0) \\ f(1) = 2 \notin S & \\ f(2) = 4 \in S &\Rightarrow g(1) = f(2) \\ f(3) = 6 \notin S & \\ f(4) = 8 \in S &\Rightarrow g(2) = f(4) \\ &\vdots \end{aligned}$$

□

Wir kennen auch andere unendliche Mengen:

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}.$$

Sind diese Mengen größer als \mathbb{N} oder sind sie abzählbar, d.h. gleich gross wie \mathbb{N} ?

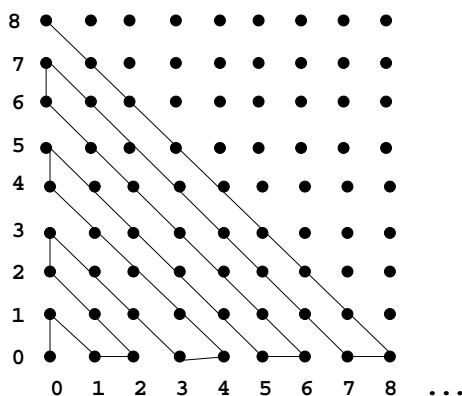
Auf dem ersten Blick scheinen diese Mengen viel größer als \mathbb{N} zu sein. Insbesondere die Menge \mathbb{Q} , da sie sehr "dicht" ist: zwischen zwei beliebigen rationalen Zahlen liegt mindestens eine rationale Zahl. Die Intuition ist aber trügerisch: Die ersten zwei Mengen \mathbb{N}, \mathbb{Z} und \mathbb{Q} sind gleich gross (sind abzählbar)! Nur die letzte Menge \mathbb{R} ist "echt größer" als \mathbb{N} (ist überzählbar).

► *Beispiel 1.9*: \mathbb{Z} ist abzählbar. Die entsprechende Injektion $f : \mathbb{Z} \rightarrow \mathbb{N}$ kann man zum Beispiel durch

$$f(x) = \begin{cases} 2x - 1 & \text{falls } x > 0 \\ -2x & \text{falls } x \leq 0 \end{cases}$$

definieren, d.h. positive Zahlen bekommen ungeraden und negative ungeraden Nummern.

► *Beispiel 1.10*: Ist $\mathbb{N} \times \mathbb{N}$ abzählbar? Ja:



Daraus folgt auch, dass \mathbb{Q} abzählbar ist.

Welche Mengen dann sind überzählbar? Wenn wir uns einen unendlichen Bus mit den Sitzen $0, 1, 2, \dots$ vorstellen, dann ist die Menge (der Fahrgäste) A abzählbar genau dann, wenn jeder Fahrgast $a \in A$ einen eigenen Sitzplatz bekommt. Ist die Menge A der Fahrgäste unendlich, so muss auch jeder Sitzplatz auch tatsächlich besetzt sein, D.h. es muss dann eine *surjektive* Abbildung $f : \mathbb{N} \rightarrow A$ geben.

► *Beispiel 1.11*: (**Cantor's Diagonalisierungsmethode**) Wir haben bereits gesehen, dass die Menge \mathbb{Z} der ganzen Zahlen wie auch die Menge \mathbb{Q} der rationalen Zahlen abzählbar sind. Ist auch die Menge \mathbb{R} der reellen Zahlen abzählbar? Die Antwort ist nein – \mathbb{R} ist überzählbar. Es gibt bereits Teilmengen der reellen Zahlen, die nicht abzählbar sind: Wäre etwa die Menge derjenigen Zahlen abzählbar, die zwischen 0 und 1 liegen, dann könnte man sogar die Menge der abzählbar unendlichen Dezimalbruchentwicklungen der Form $0, x_1 x_2 \dots$ mit $x_i \in \{0, 1, \dots, 9\}$ abzählen und wie folgt auflisten:

$$\begin{array}{cccccc} 0, & \mathbf{a_{11}} & a_{12} & a_{13} & a_{14} & \dots \\ 0, & a_{21} & \mathbf{a_{22}} & a_{23} & a_{24} & \dots \\ 0, & a_{31} & a_{32} & \mathbf{a_{33}} & a_{34} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

a_{11} wäre also die erste Kommastelle der Dezimalentwicklung, welche die kleinste Nummer in der Abzählung erhält, a_{12} die zweite Kommastelle, usw. Wir konstruieren jetzt $0, b_1 b_2 b_3 \dots$ wie folgt:

$$b_1 \neq a_{11}, b_2 \neq a_{22}, \dots, b_b \neq a_{kk}, \dots$$

und erhalten damit eine Dezimalbruchentwicklung, die *nicht* in der obigen Liste vorkommen kann, d.h. wir bekommen einen Fahrgast (Dezimalbruchentwicklung) der keinen Sitzplatz in dem Buss \mathbb{N} bekommt. Aus diesem Widerspruch folgt, dass man auch die reellen Zahlen zwischen 0 und 1 nicht abzählen kann. Das Wort "Diagonale" erklärt sich von selbst: Wir konstruieren die Dezimalbruchentwicklung $0, b_1 b_2 b_3 \dots$ indem wir die Diagonalelemente a_{kk} vertauschen. D.h. der Widerspruch befindet sich auf der Diagonale.

Satz 1.12. (Cantor) Sei A eine beliebige Menge und $2^A = \{X : X \subseteq A\}$ ihre Potenzmenge. Dann existiert keine Surjektion $f : A \rightarrow 2^A$. Insbesondere ist $2^{\mathbb{N}}$ überzählbar.

Beweis. Ein Widerspruchsbeweis. Angenommen $f : A \rightarrow 2^A$ ist surjektiv. Setze

$$D := \{a \in A : a \notin f(a)\}.$$

Weil f surjektiv ist, existiert ein $a_0 \in A$ mit $f(a_0) = D$. Es gilt: entweder $a_0 \in D$ oder $a_0 \notin D$. Aber nach der Definition von D gilt: $a_0 \in D \iff a_0 \notin f(a_0) = D$, ein Widerspruch.

Wäre $2^{\mathbb{N}}$ abzählbar, so gäbe es eine Bijektion $\mathbb{N} \rightarrow 2^{\mathbb{N}}$, aber (wie wir bereits gezeigt haben) es gibt nicht einmal eine Surjektion. \square

Damit haben wir noch eine überzählbare Menge $\mathcal{P}(\mathbb{N}) = \{X : X \subseteq \mathbb{N}\}$ gefunden! Beachte aber, dass die Menge

$$\mathcal{E}(\mathbb{N}) = \{X : X \subseteq \mathbb{N}, X \text{ endlich}\}$$

bereits abzählbar ist! (Siehe Aufgabe 9)

Der größte Unterschied zwischen der Mathematik und der Informatik ist, dass man sich in der Informatik nur mit solchen Funktionen beschäftigt, deren Werte mit einer Program berechnet werden können. Der folgende Korollar zeigt, dass damit sehr viele Funktionen ausser der Interesse von Informatikern bleiben.

Korollar 1.13. Es gibt (überzählbar viele) Abbildungen $f : \mathbb{N} \rightarrow \{0, 1\}$, die nicht durch ein Programm berechnet werden können.

Beweis. Jedes Programm, egal in welcher Programmiersprache, ist eine *endliche* Folge von Symbolen aus einer *endlichen* Menge (Alphabet + Sonderzeichen). Daher ist die Menge aller Programme abzählbar. Demnach ist die Menge $B \subseteq \{0, 1\}^{\mathbb{N}}$ aller Abbildungen $f : \mathbb{N} \rightarrow \{0, 1\}$, welche von einem Programm berechnet werden können, abzählbar. Nach dem Satz 1.12 ist $\{0, 1\}^{\mathbb{N}}$ überzählbar. Deshalb ist $\{0, 1\}^{\mathbb{N}} \setminus B$ nicht leer, sogar überzählbar. \square

1.3 Aussagenlogik und Beweismethoden

Wir haben bereits einige Aussagen bewiesen, ohne die *logische Struktur* der Beweise zu betonen. Diese Struktur in der Beweisführung ist aber äußerst wichtig, da jeder Schritt muss logisch korrekt sein. Deswegen lohnt es sich, die logische Struktur der Beweise genauer anzuschauen. Zunächst müssen wir aber genauer die Struktur des Grundobjekts der mathematischen Beweise – der Aussage – betrachten.



Aristoteles (384–322 vor Christus):

Eine *Aussage* ist ein sprachliches Gebilde, von dem es sinnvoll ist, zu sagen, es sei *wahr* oder *falsch*.

Eine Aussage also ist ein Satz, der entweder wahr oder falsch ist, aber nie beides zugleich. Wahre Aussagen haben den *Wahrheitswert 1* und falsche Aussagen den Wahrheitswert **0**.

Zum Beispiel die Aussage

$$A = \text{“15 ist durch 3 teilbar”}$$

ist wahr (also hat A den Wahrheitswert **1**) aber die Aussage

$$B = \text{“16 ist kleiner 12”}$$

ist falsch (und deshalb hat B den Wahrheitswert **0**).



Nicht jeder Satz ist eine Aussage! Um ein Beispiel für einen Satz vorzustellen, der weder wahr noch falsch sein kann und deshalb keine Aussage ist, sehen wir uns Russells Paradoxon „*Dieser Satz ist falsch.*“ an. Angenommen, der Satz wäre wahr, dann müsste er falsch sein. Gehen wir aber davon aus, dass der Satz falsch ist, dann müsste er wahr sein. ^a

^aÜbrigens hat Russell mit diesem Paradoxon sehr prägnant die tief verwurzelte Annahme der Mathematik, dass allen Sätzen ein Wahrheitswert zugeordnet sei, erschüttert und eine tiefe Grundlagenkrise der Mathematik zu Beginn des 20. Jahrhunderts ausgelöst.

Als nächstes schauen wir an, wie man einfachste (atomare) Aussagen in kompliziertere Aussagen umwandeln kann.

Folgende Verknüpfungen von Aussagen werden häufig benutzt; dabei bezeichnen A und B zwei Aussagen:

A	B	$A \wedge B$	$A \vee B$	$A \leftrightarrow B$	$A \oplus B$
0	0	0	0	1	0
0	1	0	1	0	1
1	0	0	1	0	1
1	1	1	1	1	0

A	$\neg A$
1	0
0	1

- $A \wedge B$ (lies A und B) ist genau dann wahr, wenn A und B beide wahr sind.
- $A \vee B$ (lies A oder B) ist genau dann wahr, wenn mindestens eine der Aussagen A, B wahr ist.

- $A \leftrightarrow B$ (lies *A und B sind äquivalent*) ist genau dann wahr, wenn A und B beide den gleichen Wahrheitswert haben.
- $A \oplus B$ (lies *entweder A oder B*) ist genau dann wahr, wenn genau eine der Aussagen A, B wahr ist. (Oft schreibt man $A\Delta B$ statt $A \oplus B$.)
- $\neg A$ (lies *nicht A*) hat den umgekehrten Wahrheitswert wie A .

Eine wichtige Verknüpfung ist die *Implikation* $A \rightarrow B$ (lies *wenn A, dann B*; ist A wahr, dann auch B):

A	B	$A \rightarrow B$
0	0	1
0	1	1
1	0	0
1	1	1

Insbesondere ist $A \rightarrow B$ immer wahr, wenn A falsch ist! Ist das sinnvoll? Dazu ein Beispiel von Bertrand Russel.

“Wenn $1 = 0$ ist, bin ich der Papst!”

Beweis: Aus $1 = 0$ folgt $2 = 1$. Da der Papst und ich 2 Personen sind, sind wir 1 Person. □



Die Implikation $A \rightarrow B$ sagt also nur, dass B wahr sein muss, *falls die Aussage A richtig ist*. Sie sagt aber nicht, dass A auch *tatsächlich* wahr ist! Die Implikation ist eine der wichtigsten logischen Verknüpfungen in der Mathematik: Sie erlaubt logisch konsistente Theorien bilden, ohne sich um die (tatsächliche) Richtigkeit der ursprünglichen Aussagen (sogenannten “Axiomen”) zu kümmern!

Um nicht so viele Klammern benutzen zu müssen, legt man die “Stärke” der Bindung der Operationen fest:

\neg stärker als \wedge
 \wedge stärker als \vee
 \vee stärker als \rightarrow
 \rightarrow stärker als \leftrightarrow

Die Formel $\neg(A \wedge B \vee \neg C) \rightarrow \neg C$ bedeutet also $(\neg((A \wedge B) \vee (\neg C))) \rightarrow (\neg C)$

Ordnet man den Aussagenvariablen – in diesem Kontext zusammen mit den Wahrheitswerten **1** und **0** auch *atomare Formeln* genannt – Wahrheitswerte zu, so erhält man aus der aussagenlogischen Formel eine Aussage. Zur Ermittlung des Wahrheitswerteverlaufs der Formel kann eine Wahrheitstafel aufgestellt werden, in der sich für alle möglichen Kombinationen von Wahrheitswerten, die die einzelnen Aussagenvariablen der Formel annehmen können, der Wahrheitswert der Formel ablesen lässt.

▷ *Beispiel 1.14*: Für die Formel $\neg(A \wedge B \vee \neg C) \rightarrow \neg C$ ergibt sich die folgende Wahrheitstafel.

A	B	C	$A \wedge B$	$\neg C$	$A \wedge B \vee \neg C$	$\neg(A \wedge B \vee \neg C)$	$\neg(A \wedge B \vee \neg C) \rightarrow \neg C$
1	1	1	1	0	1	0	1
1	1	0	1	1	1	0	1
1	0	1	0	0	0	1	0
1	0	0	0	1	1	0	1
0	1	1	0	0	0	1	0
0	1	0	0	1	1	0	1
0	0	1	0	0	0	1	0
0	0	0	0	1	1	0	1

Jede Zeile enthält eine *Belegung* der beteiligten Aussagenvariablen mit einem der beiden Wahrheitswerte **1** und **0**. Dabei kommt jede mögliche Belegung genau einmal vor. Die letzte Spalte der Wahrheitstafel gibt den *Wahrheitswerteverlauf* der Formel an. Man sieht, dass der Wahrheitswerteverlauf eindeutig durch die Struktur der Formel und ihren Aufbau aus den verknüpften Teilformeln festgelegt wird.

Aussagenlogische Formeln kann man als Schaltkreise (oder Schaltpläne) darstellen:

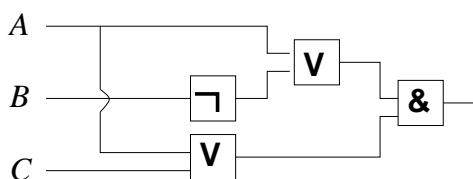


Abbildung 1.6: Darstellung von $(A \vee \neg B) \wedge (A \vee C)$

Aussagenlogische Formeln können *erfüllbar* sein (d.h. wahr bei geeigneter Wahl der Wahrheitswerte der atomaren Aussagen) aber auch nicht, z.B. $A \wedge \neg A$; dann heißen sie *Kontradiktionen* (Widersprüche). *Tautologien* sind immer wahr, z.B. $A \vee \neg A$.

Zwei aussagenlogischen Formeln F und H heißen (logisch) *äquivalent* (Bezeichnung $F \iff H$), wenn sie bei jeder Belegung der in ihnen vorkommenden Aussagenvariablen denselben Wert annehmen, d.h. wenn $F \leftrightarrow H$ eine Tautologie ist.

Doppelnegation	$\neg(\neg A) \iff A$
deMorgans Regeln	$\neg(A \vee B) \iff \neg A \wedge \neg B$ $\neg(A \wedge B) \iff \neg A \vee \neg B$
Kontrapositionsgesetz	$A \rightarrow B \iff \neg B \rightarrow \neg A$
Prinzip des indirekten Beweises oder Widerspruchsbeweis	$(A \rightarrow B) \iff (A \wedge \neg B) \rightarrow \neg A$ $(A \rightarrow B) \iff (A \wedge \neg B) \rightarrow B$ $(A \rightarrow B) \iff (A \wedge \neg B) \rightarrow \mathbf{0}$
Distributivität	$A \wedge (B \vee C) \iff (A \wedge B) \vee (A \wedge C)$ $A \vee (B \wedge C) \iff (A \vee B) \wedge (A \vee C)$
Prinzip vom ausgeschlossenen Dritten	$A \vee \neg A \iff \mathbf{1}$ $A \wedge \neg A \iff \mathbf{0}$
Idempotenz	$A \wedge A \iff A$ $A \vee A \iff A$
Kontradiktionsregeln	$A \vee \mathbf{0} \iff A$ $A \wedge \mathbf{0} \iff \mathbf{0}$

Eine nützliche Merkregel ist, dass man die logische Implikation $A \rightarrow B$ mittels \neg und \vee auffassen kann:

$$A \rightarrow B \iff \neg A \vee B$$

► *Beispiel 1.15*: Um an einem kleinen Beispiel zu demonstrieren, wie man mit Hilfe der aufgelisteten logischen Äquivalenzen tatsächlich zu Vereinfachungen kommen kann, betrachten wir die Formel $\neg(\neg A \wedge B) \wedge (A \vee B)$.

$$\begin{aligned}
 &\neg(\neg A \wedge B) \wedge (A \vee B) \\
 &(\neg(\neg A) \vee (\neg B)) \wedge (A \vee B) && \text{(deMorgans Regel)} \\
 &(A \vee \neg B) \wedge (A \vee B) && \text{(Doppelnegation)} \\
 &A \vee (\neg B \wedge B) && \text{(Distributivität)} \\
 &A \vee \mathbf{0} && \text{(Prinzip vom ausgeschlossenen Widerspr.)} \\
 &A && \text{(Kontradiktionsregel)}
 \end{aligned}$$

Aufgrund der mit dieser Umformung nachgewiesenen logischen Äquivalenz der beiden Formeln kann nun anstelle der komplizierten Formel $\neg(\neg A \wedge B) \wedge (A \vee B)$ stets die einfacher strukturierte atomare Formel A zur logischen Beschreibung des Sachverhalts benutzt werden.

Mit Mengen kann man genauso “rechnen” wie mit Aussagen!

Mengen	Aussagen
$A \cup B$	$A \vee B$
$A \cap B$	$A \wedge B$
$A \subseteq B$	$A \rightarrow B$
\overline{A}	$\neg A$

So ist zum Beispiel $\overline{A \cup B} = \overline{A} \cap \overline{B}$ (deMorgans Regel), usw.

1.3.1 Aussageformen (Prädikate)

Die bisher betrachtete Aussagenlogik erweist sich als nicht ausreichend, um allgemeine logische Aussagen zu treffen. So wollen wir beispielsweise Aussagen über Elemente von Mengen treffen. Dazu benutzt man sogenannte “Prädikate”.

Ein n -stelliges Prädikat über M ist einfach eine n -stellige Abbildung $P : M^n \rightarrow \{0, 1\}$. Ein Prädikat $P(x_1, \dots, x_n)$ nimmt also in Abhängigkeit von der Belegung der x_i mit Elementen aus M den Wert “wahr” oder “falsch” an.

Um Prädikate in Aussagen umzuwandeln, benutzt man so genannte *Quantoren*: den *Allquantor* $\forall x$ (für alle $x \in M$) und den *Existenzquantor* $\exists x$ (es gibt ein $x \in M$). Jeder Quantor *bindet* das freie Vorkommen der Variablen, die er quantifiziert. Die mit dem Allquantor gebildeten Aussagen werden *Allaussagen* genannt, und die mit dem Existenzquantor gebildeten Aussagen heißen *Existenzaussagen*.

Sei $P(x)$ eine Aussageform über dem Universum M .

- Die Aussage $\exists x P(x)$ ist genau dann wahr, wenn es mindestens ein a in M existiert, so dass $P(a)$ wahr ist.
- Die Aussage $\forall x P(x)$ ist genau dann wahr, wenn $P(a)$ für jedes a aus M wahr ist.

▷ *Beispiel 1.16*: Sei $S(x)$ die Aussageform “ $x \leq x + 1$ ” über \mathbb{N} . Dann stellt $\forall x : S(x)$ die Aussage „Für jedes n in \mathbb{N} gilt $n \leq n + 1$ “ dar. Diese Aussage ist wahr, da $S(n)$ für jede natürliche Zahl n eine Aussage mit dem Wahrheitswert **1** liefert.

Sei $S(x)$ wie oben. Dann ist $\exists x : S(x)$ die Aussage „Es gibt ein n in \mathbb{N} , für das $(n \leq n + 1)$ gilt“. Diese Aussage ist ebenfalls wahr, da z.B. $S(5)$ wahr ist.

Sei $P(x)$ die Aussageform „ x ist eine Primzahl“ über \mathbb{N} . Dann stellt $\forall x : P(x)$ die Aussage „Für jedes n aus \mathbb{N} gilt: n ist eine Primzahl“ dar. Diese Aussage ist falsch, da z.B. 4 keine Primzahl ist. Also ist $P(4)$ nicht wahr, und $P(n)$ damit nicht für jedes n aus \mathbb{N} wahr.

Sei $P(x)$ wie in oben definiert. Dann ist $\exists x : P(x)$ die Aussage „Es gibt eine natürliche Zahl n , die eine Primzahl ist“. Diese Aussage ist wahr, da z.B. 3 eine Primzahl ist und damit $P(3)$ eine wahre Aussage.

Die Quantoren \forall und \exists können nun auf mehrstellige Prädikate angewendet werden: Sei P ein n -stelliges Prädikat, dann ist $\forall x_i P(x_1, \dots, x_n)$ für $x_i, 1 \leq i \leq n$, wieder ein Prädikat, bei dem eine Variable, die Variable x_i , “gebunden” ist. $\forall x_i P(x_1, \dots, x_n)$ ist also nun ein $(n - 1)$ -stelliges Prädikat, auf das wieder ein Quantor angewandt werden kann. Analoges gilt für $\exists x_i P(x_1, \dots, x_n)$. Dabei heißt x_i *gebundene Variable*, alle anderen heißen *freie Variablen*.

Um aus einer Aussageform mit mehreren Variablen durch Quantifizierung Aussagen (d.h. nullstellige Prädikate) zu erhalten, muss *jede* freie Variable durch einen gesonderten Quantor gebunden werden!

Alle innerhalb der Aussagenlogik gültigen Äquivalenzen gelten auch in der Prädikatenlogik. Darüber hinaus existieren noch weitere Äquivalenzen, welche die Quantoren miteinbeziehen:

Negationsregeln :	$\forall x : P(x) \iff \neg(\exists x : \neg P(x))$ $\neg(\exists x : P(x)) \iff \forall x : \neg P(x)$ $\neg(\forall x : P(x)) \iff \exists x : \neg P(x)$
Ausklammerungsregeln :	$(\forall x : P(x)) \wedge (\forall x : Q(x)) \iff \forall x : P(x) \wedge Q(x)$ $(\exists x : P(x)) \vee (\exists x : Q(x)) \iff \exists x : P(x) \vee Q(x)$
Vertauschungsregel :	$\forall x \forall y : P(x, y) \iff \forall y \forall x : P(x, y)$ $\exists x \exists y : P(x, y) \iff \exists y \exists x : P(x, y)$

Zur Schreibweise: Man schreibt

$$\forall x \in M P(x) \quad \text{anstatt} \quad \forall x(x \in M \rightarrow P(x))$$

$$\exists x \in M P(x) \quad \text{anstatt} \quad \exists x(x \in M \wedge P(x))$$

Man schreibt auch $\exists x! P(x)$ für “es gibt *genau ein* x mit $P(x) = 1$ ”.

Hier sind ein paar häufiger Fehler, die man mit dem Umgang mit Quantoren macht.



Bei verschiedenen Quantoren kommt es auf die *Reihenfolge* der Quantoren an! Z.B. bezeichne $P(x, y)$ die Aussageform “ $x \geq y$ ”, x und y seien Variablen über den Universum \mathbb{N} . Dann ist $\forall y \exists x P(x, y)$ wahr. Aber $\exists x \forall y P(x, y)$ ist falsch.



Die folgenden Formelpaare sind *nicht* äquivalent, obwohl sie den Ausklammerungsregeln sehr ähnlich sind:

$$(\forall x : P(x)) \vee (\forall x : Q(x)) \quad \text{mit} \quad \forall x : P(x) \vee Q(x)$$


$$(\exists x : P(x)) \wedge (\exists x : Q(x)) \quad \text{mit} \quad \exists x : P(x) \wedge Q(x)$$





Falsche Übersetzung von umgangssprachlichen Implikationen. Wenn A = “ich bin mit meiner Hausaufgaben fertig” und B = “ich werde ins Kino gehen”, dann besagt $A \rightarrow B$ “ich werde ins Kino gehen, wenn ich mit meiner Hausaufgaben fertig bin”, wobei $B \rightarrow A$ besagt: “ich werde ins Kino gehen, nur wenn ich mit meiner Hausaufgaben fertig bin”.



Falsche Negierung von Aussagen ohne deMorgans Regeln zu benutzen: $\neg(A \vee B)$ und $\neg A \vee \neg B$ sind *nicht* äquivalent! Das Gleiche gilt für Mengenoperationen: so ist z.B. $\overline{A \cup B} = \overline{A} \cap \overline{B}$, nicht aber $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

 Falsche Beschreibung einer Existenzaussage als $\exists x(A(x) \rightarrow B(x))$ anstatt $\exists x(A(x) \wedge B(x))$. Zum Beispiel die Aussage “Es gibt eine ungerade Zahl, die prim ist” hat die Form $\exists x(U(x) \wedge P(x))$, nicht $\exists x(U(x) \rightarrow P(x))$. Allgemeine Regel: Nach einem \exists -Quantor folgt normalerweise ein UND, nicht die Implikation.

 Falsche Beschreibung einer “Für-alle-Aussage” als $\forall x(A(x) \wedge B(x))$ anstatt $\forall x(A(x) \rightarrow B(x))$. Zum Beispiel die (falsche!) Aussage “Jede gerade Zahl ist prim” hat die Form $\forall x(G(x) \rightarrow P(x))$, nicht $\forall x(G(x) \wedge P(x))$. Allgemeine Regel: Nach einem \forall -Quantor folgt normalerweise eine Implikation, nicht UND.

 Negation von Aussagen mit Quantoren, insbesondere in Umgangssprache. Zum Beispiel, Negation von “manche Katzen mögen Wurst” ist *nicht* die Aussage, dass “manche Katzen hassen Wurst”, sondern “keine Katzen mögen Wurst” oder “alle Katzen hassen Wurst”.

1.3.2 Logische Beweisregeln

Es gibt ein Paar grundlegender logischer Regeln, wie man eine neue wahre Aussage aus bereits als wahr bekannten Aussagen ableiten kann. Diese Regeln haben die Form

$$\frac{A_1, A_2, \dots, A_n}{B}$$

und ihre Bedeutung ist: *falls alle Aussagen A_1, A_2, \dots, A_n wahr sind, dann ist auch die Aussage B wahr.*

Hier sind die wichtigsten Regeln.

Modus ponens:

$$\frac{A, A \rightarrow B}{B}$$

Ist A wahr und folgt B aus A , dann ist auch B wahr.

Logische Schlusskette:

$$\frac{A \rightarrow B, B \rightarrow C}{A \rightarrow C}$$

Folgt B aus A und C aus B , dann folgt auch C aus A .

Kontrapositionsregel:

$$\frac{\neg B \rightarrow \neg A}{A \rightarrow B}$$

Folgt $\neg A$ aus $\neg B$, dann muss auch B aus A folgen.

Reductio ad absurdum - Widerspruchsregel:

$$\frac{\neg B, \quad \neg A \rightarrow B}{A}$$

Wir wissen, dass B falsch ist und dass aus $\neg A$ das Gegenteil folgen würde. Also ist die Aussage A wahr.

► **Beispiel 1.17: (Widerspruchsregel)** Eine falsche Behauptung: *1 ist die größte reelle Zahl.*

Beweis: Angenommen, es gibt eine andere größte Zahl y . Es ist nun $1 < y$, also ist y insbesondere eine positive Zahl, d.h wir können die Ungleichung mit y multiplizieren und erhalten $y < y^2$. Das ist aber ein Widerspruch zur Annahme, dass y die grösste reelle Zahl ist, also folgt die Behauptung und somit ist 1 die grösste reelle Zahl.

Wo liegt der Fehler? In der Negation der Behauptung! Die richtige Negation müsste lauten: 1 ist nicht die grösste reelle Zahl.

► **Beispiel 1.18: (Widerspruchsregel)** Wir wollen die Aussage $A =$ "es gibt unendlich viele Primzahlen" beweisen. Dazu nehmen wir das Gegenteil (also $\neg A$ ist wahr) an und sei etwa $\{p_1, p_2, \dots, p_n\}$ die endliche Menge der Primzahlen und sei $P = \prod_{i=1}^n p_i$. Dann ist $P + 1$ keine Primzahl, wird also von einer Primzahl, etwa p_{i_0} echt geteilt. Also sollte die Aussage B :

$$\frac{P + 1}{p_{i_0}} = \prod_{\substack{i=1 \\ i \neq i_0}}^n p_i + \frac{1}{p_{i_0}} \quad \text{ist eine natürliche Zahl}$$

wahr sein, was offensichtlich Unsinn ist. Also muss die Aussage A richtig sein (nach der Widerspruchsregel).

► **Beispiel 1.19: (Kontrapositionsregel)** Sei a eine beliebige ganze Zahl. Wir wollen die Aussage

$$\text{wenn } a^2 \text{ eine ungerade Zahl ist, dann ist } a \text{ ungerade} \quad (*)$$

beweisen. Diese Aussage hat die Form $A \rightarrow B$, wobei $A =$ " a^2 ist ungerade" und $B =$ " a ist ungerade". Wir führen einen Beweis durch Kontraposition. Sei also angenommen, dass a gerade ist ($\neg B$ gilt). Dann ist $a = 2 \cdot k$ für eine ganze Zahl k . Deshalb ist $a^2 = (2 \cdot k) \cdot a = 2 \cdot (k \cdot a)$. Weil $k \cdot a$ eine ganze Zahl ist, folgt schließlich $a^2 = 2 \cdot k'$ für eine ganze Zahl k' . Also ist a^2 gerade ($\neg A$ gilt). Damit haben wir $\neg B \rightarrow \neg A$ gezeigt und damit auch, dass die ursprüngliche Aussage $A \rightarrow B$ gelten muss (Kontrapositionsregel). Genauso beweist man die Aussage

$$\text{wenn } a^2 \text{ eine gerade Zahl ist, dann ist } a \text{ gerade} \quad (**)$$

Wir zeigen die kontrapositive Aussage: wenn a ungerade ist, dann ist auch a^2 ungerade. Ist a ungerade, so ist $a = 2k + 1$ für eine ganze Zahl k . Da $x = y \implies x^2 = y^2$, haben wir, dass

$$a^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$$

Da k eine ganze Zahl ist, ist auch $(2k^2 + 2k)$ eine ganze Zahl. Also ist a^2 ungerade, wie behauptet.

Die beiden Aussagen (*) und (**) zusammen liefern die Aussage

$$a^2 \text{ ist gerade} \iff a \text{ ist gerade} \quad (1.1)$$

► **Beispiel 1.20 : (Widerspruchsregel)** Eine reelle Zahl x genau dann *rational* ist, wenn es zwei ganze Zahlen a und b mit der Eigenschaft

$$x = \frac{a}{b} \text{ und die Zahlen } a, b \text{ haben}^3 \text{ keinen gemeinsamen Teiler } k \geq 2$$

Wir wollen die Aussage

$$A = \text{“}\sqrt{2} \text{ ist irrational”}$$

beweisen. Dazu nehmen wir an, dass $\sqrt{2}$ eine rationale Zahl ist. Dann kann man diese Zahl als Quotient a/b zweier ganzen Zahlen a und b darstellen, wobei a und b keinen gemeinsamen Teiler $k \geq 2$ haben. Sei nun B die Aussage

$$B = \text{“}a \text{ und } b \text{ haben einen gemeinsamen Teiler } k \geq 2\text{”}$$

Wir wissen (nach Annahme), dass B falsch ist. Also, um die Aussage A zu beweisen, reicht es zu zeigen, dass $\neg A \rightarrow B$ wahr ist. Dazu nehmen wir an, dass $\neg A$ wahr ist. Dann gilt:

1. $\sqrt{2} = a/b$.
2. Für beliebige Zahlen x, y mit $x = y$ gilt $x^2 = y^2$. Also gilt $2 = a^2/b^2$.
3. Multiplizieren wir beide Seiten mit b^2 so erhalten wir $a^2 = 2b^2$.
4. Da a^2 gleich zwei mal eine ganze Zahl ist, ist a^2 gerade.
5. Nach (1.1) ist auch a gerade.
6. Nach der Definition von geraden Zahlen, muss es eine ganze Zahl c mit $a = 2c$ geben.
7. Aus $2b^2 = a^2$ und $a = 2c$ folgt $2b^2 = 4c^2$ und damit folgt $b^2 = 2c^2$.
8. Also ist b^2 eine gerade Zahl und, wieder nach (1.1), ist auch b gerade.
9. Da beide Zahlen a und b gerade sind, ist 2 ein gemeinsamer Teiler von a und b , d.h. die Aussage B ist wahr.

Da B falsch ist und (wie wir gerade gezeigt haben) $\neg A \rightarrow B$ gilt, muss (nach der Widerspruchsregel) die Aussage A falsch sein. Also ist $\sqrt{2}$ eine irrationale Zahl, wie behauptet.



Jeder einzelner Schritt im Beweis muss korrekt begründet werden! Zum Beispiel schreibt man

$$1 = \sqrt{1} = \sqrt{(-1)(-1)} \stackrel{?}{=} \sqrt{-1} \sqrt{-1} = (\sqrt{-1})^2 = -1$$

als “Beweis” für $1 = -1$. Aber die Aussage, dass $\sqrt{xy} = \sqrt{x} \sqrt{y}$ für alle Zahlen x und y gelten muss, ist falsch!



Man muss vorsichtig sein, wenn man beide Seiten einer Gleichung durch einer Variable dividiert $ax = xb \implies a = b$ nur wenn $x \neq 0$. Um $x \neq 0$ nachzuweisen, muss man zusätzliche Arbeit leisten.



Man behauptet: aus $ax < bx$ folgt $a < b$. Das ist noch schlimmer! Die Folgerung $a < b$ ist einfach falsch, wenn $x < 0$ ist, und ist unbewiesen, wenn $x = 0$ ist.

1.4 Mathematische Induktion: Beweis von Aussagen $\forall x P(x)$

Als nächstes werden wir ein allgemeines (und überraschend mächtiges) Prinzip – das *Induktionsprinzip* – kennen lernen. Dieses Prinzip erlaubt, Aussagen der Form „ $\forall n \in \mathbb{N} : P(n)$ “ zu beweisen.⁴

Nicht immer lässt sich solche Aussagen einfach dadurch beweisen, dass man eine beliebige natürliche Zahl a wählt und dann einen Beweis für $P(a)$ führt. Zum Beispiel ist ein solcher Beweis für die Behauptung

Jeder Geldbetrag von mindestens 4 Pfennigen lässt sich allein mit Zwei- und Fünfpfennigstücken bezahlen

etwas umständlich, falls wir a Pfennige bezahlen müssen und sonst nichts über a wissen. Wenn wir jedoch wissen, in welcher Stückelung a Pfennige zu bezahlen sind, können wir daraus recht leicht schließen, wie eine Stückelung für $a + 1$ Pfennige aussehen kann: Man stelle sich dazu den Münzenhaufen für a Pfennige vor.

1. Wenn er zwei Zweipfennigstücke enthält, nehmen wir sie fort und legen dafür ein Fünfpfennigstück hinzu
2. Wenn er ein Fünfpfennigstück enthält, nehmen wir es fort und legen dafür drei Zweipfennigstücke hinzu.

Im ersten Fall enthält der Münzenhaufen $a - 2 \cdot 2 + 5 = a + 1$ Pfennige, und im zweiten Fall $a - 5 + 3 \cdot 2 = a + 1$ Pfennige. Da jeder solcher Münzenhaufen von mindestens 4 Pfennigen entweder ein Fünfpfennigstück oder zwei Zweipfennigstücke enthalten muss, ist stets eine der beiden Regeln anwendbar.

Da sich 4 Pfennige mit zwei Zweipfennigstücken bezahlen lassen, können wir jetzt mit Hilfe der beiden Umtauschregeln für jeden größeren Betrag eine Stückelung in Zwei- und Fünfpfennigstücken angeben.

In der Beweisargumentation haben wir übrigens nicht gezeigt, wie man einen beliebigen Betrag stückeln kann. Wir haben „lediglich“ gezeigt, dass man 4 Pfennige stückeln kann, und wie man aus der Stückelung eines Betrages von a Pfennigen – für ein beliebiges $a \geq 4$ – die Stückelung von $a + 1$ Pfennigen ableiten kann. Da jede natürliche Zahl $a \geq 4$ durch wiederholte Addition von 1 erhalten werden kann, erhalten wir für jedes a auch eine Stückelung. Diese Vorgehensweise nennt man (mathematische) *Induktion*.⁵

⁴Für den Beweis der Existenzaussagen $\exists x \in M : P(x)$ gibt es das sogenannte *Taubenschlagprinzip*, das wir später kennenlernen werden. Dieses Prinzip hat auch eine weitgehende (und in der Mathematik wie auch in der Informatik häufig anwendbare) Erweiterung – die sogenannte *probabilistische Methode*. Aus Zeitgründen können wir diese Methode nicht ausführlich betrachten und werden diese Methode später nur auf einigen wenigen Beispielen veranschaulichen.

⁵In deutschsprachiger Literatur wird oft die Name “vollständige Induktion” benutzt. Ich sehe aber keinen Grund, warum man hier noch das Wort “vollständige” benutzen sollte – es ist doch keine Induktion bekannt, die “nicht vollständig” ist! Das Wort “mathematische” kann man notfalls benutzen, wenn man unbedingt den Unterschied zur Induktion in der Elektrotechnik unterstreichen will. Wir werden dies aber nicht tun.

Das Induktionsprinzip

Die Grundidee der Induktion⁶ beruht auf dem axiomatischen Aufbau der natürlichen Zahlen nach Peano: Man kann jede natürliche Zahl dadurch erhalten, indem man, beginnend mit der 0, wiederholt 1 addiert. Entsprechend beweist man eine Eigenschaft $P(n)$ für jede natürliche Zahl n , indem man zuerst die Eigenschaft $P(0)$ – die so genannte *Induktionsbasis* oder *Verankerung* – beweist, und anschließend zeigt, dass für beliebige natürliche Zahlen a aus $P(a)$ auch $P(a+1)$ folgt – der so genannte *Induktionsschritt*.

Definition: (Das Induktionsprinzip)

Sei $P(n)$ ein Prädikat über dem Universum \mathbb{N} .

1. Basis: Zeige, dass $P(0)$ wahr ist.
2. Induktionsschritt $n \mapsto n + 1$: Für beliebiges $n \in \mathbb{N}$ zeige, dass $P(n) \rightarrow P(n + 1)$ wahr ist.

Dann gilt auch die Aussage $\forall n \in \mathbb{N} : P(n)$.

Beweis. Durch Kontraposition. Nehmen wir an, dass die Aussage $\forall n \in \mathbb{N} : P(n)$ gilt nicht. D.h. es gibt ein $n_0 \in \mathbb{N}$ mit $\neg P(n_0)$. Da aber $P(n_0 - 1) \rightarrow P(n_0)$, impliziert das $\neg P(n_0 - 1)$; hier haben wir die Widerspruchsregel

$$\frac{\neg P(n_0), \quad P(n_0 - 1) \rightarrow P(n_0)}{\neg P(n_0 - 1)}$$

benutzt. Damit bekommen wir nach n_0 Schritten, dass $\neg P(0)$ wahr sein muss, ein Widerspruch. \square

Man muss nicht unbedingt von Null starten! Will man eine Aussage von der Form $\forall n \geq n_0 : P(n)$ beweisen, so ist $P(n_0)$ das Induktionsbasis.

Nicht immer ist es im Induktionsschritt einfach, alleine von $P(n)$ auf $P(n + 1)$ zu schließen. Betrachtet man den Induktionsschritt genauer, so sieht man, dass man eigentlich sogar die Gültigkeit von $P(0) \wedge \dots \wedge P(n)$ als Voraussetzung nutzen kann. Damit ergibt sich das Prinzip der *verallgemeinerten Induktion*:

Gelten die beiden Aussagen $P(0)$ und $\forall n \in \mathbb{N} : (P(0) \wedge \dots \wedge P(n)) \rightarrow P(n + 1)$, dann gilt die Aussage $\forall n \in \mathbb{N} : P(n)$.

Einige falsche Anwendungen

Zuerst geben wir ein Paar falsche(!) Anwendungen.⁷ Die erste (und nicht so ganz ernst gemeinte) Behauptung ist:

In einen Koffer passen unendlich viele Paare von Socken.

“Beweis” mit Induktion: Induktionsbasis: $n = 1$. Ein Paar Socken passt in einen leeren Koffer.

Induktionsschritt $n \mapsto n + 1$: In einem Koffer sind n Paar Socken. Ein Paar Socken passt immer noch rein, dies ist eine allgemeingültige Erfahrung. Also sind nun $n + 1$ Paar Socken in dem Koffer. \square

⁶Wer hat die Induktion erfunden? Das ist nicht klar. Klar ist nur, dass Francesco Maurolico die Induktion in seinem Buch *Arithmetorum Libri Due* (1575) benutzt hat (um zu zeigen, dass die Summe der ersten n ungeraden Zahlen gleich n^2 ist; siehe Aufgabe 21).

⁷Am besten lernt man aus den Fehlern ...

Wo ist der Fehler? Die Induktion ist ein *konstruktives* Beweisverfahren und solche Beweise erfordern auch konstruktive Argumente. Im Sockenbeispiel war das Argument “die Erfahrung sagt, dass immer noch ein Paar Socken mehr in den Koffer paßt” nicht konstruktiv. Ein konstruktives Argument sollte genau sagen *wo die Lücke für das weitere Paar Socken sein wird!*

Noch eine falsche Behauptung:


Alle natürliche Zahlen sind gleich.


“Beweis” mit Induktion. Es reicht zu zeigen, dass $a = b$ für alle $a, b \in \mathbb{N}$ gilt. Wir “beweisen” das mit Induktion über $n = \max\{a, b\}$. Dazu betrachten wir die Aussage

$$P(n) = \text{“für alle } a, b, \in \mathbb{N} \text{ mit } \max\{a, b\} = n \text{ gilt } a = b\text{”}$$

Basis $n = 0$ ist richtig, denn aus $a, b \in \mathbb{N}$ und $\max\{a, b\} = 0$ folgt $a = 0$ und $b = 0$.

Induktionsschritt: $n \mapsto n + 1$. Nehmen wir an, dass $P(n)$ gilt und betrachten ein beliebiges Paar von Zahlen $a, b \in \mathbb{N}$ mit $\max\{a, b\} = n + 1$. Dann ist $\max\{a - 1, b - 1\} = n$ und, nach der Induktionsannahme, gilt $a - 1 = b - 1$ und damit auch $a = b$. Fertig. \square

 Wo ist der Fehler? Aus $\max\{a, b\} = n + 1$ folgt zwar immer, dass dann auch $\max\{a - 1, b - 1\} = n$ gelten muss. Aber die Induktionsvoraussetzung wird damit nicht unbedingt erfüllt: Ist z.B. $a = 0$, dann gehört $a - 1 = -1$ nicht mehr zum \mathbb{N} , und die Aussage $P(n)$ spricht nur über natürlichen Zahlen!

 Man vergisst oft, die Induktionsbasis zu verifizieren. Z.B. sei $P(n)$ die Aussage “ $\forall n \in \mathbb{N} : n = n + 1$ ”. Man wählt ein beliebiges $n \in \mathbb{N}$ und zeigt, dass $P(n) \rightarrow P(n + 1)$, was richtig ist, da $n = n + 1 \Rightarrow (n + 1) = (n + 1) + 1$. Aber $P(0)$ ist falsch, da $0 = 1$ nicht gilt.

Noch eine falsche Behauptung:

Wenn sich unter n Tieren ein Elefant befindet, dann sind alle diese Tiere Elefanten.

“Beweis” mit Induktion. Induktionsanfang: $n = 1$: Wenn von einem Tier eines ein Elefant ist, dann sind alle diese Tiere Elefanten.

Induktionsvoraussetzung: Die Behauptung sei richtig für alle natürlichen Zahlen kleiner oder gleich n .

Induktionsschluß: Sei unter $n + 1$ Tieren eines ein Elefant. Wir stellen die Tiere so in eine Reihe, dass sich dieser Elefant unter den ersten n Tieren befindet. Nach Induktionsannahme sind dann alle diese ersten n Tiere Elefanten. Damit befindet sich aber auch unter den letzten n Tieren ein Elefant, womit diese auch alle Elefanten sein müssen. Also sind alle $n + 1$ Tiere Elefanten. \square

Wo ist das Argument falsch? Im Fall $n + 1 = 2$ kann man den Elefanten zwar so stellen, dass er bei den ersten $n = 1$ Tieren steht. Folglich sind alle Tiere unter den ersten $n = 1$ Tieren Elefanten. Aber deshalb befinden sich unter den “letzten” n Tieren nicht notwendig Elefanten.

Einige richtige Anwendungen

Nun (endlich) folgen einige Beispiele für Sätze, die man gut (und richtig!) mittels Induktion beweisen kann.⁸

⁸Im Laufe der Vorlesung werden wir auch andere Induktionsbeweise sehen.

Satz 1.21. (Bernoulli-Ungleichung) Ist $x \in \mathbb{R}$ und $x \geq -1$, so gilt für alle $n \in \mathbb{N}_+$

$$(1 + x)^n \geq 1 + nx$$

Beweis. Wir führen den Beweis mittels Induktion über n .

Basis: Für $n = 1$ gilt $1 + x \geq 1 + x$.

Induktionsschritt $n \rightarrow n + 1$:

$$\begin{aligned} (1 + x)^{n+1} &= (1 + x)^n \cdot (1 + x) \quad (\text{Bemerkung: } 1 + x \geq 0) \\ &\geq (1 + nx) \cdot (1 + x) \quad (\text{nach Induktionsvoraussetzung}) \\ &= 1 + nx + x + nx^2 \\ &= 1 + (n + 1)x + nx^2 \\ &\geq 1 + (n + 1)x. \end{aligned}$$

□

Wir zeigen nun mittels verallgemeinerter Induktion, dass jede natürliche Zahl, die größer oder gleich 2 ist, als Produkt von Primzahlen dargestellt werden kann.⁹ Es gilt zum Beispiel

$$1815 = 3 \cdot 5 \cdot 11 \cdot 11.$$

Aus dieser Faktorisierung von 1815 lässt sich nicht unmittelbar auf die Primzahlzerlegung von $1815 + 1 = 1816$ schließen, wie es bei der bisher betrachteten Induktion notwendig gewesen wäre.

Satz 1.22. (Primzahldarstellung) Sei n eine natürliche Zahl, und $n \geq 2$. Dann ist n Produkt von Primzahlen.

Beweis. Wir führen den Beweis mittels verallgemeinerter Induktion.

Basis: Da 2 eine Primzahl ist, ist 2 triviales Produkt von sich selbst, also Produkt einer Primzahl.

Schritt $n \rightarrow n + 1$: Sei n beliebig und nehmen wir an, dass *alle* Zahlen von 2 bis n sich als Produkte von Primzahlen schreiben lassen. Wir zeigen, dass dann $n + 1$ ebenfalls ein Produkt von Primzahlen ist. Dazu machen wir eine Fallunterscheidung.

Fall 1: $n + 1$ ist eine Primzahl. Dann ist $n + 1$ die einzige Primzahl, aus der das Produkt $n + 1$ besteht.

Fall 2: $n + 1$ ist keine Primzahl. Dann gibt es zwei echte Teiler von $n + 1$, also natürliche Zahlen a und b mit $2 \leq a, b < n + 1$, so dass $n + 1 = ab$. Da a und b beide kleiner als $n + 1$ sind, können wir die Induktionsvoraussetzung nutzen und die Zahlen a und b als Produkte von Primzahlen schreiben. Dann ist $n + 1 = ab$ auch ein Produkt von Primzahlen. □

Nun benutzen wir die Induktion, um noch eine nützliche Ungleichung zu beweisen.

Eine reellwertige Funktion $f(x)$ heißt *konvex*, falls für alle reelle Zahlen λ zwischen 0 und 1 gilt:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

⁹Zur Erinnerung: $p \in \mathbb{N}$ ist eine *Primzahl* genau dann, wenn $p \geq 2$ und p ist *nur* durch 1 und p teilbar. Achtung: 1 ist also *keine* Primzahl!

Geometrisch gesehen, ist f konvex, falls die Gerade ℓ , die die Punkte $(x, f(x))$ und $(y, f(y))$ verbindet, oberhalb der Kurve $f(z)$ liegt. Ein einfaches Konvexitätskriterium ist $f''(x) \geq 0$ (siehe Abschnitt 3.8).

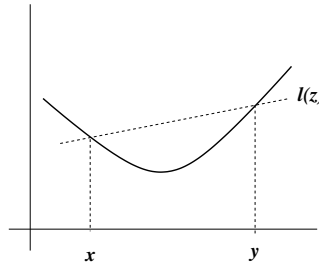


Abbildung 1.7: Eine konvexe Funktion

Satz 1.23. (Jensen's Ungleichung) Seien $0 \leq \lambda_i \leq 1$ mit $\sum_{i=1}^r \lambda_i = 1$. Ist f konvex, so gilt:

$$f\left(\sum_{i=1}^r \lambda_i x_i\right) \leq \sum_{i=1}^r \lambda_i f(x_i).$$

Beweis. Induktion über r . Für $r = 1$ ist die Ungleichung offensichtlich richtig, da dann $\lambda_1 = 1$ gelten muss. Für $r = 2$ ist die Ungleichung auch richtig wegen der Konvexität. Nehmen wir also an, dass die Ungleichung für r Summanden richtig ist, und zeigen, dass sie dann auch für $r + 1$ Summanden richtig bleibt. Dazu reicht es, die Summe der ersten zwei Terme in $\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_{r+1} x_{r+1}$ durch den Term μy , wobei

$$\mu = \lambda_1 + \lambda_2 \quad \text{und} \quad y = \frac{\lambda_1}{\lambda_1 + \lambda_2} x_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} x_2.$$

Da $0 \leq \mu \leq 1$ und $\mu + \sum_{i=3}^r \lambda_i = \sum_{i=1}^r \lambda_i = 1$, können wir die Induktionsannahme anwenden, woraus

$$f\left(\sum_{i=1}^{r+1} \lambda_i x_i\right) = f\left(\mu y + \sum_{i=3}^{r+1} \lambda_i x_i\right) \leq \mu f(y) + \sum_{i=3}^{r+1} \lambda_i f(x_i)$$

folgt. Nun benutzen wir wieder die Kovexität von f und erhalten

$$\begin{aligned} \mu f(y) &= (\lambda_1 + \lambda_2) f\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} x_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} x_2\right) \\ &\leq (\lambda_1 + \lambda_2) \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} f(x_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2} f(x_2)\right) \\ &= \lambda_1 f(x_1) + \lambda_2 f(x_2). \end{aligned}$$

□

Obwohl einfach, ist Jensen's Ungleichung¹⁰ in vielen Fällen sehr nützlich. Sie liefert uns sofort, zum Beispiel, die folgende Ungleichung zwischen arithmetischem und geometrischem Mittel.

¹⁰Genau wie sogenannte Cauchy-Schwarz Ungleichung

$$\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right) \left(\sum_{i=1}^n y_i^2\right),$$

Satz 1.24. (Geometrisches und arithmetisches Mittel) Seien a_1, \dots, a_n nicht-negative Zahlen. Dann gilt

$$\sqrt[n]{a_1 \cdot a_2 \cdot \dots \cdot a_n} \leq \frac{a_1 + a_2 + \dots + a_n}{n}. \quad (1.2)$$

Beweis. Sei $f(x) = e^x$, $\lambda_1 = \dots = \lambda_n = 1/n$ und $x_i = \ln a_i$, für alle $i = 1, \dots, n$. Nach der Jensen's Ungleichung gilt:

$$\frac{1}{n} \sum_{i=1}^n a_i = \sum_{i=1}^n \lambda_i f(x_i) \geq f\left(\sum_{i=1}^n \lambda_i x_i\right) = e^{(\sum_{i=1}^n x_i)/n} = \left(\prod_{i=1}^n e^{\ln a_i}\right)^{1/n} = \left(\prod_{i=1}^n a_i\right)^{1/n}.$$

□

Ist Induktion nur etwas für Zahlen? Nein, in $P(n)$ ist n nur ein *Parameter*, die Aussage “ $P(n)$ ist wahr” muss *nicht* unbedingt eine Aussage über die Zahl n sein. Z.B. kann man auch solche Aussagen $P(n)$ nehmen: “Jedes einfache Polygon mit n Ecken lässt sich durch $n - 3$ Diagonalen triangulieren.” Das allgemeine Schema ist wie folgt. Man hat eine Menge S und ein Prädikat $Q(s)$ über S . Man will zeigen, dass $\forall s \in S : Q(s)$ gilt. Dafür wählt man eine passende Abbildung

$$S \ni s \mapsto L(s) \in \mathbb{N},$$

die zu jedem Element $s \in S$ seine “Länge” $L(s)$ zuweist.¹¹ Dann betrachtet man das Prädikat $P(n) =$ “für alle $s \in S$ mit $L(s) = n$ gilt $Q(s)$ ” und zeigt mittels Induktion, dass $\forall n \in \mathbb{N} P(n)$ gilt.

Um diese Verallgemeinerung zu demonstrieren, betrachten wir die Anzahl der Elemente in dem kartesischen Produkt zweier Mengen.

Behauptung 1.25. Für endliche Mengen A, B gilt: $|A \times B| = |A| \cdot |B|$.

Beweis. Durch Induktion nach $n := |A|$ (dabei sei B beliebig aber fest).

Induktionsanfang: für $n = 0$ ist $A = \emptyset$, also auch $A \times B = \emptyset$.

Induktionsschritt: $n \mapsto n + 1$. Sei A eine *beliebige* Menge mit $|A| = n + 1$ Elementen. Wegen $n + 1 \geq 1$ existiert ein Element $a \in A$. Wir betrachten die Menge $X := A \setminus \{a\}$. Dann hat X nur n Elemente, erfüllt also $|X \times B| = n \cdot |B|$ nach Induktionsvoraussetzung. Wegen

$$A \times B = (\{a\} \times B) \cup (X \times B) \quad \text{und} \quad (\{a\} \times B) \cap (X \times B) = \emptyset$$

folgt

$$|A \times B| = |\{a\} \times B| + |X \times B| = |B| + n \cdot |B| = (n + 1) \cdot |B| = |A| \cdot |B|.$$

□

Ein *Geradenarrangement* besteht aus endlich vielen Geraden in der Ebene. Dabei habe eine Gerade kein Ende und keinen Anfang. Die Geraden unterteilen die Ebene in verschiedene Flächen.

die wir erst später beweisen werden (siehe Satz 5.6).

¹¹Dieser Schritt, den man als *Parametrisierungsschritt* bezeichnen kann, ist wichtig. Hat ein Objekt s zwei (oder mehrere) Merkmale $a, b \in \mathbb{N}$, so kann man $L(s) = a + b$ aber auch $L(s) = a \cdot b$ oder auch $L(s) = \max\{a, b\}$ usw. wählen. Nicht jede Auswahl wird jedoch zum Erfolg führen. Deshalb muss man an dieser Stelle genauer überlegen, mit welcher Funktion $L(s) = f(a, b)$ es am einfachsten wird, den Induktionsschritt durchzuführen.

Satz 1.26. (Färbungen von Flächen) Bei beliebiger Anzahl der Geraden ist es möglich die entstehenden Flächen mit nur zwei Farben so zu färben, dass keine zwei benachbarte Flächen dieselbe Farbe tragen werden.

Beweis. Induktion über die Anzahl der Geraden n . Basis $n = 1$ ist richtig, da wir dann nur zwei Flächen haben können.

Induktionsschritt: $n - 1 \mapsto n$. Sei die Aussage nun für $n - 1$ Geraden richtig. Wir wollen zeigen, dass dann die Aussage auch für n Geraden richtig ist. Dazu nehmen wir ein *beliebiges* Geradenarrangement mit n Geraden und entfernen eine (auch beliebige) Gerade g . Nach der Induktionsannahme, muss eine Zweifärbung der Flächen in dem verbleibenden Arrangement von $n - 1$ Geraden möglich sein. Nun nehmen wir die Gerade g wieder dazu. Nach der Hinzunahme von g werden manche der alten Flächen in zwei neue Flächen geteilt. Nun kommt der Trick: Wir *behalten* die Farben für (neue) Flächen, die auf einer Seite von g liegen, und *kippen* die Farben der verbleibenden Flächen *um*. \square

In vielen Anwendungen ist ein Knoten des Baumes B als Startknoten ausgezeichnet; dann spricht man über einen *Wurzelbaum*. In einem solchen Baum kann das Verhältnis der einzelnen Knoten des Baumes zueinander begrifflich gut beschrieben werden. Dazu benutzt man den Begriff der *Tiefe* eines Knotens.

Als *Tiefe* eines Knotens v von B wird der Abstand von v zur Wurzel (d.h. die Anzahl der Kanten in dem einzigen Weg von v zur Wurzel) bezeichnet. Die *Tiefe* von B ist die maximale Tiefe eines Knotens von T . Alle Knoten gleicher Tiefe bilden ein *Knotenniveau*. Als *Kinder* eines Knotens v von B werden sämtliche Knoten bezeichnet, die zu v benachbart sind und deren Tiefe die von v um eins übersteigt. v heißt *Vater* seiner Kinder. In einem *binären Baum* hat jeder innere Knoten höchstens zwei Kinder. Knoten ohne Kinder heißen *Blätter* (siehe Abb. 1.8).

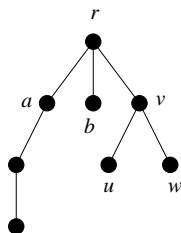


Abbildung 1.8: Dieser Wurzelbaum mit der Wurzel r hat die Tiefe 3. u und w sind Kinder des Knotens v , und v besitzt die Tiefe 1. a , b und v bilden ein Knotenniveau.

Für ein Baum B sei $t(B)$ seine Tiefe und $|B|$ die Anzahl der Blätter. Weiss man $|B|$, was kann man dann über die Tiefe $t(B)$ sagen? Mann kann aber zeigen, dass $\log_2 |B|$ bereits die untere Grenze für die Tiefe ist: Kein binärer Baum B kann kleinere Tiefe als $\log_2 |B|$ haben.

Satz 1.27. Für jeden binären Baum B gilt: $t(B) \geq \log_2 |B|$, d.h.

$$\text{Tiefe}(B) \geq \log_2(\text{Anzahl der Blätter}).$$

Beweis. Wir führen den Beweis mittels Induktion über die Tiefe $t = t(B)$.

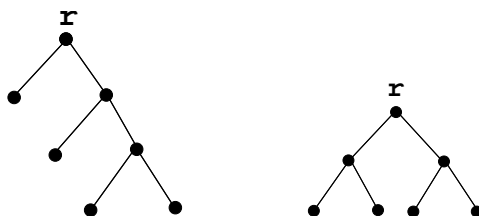


Abbildung 1.9: Diese zwei Beispiele zeigen, dass es sowohl Bäume mit $t(B) = |B| - 1$ wie auch Bäume mit $t(B) = \log_2 |B|$ geben kann.

Basis $t = 0$: In diesem Fall besteht B nur aus einem Knoten. Dieser Knoten ist ein Blatt, es gilt also $|B| = 1$. Da $1 \leq 2^0$ bzw. $\log_2 1 \leq 0$ ist die Behauptung für $d = 0$ wahr.

Induktionsschritt $t - 1 \mapsto t$: Sei die Behauptung bereits für alle binären Bäume der Tiefe $\leq t - 1$ bewiesen und sei B ein beliebiger binärer Baum der Tiefe $t = t(B)$. Wir wollen zeigen, dass $|B| \leq 2^{t(B)}$ gilt (was äquivalent zu $t(B) \geq \log_2 |B|$ ist).

Sei r die Wurzel von B und seien (r, u) und (r, v) die beiden mit r inzidenten Kanten. Wir betrachten die beiden in u und v wurzelnden Unterbäume B_u und B_v von B . Beide diese Teilbäume haben Tiefe höchstens $t - 1$ und für die Anzahl der Blätter gilt: $|B_u| + |B_v| = |B|$. Nach Induktionsannahme gilt also

$$|B_u| \leq 2^{t(B_u)} \quad \text{und} \quad |B_v| \leq 2^{t(B_v)}.$$

Wir erhalten damit

$$|B| = |B_u| + |B_v| \leq 2^{t(B_u)} + 2^{t(B_v)} \leq 2 \cdot 2^{t-1} \leq 2^t.$$

□

1.4.1 Induktion und Entwurf von Algorithmen

Anwendungen der Induktion findet man in allen mathematischen Gebieten, von Mengenlehre bis Geometrie, von Differentialrechnung bis Zahlentheorie. Sogar der Beweis des großen Fermatschen Satzes¹² (Andrew Wiles, 1993) verwendet die Induktion (neben vielen anderen Argumenten). Die Induktion ist auch in der Informatik wichtig, denn rekursive Algorithmen sind *induktive* Beschreibungen von Objekten.

Viele Probleme, Modelle oder Phänomene haben eine sich selbst referenzierende Form, bei der die eigene Struktur immer wieder in unterschiedlichen Varianten enthalten ist. Wenn diese Strukturen in eine mathematische Definition, einen Algorithmus oder eine Datenstruktur übernommen werden, wird von Rekursion gesprochen. Rekursive Definitionen sind jedoch nur sinnvoll, wenn etwas immer durch einfachere Versionen seiner selbst definiert wird, wobei im Grenzfall ein Trivialfall gegeben ist, der keine Rekursion benötigt. Rekursion hat also was mit der Induktion zu tun – man kann die Rekursion als Induktion “in umgekehrter Richtung” betrachten.

Deshalb ist Induktion nicht nur für den *Beweis*, dass ein gegebener Algorithmus korrekt ist, geeignet – man kann sie auch für den *Entwurf* der Algorithmen benutzen! Das allgemeine Schema – als *dynamisches Programmieren* bekannt – ist folgendes:

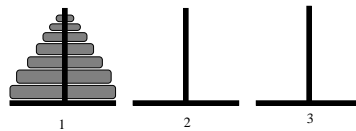
¹²Dieser Satz besagt, dass es keine natürlichen Zahlen x, y, z gibt, so dass für n größer als 2 die folgende Relation gilt: $x^n + y^n = z^n$.

Um ein Problem zu lösen, löst man zuerst *Teilprobleme* und die Lösungen der Teilprobleme zu einer Lösung des ursprünglichen Problems kombiniert.

Rekursive Programme sind ein Spezialfall dynamischen Programme: Die Lösungen der Teilprobleme müssen nicht abgespeichert werden.

Das folgende Problem wurde 1883 von Edouard Lucas erfunden.

- Gegeben sind drei Plätze zum Stapeln und n Scheiben, die zu Beginn alle auf dem 1. Stapel liegen.
- Alle Scheiben sind unterschiedlich groß und sie müssen auf einem Stapel in geordneter Weise liegen, so daß nicht größere Scheiben auf kleineren liegen.
- Jede Scheibe ist beweglich und kann von einem Stapel auf einen anderen getragen werden. Es dürfen jedoch nicht mehrere Scheiben auf einmal bewegt werden.
- Aufgabenstellung: Alle Scheiben sind von dem 1. Stapel auf den 2. Stapel zu befördern.

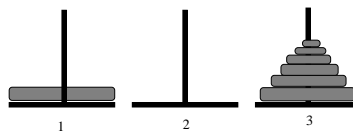


Man kann dieses Problem mittels Induktion lösen. Sei $P(n)$ die Aussage: “Wir haben einen Algorithmus, der das Problem für n Scheiben löst”.

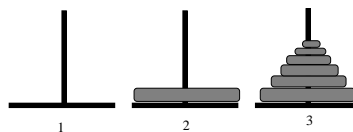
Induktionsbasis: $P(0)$ ist wahr, da wir keine Scheiben überhaupt haben.

Induktionsschritt: Wir wissen wie man das Problem für $n - 1$ Scheiben lösen kann und wollen einen Algorithmus für n Scheiben entwerfen. Dazu beobachten wir, dass das Problem, n Scheiben vom Stapel 1 zum Stapel 2 zu befördern, lässt sich lösen, wenn

1. zunächst die obersten $n - 1$ Scheiben von Stapel 1 zum als Hilfsstapel genutzten Stapel 3 verlegt werden,

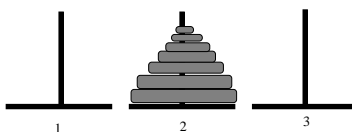


2. dann die jetzt zuoberst liegende Scheibe auf Stapel 1 nach Stapel 2 verlegt wird



und abschließend

3. wieder $n - 1$ Scheiben vom Hilfsstapel 3 nach Stapel 2 verlagert werden.



Zu beachten ist dabei, dass die Rollen der drei Stapel (Ausgangs-, Ziel- und Hilfsstapel) ständig wechseln.

Damit können wir ein rekursives Programm $\text{Hanoi}(n; 1, 2, 3)$, das die gegebenen n Scheiben von 1. Stapel auf 3. Stapel befördert, wie folgt beschreiben:

Algorithmus $\text{Hanoi}(n; 1, 2, 3)$

While $n > 0$ do

Rufe $\text{Hanoi}(n - 1; 1, 3, 2)$ auf [die obersten $n - 1$ Scheiben von Stapel 1 zum Stapel 2]

Verlege die zuoberst liegende Scheibe auf 1 nach 3

Rufe $\text{Hanoi}(n - 1; 2, 1, 3)$ auf [verlege $n - 1$ Scheiben vom Hilfsstapel 2 nach 3]

1.5 Das Taubenschlagprinzip: Beweis von Aussagen $\exists x P(x)$

Das sogenannte *Taubenschlagprinzip*, in der englischsprachiger Literatur auch als *Pigeonhol Principle* bezeichnet, geht auf den Mathematiker G. L. Dirichlet zurück. Dieses Prinzip erlaubt Existenzaussagen $\exists x P(x)$ über endliche Universen M zu beweisen, ohne ein konkretes Element $a \in M$, für das $P(a)$ gilt, anzugeben! Das Prinzip selbst basiert sich auf folgender einfacher Beobachtung:

Halten sich $r + 1$ Tauben in r Taubenlöcher auf, so gibt es mindestens einen Taubenlöch, in dem sich wenigstens zwei Tauben befinden.

Ist das Verhältnis von Taubenlöcher zu Tauben nicht nur $k + 1$ zu k sondern zum Beispiel $2k + 1$ zu k , so kann man sogar schließen, dass in einem der Taubenlöcher mindestens 3 Tauben sitzen müssen.



Taubenschlagprinzip:

Halten sich $sr + 1$ Tauben in r Taubenlöcher auf, so gibt es mindestens einen Taubenloch, in dem sich wenigstens $s + 1$ Taube befinden.

Wäre das nämlich nicht der Fall, so hätten wir höchstens sr Tauben insgesamt.

▷ *Beispiel 1.28*: In einer Gruppe von 8 Leuten haben (mindestens) zwei am gleichen Wochentag Geburtstag. Warum? Seien die Leute "Tauben" und die Wochentage "Taubenlöcher". Wir haben also $r = 7$ Taubenlöcher und $r + 1$ Taube. Das Taubenschlagprinzip (mit $s = 1$) garantiert in dieser Situation die Existenz eines Wochentages, an dem also mindestens $s + 1 = 2$ Leute der Gruppe Geburtstag haben.

► *Beispiel 1.29* : Wir sagen, dass zwei Leute befeindet sind, falls sie nicht befreundet sind. Behauptung:

- (*) In jeder Gruppe von 6 Leuten gibt es 3 Leute, die paarweise befreundet oder paarweise verfeindet sind.¹³

Dieses Problem lässt sich auch als Graphenproblem stellen. Die Gruppe von Leuten steht für eine Menge von Knoten. Sind zwei Leute befreundet, so wird eine Kante zwischen den beiden entsprechenden Knoten erstellt, d.h. die beiden Knoten sind benachbart; sind zwei Leute verfeindet, so liegt keine Kante zwischen den entsprechenden Knoten. Damit ist die Kantenmenge bestimmt. Die Behauptung liest sich dann folgendermaßen: jeder Graph mit 6 Knoten enthält 3 Knoten, die entweder paarweise benachbart oder die paarweise nicht benachbart sind. Den Beweis kann man leicht aus der Abbildung 1.10 ablesen.

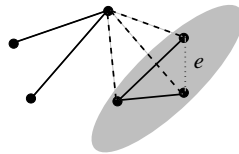


Abbildung 1.10: Ist das Paar e befreundet oder befeindet?

Frage: Gilt die Behauptung (*) mit 5 statt 6 Leuten in der Gruppe?

► *Beispiel 1.30* : Ein Sandkasten hat die Form eines gleichseitigen Dreiecks mit Seiten der Länge 2 Meter (Abb. 1.11(A)).

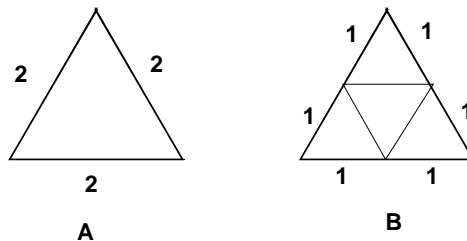


Abbildung 1.11: Die Sandkasten und ihre Aufteilung.

In diesem Sandkasten wollen 5 Kinder spielen. Das Problem ist nur, dass die Kinder um mehr als einem Meter von einander entfernt sein sollen, da sonst werden sie kräftig streiten. Ist es möglich, die 5 Kinder in den Sandkasten so zu verteilen, dass endlich Ruhe herrschen wird?

Antwort ist *nein*. Dazu teile die Sandkasten in 4 Teilen auf, wie in Abb. 1.11(B) gezeigt ist (das sind unsere "Taubenlöcher"). Da wir mehr als 4 Kinder haben, werden mindestens 2 Kinder in einem Taubenloch (=kleinem Dreieck) sitzen, und die Abstand zwischen den Kindern wird höchstens 1 Meter sein.

¹³Das ist die einfachste Form ("Kindergartenform") des in 1930 bewiesenen berühmten Satzes von Frank Plumpton Ramsey, der zur Geburt der *Ramseytheorie* – einem Gebiet der diskreten Mathematik – geführt hat.

► *Beispiel 1.31* : Behauptung: In *jeder* Menge $S \subseteq \mathbb{Z}$ von $|S| = 1000$ ganzen Zahlen gibt es zwei Zahlen $x \neq y$, so dass $x - y$ durch 573 teilbar ist.

Auf erstem Blick sieht das Problem sehr schwer aus – diese 1000 Zahlen können doch beliebig sein! Auch unsere alte Freundinnen – die Induktion – kann hier nur wenig helfen. Andererseits, können wir die Behauptung mit dem Taubenschlagprinzip in ein paar Zeilen beweisen.

Beweis: Als Tauben nehmen wir die Elemente von S und als Taubenlöcher die Elementen von $R = \{0, 1, \dots, 572\}$. Wir setzen die Taube $x \in S$ in das Taubenloch $r \in R$ genau dann, wenn x geteilt durch 573 den Rest r ergibt. Da $|S| > |R|$, müssen mindestens zwei Tauben x und y in einem Taubenloch r aufhalten. Das bedeutet, dass x und y geteilt durch 573 den selben Rest r ergeben, und damit muss $x - y$ durch 573 teilbar sein.

Beachte, dass die Wahl der Zahlen (1000 und 573) hier unwichtig ist – wichtig ist nur, dass $|S| > |R|$ gilt. Ist $n > m$, so muss jede Menge aus n Zahlen zwei Zahlen x und y enthalten, deren Differenz $x - y$ durch m teilbar ist.

Auf dem ersten Blick scheint das Taubenschlagprinzip nur einfache Schlussfolgerungen zuzulassen kann. Wie die folgenden Anwendungen zeigen, trügt der Schein! Man kann nämlich mit diesem Prinzip einige klassische Resultate beweisen. Hier beschränken wir uns mit einigen Beispielen.

In folgenden Satz geht es um die Existenz rationaler Approximationen von irrationalen Zahlen. Der Beweis war die erste nicht triviale Anwendung des Taubenschlagsprinzips in der Mathematik überhaupt! Diese Anwendung hat Dirichlet gemacht, deshalb nennt man das Prinzip oft *Dirichlet's Prinzip*.



Dirichlet 1879:

Sei x eine reelle Zahl und nicht rational. Für jede natürliche Zahl n existiert eine rationale Zahl p/q mit $1 \leq q \leq n$ und

$$\left| x - \frac{p}{q} \right| < \frac{1}{nq} \leq \frac{1}{q^2}.$$

Beweis. Ist x eine reelle Zahl, so definieren wir $L(x)$ als die Abstand zwischen x und der *ersten* ganzen Zahl, die links von x steht: $L(x) := x - \lfloor x \rfloor$. Es ist klar, dass dieser Abstand zwischen 0 und 1 liegen muss.

Sei $x \in \mathbb{R} \setminus \mathbb{Q}$ eine reelle aber irrationale Zahl. Als “Tauben” nehmen wir $n + 1$ Zahlen $L(ax)$ für $a = 1, 2, \dots, n + 1$ und sortieren sie in folgende n Intervalle ein, die die “Taubenlöcher” darstellen:

$$\left(0, \frac{1}{n}\right), \left(\frac{1}{n}, \frac{2}{n}\right), \dots, \left(\frac{n-1}{n}, 1\right).$$

Die Ränder der Intervalle werden *nicht* angenommen, da x reell und nicht rational ist. Da wir mehr Tauben als Taubenlöcher haben, müssen nach dem Taubenschlagsprinzip mindestens zwei Zahlen – dies seien $L(ax)$ und $L(bx)$ – in einem Intervall liegen. Da jedes der Intervalle kürze als $1/n$ ist, muss der Abstand zwischen diesen zwei Zahlen kleiner als $1/n$ sein:

$$|L(ax) - L(bx)| = |(ax - \lfloor ax \rfloor) - (bx - \lfloor bx \rfloor)| = \underbrace{|(a - b)x|}_{q} - \underbrace{|\lfloor ax \rfloor - \lfloor bx \rfloor|}_{p} < \frac{1}{n}.$$

Damit haben wir zwei ganze Zahlen p und q gefunden, für die die Ungleichung $|qx - p| < 1/n$ und damit auch die Ungleichung $|x - p/q| \leq \frac{1}{nq}$ gilt. Da q die Differenz zweier ganzen Zahlen ist, welche sich im Bereich $1, 2, \dots, n + 1$ bewegen, gilt auch $q \leq n$. \square

Als nächstes betrachten wir ein (auch klassisches) Resultat über Teilfolgen einer Zahlenfolge. Für eine Folge a_1, a_2, \dots, a_n von Zahlen ist $a_{i_1}, a_{i_2}, \dots, a_{i_r}$ eine *Teilfolge* der Länge r , falls $1 \leq i_1 < i_2 < \dots < i_r \leq n$ gilt. D.h. Eine Teilfolge der Länge r einer Folge entsteht, wenn wir alle ausser r Folgenglieder weg lassen. Die Teilfolge ist *monoton steigend* bzw. *monoton fallend* falls $a_{i_1} < a_{i_2} < \dots < a_{i_r}$ bzw. $a_{i_1} > a_{i_2} > \dots > a_{i_r}$ gilt. Die Teilfolge ist *monoton*, falls sie monoton steigend oder monoton fallend ist.

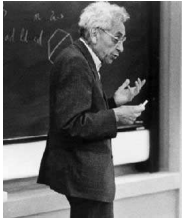
Zum Beispiel besitzt die Folge

$$7, 6, 11, 13, 5, 2, 4, 1, 9, 8$$

der Länge $n = 10$ die monoton fallende Teilfolge

$$7, 6, 2, 1$$

der Länge $r = 4$.



Erdős^a–Szekeres 1935:

Jede Folge von n verschiedenen Zahlen enthält eine monotone Teilfolge der Länge mindestens \sqrt{n} .

^aPaul Erdős 1913–1996. Einer der größten Mathematiker des letzten Jahrhunderts.

Bemerkung 1.32. Wichtig hier ist, dass das Ergebnis für *alle* Folgen gilt. D.h. egal aus welchen Zahlen eine beliebige Folge der Länge n besteht, muss sie eine monotone Teilfolge der Länge \sqrt{n} enthalten.

Beweis. Um die Notation zu vereinfachen, nehmen wir an, dass n ein Quadrat ist, d.h. $n = r^2$ für eine natürliche Zahl r gilt. Sei nun a_1, \dots, a_n eine beliebige Folge von n verschiedenen Zahlen. Wir ordnen a_i das Paar (F_i, W_i) zu, wobei

F_i = die Länge der längsten monoton *fallenden* Teilfolge steht, die in a_i endet,

W_i = die Länge der längsten monoton *wachsenden* Teilfolge, die in a_i startet.

Angenommen, a_1, \dots, a_n besitzt weder eine monoton wachsende noch eine monoton fallende Teilfolge der Länge r . Dann gilt $F_i, W_i < r$ für alle $i = 1, \dots, n$. Die Zahlen F_i und W_i liegen also zwischen 1 und $r - 1$ und es kann deshalb höchstens $(r - 1)^2$ verschiedene Paare (F_i, W_i) , $i = 1, \dots, n$ geben.

Wir betrachten die Zahlen a_1, \dots, a_n als ‘‘Tauben’’ und die Paare (x, y) mit $x, y \in \{1, \dots, r - 1\}$ als ‘‘Taubenlöcher’’. Wir setzen die i -te Taube in Taubenloch (x, y) genau dann, wenn $F_i = x$ und $W_i = y$ gilt. Da wir n Tauben und nur $(r - 1)^2 < r^2 = n$ Taubenlöcher haben, müssen zwei Tauben, seien es a_i und a_j mit $i < j$, in einem Taubenloch (x, y) sitzen, d.h. es muss $F_i = F_j = x$ und $W_i = W_j = y$ gelten. Das ist jedoch unmöglich, da $a_i > a_j$ ein Widerspruch zu $F_i = F_j$ und $a_i < a_j$ einen Widerspruch zu $W_i = W_j$ liefert. In erstem Fall ($a_i > a_j$) gilt $F_j \geq F_i + 1$ (die fallende Folge zu a_i kann man um ein Element a_j erweitern) und in zweiten Fall ($a_i < a_j$) gilt $W_i \geq W_j + 1$ (die steigende Folge zu a_j kann man um ein Element a_i erweitern). \square

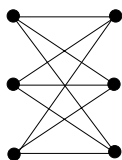
Um kompliziertere Existenzaussagen von der Form

$$\forall x(P(x) \rightarrow \exists yQ(x, y))$$

zu beweisen, lohnt es sich oft die beide mächtige Beweisprinzipien – Induktion und das Taubenschlagprinzip – zu kombinieren. Wir beschränken uns nur auf einem Beispiel. In diesem Beispiel geht es um Eigenschaften von Graphen.

Ein *Dreieck* in einem ungerichteten Graphen besteht aus drei benachbarten Knoten. Uns interessiert nun die Frage: Wieviel Kanten kann ein Graph haben ohne dass er einen Dreieck enthält?

In Bild unten erkennt man einen vollständigen bipartiten Graphen mit $2n$ Knoten für $n = 3$, also insgesamt 6 Knoten.

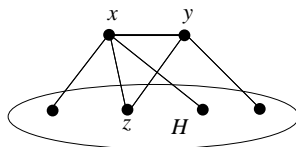


Dieser Graph enthält *keine* Dreiecke. Wie man unschwer erkennen kann, würde aber die Hinzunahme einer beliebigen der verbleibenden möglichen 6 Kanten jeweils 3 Dreiecke schließen. Somit scheint bei einer Knotenanzahl von $2n$, n^2 der höchste Wert zu sein, für den sich ein dreiecksfreier Graph noch konstruieren lässt. Diese ganzen Überlegungen münden nun in den

Satz 1.33. (Mantel 1907) Wenn ein Graph G mit $2n$ Knoten mindestens $n^2 + 1$ Kanten besitzt, dann besitzt G ein Dreieck.

Beweis. Der Beweis erfolgt per Induktion nach n . Induktionsbasis $n = 1$: In diesem Fall ist die Behauptung wahr, denn beide Seiten der Implikation sind falsch, (ein Graph mit 2 Knoten keine 2 Kanten besitzen kann).

Induktionsschritt $n \mapsto n + 1$: Wir nehmen also an, dass die Behauptung für n gilt und betrachten jetzt einen *beliebigen* Graphen $G = (V, E)$ mit $|V| = 2(n + 1)$ Knoten und $|E| = (n + 1)^2 + 1$ Kanten. Die beiden Knoten x und y seien adjazent in G , d.h. $xy \in E$. Mit H wird der, durch die Wahl von x und y definierte Teilgraph, bezeichnet:¹⁴



Der Graph H besitzt folglich $2n$ Knoten. Sollte H *mehr* als n^2 Kanten haben, gilt der Satz aufgrund unserer Induktionsvoraussetzung, denn dann gäbe es ein Dreieck im Teilgraphen H und damit auch in G . Also wird nun angenommen, dass der Teilgraph H *höchstens* n^2 Kanten hat. Sei F die Menge aller Kanten, die von den Knoten x und y zu Knoten in H führen. Insgesamt haben wir¹⁵

$$|F| \geq |E| - n^2 - 1 = ((n + 1)^2 + 1) - n^2 - 1 = 2n + 1$$

solchen Kanten. Wir haben also $2n + 1$ Kanten die von x bzw. y in den Teilgraphen H mit $2n$ Knoten führen. Diese Kanten betrachten wir nun als ‘‘Tauben’’ und Knoten von H als ‘‘Taubenlöcher’’. Nach dem Taubenschlagprinzip muss in H ein Knoten z existieren, der mit x und y jeweils eine Kante hat. Folglich besitzt G das Dreieck $\{x, y, z\}$. \square

¹⁴Um H zu bekommen, entfernen wir also aus G die Knoten x und y mit der entsprechenden Kanten.

¹⁵Wir müssen aus $|E| - n^2$ noch 1 abziehen, da $xy \in E$ gehört.

1.6 Kombinatorische Abzählargumente

Kombinatorik (wie die Diskrete Mathematik) beschäftigt sich vor allem mit *endlichen* Mengen. In klassischer Kombinatorik geht es hauptsächlich um Fragen vom Typ: “Auf wie viele Arten kann man ...”, was im Extremfall heißen soll “Kann man ... überhaupt?” Um vernünftig über solche Fragen reden zu können, bilden wir die Menge der Objekte, die uns interessieren, und fragen nach ihrer Mächtigkeit. In Kombinatorik haben sich einige spezielle Regeln herausgebildet, die alle ganz klar sind, sobald man sie einmal gesehen hat, auf die man aber erst einmal kommen muss.

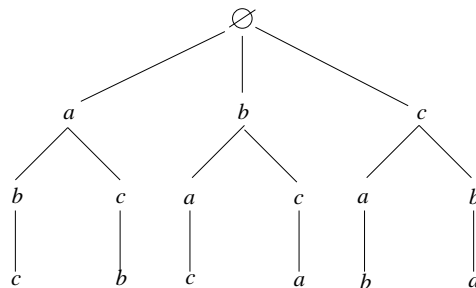
Satz 1.34. Kombinatorische Abzählregeln:

1. *Gleichheitsregel:* Wenn zwischen Mengen A und B eine bijektive Abbildung existiert, dann ist $|A| = |B|$.
2. *Summenregel:* Ist A die Vereinigung von k paarweise *disjunkten*, endlichen Mengen A_1, \dots, A_k , dann ist $|A| = |A_1| + \dots + |A_k|$.
3. *Produktregel:* Ist $A = A_1 \times A_2 \times \dots \times A_k$ das kartesische Produkt endlicher Mengen A_1, \dots, A_k , dann gilt $|A| = |A_1| \cdot |A_2| \cdot \dots \cdot |A_k|$.
4. *Zerlegungsregel:* Ist $f : A \rightarrow B$ eine Abbildung, dann gilt:

$$|A| = \sum_{b \in B} |f^{-1}(b)|$$

Beweis. Zu (3): siehe Behauptung 1.25. Zu (4): Sind $b_1, b_2 \in B$ und $b_1 \neq b_2$, so gilt $f^{-1}(b_1) \cap f^{-1}(b_2) = \emptyset$. □

► *Beispiel 1.35:* (Produktregel) Wenn man ein Objekt in k Schritten konstruiert, und man im i -ten Schritt die Wahl zwischen s_i Möglichkeiten hat, dann kann man das Objekt insgesamt auf $s_1 \cdot s_2 \cdot \dots \cdot s_k$ Arten konstruieren. Man kann sich diese Regel sehr schön mit einem Baumdiagramm veranschaulichen. Das folgende Diagramm veranschaulicht die Konstruktion der sechs Permutationen von $\{a, b, c\}$:



1.6.1 Prinzip der doppelten Abzählung

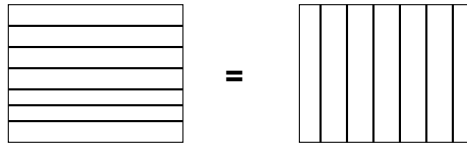
Das sogenannte *Prinzip der doppelten Abzählung* ist die folgende “offensichtliche” Aussage: Wenn wir die Elemente einer Menge in zwei verschiedenen Reihenfolgen abzählen (und dabei jeweils keine Fehler machen), dann werden wir die gleichen Antworten bekommen.

Prinzip der doppelten Abzählung:

Sei A eine Tabelle mit m Zeilen, n Spalten und mit Elementen 0 und 1. Sei z_i die Anzahl der Einsen in der i -ten Zeile, und s_j die Anzahl der Einsen in der j -ten Spalte. Dann gilt

$$\sum_{i=1}^m z_i = \sum_{j=1}^n s_j = \text{Gesamtzahl der Einsen in } A$$

Anschaulich:



► *Beispiel 1.36*: In einer Vorlesung sind 32 der Hörer männlich. Dabei ist jeder Student mit 5 Studentinnen und jede Studentin mit 8 Studenten befreundet.

Frage: Wie viele Studentinnen besuchen die Vorlesung?

Diese Frage lässt sich mit dem Prinzip der doppelten Abzählung beantworten. Sei n die (bis jetzt unbekannte) Anzahl der Studentinnen in der Vorlesung, und betrachte die "Freundschaftstabelle" $A = (a_{ij})$ wobei $a_{ij} = 1$ falls der i -te Student mit der j -ten Studentinnen befreundet ist, und $a_{ij} = 0$ sonst. Die Tabelle hat also $m = 32$ Zeilen und n Spalten. Jede Zeile hat 5 Einseinträge ($z_i = 5$) und jede Spalte hat 8 Einseinträge ($s_j = 8$). Also folgt durch zeilen- und spaltenweises Abzählen

$$32 \cdot 5 = \sum_{i=1}^{32} z_i = \sum_{j=1}^n s_j = 8 \cdot n$$

Also besuchen $n = 20$ Studentinnen die Vorlesung.

Im folgenden Satz ist $d(u) = |\{e \in E : u \in e\}|$ der Grad von Knoten u (=die Anzahl der zu u inzidenten Kanten).

Satz 1.37. (Euler 1736) Sei $G = (V, E)$ ein ungerichteter Graph. Dann gilt

$$\sum_{u \in V} d(u) = 2 \cdot |E|.$$

Beweis. Betrachte die Tabelle M , die aus $n = |V|$ Zeilen und $m = |E|$ Spalten besteht und deren Einträge wie folgt definiert sind:

$$M(u, e) = \begin{cases} 1 & \text{falls } u \in e \\ 0 & \text{sonst.} \end{cases}$$

Dann hat die Zeile zu Knoten u genau $d(u)$ Einsen und die Spalte zu Kante e genau 2 Einsen. Die Behauptung folgt also direkt aus dem Prinzip der doppelten Abzählung. \square

1.6.2 Binomialkoeffizienten

Sei X eine endliche Menge mit $n = |X|$ Elementen und $k \leq n$.

Eine k -Teilmenge von X ist eine Menge $\{x_1, x_2, \dots, x_k\}$ aus k verschiedenen Elementen von X (Ordnung hier ist unwichtig!). Die Anzahl $C(n, k)$ solchen Teilmengen bezeichnet man mit

$$\binom{n}{k}$$

und nennt diese Zahl *binomischer Koeffizient* (oder *Binomialkoeffizient*). Beachte, dass $\binom{n}{k}$ genau die Anzahl der 0-1 Folgen der Länge n mit k Einsen ist; eine Eins bzw. Null in der Position i sagt uns, ob das i -te Element von X gewählt bzw. nicht gewählt wird. Also gilt:

$$\begin{aligned} \binom{n}{k} &= \text{Anzahl der } k\text{-elementigen Teilmengen einer Menge aus } n \text{ Elementen} \\ &= \text{Anzahl der 0-1 Folgen der Länge } n \text{ mit genau } k \text{ Einsen} \end{aligned}$$



Das ist die Definition von $\binom{n}{k}$! Nicht (wie man üblicherweise behauptet) die Gleichung $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Diese Gleichung leitet man aus der Definition ab! (Siehe Satz 1.39.)

Eine k -Permutation von X ist eine geordnete Folge (x_1, x_2, \dots, x_k) aus k verschiedenen Elementen von X . (Ordnung ist hier wichtig!) Die Anzahl $P(n, k)$ solcher Folgen bezeichnet man mit

$$(n)_k$$

Für $k = n$ schreibt man $n!$ statt $(n)_n$ (gesprochen: n Fakultät); man setzt auch $0! = 1$ (nur eine Vereinbarung).

Viele einfache kombinatorische Fragestellungen lassen sich auf gewisse Grundaufgaben zurückzuführen, nämlich aus einer gegebenen endlichen Menge bestimmte Wahlen zu treffen. Je nachdem, ob wir es zulassen, dass dabei ein und dasselbe Element mehrfach ausgewählt wird (ohne oder mit Wiederholungen), und ob wir die Auswahl als geordnet betrachten (1. Element, 2. Element, ...) oder als ungeordnet, ergeben sich vier verschiedene Anzahlen:

Satz 1.38. Sei X eine endliche Menge mit n Elementen und sei k eine natürliche Zahl. Dann wird die Anzahl der Möglichkeiten für die Auswahl von k Elementen aus dieser n -elementigen Menge X gegeben durch:

	ungeordnet	geordnet
ohne Wiederholungen	$\binom{n}{k}$	$(n)_k = \binom{n}{k} \cdot k!$
mit Wiederholungen	$\binom{n+k-1}{k}$	n^k

Beweis. Die Einträge der ersten Zeile in dieser Tabelle entsprechen einfach den Definitionen. Es bleibt also die beide Einträge in der zweiten Zeile zu begründen.

Fall 1: mit Wiederholungen + geordnet. Es gibt n Möglichkeiten das erste Element x_1 auszuwählen. Da wir Wiederholungen von Elementen in der Auswahl erlauben, gibt es danach immer noch n Möglichkeiten das zweite Element x_2 auszuwählen, usw. Nach der Produktregel gibt es also $\widehat{n} \cdot \widehat{n} \cdot \dots \cdot \widehat{n} = n^k$ Möglichkeiten eine geordnete Folge (x_1, x_2, \dots, x_k) aus k Elementen von X auszuwählen.

Fall 2: mit Wiederholungen + ungeordnet. In diesem Fall ist die gesuchte Zahl $S(n, k)$ die Anzahl der ganzzahligen Lösungen (a_1, \dots, a_n) für die Gleichung

$$a_1 + \dots + a_n = k \quad (*)$$

unter der Bedingung, dass alle $a_i \geq 0$ sind: es reicht jedes a_i als die Anzahl der Vorkommen des i -ten Elements von M in der gewählten Teilmenge zu interpretieren. Jede Lösung (a_1, \dots, a_n) der Gleichung $(*)$ entspricht genau einer Folge aus Nullen und Einsen der Länge $k + (n - 1) = n + k - 1$ mit $n - 1$ Einsen:

$$\underbrace{0 \dots 0}_1 1 \underbrace{0 \dots 0}_2 1 \underbrace{0 \dots 0}_3 1 \dots \dots 1 \underbrace{0 \dots 0}_{a_n}$$

Da es genau¹⁶ $\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$ solche Folgen gibt, sind wir fertig. \square

Numerisch lassen sich die Binomialkoeffizienten und Fakultäten wie folgt berechnen:

Satz 1.39. Es gilt:

$$(n)_k = n(n-1) \dots (n-k+1)$$

und

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!} = \frac{n}{1} \cdot \frac{n-1}{2} \cdot \frac{n-2}{3} \dots \frac{n-k+1}{k}.$$

Beweis. Die erste Gleichheit ist einfach: es gibt $|X| = n$ Möglichkeiten das erste Element x_1 auszuwählen; danach gibt es $|X \setminus \{x_1\}| = n - 1$ Möglichkeiten das zweite Element x_2 auszuwählen, usw. Letztendlich gibt es $|X \setminus \{x_1, \dots, x_{k-1}\}| = n - (k - 1) = n - k + 1$ Möglichkeiten das k -te Element x_k auszuwählen. Insgesamt gibt es also $n(n-1) \dots (n-k+1)$ Möglichkeiten eine geordnete Folge (x_1, x_2, \dots, x_k) aus k verschiedenen Elementen von X auszuwählen.

Nun beweisen wir die zweite Gleichung. Wir können alle geordnete Folgen (x_1, x_2, \dots, x_k) aus k verschiedenen Elementen von X wie folgt erzeugen: zuerst wählen wir eine k -Teilmenge $\{x_1, x_2, \dots, x_k\}$ aus k verschiedenen Elementen von X ; hier haben wir genau $\binom{n}{k}$ Möglichkeiten. Wenn die Teilmenge $\{x_1, x_2, \dots, x_k\}$ bereits gewählt ist, bleiben genau $k!$ Möglichkeiten diese Teilmenge zu permutieren. Also haben wir insgesamt $\binom{n}{k} \cdot k!$ geordneten Folgen (x_1, x_2, \dots, x_k) . Da nach Definition diese Zahl gleich $(n)_k$ ist, es folgt

$$\binom{n}{k} \cdot k! = (n)_k$$

oder äquivalent

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!}$$

\square

¹⁶Hier haben wir die Formel $\binom{n}{n-r} = \binom{n}{r}$ benutzt (siehe 1.3).

Es gibt viele nützliche Gleichungen, die die Arbeit mit Binomialkoeffizienten erleichtern. Um solche Gleichungen zu erhalten, reicht es in den meisten Fällen die *kombinatorische* (nicht die *numerische*, wie vom Satz 1.39 gegeben) Natur der Binomialkoeffizienten auszunutzen. Um das zu demonstrieren, beachten wir, dass eine Teilmenge durch sein Komplement *eindeutig* bestimmt ist. Diese einfache Beobachtung liefert uns sofort die Gleichung:

$$\binom{n}{n-k} = \binom{n}{k}. \quad (1.3)$$

In ähnlicher Weise kann man auch andere Gleichungen beweisen.

Satz 1.40. (Pascal'scher Rekurrenzsatz für Binomialkoeffizienten)

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Beweis. Sei X eine Menge mit $n = |X|$ Elementen. Fixiere ein beliebiges $x \in X$. Die Anzahl der k -Teilmengen von X , die das Element x enthalten, ist $\binom{n-1}{k-1}$ und die Anzahl der k -Teilmengen von X , die das Element x vermeiden (d.h. nicht enthalten), ist $\binom{n-1}{k}$. Da es keine anderen k -Teilmengen in X gibt, folgt die Behauptung. \square



Der folgender einfacher (aber sehr nützlicher) Satz ist vom Sir Isaac Newton in ca. 1666 bewiesen worden. Dieser Satz erklärt den Namen: Binomialkoeffizienten sind die Koeffizienten in der Berechnung des "binomischen Ausdrucks" $(x + y)^n$.

Satz 1.41. (Binomischer Lehrsatz) Sei n eine positive ganze Zahl. Dann gilt für alle reelle Zahlen x and y :

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

Beweis. (Kombinatorischer Beweis) Wenn wir die Terme

$$(x + y)^n = \underbrace{(x + y) \cdot (x + y) \cdot \dots \cdot (x + y)}_{n\text{-mal}}$$

ausmultiplizieren, dann gibt es genau $\binom{n}{k}$ Möglichkeiten den Term $x^k y^{n-k}$ zu erhalten. Warum? Wenn wir die Terme multiplizieren, wählen wir aus jedem Term entweder x oder y . Sei $a = (a_1, \dots, a_n)$ die (diese Auswahl) entsprechende 0-1 Folge mit $a_i = 1$ genau dann, wenn aus dem i -ten Term $(x + y)$ die erste Zahl x ausgewählt würde. Sei $N(k)$ die Anzahl der Vorkommen von $x^k y^{n-k}$ in dem Produkt $(x + y)^n$. Dann ist

$$N(k) = \text{Anzahl der 0-1 Folgen der Länge } n \text{ mit genau } k \text{ Einsen} = \binom{n}{k}.$$

\square

Beweis. (Induktiver Beweis) Für $n = 0$ ist die linke Seite gleich $(x + y)^0 = 1$ und die rechte ist auch $\binom{n}{0}x^0y^{0-0} = 1$.

Induktionsschritt: $n \mapsto n + 1$.

$$\begin{aligned}
 (x + y)^{n+1} &= (x + y) \cdot (x + y)^n \\
 &= (x + y) \cdot \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} && \text{Induktionsannahme} \\
 &= x \cdot \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} + y \cdot \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} \\
 &= \sum_{k=0}^n \binom{n}{k} x^{k+1} y^{n-k} + \sum_{k=0}^n \binom{n}{k} x^k y^{n-k+1} \\
 &= \sum_{r=1}^n \binom{n}{r-1} x^r y^{n-r+1} + \sum_{k=0}^n \binom{n}{k} x^k y^{n-k+1} && (r := k + 1) \\
 &= \sum_{r=1}^n \left[\binom{n}{r-1} + \binom{n}{r} \right] x^r y^{n-r+1} + \binom{n}{0} x^0 y^{n+1} + \binom{n}{n} x^{n+1} y^0 \\
 &= \sum_{r=1}^n \binom{n+1}{r} x^r y^{n-r+1} + \binom{n+1}{0} x^0 y^{n+1} + \binom{n+1}{n+1} x^{n+1} y^0 && (\text{Satz 1.40}) \\
 &= \sum_{r=0}^{n+1} \binom{n+1}{r} x^r y^{(n+1)-r}.
 \end{aligned}$$

□

► *Beispiel 1.42* : Mit binomischem Lehrsatz kann man viele nützliche Sachen beweisen. Zum Beispiel so kann man die folgende Eigenschaft der ganzen Zahlen zeigen. Für jede ganze Zahl m und für jede natürliche Zahl $k \geq 1$ gilt:

$$m^k \text{ ist ungerade} \iff m \text{ ist ungerade.}$$

Beweis: Die Richtung \Rightarrow ist trivial: wäre nämlich m gerade d.h. $m = 2j$ für eine ganze Zahl j , so wäre auch $m^k = 2^k(j^k)$ eine gerade Zahl. Um die andere Richtung \Leftarrow zu beweisen, nehmen wir an, dass m ungerade ist, d.h. m hat die Form $m = 2j + 1$ für eine ganze Zahl j . Setze $x = 2j$ und $y = 1$ und wende den binomischen Satz an:

$$m^k = (2j + 1)^k = 1 + (2j)^1 \binom{k}{1} + (2j)^2 \binom{k}{2} + \cdots + (2j)^k \binom{k}{k}.$$

Also hat unswere Zahl m^k die Form 1 plus eine gerade Zahl, und muss deshalb ungerade sein. □

Für wachsenden n und k ist es schwer, den Binomialkoeffizient $\binom{n}{k}$ exakt auszurechnen. Andererseits reicht es für viele Anwendungen, nur die Zuwachsrate (wie schnell oder wie langsam $\binom{n}{k}$ wächst) zu wissen. Oft reicht bereits die folgende Abschätzung:

Lemma 1.43.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} < \left(\frac{en}{k}\right)^k.$$

Beweis. Untere Schranke:

$$\binom{n}{k}^k = \frac{n}{k} \cdot \frac{n}{k} \cdots \frac{n}{k} \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{1} = \binom{n}{k}.$$

Obere Schranke. Nach (1) und binomischem Lehrsatz,

$$e^{nt} > (1+t)^n = \sum_{i=0}^n \binom{n}{i} t^i > \binom{n}{k} t^k.$$

Für $t = k/n$ bekommt man

$$e^k > \binom{n}{k} \left(\frac{k}{n}\right)^k,$$

wie erwünscht. □

Für Fakultäten oft reichen die folgenden triviale Abschätzungen:

$$\left(\frac{n}{2}\right)^{n/2} \leq n! \leq n^n$$

Viel bessere Abschätzung ist durch die berühmte *Stirling-Formel*¹⁷ gegeben:

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \cdot e^{1/12n}. \quad (1.4)$$

Mit dieser Abschätzung von $n!$ kann man die folgende Abschätzung für Binomialkoeffizienten zeigen (Übungsaufgabe ¹⁸) Sei $0 < \alpha < 1$ und sei αn eine ganze Zahl. Dann gilt:

$$\binom{n}{\alpha n} = \frac{1 + o(1)}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot 2^{n \cdot H(\alpha)}, \quad (1.5)$$

wobei

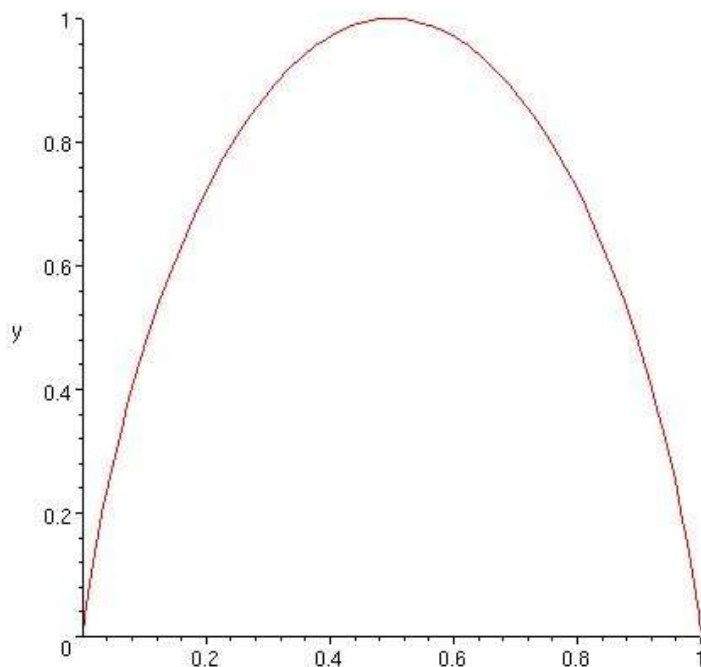
$$H(\alpha) := -(\alpha \log_2 \alpha + (1-\alpha) \log_2(1-\alpha))$$

die sogenannte *binäre Entropie-Funktion* ist. ¹⁹ Graphisch sieht diese Funktion so aus:

¹⁷James Stirling *Methodus Differentialis*, 1730.

¹⁸Hinweis: $H(\alpha) = \log_2 h(\alpha)$ mit $h(\alpha) = \alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}$.

¹⁹Der Term “ $o(1)$ ” hier ist eine funktion, die für wachsendes n gegen 0 strebt. Wir werden diese “klein- o ” Notation in Abschnitt 3.9 genauer betrachten.



1.6.3 Prinzip von Inklusion and Exklusion

Das *Prinzip von Inklusion und Exklusion* (bekannt auch als das *Sieb von Eratosthenes*) ist ein mächtiges kombinatorisches Abzählprinzip.

Für zwei beliebige endlichen Mengen A und B gilt

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Für drei Mengen A, B und C gilt (siehe Abb. 1.12):

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

Im Allgemeinen wollen wir für n gegebene Teilmengen A_1, \dots, A_n von X die Anzahl $|A_1 \cup \dots \cup A_n|$ der Elemente in der Vereinigung bestimmen. Als die erste Approximation können wir die Summe

$$|A_1| + \dots + |A_n| \tag{1.6}$$

nehmen. Jedoch wird diese Zahl im allgemeinen zu groß sein: Wenn $A_i \cap A_j \neq \emptyset$, dann wird jedes Element von $A_i \cap A_j$ in (1.6) zweimal gezählt, einmal in $|A_i|$ und ein zweites Mal in $|A_j|$. Wir können die Situation korrigieren, indem wir aus (1.6) die Summe

$$\sum_{1 \leq i < j \leq n} |A_i \cap A_j| \tag{1.7}$$

subtrahieren. Aber dann wird die Zahl zu klein sein, da jedes Element von $A_i \cap A_j \cap A_k \neq \emptyset$ drei Mal in (1.7) gezählt ist: einmal in $|A_i \cap A_j|$, ein zweites Mal in $|A_j \cap A_k|$, und ein drittes Mal in $|A_i \cap A_k|$. Wir probieren noch mal die Situation zu korrigieren, indem wir die Summe

$$\sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k|,$$

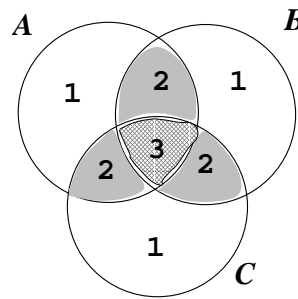


Abbildung 1.12: In $|A| + |B| + |C|$ wird jedes Element aus $A \setminus (B \cup C)$, $B \setminus (A \cup C)$ und $C \setminus (A \cup B)$ nur einmal gezählt, aber jedes Element aus $A \cap B$, $B \cap C$ und $A \cap C$ wird zweimal gezählt; deshalb muss man diese Zahlen abziehen. Aber dann wird jedes Element aus $A \cap B \cap C$ zweimal abgezogen (anstatt einmal); also müssen wir für jedes solches Element noch 1 zu addieren.

dazu addieren. Dann wird aber die Zahl wieder zu groß sein, usw. Dennoch werden wir nach n Schritten bereits die richtige Zahl finden.

Satz 1.44. (Prinzip von Inklusion and Exklusion, Siebformel) Seien A_1, \dots, A_n endliche Teilmengen eines gemeinsamen Universums X . Dann gilt:

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{1 \leq i \leq n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| + \dots \\ \dots + (-1)^{n+1} |A_1 \cap A_2 \cap \dots \cap A_n|$$

Beweis. (mittels Abzählen): Für eine Indexmenge $I \subseteq \{1, \dots, n\}$ und ein Element $x \in X$ sei

$$f_I(x) := \begin{cases} (-1)^{|I|+1} & \text{falls } x \in \bigcap_{i \in I} A_i \\ 0 & \text{sonst.} \end{cases}$$

Dann können wir die rechte Summe so umschreiben:

$$\sum_{I \neq \emptyset} (-1)^{|I|+1} \left| \bigcap_{i \in I} A_i \right| = \sum_{I \neq \emptyset} \sum_{x \in X} f_I(x) \\ = \sum_{x \in X} \sum_{I \neq \emptyset} \widehat{f_I(x)} \quad (\text{doppeltes Zählen}) \\ = \sum_{x \in X} S(x).$$

Es reicht deshalb zu zeigen, dass $S(x) = 1$ für jedes $x \in A_1 \cup \dots \cup A_n$, und $S(x) = 0$ für alle anderen Elemente x gilt. Um das zu zeigen, nehmen wir ein beliebiges (aber festes) $x \in X$.

Fall 1: $x \notin A_1 \cup \dots \cup A_n$. Dann sind alle $f_I(x) = 0$ und damit ist auch $S(x) = 0$.

Fall 2: $x \in A_1 \cup \dots \cup A_n$. Dann ist die Menge $J := \{i : x \in A_i\}$ nicht leer, und $f_I(x) \neq 0$ genau

dann, wenn $I \subseteq J$ und $I \neq \emptyset$ gilt. Sei $m := |J|$. Dann gilt

$$S(x) = \sum_{\emptyset \neq I \subseteq J} f_I(x) = \sum_{\emptyset \neq I \subseteq J} (-1)^{|I|+1} = \sum_{k=1}^m \binom{m}{k} (-1)^{k+1}$$

oder äquivalent

$$-S(x) = \sum_{k=1}^m \binom{m}{k} (-1)^k = \sum_{k=0}^m \binom{m}{k} (-1)^k - \binom{m}{0} \\ = \underbrace{(-1-1)^m}_{=0} - 1 = -1$$

Also haben wir in diesem Fall $S(x) = \binom{m}{0} = 1$, wie erwünscht. \square

► *Beispiel 1.45*: Eine Sekretärin hat n Briefe an n verschiedene Empfänger geschrieben und die Kuverts adressiert. Nun steckt sie die Briefe blindlings in die Kuverts. Wie wahrscheinlich ist es, dass *kein* Brief im richtig adressierten Kuvert steckt?

An der Theatergarderobe gaben n Herrn ihre Hüte ab. Nach der Vorstellung gibt die Garderobefrau die Hüte wahllos und zufällig zurück. Wie wahrscheinlich ist es, dass *kein einziger* Herr seinen eigenen Hut erhält?

Mit dem Prinzip von Inklusion and Exklusion lässt sich zeigen, dass diese Wahrscheinlichkeit überraschen groß ist: sie liegt sehr nah zu $e^{-1} = 0.3678\dots$

Das Problem lässt sich wie folgt formalisieren. Die Menge $[n] = \{1, 2, 3, 4, \dots, n\}$ wird bijektiv auf sich selbst abgebildet. Wie wahrscheinlich ist eine fixpunktfreie Permutation von $[n]$? Eine Permutation $f : [n] \rightarrow [n]$ ist *fixpunktfrei*, wenn $f(x) \neq x$ für alle $x \in [n]$.

Behauptung: Die Anzahl der fixpunktfreien Permutationen von $\{1, \dots, n\}$ ist gleich

$$\sum_{i=0}^n (-1)^i \binom{n}{i} (n-i)! = n! \sum_{i=0}^n \frac{(-1)^i}{i!}.$$

Die Summe $\sum_{i=0}^n \frac{(-1)^i}{i!}$ ist der Anfangsterm der Taylor-Reihe²⁰ für e^{-1} . Die Anzahl der fixpunktfreien Permutationen stimmt also asymptotisch mit $n!/e$ überein.

Beweis. Sei X die Menge aller $n!$ Permutationen $f : [n] \rightarrow [n]$, und sei $A_x \subseteq X$ die Menge aller Permutationen f für die gilt: $f(x) = x$. Dann ist $|A_x| = (n-1)!$, und allgemeiner, $|\bigcap_{x \in I} A_x| = (n-|I|)!$, da die Permutationen in $\bigcap_{x \in I} A_x$ *alle* Punkte $x \in I$ fixieren müssen, und den Rest permutieren müssen. Eine Permutation ist also fixpunktfrei genau dann, wenn sie in *keiner* der Mengen A_1, \dots, A_n liegt. Nach dem Prinzip von Inklusion and Exklusion gilt:

$$|X \setminus (A_1 \cup \dots \cup A_n)| = \sum_{I \subseteq \{1, \dots, n\}} (-1)^{|I|} (n-|I|)! = \sum_{i=0}^n (-1)^i \binom{n}{i} (n-i)!$$

\square

²⁰Mehr über Taylorentwicklung von Funktionen kann man in Abschnitt 3.7 finden.

Eine direkte (aber oft sehr nützliche) Folgerung aus dem Prinzip von Inklusion and Exklusion ist die folgende Abschätzung.

Korollar 1.46. (Boole-Ungleichungen) Seien A_1, \dots, A_n endliche Teilmengen. Dann gilt:

$$\sum_{1 \leq i \leq n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| \leq \left| \bigcup_{i=1}^n A_i \right| \leq \sum_{1 \leq i \leq n} |A_i|$$

1.7 Aufgaben

1.1. Vereinfache folgenden Ausdrücke durch Betrachten eines Mengendiagramms:

- (a) $A \cup (A \setminus B)$, (b) $A \cap (A \setminus B)$, (c) $A \setminus (A \cup B)$, (d) $B \cup (A \setminus B)$, (e) $A \setminus (B \setminus A)$,
 (f) $A \setminus (A \setminus B)$.

1.2. Zeichne Mengendiagramme für folgenden Mengen: (a) $A \cap \overline{B}$, (b) $\overline{A} \cup \overline{B}$, (c) $\overline{A} \cap \overline{B}$, (d) $\overline{A} \cup B$.

1.3. Stelle fest, welche der folgenden Relationen R auf der Menge der Menschen reflexiv, symmetrisch, antisymmetrisch, asymmetrisch oder transitiv sind, wobei $(a, b) \in R$ genau dann, wenn

1. a ist größer als b
2. a und b wurden am selben Tag geboren
3. a hat denselben Vornamen wie b
4. a und b haben eine gemeinsame Großmutter.

Stelle fest, ob die folgenden Relationen R auf der Menge der ganzen Zahlen reflexiv, symmetrisch, antisymmetrisch, asymmetrisch oder transitiv sind, wobei $(x, y) \in R$ genau dann, wenn

- (a) $x \neq y$
- (b) $xy \geq 1$
- (c) $x = y + 1$ oder $x = y - 1$
- (d) $x \equiv y \pmod{7}$ (geteilt durch 7 ergeben beide Zahlen den selben Rest)
- (e) x ist ein Vielfaches von y
- (f) x und y sind beide negativ oder beide nicht-negativ
- (g) $x = y^2$
- (h) $x \geq y^2$

1.4. Seien $f : A \rightarrow B$ und $g : B \rightarrow C$ Abbildungen. Man zeige

- (a) Sind f und g injektiv, so auch $f \circ g$.
- (b) Sind f und g surjektiv, so auch $f \circ g$.
- (b) Sind f und g bijektiv, so auch $f \circ g$.

1.5. Für zwei ganzen Zahlen x und y schreibt man $x|y$ genau dann, wenn x die Zahl y ohne Rest teilt, d.h. wenn es ein $z \in \mathbb{Z}$ mit $y = x \cdot z$ gibt. Seien R und S die folgenden beiden Relationen über \mathbb{Z} :

$$R = \{(x, y) \in \mathbb{Z}^2 : x|y\} \quad \text{und} \quad S = \{(y, z) \in \mathbb{Z}^2 : 2|(y+z)\}.$$

Zeige, dass dann

$$R \circ S = \{(x, z) \in \mathbb{Z}^2 : x \text{ ist ungerade oder } z \text{ ist gerade}\}.$$

1.6. Sei A eine beliebige Menge, und seien R und S Relationen über A . Zeige, dass dann:

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}.$$

1.7. Seien R und S zwei Äquivalenzrelationen über A . Zeige, dass $R \circ S$ genau dann eine Äquivalenzrelation über A ist, wenn $R \circ S = S \circ R$ gilt.

1.8. Interessanterweise sind die Eigenschaften injektiv, surjektiv und bijektiv bei Abbildungen endlicher Mengen aus kombinatorischen Gründen gleichwertig. Sei A eine endliche Menge und $f : A \rightarrow A$ eine Abbildung. Zeige, dass die folgenden Aussagen äquivalent sind:

- (1) f ist surjektiv.
- (2) f ist injektiv.
- (3) f ist bijektiv.

Hinweis: $|A| = \sum_{a \in A} |f^{-1}(a)|$.

1.9. Sei $\mathcal{E}(\mathbb{N}) = \{X : X \subseteq \mathbb{N}, X \text{ endlich}\}$ die Menge aller endlichen Teilmengen von \mathbb{N} . Wir definieren $f : \mathcal{E} \rightarrow \mathbb{N}$ durch $f(\emptyset) = 0$ und $f(X) = \sum_{x \in X} 2^x$ für $\emptyset \neq X \in \mathcal{E}$. Zeige, dass f eine Bijektion von $\mathcal{E}(\mathbb{N})$ auf \mathbb{N} ist. *Hinweis:* Warum ist die Binärdarstellung einer natürlichen Zahl eindeutig?

1.10. Zeige, dass der n -dimensionale Würfel Q_n bipartit ist. *Hinweis:* Sei U die Menge aller Strings, die eine ungerade Anzahl von Einsen haben.

1.11. Von den folgenden drei Aussagen ist genau eine richtig:

- (a) Fritz hat über tausend Bücher.
- (b) Fritz hat weniger als tausend Bücher.
- (c) Fritz hat mindestens ein Büch.

Wieviele Bücher hat Fritz?

1.12. Auf einem Bauernhof in der Nähe von Frankfurt gibt es sowohl gefräßige als auch nicht gefräßige Schweine. Es ist bekannt, dass jedes alte Schwein gefräßig ist und dass jedes gesunde Schwein gefräßig ist.

Welche der nachstehenden Folgerungen sind dann zulässig?

- (a) Es gibt sowohl alte als auch junge Schweine auf dem Hof.
- (b) Es gibt junge Schweine auf dem Hof.
- (c) Alle nicht gefräßigen Schweine sind jung.
- (d) Einige junge Schweine sind krank.
- (e) Alle junge Schweine sind krank.

1.13.

- (a) Gilt die Aussage: $A \subseteq B \Rightarrow A \cap B = A$?
- (b) Beweise: $A \cap B = A \setminus (A \setminus B)$.
- (c) Das kartesische Produkt $A \times A$ einer Menge A mit sich selbst lässt sich formal auch mit Hilfe der Potenzmenge $\mathcal{P}(A)$ definieren. Seien a, b Elemente von A . Dann soll das geordnete Paar (a, b) definiert sein als

die Teilmenge von $\mathcal{P}(A)$, die genau die Mengen $\{a\}$ und $\{a, b\}$ enthält, also kurz geschrieben:

$$(a, b) := \left\{ \{a\}, \{a, b\} \right\}$$

Seien $a, b, c, d \in A$. Zeige, dass folgende Äquivalenz gilt:

$$(a, b) = (c, d) \iff a = c \text{ und } b = d$$

1.14. Für $f : A \rightarrow B$ und $g : B \rightarrow A$ sei die Komposition $(g \circ f)(x) = g(f(x))$ die identische Abbildung, d.h. $(g \circ f)(x) = x$ für alle $x \in A$. Zeige, dass dann f injektiv und g surjektiv sind. Gebe ein Beispiel an, in dem f nicht surjektiv und g nicht injektiv ist.

1.15. Wir betrachten die Abbildung $f : A \rightarrow B$. Seien $U, V \subseteq B$. Zeige: Es gilt

$$f^{-1}(B \setminus U) = A \setminus f^{-1}(U)$$

und

$$f^{-1}(U \cap V) = f^{-1}(U) \cap f^{-1}(V).$$

Hinweis: Um die Gleichheit zweier Mengen X und Y zu beweisen, muss man folgendes zeigen: Für jedes $x \in X$ gilt $x \in Y$ und für jedes $y \in Y$ gilt $y \in X$.

1.16. Wir betrachten die Abbildung $f : X \rightarrow Y$. Seien $A, B \subseteq X$ bzw. $U, V \subseteq Y$.

(a) Zeige: Es ist $f(A \cap B) \subseteq (f(A) \cap f(B))$. Gilt die umgekehrte Richtung?

(b) Zeige: Es gilt $f^{-1}(f(A)) \supseteq A$ und $f^{-1}(f^{-1}(U)) \supseteq U$. Zeige, dass i.A. keine Gleichheit gilt.

1.17.

(a) Zeige die logische Äquivalenz von $A \leftrightarrow B$ und $(A \wedge B) \vee (\neg A \wedge \neg B)$.

(b) Zeige, dass die folgenden Formelpaare *nicht* äquivalent sind:

$$\begin{aligned} (\forall x : P(x)) \vee (\forall x : Q(x)) & \text{ mit } \forall x : P(x) \vee Q(x) \\ (\exists x : P(x)) \wedge (\exists x : Q(x)) & \text{ mit } \exists x : P(x) \wedge Q(x) \end{aligned}$$

Hinweis: Um zu zeigen, dass zwei Formelpaare A und B nicht äquivalent sind, reicht es ein Gegenbeispiel anzugeben. D.h. es reicht ein Universum M und die Prädikate $P(x)$ und $Q(x)$ auf M anzugeben, für die die Äquivalenz $A \iff B$ nicht gilt.

1.18. Gebe einen Widerspruchs-Beweis für die folgende Aussage an. Für jede natürliche Zahl t und für jede natürliche Zahl n gilt:

Wenn $t \geq 2$ und t ein Teiler von n ist, dann ist t kein Teiler von $n + 1$.

1.19. Eine Schnecke kriecht tagsüber an einer Mauer nach oben, rutscht aber nachts um die Hälfte der erreichten Höhe wieder nach unten. Bezeichnen wir mit h_n die am n -ten Abend erreichte Höhe, so ist $h_{n+1} = \frac{1}{2}h_n + 1$. Zeige, dass

$$h_n = 2 - \frac{1}{2^{n-1}}$$

für alle $n = 1, 2, \dots$ gilt.

1.20. Zeige, dass eine Menge von n Elementen 2^n Teilmengen besitzt. *Hinweis:* Induktion über n . Für ein festes Element a und jede Teilmenge S entweder $a \in S$ oder $a \notin S$ gilt.

1.21. Zeige, dass die Summe $1 + 3 + 5 + \dots + (2n - 1)$ der ersten n ungeraden natürlichen Zahlen gleich n^2 ist. *Hinweis:* Induktion.

1.22. Beweise die folgenden Aussagen durch Induktion:

1. Für jede natürliche Zahl n ist sind $n^3 + 2n$ und $n^3 - n$ durch 3 teilbar.
2. Für jede natürliche Zahl n gilt $\sum_{k=0}^n k^3 = \frac{n^2(n+1)^2}{4}$.
3. Für jede natürliche Zahl $n \geq 1$ gilt $\sum_{k=1}^n \frac{1}{\sqrt{k}} > 2(\sqrt{n+1} - 1)$. *Hinweis:* Nach Anwendung der Induktionsvoraussetzung reduziert sich das Problem auf eine Ungleichung, deren Gültigkeit man durch Quadrieren nachweisen kann.
4. Für jede natürliche Zahl n gilt: $\sum_{i=0}^n 2^i = 2^{n+1} - 1$.
5. Die Folge der so genannten *harmonischen Zahlen* H_1, H_2, H_3, \dots ist definiert durch

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}.$$

Zeige, dass für jede natürliche Zahlen n gilt:

$$H_{2^n} \geq 1 + \frac{n}{2}.$$

6. Für alle natürlichen Zahlen $n, k \geq 1$ gilt: $1 + k/n \leq (1 + 1/n)^k$.

1.23. Ein *Polygon* ist eine geschlossene zusammenhängende Folge von Strecken, wobei keine zwei Strecken auf einer Gerade liegen. Ein Polygon wird *einfach* genannt, wenn sich keine Strecken schneiden. Die Treffpunkte der Strecken heißen *Ecken*. Ein einfaches Polygon mit n Ecken wird auch als ein *n-Eck* genannt. D.h. ein einfaches Polygon ist ein von einem geschlossenen, sich nicht schneidenden, Streckenzug begrenztes, ebenes geometrisches Objekt. Eine *Diagonale* ist eine Strecke, die zwei nicht benachbarte Ecken verbindet. Ein einfaches Polygon S heißt *konvex*, wenn für alle nicht benachbarte Ecken $a, b \in S$ gilt, dass alle Punkte der Diagonale zwischen a und b ebenfalls in S liegen.

Zeige folgendes: Die Zahl der Diagonalen in *jedem* ebenen, *konvexen* n -Eck mit $n \geq 3$ ist gleich

$$D(n) = \frac{n(n-3)}{2}$$

Hinweis: Induktion über die Anzahl der Ecken könnte hilfreich sein.

1.24. Wir betrachten das folgende Maximierungsproblem. Gegeben sind eine endliche Menge A und eine Abbildung $f : A \rightarrow A$. Das Ziel ist, eine größtmögliche Teilmenge $S \subseteq A$ zu finden, so dass die Einschränkung f_S von f auf S eine Bijektion ist, d.h. es muss folgendes gelten:

- (a) $f(S) \subseteq S$,
- (b) für alle $x \in S$ gibt es genau ein $y \in S$ mit $f(x) = y$.

Zeige, wie man dieses Problem mit Hilfe der Induktion effizient lösen kann. *Hinweis:* Reduziere das Problem für Mengen A mit n Elementen auf dasselbe Problem für Mengen mit $n - 1$ Elementen. Dafür entferne aus A ein Element x_0 mit $f^{-1}(x_0) = \emptyset$, falls es ein solches Element gibt. Was passiert, wenn es kein solches Element x_0 gibt?

1.25. (Das "Berühmtheits-Problem") In einer Party nehmen n Personen teil. Eine *Berühmtheit* ist eine Person X , die keine andere der $n - 1$ Teilnehmer kennt, aber die allen anderen Personen bekannt ist. Sie kommen in eine solche Party und wollen wissen, ob es eine Berühmtheit gibt und, falls ja, diese Berühmtheit auch herausfinden. Sie können nur die Fragen stellen, ob eine Person eine andere Person kennt oder nicht. Insgesamt gibt es also $n(n - 1)/2$ mögliche Fragen.

Zeige, wie man dieses Problem mit nur $3(n - 1)$ Fragen lösen kann. *Hinweis:* Benutze Induktion, d.h. reduziere das Problem für n Personen auf ein Problem für weniger Personen.

1.26. Ein *Geradenarrangement in allgemeiner Lage* (engl. lines in general position) ist ein Arrangement aus endlich vielen, paarweise nicht parallelen Geraden in der Ebene, von denen keine drei einen Punkt gemeinsam

haben. Dabei habe eine Gerade kein Ende und keinen Anfang. Ein solches Arrangement unterteilt die Ebene in verschiedene Flächen.

Zeige: Jedes Geradenarrangement in allgemeiner Lage mit mindestens 3 Geraden besitzt stets ein Dreieck als Fläche.

1.27. Zeige mit Hilfe des binomischen Lehrsatzes, dass

(a) der Wert $(1 - \sqrt{5})^n + (1 + \sqrt{5})^n$ für jedes $n \in \mathbb{N}$ ganzzahlig ist.

(b) der Wert $(\sqrt{2} + \sqrt{3})^n + (\sqrt{2} - \sqrt{3})^n$ für jedes gerade $n \in \mathbb{N}$ ganzzahlig ist.

1.28. Wie viele Möglichkeiten gibt es, k Kugeln in n ($n \geq k$) Kisten zu verteilen, so dass keine Kiste mehr als eine Kugel enthält?

1.29. Zeige, dass $\binom{n}{k+1} = \binom{n}{k} \frac{n-k}{k+1}$ gilt.

1.30. Zeige, dass für jedes k das Produkt von k aufeinanderfolgenden natürlichen Zahlen durch $k!$ teilbar ist. *Hinweis:* Betrachte $\binom{n+k}{k}$.

1.31. Zeige die folgende Rekursionsgleichung:

$$\binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1}.$$

Hinweis: Benutze das Prinzip der doppelten Abzählung um $k \cdot \binom{n}{k} = n \cdot \binom{n-1}{k-1}$ zu zeigen. Dafür zähle, wie viele Paare (x, M) es gibt, wobei M eine k -elementige Teilmenge von $\{1, \dots, n\}$ ist und $x \in M$.

1.32. Seien $0 \leq l \leq k \leq n$. Zeige, dass

$$\binom{n}{k} \binom{k}{l} = \binom{n}{l} \binom{n-l}{k-l}.$$

Hinweis: Benutze das Prinzip der doppelten Abzählung, um die Anzahl aller Paare (L, K) von Teilmengen aus $\{1, \dots, n\}$ mit $L \subseteq K$, $|L| = l$ und $|K| = k$ zu bestimmen.

1.33. Beweise die Gleichung: $\sum_{i=0}^n \binom{n}{i}^2 = \binom{2n}{n}$. *Hinweis:* $\binom{n}{i}^2 = \binom{n}{i} \binom{n}{n-i}$. Zeige, dass es genau $\sum_{i=0}^n \binom{n}{i} \binom{n}{n-i}$ Möglichkeiten gibt, n -elementige Teilmenge aus $\{1, 2, \dots, 2n\}$ auszuwählen. Alternativ: Wende den binomischen Lehrsatz auf $(a+x)^{2n}$ an.

1.34. Berechne die Anzahl $W(n, k)$ der kürzesten Wege im 2-dimensionalen Gitter vom Punkt $A = (0, 0)$ zum Punkt $B = (n, k)$. *Hinweis:* Jeden Weg kann man als eine Folge der Buchstaben h (für "hoch") und r (für "rechts") beschreiben.

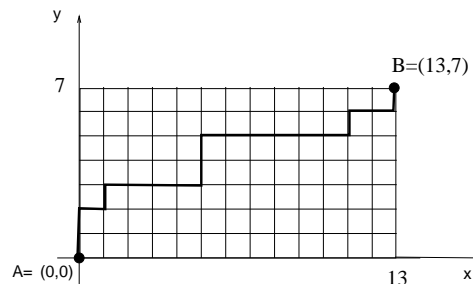


Abbildung 1.13: Ein kürzester Weg von A nach B

1.35. Gebe eine geschlossene Form für die Summe $\sum_{i=0}^n \binom{n}{i} 2^i$ an.

1.36. Sei K das durchschnittliche Kapital (in Euro) eines deutschen Bürgers. Betrachte eine Person als "reich", falls sie mindestens $K/2$ Euro besitzt. Zeige, dass die reichen Personen mindestens die Hälfte aller Gelder haben.

1.37. Ein Affe wurde erfolgreich darauf dressiert, Klötzchen aus einer Urne (ohne Zurücklegen) zu ziehen und sie in einer Reihe von links nach rechts aufzustellen.

In die Urne wurden nun sechs Klötzchen gelegt, wovon drei mit dem Buchstaben A, zwei mit dem Buchstaben N und eins mit dem Buchstaben B markiert wurden.

Wie groß ist die Wahrscheinlichkeit dafür, dass die vom Affen aufgestellte Reihe die Buchstabenfolge „BANANA“ ergibt?

Hinweis: Der Affe kann nicht lesen. Da der Affe die Klötzchen rein zufällig zieht, ist die Wahrscheinlichkeit gleich:

$$\frac{\text{Anzahl der günstigen Versuche}}{\text{Anzahl aller Versuche}}$$

1.38. Wir haben eine Menge X der zu erledigenden Jobs, die wir auf m Prozessoren beliebig verteilen: $X = X_1 \cup X_2 \cup \dots \cup X_m$, wobei X_i die Menge der dem Prozessor i zugewiesenen Jobs ist. Sei

$$L := \frac{1}{m} \sum_{i=1}^m |X_i|$$

die durchschnittliche Belastung eines Prozessors. Wir sagen, dass ein Prozessor i *belastet* ist, falls $|X_i| \geq L/2$ gilt. Zeige, dass (egal wie wir die Jobs verteilen!) mindestens die Hälfte aller Jobs von belasteten Prozessoren ausgeführt wird.

1.39. Zeige, dass in jedem endlichen ungerichteten Graphen mindestens zwei Knoten denselben Grad haben müssen. *Hinweis:* Taubenschlagprinzip.

1.40. Sei $S \subset \{1, 2, \dots, 2n\}$ mit $|S| = n + 1$. Zeige:

- Es gibt zwei Zahlen a, b in S , so dass $b = a + 1$. *Hinweis:* Betrachte die Taubenlöcher $(i, i + 1)$.
- Es gibt zwei Zahlen a, b in S , so dass $a + b = 2n + 1$. *Hinweis:* Betrachte die Taubenlöcher $(i, 2n - i + 1)$, $i = 1, 2, \dots, n + 1$.
- Es gibt zwei Zahlen $a \neq b$ in S , so dass a ein Teiler von b ist. *Hinweis:* Jede Zahl $x \in \{1, 2, \dots, 2n\}$ lässt sich als Produkt $x = k(x) \cdot 2^a$ eindeutig darstellen, wobei $k(x)$ eine ungerade Zahl zwischen 1 und $2n - 1$ ist. Betrachte nun die Zahlen $x \in S$ als Tauben und nimm die Taubenlöcher $1, 3, 5, \dots, 2n - 1$. Setze die Taube x ins Taubenloch $k(x)$. Zeige, dass dann mindestens ein Taubenloch zwei Tauben (Zahlen) $x < y$ enthalten muss.

1.41. Ein *Matching* ist eine Menge paarweise disjunkter Kanten. Ein *Stern* ist eine Menge von Kanten, die einen Knoten gemeinsam haben. Zeige: Jeder ungerichtete Graph mit $2(k - 1)^2 + 1$ Kanten enthält ein Matching oder einen Stern mit k Kanten.

1.42. Ein *vollständiger binärer Baum* ist ein binärer Baum bei dem alle Blätter die gleiche Tiefe haben. Offenbar sind vollständige binäre Bäume besonders kompakt gebaute Bäume.

Zeige: jeder vollständiger binärer Baum der Tiefe d besitzt insgesamt $2^{d+1} - 1$ Knoten.

Kapitel 2

Algebra und Elementare Zahlentheorie

Contents

2.1	Division mit Rest	58
2.2	Euklidischer Algorithmus	64
2.3	Primzahlen	65
2.4	Kleiner Satz von Fermat	67
2.4.1	Anwendung in der Kryptographie: RSA-Codes*	68
2.5	Chinesischer Restsatz	72
2.5.1	Anwendung: Schneller Gleichheitstest*	74
2.6	Gruppen	75
2.6.1	Zyklische Gruppen	78
2.7	Ringe und Körper	80
2.7.1	Polynomring	81
2.7.2	Komplexe Zahlen*	84
2.8	Allgemeine Vektorräume	85
2.9	Aufgaben	86

In der Algebra geht es um Rechnen mit Zahlen und Verallgemeinerungen von Zahlen. Wir wissen aus der Schule, wie man in manchen *unendlichen* Mengen, wie \mathbb{N} , \mathbb{Q} oder \mathbb{R} , rechnen kann. Jeder (klassischer) Computer kann aber nur in einer *endlichen* Menge der Zahlen

$$\mathbb{Z}_n = \{0, 1, \dots, n - 1\}$$

rechnen, wobei n durch den Speicherkapazität beschränkt ist. Wie kann er nun in einer solchen Menge vernünftig addieren/subtrahieren oder multiplizieren/dividieren? Die Antwort ist einfach: Rechne *zyklisch*! Im täglichen Leben rechnen wir ständig “zyklisch”, besonders offensichtlich in der Zeitrechnung. Addieren und subtrahieren kann man hier sehr einfach. Es ist 9 Uhr. Wie spät ist es 5 Stunden später? Wie spät war es vor 3 Stunden? Die Frage also ist, wie soll man zyklisch multiplizieren und dividieren? Dies ist die wichtigste Frage der sogenannten *modularen Arithmetik*, und wir werden in diesem Kapitel diese Frage beantworten.

2.1 Division mit Rest

In diesem Abschnitt betrachten wir nur die ganze Zahlen.

Sind $a \in \mathbb{Z}$ und $n \in \mathbb{N}_+$, so kann man a durch n teilen (durch mehrfaches Abziehen von n aus a) bis ein Rest $0 \leq r < n$ bleibt. Diesen Rest r bezeichnet man mit

$$a \bmod n$$

und nennt den *Rest von a modulo n* . D.h.¹

$$a \bmod n = a - n \cdot \left\lfloor \frac{a}{n} \right\rfloor$$

Die Zahl n heißt *Teiler* von a , in Zeichen $n \mid a$, wenn es ein $q \in \mathbb{Z}$ mit $a = qn$ gibt. D.h.

$$n \mid a \iff \exists q : a = qn \iff a \bmod n = 0.$$

Lemma 2.1. Teilbarkeits-Regeln:

1. Aus $a \mid b$ folgt $a \mid bc$ für alle c .
2. Aus $a \mid b$ und $b \mid c$ folgt $a \mid c$ (Transitivität).
3. Aus $a \mid b$ und $a \mid c$ folgt $a \mid sb + tb$ für alle s und t .
4. Ist $c \neq 0$, so gilt $a \mid b \iff ac \mid bc$.

Beweis. Wir beweisen nur (2) (alle andere Behauptungen kann man analog zeigen). Aus $a \mid b$ und $b \mid c$ folgt $b = sa$ und $c = tb$ für bestimmte s und t . Daraus $c = tb = t(sa) = (ts)a$ und damit auch $a \mid c$ folgt. \square

Zwei ganze Zahlen a und b , die den gleichen Rest bezüglich n haben, werden *kongruent modulo n* genannt. Man schreibe

$$a \equiv b \pmod{n}.$$



In modularer Arithmetik muss man klar zwischen Bezeichnungen " $a \equiv b \pmod{n}$ " und " $a = b \pmod{n}$ " unterscheiden. Die erste Bezeichnung sagt, dass die Reste $a \bmod n$ und $b \bmod n$ gleich sind, während die zweite sagt, dass a der Rest von b modulo n ist (und damit auch zwangsweise $0 \leq a \leq n - 1$ gelten muss):

$$\begin{aligned} a \equiv b \pmod{n} &\iff \text{geteilt durch } n, \text{ beide } a \text{ und } b \text{ denselben Rest besitzen} \\ a = b \pmod{n} &\iff a \text{ ist der Rest von } b \text{ geteilt durch } n \end{aligned}$$

Die folgende Eigenschaft der Kongruenzen werden werden wir oft (ohne das explizit zu erwähnen) benutzen.

Lemma 2.2. Für $a, b \in \mathbb{Z}$ gilt genau dann $a \equiv b \pmod{n}$, wenn $n \mid (a - b)$.

¹Zur Erinnerung: $\lfloor x \rfloor$ ist die größte ganze Zahl, die nicht größer als x ist. So ist z.B. $\lfloor 3/2 \rfloor = 1$ und $\lfloor -3/2 \rfloor = -2$.

Beweis. Wir schreiben $a = q_1n + r_1$ und $b = q_2n + r_2$ mit $0 \leq r_1, r_2 < n$. Das Lemma sagt nun: $r_1 = r_2 \iff n \mid (q_1 - q_2)n + (r_1 - r_2)$. Die Richtung \Rightarrow ist völlig klar. Die andere Richtung \Leftarrow gilt, weil $|r_1 - r_2| < n$ gilt. \square

Lemma 2.3. Seien $x \equiv y \pmod{n}$ und $a \equiv b \pmod{n}$. Dann gilt:

1. $x + a \equiv y + b \pmod{n}$.
2. $x - a \equiv y - b \pmod{n}$.
3. $xa \equiv yb \pmod{n}$.
4. $x^d \equiv y^d \pmod{n}$.

Beweis. Übungsaufgabe. \square

Man fasst alle ganzen Zahlen, die bei Division durch m denselben Rest haben, d.h. die paarweise kongruent modulo m sind, zu einer sogenannten *Restklasse* zusammen.

Ist $m > 0$, so heißt für alle $a \in \mathbb{Z}$

$$[a]_m := \{n \in \mathbb{Z} : n \equiv a \pmod{m}\}$$

die *Restklasse* von a modulo m . Die Menge aller Restklassen bezeichnet man mit

$$\mathbb{Z}/\mathbb{Z}_m = \{[a]_m : a \in \mathbb{Z}\}.$$

Zum Beispiel für $m = 5$ bildet jede der fünf Zeilen eine Restklasse modulo 5:

...	-10	-5	0	5	10	15	20	...
...	-9	-4	1	6	11	16	21	...
...	-8	-3	2	7	12	17	22	...
...	-7	-2	3	8	13	18	23	...
...	-6	-1	4	9	14	19	24	...

Jede Restklasse $[a]_m$ enthält genau eine Zahl $r \in \{0, 1, \dots, m-1\}$ mit $[r]_m = [a]_m$ und diese Zahl nennt man den *Repräsentanten* dieser Klasse. Die Menge dieser Repräsentanten bezeichnet man mit \mathbb{Z}_m , d.h.

$$\mathbb{Z}_m := \{0, 1, \dots, m-1\}.$$

In \mathbb{Z}_m kann man genauso wie in \mathbb{Z} addieren und multiplizieren:

$$\begin{aligned} a + b &:= (a + b) \pmod{m} \\ a \cdot b &:= (a \cdot b) \pmod{m} \end{aligned}$$

D.h.

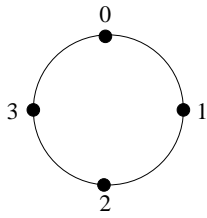
$$\begin{aligned} a + b \text{ in } \mathbb{Z}_m &= \text{Rest von } a + b \text{ geteilt durch } m \text{ in } \mathbb{Z} \\ a \cdot b \text{ in } \mathbb{Z}_m &= \text{Rest von } a \cdot b \text{ geteilt durch } m \text{ in } \mathbb{Z} \end{aligned}$$

Die zwei anderen Operationen—die Differenz und die Division—sind in allen algebraischen Strukturen durch die Addition und die Multiplikation definiert:

$$\begin{aligned}x - a &:= x + y \text{ wobei } y \text{ die Lösung von } y + a = 0 \text{ ist} \\x/a &:= x \cdot z \text{ wobei } z \text{ die Lösung von } z \cdot a = 1 \text{ ist.}\end{aligned}$$

Die Zahlen y und z nennt man dann entsprechend die *additive* und *multiplikative* Inversen von a und sind durch $y = -a$ und $z = a^{-1}$ bezeichnet. In der Struktur $(\mathbb{Z}_m, +, \cdot)$ müssen also diese Inversen die folgenden Gleichungen erfüllen:

$$\begin{aligned}a + (-a) &\equiv 0 \pmod{m} \\a \cdot a^{-1} &\equiv 1 \pmod{m}.\end{aligned}$$



Die Menge \mathbb{Z}_m kann man als einen Kreis vorstellen. Dann sind die Operationen $a + b$ und $a - b$ besonders einfach auszuführen. Will man $a + b$ (bzw. $a - b$) berechnen, so startet man im Punkt a und läuft den Kreis b Schritte vorwärts (bzw. rückwärts). So bekommt man z.B. in \mathbb{Z}_4 , dass $1 + 3 = 0$ und $1 - 3 = 2$ gilt.

Aber vorsichtig: Dividieren in \mathbb{Z}_m kann man ohne weiteres nicht! Für die Division muss folgendes gelten:

$$x/a = b \iff x = ab.$$

D.h. x/a bezeichnet die *einzig*(!) Zahl b , für die die Gleichheit $x = ab$ gilt. Insbesondere muss $0/a = 0$ für alle $a \neq 0$ gelten. Diese Eigenschaft ist bereits in \mathbb{Z}_4 verletzt: Hier gilt $0 = 2 \cdot 2$ und damit auch $0/2 = 2 \neq 0$.

Um durch $a \in \mathbb{Z}_m$ dividieren zu können, muss also \mathbb{Z}_m die multiplikative Inverse enthalten, d.h. es muss eine Zahl $x = a^{-1} \in \{1, 2, \dots, m-1\}$ mit der Eigenschaft $ax \equiv 1 \pmod{m}$ geben. Dann ist die Division x/a von x durch a modulo m nichts anderes als die *Multiplikation* von x mit a^{-1} . D.h.

$$x/a := xa^{-1}$$

Deshalb kann man in \mathbb{Z}_m nur durch solche Zahlen $a \in \mathbb{Z}_m \setminus \{0\}$ dividieren, die eine multiplikative Inverse a^{-1} modulo m besitzen.

► *Beispiel 2.4*: Wir betrachten die Strukturen $(\mathbb{Z}_m, +, \cdot)$ für $m = 2, 3, 4, 5$, wobei die Operationen $+$ und \cdot modulo m definiert sind.

1. $\mathbb{Z}_2 = \{0, 1\}$

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

1 hat die Inverse \Rightarrow division möglich!

2. $\mathbb{Z}_3 = \{0, 1, 2\}$

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

·	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

1 und 2 sind die Inversen von sich selbst \Rightarrow division möglich!

3. $\mathbb{Z}_4 = \{0, 1, 2, 3\}$

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

·	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

$a = 2$ hat *keine* Inverse $a^{-1} \Rightarrow$ Division durch 2 unmöglich!

4. $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Alle $a \in \mathbb{Z}_5$ haben Inversen \Rightarrow division möglich! Was ist z.B. die multiplikative Inverse von 2? Antwort: Wenn wir die Zeile zu 2 in der Multiplikationstabelle anschauen, finden wir eine 1 in der Spalte zu 3. Also ist $2 \cdot 3 \equiv 1 \pmod{5}$ und damit ist 3 die multiplikative Inverse von 2 modulo 5. So ist

Division durch 2 = Multiplikation mit 3

Division durch 3 = Multiplikation mit 2

Division durch 4 = Multiplikation mit 4

5. Sei $a = 8$ und $m = 15$. Dann ist $2a = 16 \equiv 1 \pmod{15}$. Also ist $x = 2$ eine multiplikative Inverse von 8 modulo m .

6. Sei $a = 12$ und $m = 15$. Dann ist die Folge $(ax \pmod{m} : x = 0, 1, 2, \dots)$ periodisch und nimmt die Werte aus $\{0, 12, 9, 6, 3\}$ (nachrechnen!). Also hat die Zahl 12 keine multiplikative Inverse modulo 15.

Eine natürliche Frage deshalb ist

Durch welche Zahlen $a \in \mathbb{Z}_m$ kann man dividieren?

Oder äquivalent

Welche Zahlen $a \in \mathbb{Z}_m$ besitzen ihre multiplikative Inversen a^{-1} ?

Diese Fragen haben eine sehr elegante Antwort. Zwei Zahlen a und b heißen *teilerfremd* (oder *relativ prim*), falls sie keinen gemeinsamen Teiler (außer 1) haben, d.h. aus $x \mid a$ und $x \mid b$ folgt $x = 1$.

Satz 2.5. (Existenz von multiplikativen Inversen)

$a \in \mathbb{Z}_m$ hat die multiplikative Inverse $a^{-1} \iff a$ und m teilerfremd sind.

Damit haben genau die Zahlen aus

$$\mathbb{Z}_m^* := \{a \in \mathbb{Z}_m : a \neq 0 \text{ und } \text{ggT}(a, m) = 1\}$$

ihre multiplikative Inversen. D.h. wir können jede Zahl in \mathbb{Z}_m durch jede der Zahlen aus \mathbb{Z}_m^* dividieren.

Um Satz 2.5 zu beweisen, brauchen wir das Konzept des “größten gemeinsamen Teilers”.

Der *größter gemeinsamer Teiler* von a und b ist definiert als

$$\text{ggT}(a, b) = \max\{d : d \text{ teilt } a \text{ und } b\};$$

manchmal ist diese Zahl mit $\text{ggT}(a, b)$ bezeichnet. Die Zahlen a und b sind also *teilerfremd* (oder *relativ prim*), falls $\text{ggT}(a, b) = 1$ gilt.

Linearkombinationen von zwei Zahlen $a, b \in \mathbb{Z}$ sind alle Zahlen von der Form $ax + by$ mit $x, y \in \mathbb{Z}$. Eine Linearkombination $ax + by$ ist *positiv*, falls $ax + by \geq 1$ gilt.

Satz 2.6. $\text{ggT}(a, b)$ = die kleinste positive Linearkombination von a und b .

Beweis. Sei $d = \text{ggT}(a, b)$ und sei t die *kleinste Zahl*² in der Menge

$$A = \{ax + by : x, y \in \mathbb{Z}\} \cap \mathbb{N}_+.$$

Aus $d|a$ und $d|b$ folgt $d|t$. Wir wollen zeigen, dass auch $t|a$ und $t|b$ gilt, woraus $t | d$ und damit auch $t = d$ folgt.

Um $t|a$ zu zeigen, schreiben wir $a = qt + r$ mit $q \in \mathbb{Z}$ und $0 \leq r < t$, woraus $r = a - qt$ folgt. Wir wissen, dass t die Form $ax + by$ mit $x, y \in \mathbb{Z}$ hat. Deshalb ist die Zahl

$$r = a - qt = a - q(ax + by) = a(1 - qx) + b(-qy)$$

auch eine nicht negative (denn $r \geq 0$) Linearkombination von a und b . Da $0 \leq r < t$ und t als die *kleinste* positive Linearkombination von a und b gewählt war, ist das nur dann möglich, wenn $r = 0$ gilt.

Die Teilbarkeit von b durch t folgt mit dem selben Argument. □

Satz 2.6 liefert uns die folgenden wichtigsten Eigenschaften des größten gemeinsamen Teilers.

Lemma 2.7. Für alle ganze Zahlen $a, b \in \mathbb{Z}$ und $n \geq 1$ gilt: $\text{ggT}(an, bn) = n \cdot \text{ggT}(a, b)$.

Beweis. Nach Satz 2.6 gilt:

$$\begin{aligned} \text{ggT}(an, bn) &= \min\{anx + bny : x, y \in \mathbb{Z}\} \quad (\text{Minimum in } \mathbb{N}_+) \\ &= n \cdot \min\{ax + by : x, y \in \mathbb{Z}\} \quad (\text{Minimum in } \mathbb{N}_+) \\ &= n \cdot \text{ggT}(a, b). \end{aligned}$$

□

²Beachte, dass A nicht leer ist.

Der folgende Fakt gibt uns eine der wichtigsten Eigenschaften der teilerfremden Zahlen.

Lemma 2.8. Aus $m \mid ab$ und $\text{ggT}(a, m) = 1$ folgt: $m \mid b$.

Beweis. Nach dem Satz 2.6 können wir $\text{ggT}(m, a)$ als Linearkombination $1 = \text{ggT}(m, a) = mx + ay$ darstellen. Dann gilt auch $b = bmx + bay$. Da m beide Summanden bmx und bay teilt, muss m auch ihre Summe b teilen. \square

Für $a \in \mathbb{Z}_m$ sei

$$a\mathbb{Z}_m := \{ax \bmod m : x = 0, 1, 2, \dots, m-1\} \subseteq \mathbb{Z}_m$$

die Menge aller *verschiedenen* Zahlen $0, a, 2a, \dots, (m-1)a$ modulo m . Im Allgemeinen kann es passieren, dass $ax \equiv ay \pmod m$ für einige Zahlen $x \neq y \in \mathbb{Z}_m$ gilt (z.B. $2 \cdot 1 \equiv 2 \cdot 3 \pmod 4$); dann ist $a\mathbb{Z}_m$ eine *echte* Teilmenge von \mathbb{Z}_m . Ist aber a relativ prim zu m , dann sagt der folgende Satz, dass $a\mathbb{Z}_m = \mathbb{Z}_m$ gelten muss, d.h. in diesem Fall muss $a\mathbb{Z}_m$ einfach eine Permutation von $\mathbb{Z}_m = \{0, 1, \dots, m\}$ sein.

Satz 2.9. Ist $\text{ggT}(a, m) = 1$, so gilt

$$a\mathbb{Z}_m = \mathbb{Z}_m.$$

Insbesondere hat dann die Gleichung $ax \equiv b \pmod m$ genau eine Lösung in \mathbb{Z}_m .

Beweis. Um die Behauptung zu verifizieren, nehmen wir an, dass es zwei verschiedene Zahlen $0 \leq x \neq y \leq m-1$ mit $ax \equiv ay \pmod m$ gibt. Dann muss auch $ax - ay = a(x - y)$ durch m teilbar sein, d.h. es muss ein $q \in \mathbb{Z}$ mit $(x - y)a = qm$ geben. Da aber a und m relativ prim sind, muss $x - y$ durch m teilbar sein (Lemma 2.8). Das ist aber unmöglich, da beide Zahlen x und y nicht negativ und kleiner als m sind. Ein Widerspruch.

Nun betrachten wir die (modulare) Gleichung $ax \equiv b \pmod m$. Da der Rest $r = b \bmod m$ in \mathbb{Z}_m liegt und $\mathbb{Z}_m = a\mathbb{Z}_m$ gilt, muss es ein einziges $x \in \mathbb{Z}_m$ mit $xa \bmod m = r$ geben. \square

Damit haben wir eine Richtung des Satzes 2.5 bewiesen: Ist $\text{ggT}(a, m) = 1$, so hat die Gleichung $ax \equiv 1 \pmod m$ genau eine Lösung $x \in \mathbb{Z}_m$, und diese Lösung ist genau die multiplikative Inverse $x = a^{-1}$ von a modulo m . Die andere Richtung des Satzes 2.5 ($a \in \mathbb{Z}_m$ hat *keine* multiplikative Inverse in \mathbb{Z}_m , wenn $\text{ggT}(a, m) \neq 1$) folgt unmittelbar aus dem folgenden Fakt.

Lemma 2.10.

$$\exists x \in \mathbb{Z}_m : ax \equiv b \pmod m \iff \text{ggT}(a, m) \mid b.$$

Beweis. (\Rightarrow) Sei $d = \text{ggT}(a, m)$. Hat $ax \equiv b \pmod m$ eine Lösung x , so muss (nach Lemma 2.2) die Zahl m (und damit auch die Zahl d) die Differenz $ax - b$ teilen. Da aber a durch d teilbar ist, muss auch b durch d teilbar sein.

(\Leftarrow) Wenn $d \mid b$, dann sind a, b und m durch d teilbar und wir können die Gleichung $\frac{a}{d}x \equiv \frac{b}{d} \pmod{\frac{m}{d}}$ betrachten. Da $\frac{a}{d}$ und $\frac{m}{d}$ teilerfremd sein sind ($d = \text{ggT}(a, m)$ ist der *größter* gemeinsamer Teiler von a und m), hat diese Gleichung nach Lemma 2.9 eine Lösung $x \in \mathbb{Z}_m$. Da $m \mid (ax - b)$ genau dann, wenn $\frac{a}{d}x - \frac{b}{d}$ durch $\frac{m}{d}$ teilbar ist, ist muss x auch die Lösung von $ax \equiv b \pmod m$ sein. \square

In \mathbb{Z} kann man die Gleichung $a \cdot c = b \cdot c$ mit $c \neq 0$ durch c kürzen:

$$x \cdot \bar{a} = y \cdot \bar{a} \Rightarrow x = y \quad (\text{falls } a \neq 0).$$



In \mathbb{Z}_m kann man dies ohne weiteres *nicht* tun:

$$2 \cdot \beta \equiv 4 \cdot \beta \pmod{6} \Rightarrow 2 \equiv 4 \pmod{6} \quad \leftarrow \text{das ist falsch!}$$

Nichtdestotrotz, kann man auch die modulare Gleichung $x \cdot a = y \cdot a \pmod{m}$ durch a kürzen, falls a und m teilerfremd sind.

Lemma 2.11. (Kürzungsregel) Ist $\text{ggT}(a, m) = 1$, so kann man die beiden Seiten der Gleichung $ax \equiv ay \pmod{m}$ durch a kürzen:

$$ax \equiv ay \pmod{m} \Rightarrow x \equiv y \pmod{m}.$$

Beweis. Da $ax \equiv ay \pmod{m}$, ist $ax - ay = (x - y)a$ durch m teilbar. Da aber $\text{ggT}(a, m) = 1$ gilt, muss dann (laut Lemma 2.8) $x - y$ durch m teilbar sein, d.h. muss $x \equiv y \pmod{m}$ gelten. \square

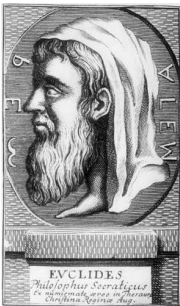
2.2 Euklidischer Algorithmus

Wie kann man den größter gemeinsamer Teiler $\text{ggT}(a, b)$ berechnet, ohne zuvor mühsam aller Teiler von a und b zu bestimmen? Dafür gibt es ein gutes altes Verfahren – der *Euklidischer Algorithmus*. Die Idee des Euklidischen Algorithmus ist es, aus zwei Zahlen den größten gemeinsamen Teiler schrittweise herauszudividieren.

Der Algorithmus basiert sich auf den folgenden zwei einfachen Beobachtungen:

1. When $b|a$, dann $\text{ggT}(a, b) = b$.
2. When $a = bt + r$, dann $\text{ggT}(a, b) = \text{ggT}(b, r)$:

Beweis: Jeder gemeinsamer Teiler von a und b muss auch $r = a - bt$ teilen, woraus $\text{ggT}(a, b) \leq \text{ggT}(b, r)$ folgt. Die andere Richtung $\text{ggT}(b, r) \leq \text{ggT}(a, b)$ ist auch richtig, da jeder Teiler von b und r muss auch $a = bt + r$ teilen.



Euklid aus Alexandria, ca. 325 - 265 J. vor Christus:

Euklid(a, b) (braucht $a \geq b \geq 0$)
 if $b = 0$
 gib a aus
 else
 gib Euklid($b, a \bmod b$) aus

► *Beispiel 2.12* : Berechne $\text{ggT}(348, 124)$:

$$\begin{aligned} \text{Euklid}(348, 124) &\Rightarrow 348 = 2 \cdot 124 + 100 \\ \text{Euklid}(124, 100) &\Rightarrow 124 = 1 \cdot 100 + 24 \\ \text{Euklid}(100, 24) &\Rightarrow 100 = 4 \cdot 24 + 4 \\ \text{Euklid}(24, 4) &\Rightarrow 24 = 6 \cdot 4 + 0 \\ \text{Euklid}(4, 0) &\Rightarrow \text{gib 4 als } \text{ggT}(348, 124) \text{ aus} \end{aligned}$$

Wir wissen bereits (Satz 2.5), dass man in \mathbb{Z}_m durch alle Zahlen $a \in \mathbb{Z}_m$, $a \neq 0$ dividieren kann, die relativ prim zu m sind. Dazu müssen wir aber die multiplikativen Inversen $a^{-1} \bmod m$ auch finden können. Und das kann man mit Hilfe von Euklidischem Algorithmus tun. Der Algorithmus liefert uns nämlich die Lösung für $ax + my = \text{ggT}(a, m) = 1$, was äquivalent zu $ax \equiv 1 \pmod m$ ist.

► *Beispiel 2.13* : Finde die multiplikative Inverse von 17 modulo 64. Zuerst wenden wir den Euklidischen Algorithmus, um $\text{ggT}(17, 64)$ zu bestimmen:

$$\begin{aligned} \text{(a)} \quad 64 &= 3 \cdot 17 + 13 \quad \rightarrow r_2 = 13 \\ \text{(b)} \quad 17 &= 1 \cdot 13 + 4 \quad \rightarrow r_3 = 4 \\ \text{(c)} \quad 13 &= 3 \cdot 4 + 1 \quad \rightarrow r_4 = 1 \\ 4 &= 4 \cdot 1 + 0 \quad \rightarrow r_5 = 0 \end{aligned}$$

Nun rechnen wir rückwärts:

$$\begin{aligned} 1 &\stackrel{\text{(c)}}{=} 13 - 3 \cdot 4 && \stackrel{\text{(b)}}{=} 13 - 3 \cdot (17 - 1 \cdot 13) = 4 \cdot 13 - 3 \cdot 17 \\ &&& \stackrel{\text{(a)}}{=} 4 \cdot (64 - 3 \cdot 17) - 3 \cdot 17 \\ &= \underbrace{4}_{x} \cdot \underbrace{64}_a - \underbrace{15}_y \cdot \underbrace{17}_b \\ &= 4 \cdot 64 + (-15) \cdot 17 \\ &\equiv 49 \cdot 17 \pmod{64} && \text{(da } -15 \equiv 49 \pmod{64} \text{)} \end{aligned}$$

Also, 49 ist die gesuchte multiplikative Inverse von 17 modulo 64, d.h. $17^{-1} \bmod 64 = 49$ gilt.

2.3 Primzahlen

Eine *Primzahl* ist eine natürliche Zahl $p \geq 2$, die nur durch 1 und sich selbst teilbar ist.



Achtung: 1 ist also *keine* Primzahl!

Satz 2.14. (Euklidischer Hilfsatz) Ist p prim, so gilt: $p \mid ab \Rightarrow p \mid a$ oder $p \mid b$.

Beweis. Angenommen, p teilt a nicht. Da p eine Primzahl ist, muss dann $\text{ggT}(a, p) = 1$ gelten. Dann muss aber nach Lemma 2.8 $p \mid b$ gelten. \square

Satz 2.15. (Fundamentalsatz der Arithmetik) Jede natürliche Zahl $n \geq 2$ lässt sich bis auf die Reihenfolge der Faktoren auf genau einer Weise als Produkt von Primzahlen schreiben:

$$n = p_1 \cdot p_2 \cdots p_k.$$



Achtung: In der Produktzerlegung $n = p_1 p_2 \cdots p_k$ kann eine Primzahl *mehrmals* vorkommen!

Beweis. Die Existenz einer solchen Primzahlzerlegung haben wir bereits in Kapitel 1 mittels Induktion bewiesen (siehe Satz 1.22). Zur Eindeutigkeit: Sind

$$n = p_1 p_2 \cdots p_s = q_1 q_2 \cdots q_r$$

zwei Primzahlzerlegungen von n , dann teilt p_1 das Produkt links. Nach Satz 2.14 muss p_1 mindestens einen der Terme q_i teilen, woraus $p_1 = p_i$ folgt. Durch Umm Nummerierung der q_i 's wird $p_1 = q_1$ erreicht. Dann bleibt $p_2 \cdots p_s = q_2 \cdots q_r$ und durch Wiederholung desselben Schlusses ergibt sich schließlich $r = s$ und $p_i = q_i$ für alle i nach eventueller Umm Nummerierung der q_i 's. \square

Mit diesem Satz können wir zum Beispiel eine weitere (nicht triviale!) Eigenschaft der teilerfremden Zahlen zeigen.

Lemma 2.16. Aus $a \mid m$ und $b \mid m$ und $\text{ggT}(a, b) = 1$ folgt: $ab \mid m$.

Beweis. Für jede natürliche Zahl $x \geq 2$ sei $P(x)$ die Menge aller *verschiedenen* Primzahlen in Primzahldarstellung von x . Aus dem Euklidischen Hilfsatz (Satz 2.14) folgt

$$x \mid y \iff P(x) \subseteq P(y),$$

denn jede Primzahl $p \in P(x)$ muss (wegen $x \mid y$) nach Satz 2.14 mindestens eine Primzahl aus $P(y)$ teilen, was nur dann möglich ist, wenn p selbst zu $P(y)$ gehört.

Aus $a \mid m$ folgt $P(a) \subseteq P(m)$, und aus $b \mid m$ folgt $P(b) \subseteq P(m)$. Damit muss $P(a) \cup P(b) \subseteq P(m)$ gelten. Ausserdem, muss $P(a) \cap P(b) = \emptyset$ gelten, da a und b teilerfremd sind. Somit gilt auch $P(a \cdot b) \subseteq P(m)$, woraus $ab \mid m$ folgt. \square

Da Primzahlen den "Skelet" aller Zahlen darstellen, haben sie die Köpfe von Menschen seit Ewigkeit beschäftigt. Es ist zum Beispiel bekannt, dass es "ungefähr" $n / \ln n$ Primzahlen in Intervall $\{2, 3, \dots, n\}$ gibt. Viele Fragen aber bleiben immer noch offen. Die bekanntesten davon sind die folgenden zwei Vermutungen.

Vermutung 2.17. (Goldbach Conjecture) Jede natürliche Zahl $n \geq 4$ ist die Summe zweier Primzahlen.

Zum Beispiel, $4 = 2 + 2$, $6 = 3 + 3$, $8 = 3 + 5$, usw. Es war bereits gezeigt, dass die Vermutung für alle Zahlen $n \leq 10^{16}$ gilt. In 1939 hat Schnirelman gezeigt, dass jede gerade Zahl eine Summe von höchstens 300000 Primzahlen ist. Das war nur ein Anfang – heute wissen wir bereits, dass jede gerade Zahl die Summe von 6 Primzahlen ist.

Vermutung 2.18. (Twin Prime Conjecture) Es gibt unendlich viele Zahlen p , so dass beide p und $p + 2$ prim sind.

In 1966 hat Chen gezeigt, dass es unendlich viele Primzahlen p gibt, so dass $p + 2$ ein Produkt von höchstens zwei Primzahlen ist. Also ist diese Vermutung “fast” richtig!

2.4 Kleiner Satz von Fermat

Nun beweisen wir den sogenannten “kleinen Satz” von Fermat³, der sich als sehr nützlich – insbesondere in der Kryptographie – erwiesen hat.



Kleiner Satz von Fermat

Ist p eine Primzahl und $a \in \mathbb{N}$, dann

$$a^p \equiv a \pmod{p}$$

Insbesondere, falls p kein Teiler von a ist, dann

$$a^{p-1} \equiv 1 \pmod{p}$$

Beweis. (Direkter Beweis) Wenn wir modulo p rechnen, so sind nach Satz 2.9 alle Zahlen

$$a, 2a, 3a, \dots, (p-1)a$$

verschieden und keiner kann gleich 0 sein: Wäre es nämlich $ka \equiv 0a \pmod{p}$, so könnten wir beide Seiten kürzen, was $k \equiv 0 \pmod{p}$ liefern würde; aber die Zahl $k \leq p-1$ zu klein dafür ist. Wenn wir also diese Zahlen modulo p nehmen, so bekommen wir genau die Zahlen $1, 2, \dots, p-1$ (vielleicht in einer anderen Reihenfolge). Deshalb muss auch das Produkt

$$a \cdot 2a \cdot 3a \cdots (p-1)a = a^{p-1} \cdot (p-1)!$$

modulo p dem Produkt

$$1 \cdot 2 \cdot 3 \cdots (p-1) = (p-1)!$$

gleich sein, d.h.

$$a^{p-1} \cdot (p-1)! \equiv (p-1)! \pmod{p}$$

gelten muss. Da p und $(p-1)!$ offenbar teilerfremd sind, können wir (nach Lemma 2.11) diese modulare Gleichung durch $(p-1)!$ kürzen, was die gewünschte Kongruenz $a^p \equiv a \pmod{p}$ liefert. \square

Beweis. (Induktiver Beweis) Induktion über a . Induktionsbasis $a = 0$ ist trivial.

Induktionsschritt: $a \mapsto a + 1$. Wir wenden den binomischen Lehrsatz an und erhalten:

$$(a+1)^p = \sum_{k=0}^p \binom{p}{k} \cdot a^k = \sum_{k=0}^p a^k \cdot \frac{p!}{k!(p-k)!} \equiv a^p + 1 \pmod{p}$$

³Nicht verwechseln das mit dem berühmten “Letzten Satz von Fermat”: Die Gleichung $x^n + y^n = z^n$ für $n \geq 2$ hat keine nicht triviale ganzzahlige Lösungen. Dieser Satz war nur in 1994 (nach mehr als 300 Jahren!) von Andrew John Weils bewiesen worden.

da in der letzten Summe für $k \notin \{0, p\}$ der k -the Term durch p teilbar ist. Da nach der Indultionsvoraussetzung $a^p \equiv a \pmod{p}$, folgt die Behauptung:

$$(a + 1)^p \equiv a^p + 1 \equiv a + 1 \pmod{p}.$$

Ist nun p kein Teiler von a , so kann man nach Lemma 2.11 (Kürzungsregel) beide Seiten der Kongruenz $a^p \equiv a \pmod{p}$ durch a dividieren, um die gewünschte Kongruenz $a^{p-1} \equiv 1 \pmod{p}$ zu erhalten. \square



Ist p prim, so kann man die multiplikative Inversen a^{-1} in \mathbb{Z}_p sehr leicht berechnen: Nimm einfach

$$a^{-1} = a^{p-2}.$$

Nun so weit so gut ... aber wie soll man denn die Potenzen a^m modulo n schnell berechnen? Ein trivialer Algorithmus $a^m = a \cdot a \cdots a$ braucht fast m Multiplikationen. Es gibt aber viel schneller Algorithmus, der nur *logarithmisch* viele (anstatt satten m) Multiplikationen braucht. um die Potenzen $a^m \pmod{n}$ auszurechnen.

Potenzierungs-Algorithmus: Zuerst bestimme die 0-1 Bits $(\epsilon_0, \epsilon_1, \dots, \epsilon_r)$ mit $r \leq \log_2(m + 1)$ in der Binärdarstellung von m , d.h. $m = \sum_{i=0}^r \epsilon_i 2^i$. Dieser Schritt ist einfach:

$$\epsilon_0 := \begin{cases} 0 & \text{falls } m \text{ gerade} \\ 1 & \text{sonst} \end{cases}$$

Danach ersetze m durch $\lceil \frac{m}{2} \rceil$ und bestimme das nächste Bit

$$\epsilon_1 := \begin{cases} 0 & \text{falls } m \text{ gerade} \\ 1 & \text{sonst} \end{cases}$$

usw. Nach $r \leq \log_2(m + 1)$ Schritten sind wir fertig. Nun ist

$$a^m = a^{\sum_{i:\epsilon_i=1} 2^i} = \prod_{i:\epsilon_i=1} a^{2^i}$$

Also reicht uns nur die $r + 1$ Zahlen

$$a, a^2, a^{2^2} = (a^2)^2, a^{2^3} = (a^{2^2})^2, \dots, a^{2^r}$$

modulo n auszurechnen, wobei jede nächste Zahl einfach Quadrat der vorigen ist!

Der kleiner Satz von Fermat sieht ziemlich einfach aus, hat aber bereits viele Abwendungen gefunden. Im nächsten Abschnitt betrachten eine Anwendung in der Kryptographie.

2.4.1 Anwendung in der Kryptographie: RSA-Codes*

Nachrichten so zu verschlüsseln, dass sie kein Unbefugter versteht, ist nicht nur der Traum von kleinen Jungs oder von Spionen – es ist mittlerweile unser Alltag geworden. Das allgemeine Model ist das folgende: Eine Nachricht besteht aus einer Zahl $a \in \mathbb{Z}_n$ (n groß genug), die der Sender Bob (der Bankkunde) der Anfängerin Alice (einer Bankangestellten) so mitteilen will, dass kein Lauscher die Nachricht versteht. Dazu wendet Bob eine Bijektion $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ auf a und sendet $f(a)$ an Alice. Sie kennt eine Funktion g mit der Eigenschaft

$$g(f(a)) = a$$

(die Inverse $g(x) = f^{-1}(x)$ von f) und kann also die Nachricht a rekonstruieren.

Eine der Schwierigkeiten von verschlüsselter Kommunikation ist die Tatsache, dass man vor dem Senden der Nachricht eine Verschlüsselungsmethode (d.h. Funktionen f und $g = f^{-1}$) verabreden muss, damit die Empfänger die Nachricht versteht.

1976 überlegten sich Rivest, Shamir und Adleman, dass die Verschlüsselungsfunktion f eigentlich nicht geheim zu sein braucht; wichtig ist nur, dass kein Unbefugter die Entschlüsselungsfunktion $g = f^{-1}$ kennt. (Lange ging man davon aus, dass deshalb auch f geheim sein muss.) Die *RSA-Codes* von Rivest, Shamir und Adleman sind so genannte *Public-Key Verfahren*.

Das wichtigste in diesem (und manchen ähnlichen) Verfahren ist, dass die Bijektion $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ die folgende "Sicherheits-Bedingung" erfüllt:

(*) Ohne die Umkehrfunktion $g = f^{-1}$ zu wissen, ist es sehr schwer aus $b = f(a)$ die Zahl a zu bestimmen.

Hat Alice eine solche (schwer umkehrbare) Funktion f , so kann sie f bekannt machen. Zum Beispiel kann sie diese Funktion auf ihrer Web-Seite angeben. Diese Funktion f ist also das "public-key". Die Umkehrfunktion $g = f^{-1}$ (das "secret-key") behält Alice für sich selbst streng geheim.

Nun kann Bob (der Kunde) seine Nachricht $a \in \mathbb{Z}_n$ (z.B. ein Überweisungsantrag) Alice so mitteilen:

- Zuerst holt er sich das Public-Key f .
- Danach berechnet er die verschlüsselte Nachricht $b = f(a)$ und verschickt b an Alice.
- Alice benutzt ihr secret-key $g = f^{-1}$, um die Nachricht zu entschlüsseln: $g(b) = f^{-1}(f(a)) = a$.

Mit einem ähnlichen Verfahren kann Alice (eine Bankangestellte) ihre Nachrichten auch unterschreiben ("digital signature"). Will nämlich Bob sicher sein, dass eine (nicht verschlüsselte) Nachricht a auch *wirklich* von Alice stammt, müssen sie beide sich so verhalten:

- Zuerst berechnet Alice $\sigma = f^{-1}(a)$, ihre "Digitale-Unterschrift".
- Dann verschickt sie $(a, \sigma) = (\text{Nachricht}, \text{Unterschrift})$ an Bob.
- Bob berechnet dann $a' = f(\sigma)$. Gilt $a' = a$, so weiss Bob, dass die Nachricht a von Alice stammt, da $f(\sigma) = f(f^{-1}(a)) = a$ gilt.

Das alles klingt sehr gut. Aber wie sollte man die Funktionen $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ mit der Eigenschaft (*) wählen? Dafür kann man die modulare Arithmetik benutzen! Nämlich kann Alice (die Bankangestellte) den folgenden Algorithmus anwenden.

Die Zahlen d und e erfüllen also die Gleichung ⁴

$$de = 1 + k(p - 1)(q - 1)$$

für ein $k \in \mathbb{Z}$. Für uns wichtig wird nur, dass

Die Zahl $de - 1$ durch **beide** Zahlen $p - 1$ und $q - 1$ teilbar ist.

⁴Als notwendige Bedingung wird heute empfohlen, p und q jeweils als 256-Bit Zahl zu wählen: der Faktorisierungsweltrekord liegt bei 512 Bit, d.h. bei Zahlen, die aus zwei Primfaktoren von je 256 Bit zusammengesetzt sind.

Tabelle 2.1: RSA-Algorithmus

-
1. Wähle rein zufällig zwei *große* Primzahlen p, q und berechne

$$n = pq \text{ wie auch } \phi(n) = (p-1)(q-1).$$

Die Nachrichten sind dann natürliche Zahlen aus $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$.

2. Wähle eine *kleine* Zahl e , die teilerfremd zu $\phi(n)$ ist (Public Key).
 3. Berechne die multiplikative Inverse $d = e^{-1} \bmod \phi(n)$ (Secret Key).
 4. Mache das Paar (n, e) der Zahlen n und e bekannt.
-

Verschlüsselungs- und Entschlüsselungsfunktionen f und g sind dann definiert durch

$$\begin{aligned} f(x) &:= x^e \bmod n \\ g(x) &:= x^d \bmod n \end{aligned}$$

Die **Sicherheit** des Verfahrens beruht sich darauf, dass es sehr schwer für den Lauscher (ohne die Zahl d zu wissen) aus $b = a^d \bmod n$ die Nachricht a hierauszukriegen ist. Warum? Da der Lauscher die Primzahlen p und q mit $p \cdot q = n$ finden muss⁵ und bisher keine effizienten (Polynomialzeit) Algorithmen für die Primzahlzerlegung bekannt sind.⁶

Und wie mit der **Korrektheit** des Verfahrens? Um die Korrektheit zu beweisen, müssen wir zeigen, dass $g(f(a)) = a$, d.h.

$$a^{ed} \equiv a \bmod n,$$

für alle $a \in \mathbb{Z}_n$ gilt. Da nach der Auswahl $ed - 1$ durch $p - 1$ wie auch durch $q - 1$ teilbar ist, folgt diese Behauptung direkt aus folgendem Fakt.

Lemma 2.19. Sei n ein Produkt von verschiedenen Primzahlen und sei b eine Zahl, so dass $b - 1$ durch $p - 1$ für jeden Primteiler p von n teilbar ist. Dann gilt für alle $a \in \mathbb{Z}$:

$$a^b \equiv a \bmod n.$$

Insbesondere gilt dann $a^b \bmod n = a$ für alle $a \in \mathbb{Z}_n$.

Beweis. Sei P die Menge aller Primzahlen, die n teilen. Nach unsere Annahme ist n das Produkt $n = \prod_{p \in P} p$ dieser Primzahlen. Aus Lemma 2.16 folgt, dass eine Zahl x genau dann durch n teilbar sein kann, wenn x durch *alle* Primzahlen $p \in P$ teilbar ist. Da wir $n \mid a^b - a$ zeigen wollen, reicht es uns also zu zeigen, dass $a^b \equiv a \bmod p$ für jede Primzahl $p \in P$ gilt.⁷ Da p eine Primzahl ist, gibt es

⁵Obwohl das nicht offensichtlich ist, man kann aber folgendes zeigen: Hat man die Zahl $d > 1$ mit $a^d \equiv 1 \bmod n$ für alle $a \in \mathbb{Z}_n$, so kann man mit großer Wahrscheinlichkeit eine Zerlegung $n = pq$ von n in Primfaktoren schnell finden.

⁶Also ist das RSA-Verfahren nicht 100% sicher! Deshalb bemühen sich viele Mathematiker einen *mathematischen Beweis* zu finden, dass ein solcher (Polynomialzeit) Primzahlzerlegungsalgorithmus wirklich *nicht existiert*. Das ist aber eines der schwierigsten Problemen der Mathematik und Theoretischen Informatik überhaupt – das berühmte **P = NP?** Problem.

⁷Zur Erinnerung: $n \mid (x - y) \iff x \equiv y \bmod n$ (siehe Lemma 2.2).

nur zwei mögliche Fälle: entweder $\text{ggT}(a, p) = p$ oder $\text{ggT}(a, p) = 1$.

Fall 1: $\text{ggT}(a, p) = p$. In diesem Fall ist a durch p teilbar, woraus $a \equiv 0 \pmod{p}$ und damit auch $a^b \equiv a \pmod{p}$ folgt.

Fall 2: $\text{ggT}(a, p) = 1$. In diesem Fall sagt uns der kleiner Satz von Fermat, dass es $a^{p-1} \equiv 1 \pmod{p}$ gelten muss. Wir wissen, dass $b - 1$ durch $p - 1$ teilbar ist, d.h. $b - 1 = k(p - 1)$ für eine Zahl $k \in \mathbb{Z}$ gilt. Nach Lemma 2.3(4) muss $(a^{p-1})^k \equiv 1^k \pmod{p}$ und damit auch $a^{b-1} \equiv 1 \pmod{p}$ gelten. Es bleibt also beide Seiten mit a zu multiplizieren (Lemma 2.3(3) erlaubt das) um die gewünschte Kongruenz $a^b \equiv a \pmod{p}$ zu erhalten. \square

Der Lauscher hat verschlüsselte Nachricht – die Zahl $b = f(a)$ – gesehen. Genau wie Bob, kennt er die Zahlen n und e und die Verschlüsselungsfunktion $f(x) := x^e \pmod{n}$. Er weiss auch, dass $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ eine *Bijektion* ist (sonst hätte das ganze Verfachen gar nicht funktioniert!). Also kann er “einfach” $f(x)$ für alle $x \in \mathbb{Z}_n$ berechnen bis er das einzige x mit $f(x) = b (= f(a))$ findet; dann muss ja $x = a$ gelten, und er hat die Nachricht geknackt! Wirklich? Ganz und gar nicht! Ein solcher Vorgang von Lauscher ist absolut hoffnungslos, da die Menge \mathbb{Z}_n zu groß ist: Da n eine mindestens 512-Bit Zahl ist, hat \mathbb{Z}_n mehr als $2^{512} > 10^{500}$ Elemente!



Wenn Alice ihre geheime Primzahlen p und q gewählt hat, dann macht sie das Produkt $n = p \cdot q$ für alle bekannt. Kann sie auch das Produkt $\phi(n) = (p-1) \cdot (q-1)$ bekannt geben? Die Antwort ist: Nein! Um das zu sehen, beachte, dass

$$\phi(n) = pq - (p + q) + 1 = n - (p + q) + 1,$$

also

$$p + q = n - \phi(n) + 1.$$

Aber dann

$$(p - q)^2 = (p + q)^2 - 4pq = (n - \phi(n) + 1)^2 - 4n.$$

Waren die beide Zahlen n und $\phi(n)$ bekannt, so könnte man leicht die beide Zahlen $p + q$ und $p - q$ berechnen, und damit auch die beide Primzahlen p und q leicht bestimmen.



Es ist zu empfehlen, die Primzahlen p, q ($p > q$) so zu wählen, dass die Differenz $p - q$ groß ist. Warum? Angenommen $p - q$ ist klein. Ist $n = pq$, so gilt

$$n = \left(\frac{p+q}{2}\right)^2 - \left(\frac{p-q}{2}\right)^2$$

Da p und q nah zu einander liegen, ist

$$s = \frac{p-q}{2}$$

klein, und

$$t = \frac{p+q}{2}$$

nicht viel größer als \sqrt{n} sein kann. Ausserdem wissen wir, dass $t^2 - n = s^2$ das Quadrat einer Zahl (diesmal s) sein muss. Also kann der Lauscher einfach alle Zahlen

$$t = \lceil \sqrt{n} \rceil, \quad t = \lceil \sqrt{n} \rceil + 1, \quad t = \lceil \sqrt{n} \rceil + 2, \dots$$

ausprobieren, bis $t^2 - n = s^2$ für ein s gilt. Dann hat er die Primzahlen p und q bereits gefunden:

$$p = t + s \quad \text{und} \quad q = t - s.$$

Wie wir bereits erwähnt haben, beruht die Sicherheit des RSA-Verfahrens auf die Schwierigkeit, eine Zahl n als Produkt $n = pq$ zweier Primzahlen darzustellen (Faktorisierungsproblem). Ein anderes Verfahren, das *Diffie-Helman public-key* Verfahren, benutzt stattdessen die Schwierigkeit, Logarithmen modulo n zu berechnen, d.h. für gegebene $a, b \in \mathbb{Z}_n^*$ eine Zahl x mit $b^x \equiv a \pmod{n}$ zu finden.

2.5 Chinesischer Restsatz

In vielen Mathematikbüchern aus alten Zeiten, angefangen bei über 2000 Jahre alten chinesischen Mathematikbüchern (Handbuch der Arithmetik von Sun-Tzun Suan-Ching), aber auch in berühmten "Liber abaci" von Leonardo von Pisa (Fibonacci), finden sich Aufgaben, in denen Zahlen gesucht werden, die bei Division durch verschiedenen andere Zahlen vorgegebene Reste lassen. Fangen wir zur Demonstration mit einem Beispiel an:

"Wie alt bist Du?" wird Daisy von Donald gefragt. "So was fragt man eine Dame doch nicht" antwortet diese. "Aber wenn Du mein Alter durch drei teilst, bleibt der Rest zwei." "Und wenn man es durch fünf teilt?" "Dann bleibt wieder der Rest zwei. Und jetzt sage ich Dir auch noch, dass bei Division durch sieben der Rest fünf bleibt. Nun müsstest Du aber wissen, wie alt ich bin."

Übersetzt in heutige mathematische Sprache lautet diese Aufgabe so: Man finde eine Zahl, die bei Division durch 3,5,7 die Reste 2,3,2 lässt. Zu lösen ist also das modulare Gleichungssystem

$$x \equiv 2 \pmod{3}$$

$$x \equiv 3 \pmod{5}$$

$$x \equiv 2 \pmod{7}$$

Ähnliche Aufgabe stammt von Sun-Tzun Suan-Ching (zwischen 280 und 473 vor Christus).

Den folgende allgemeine Satz hat Ch'in-Chiu-Shao 1247 bewiesen.

Satz 2.20. (Chinesischer Restsatz) Seien m_1, m_2, \dots, m_r paarweise teilerfremde, positive Zahlen und $M = m_1 \cdot m_2 \cdot \dots \cdot m_r$. Dann gibt es für beliebig gewählte Zahlen a_1, \dots, a_r genau eine Zahl x mit $0 \leq x < M$, die alle Kongruenzen $x \equiv a_i \pmod{m_i}$, $i = 1, \dots, r$ simultan erfüllt. Die Lösung ist durch

$$x := a_1 M_1 s_1 + a_2 M_2 s_2 + \dots + a_r M_r s_r$$

gegeben, wobei $M_i := M/m_i$ und $s_i = M_i^{-1} \pmod{m_i}$ die multiplikative Inverse von M_i modulo m_i ist.

Beweis. Setze $M_i := M/m_i$ und beachte, dass $\text{ggT}(m_i, M_i) = 1$ für alle $i = 1, \dots, r$ gilt.⁸ Also hat (nach dem Satz von Bézout) jedes M_i die multiplikative Inverse

$$s_i = M_i^{-1} \pmod{m_i}$$

modulo m_i . Wir setzen

$$x := a_1 M_1 s_1 + a_2 M_2 s_2 + \dots + a_r M_r s_r$$

und überprüfen, ob es die obigen Kongruenzen erfüllt. Zuerst stellen wir fest, dass für $i \neq j$ stets m_i ein Teiler von M_j ist, d.h. $M_j \equiv 0 \pmod{m_i}$. Daraus folgt für alle i

$$x \equiv 0 + \dots + 0 + a_i M_i s_i + 0 \dots + 0 \equiv a_i \cdot 1 \equiv a_i \pmod{m_i}.$$

Zum Beweis der Eindeutigkeit nimmt man an, dass es zwei Lösungen $0 \leq x < y < M$ gibt. Dann sind mit Lemma 2.2 alle m_i Teiler von $y - x$. Nach Voraussetzung (paarweise teilerfremd) ist auch M Teiler von $y - x$, und folglich ist $y = x$. \square

▷ *Beispiel 2.21* : Gesucht ist eine Lösung des Gleichungssystems:

$$x \equiv 2 \pmod{3}$$

$$x \equiv 3 \pmod{5}$$

$$x \equiv 2 \pmod{7}$$

Wir haben $M = 3 \cdot 5 \cdot 7 = 105$ und

$$M_1 = 105/3 = 35 \equiv 2 \pmod{3}$$

$$M_2 = 105/5 = 21 \equiv 1 \pmod{5}$$

$$M_3 = 105/7 = 15 \equiv 1 \pmod{7}$$

$$s_1 = 2^{-1} \pmod{3} = 2$$

$$s_2 = 1^{-1} \pmod{5} = 1$$

$$s_3 = 1^{-1} \pmod{7} = 1$$

⁸Das folgt aus Lemma 2.8: Ist $\text{ggT}(m, a) = \text{ggT}(m, b) = 1$, so gilt $\text{ggT}(m, ab) = 1$. Die Zahl $d = \text{ggT}(m, ab)$ muss m wie auch ab teilen. Aus $\text{ggT}(m, a) = \text{ggT}(m, b) = 1$ und $d \mid m$ folgt, dass $\text{ggT}(d, a) = \text{ggT}(d, b) = 1$ gelten muss. Da nach Lemma 2.8 d beide Zahlen a und b teilen muss, kann d nicht größer als 1 sein.

Also ist die Lösung

$$x = \overbrace{2 \cdot 35 \cdot 2}^{a_1 M_1 s_1} + \overbrace{3 \cdot 21 \cdot 1}^{a_2 M_2 s_2} + \overbrace{2 \cdot 15 \cdot 1}^{a_3 M_3 s_3} = 233 \equiv 23 \pmod{105}.$$

In Anwendungen ist das folgende Korollar von Chinesischem Restsatz oft sehr nützlich.

Korollar 2.22. Seien p_1, \dots, p_r Primzahlen und $M = \prod_{i=1}^r p_i$. Weiterhin seien $a, b < M$ beliebige ganze Zahlen. Gilt $a \equiv b \pmod{p_i}$ für alle $i = 1, \dots, r$, so gilt $a = b$.

Beweis. Da $a < M$ ist, kann nach dem Chinesischem Restsatz nur eine Zahl x mit $0 \leq x < M$ alle Kongruenzen $x \equiv a \pmod{p_i}$ simultan erfüllen, nämlich die Zahl $x = a$ selbst. \square

Bemerkung 2.23. Warum ist der Chinesische Restsatz interessant? Einfach, weil man mit ihm große Zahlen mittels viel kleineren Zahlen *eindeutig* kodieren kann. Sind z.B. p, q teilerfremd und $n = p \cdot q$, dann ist

$$\mathbb{Z}_n \ni x \mapsto (x \pmod{p}, x \pmod{q}) \in \mathbb{Z}_p \times \mathbb{Z}_q$$

eine *bijektive* Abbildung. D.h. keine zwei Zahlen in \mathbb{Z}_n können kongruent modulo p und modulo q sein!

2.5.1 Anwendung: Schneller Gleichheitstest*

Zwei Personen an den Enden eines Nachrichtenskanals wollen zwei natürliche Zahlen $a, b \leq 2^{10.000}$ auf Gleichheit hin überprüfen. Um Übertragungsfehler zu vermeiden, möchten sie die Zahlen nicht vollständig übermitteln.

Das folgende Verfahren erlaubt einen Vergleich der beiden Zahlen, dabei werden anstelle der 10.000 Bits einer Zahl nur $k \cdot 202$ Bits gesendet. Wir werden sehen, dass schon für $k = 1$ das Verfahren höchste Sicherheit garantiert.

Algorithmus (Probabilistischer Gleichheitstest)

1. Wähle zufällig Primzahlen p_1, \dots, p_k zwischen 2^{100} und 2^{101} .
2. Übertrage die Zahlen p_i und $a \pmod{p_i}$ für alle $i = 1, \dots, k$.
3. Falls $a \not\equiv b \pmod{p_i}$ für ein i , gib “ $a \neq b$ ” aus.
4. Anderfalls treffe die Entscheidung “ $a = b$ ”.

Wie in RSA-Codes, setzt das Verfahren voraus, dass man sich die benötigten Primzahlen leicht verschaffen kann. Darauf werden wir nicht eingehen.

Wir wollen nun die Wahrscheinlichkeit schätzen, dass das Verfahren zu einer Fehlentscheidung führt. Die Fehlentscheidung kann nur dann auftreten, wenn die Zahlen a und b verschieden sind und trotzdem $a \equiv b \pmod{p_i}$ für alle $i = 1, \dots, k$ gilt. Sei P die Menge aller Primzahlen zwischen 2^{100} und 2^{101} :

$$P := \{q : q \text{ ist eine Primzahl und } 2^{100} < q \leq 2^{101}\}$$

Nach dem berühmten Primzahlsatz gilt für die Anzahl $\pi(x)$ aller Primzahlen $q < x$ die asymptotische Formel $\pi(x) \sim x/\ln x$. Zwischen 2^{100} und 2^{101} gibt es daher approximativ

$$|P| = \pi(2^{101}) - \pi(2^{100}) \approx \frac{2^{101}}{\ln 2^{101}} - \frac{2^{100}}{\ln 2^{100}} \approx \frac{2^{100}}{100 \ln 2} \approx 99 \cdot 10^{26}$$

solchen Primzahlen. Sei auch

$$P(a, b) := \{q \in P : a \equiv b \pmod{q}\}.$$

Behauptung: Sind $a, b \leq 2^{10.000}$ und $a \neq b$, so gilt: $|P(a, b)| < 100$.

Um die Behauptung zu beweisen, nehmen wir an, dass $P(a, b)$ mindestens 100 Primzahlen q_1, \dots, q_{100} enthält. Aus dem Chinesischen Restsatz (siehe Korollar 2.22) folgt $a \equiv b \pmod{M}$, mit $M = q_1 \cdot \dots \cdot q_{100}$. Es gilt $M > (2^{100})^{100} = 2^{10.000}$, nach der Annahme (dass $a, b \leq 2^{10.000}$) folgt daher $a = b$.

Eine Fehlentscheidung ist also nur möglich, falls $a \neq b$ und das Verfahren zufälligerweise nur Primzahlen aus $P(a, b)$ auswählt. Bei k -facher unabhängiger Wahl einer Primzahl ist die Fehlerwahrscheinlichkeit also höchstens

$$\left(\frac{|P(a, b)|}{|P|}\right)^k \approx \left(\frac{99}{99 \cdot 10^{26}}\right)^k = 10^{-26 \cdot k}.$$

Schon für $k = 1$ ist dies ein verschwindend kleiner Wert.

2.6 Gruppen

Wir haben bereits gesehen, dass man in \mathbb{Z}_p für beliebige Primzahlen p vernünftig addieren/subtrahieren wie auch multiplizieren/dividieren kann. In diesem Fall sagt man, dass \mathbb{Z}_p ein “Körper” ist. Das Wort “vernünftig” bedeutet hier, dass die Addition wie auch Multiplikation in \mathbb{Z}_p sogenannte “Gruppen-Eigenschaften” haben.

Sei G eine nichtleere Menge und \circ eine binäre Operation (so etwas wie eine Addition $+$ oder Multiplikation \cdot). Man sagt, dass \circ eine binäre Operation *auf* G ist (oder, dass G *abgeschlossen unter* \circ ist), falls $x \circ y \in G$ für alle $x, y \in G$ gilt. Die Operation \circ ist:

- *assoziativ*, falls $(x \circ y) \circ z = x \circ (y \circ z)$;
- *kommutativ*, falls $x \circ y = y \circ x$.

Ist G unter einer assoziativen Operation \circ abgeschlossen, so nennt man (G, \circ) *Halbgruppe*.

In vielen Strukturen gibt es ein sogenanntes “neutrales” Element $e \in M$, das im gewissen Sinne “macht nichts”:

- $e \in M$ ist ein *neutrales Element*, falls $x \circ e = e \circ x = x$.

Zum Beispiel in $(\mathbb{N}, +)$ und $(\mathbb{Z}, +)$ ist $e = 0$, während in (\mathbb{N}, \cdot) und (\mathbb{Z}, \cdot) ist $e = 1$.

Hat man ein neutrales Element $e \in M$, so fragt man, ob es für jedes $x \in M$ seine “Inverse” gibt.

- $x^{-1} \in M$ ist eine *Inverse* von $x \in M$, falls $x \circ x^{-1} = x^{-1} \circ x = e$.

Zum Beispiel in $(\mathbb{Z}, +)$ ist $x^{-1} = -x$ und in (\mathbb{Q}, \cdot) ist $x^{-1} = 1/x$.



Definition:

Ist \circ eine associative binäre Operation auf einer Menge G , so ist (G, \circ) eine *Gruppe*, falls es ein neutrales Element e gibt und jedes Element $x \in G$ eine Inverse $a^{-1} \in G$ hat. Ist die Operation \circ auch kommutativ, so nennt man die Gruppe *kommutativ* oder *abelisch*^a

^aNiels Henrik Abel, 1802–1829

▷ *Beispiel 2.24*: Sei $n \in \mathbb{Z}$, $n > 1$ und $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$, und seien $+$ und \cdot die Addition und die Multiplikation modulo n .

- Ist $(\mathbb{Z}_n, +)$ eine Gruppe? Ja, das ist eine abelsche (d.h. kommutative) Gruppe mit neutralem Element 0 und mit Inversen $a^{-1} = n - a$.
- Ist (\mathbb{Z}_n, \cdot) eine Gruppe? Nein, da z.B. 0 keine Inverse hat.
- Ist dann $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine Gruppe? Nicht unbedingt! Für $n = 2$ und $n = 3$ ist das eine abelsche Gruppe, aber für $n = 4$ nicht mehr:

\cdot	1	2	3
1	1	2	3
2	2	0	2
3	3	2	1

Die Menge $\{1, 2, 3\}$ ist nicht durch \cdot abgeschlossen ($2 \cdot 2 \bmod 4 = 0 \notin \mathbb{Z}_4 \setminus \{0\}$) und 2 hat keine Inverse.

- $(\mathbb{Z}_5 \setminus \{0\}, \cdot)$ ist wiederum eine Gruppe:

\cdot	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

Beachte, dass jede Zeile und jede Spalte in der Multiplikationstabelle eine *Permutation* der Elemente $\mathbb{Z}_n \setminus \{0\} = \{1, 2, 3, 4\}$ ist. Ist das ein Zufall? Man kann zeigen (tue das!), dass dies in *jeder* endlichen Gruppe gilt.

Ist $n > 1$ keine Primzahl, also $n = ab$ für $a, b \geq 2$, so ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ mit der Multiplikation \cdot modulo n keine Gruppe mehr! Warum? Da $a \cdot b = 0 \bmod n$ gilt und 0 nicht in $\mathbb{Z}_n \setminus \{0\}$ liegt. Trotzdem auch für zusammengesetzten⁹ Zahlen n kann man in \mathbb{Z}_n eine *Teilmenge* finden, die bereits eine Gruppenstruktur hat. Diese Teilmenge ist uns bereits bekannt:

$$\mathbb{Z}_n^* := \{a \in \mathbb{Z}_n : a \neq 0 \text{ und } \text{ggT}(a, n) = 1\}$$

Satz 2.25. Für jedes $n > 1$ ist (\mathbb{Z}_n^*, \cdot) mit der Multiplikation modulo n eine kommutative Gruppe.

⁹Zusammengesetzt = nicht prim.

Beweis. Die Multiplikation \cdot ist offensichtlich kommutativ, und $1 \in \mathbb{Z}_n^*$ ist ein neutrales Element in (\mathbb{Z}_n^*, \cdot) . Ausserdem, hat nach dem Satz von Bézout jedes Element $a \in \mathbb{Z}_n^*$ seine multiplikative Inverse a^{-1} , die auch ein Element von \mathbb{Z}_n^* ist.¹⁰

Es bleibt also zu zeigen, dass die Menge \mathbb{Z}_n^* überhaupt unter der Multiplikation modulo n abgeschlossen ist.

Wir führen einen Widerspruchsbeweis durch. Seien $a, b \in \mathbb{Z}_n^*$ und nehmen wir an, dass $a \cdot b \notin \mathbb{Z}_n^*$, d.h. $\text{ggT}(a \cdot b, n) > 1$ ist. Dann gibt es eine Primzahl p mit $p \mid n$ und $p \mid a \cdot b$. Ausserdem muss $\text{ggT}(p, a) = 1$ gelten, da andererseits p die Zahl a teilen würde, was zusammen mit $p \mid n$ einen Widerspruch mit $\text{ggT}(a, n) = 1$ liefern würde. Da nun $p \mid ab$ und $\text{ggT}(p, a) = 1$, so muss $p \mid b$ gelten (siehe Lemma 2.8). Aber zusammen mit $p \mid n$ wiederum einen Widerspruch mit $\text{ggT}(b, n) = 1$. \square

Sei (G, \circ) eine beliebige Gruppe. Eine Teilmenge $U \subseteq G$ heißt *Untergruppe* von G , wenn (U, \circ) eine Gruppe ist.

► *Beispiel 2.26:* - Für jede Gruppe G sind $U = G$ und $U = \{e\}$ Untergruppen.

- Sei $m \in \mathbb{N}$, dann bildet die Menge $m\mathbb{Z} = \{x \in \mathbb{Z} : m \mid x\}$ aller durch m teilbaren Zahlen eine Untergruppe von $(\mathbb{Z}, +)$.
- \mathbb{Z} und \mathbb{Q} sind Untergruppen von $(\mathbb{R}, +)$.
- $(\mathbb{Q} \setminus \{0\}, \cdot)$ ist Untergruppe von $(\mathbb{R} \setminus \{0\}, \cdot)$.
- Betrachte die Gruppe $\mathbb{Z}_{10} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ mit der Addition $+$ modulo 10. Dann ist $H = \{0, 2, 4, 6, 8\}$ eine Untergruppe von \mathbb{Z}_{10} .

Für jede Teilmenge $U \subseteq G$ und für jedes Gruppenelement $a \in G$ wird die Menge

$$aU = \{a \circ u : u \in U\}$$

Nebenklasse (oder *Linksnebenklasse*) von a bezüglich U genannt.¹¹

► *Beispiel 2.27:* Wir betrachten die Untergruppe $U = 5\mathbb{Z} := \{5x : x \in \mathbb{Z}\}$ der Gruppe $(\mathbb{Z}, +)$. Die Nebenklassen $2+U = \{\dots, -8, -3, 2, 7, 12, \dots\}$ und $7+U$ sind gleich. Es gibt genau 5 voneinander verschiedene Nebenklassen, nämlich die Äquivalenzklassen der Relation $\equiv \pmod{5}$.

► *Beispiel 2.28:* In der Untergruppe $U = \{x \in \mathbb{R} : x > 0\}$ der Gruppe $(\mathbb{R} \setminus \{0\}, \cdot)$ gibt es genau zwei voneinander verschiedene Nebenklassen, die Menge der positiven und die Menge der negativen Zahlen.

Die wichtigste Eigenschaft der Nebenklassen ist, dass sie eine *disjunkte* Zerlegung von Gruppen darstellen.

Satz 2.29. Sei U eine Untergruppe der Gruppe (G, \circ) . Dann gehört jedes Element $x \in G$ zu *genau einer* Nebenklasse von U .

¹⁰Warum?! Sei $d = \text{ggT}(a^{-1}, n)$. Aus $a \circ a^{-1} = 1 + kn$, $d \mid a^{-1}$ und $d \mid n$ folgt $d \mid 1$, also $d = 1$ gelten muss.

¹¹Die *Rechtsnebenklasse* von U ist als $Ua = \{u \circ a : u \in U\}$ definiert.

Beweis. Da $e \in U$, gehört jedes $x \in G$ zu xU . Es bleibt also zu zeigen, dass kein x zu zwei *verschiedenen* Nebenklassen gehören kann.

Nehmen wir deshalb an, dass $aU \cap bU \neq \emptyset$. Dann gilt $aw = bv$ für geeignete $w, v \in U$; also ist $a = bvw^{-1}$ und damit auch folgt

$$a \circ u = b \circ \overbrace{v \circ w^{-1}}^{\text{liegt in } U} \circ u$$

für alle $u \in G$. Deswegen gilt $aU \subseteq bU$, und analog $bU \subseteq aU$. □

Nun können wir einen der Hauptsätze der Gruppentheorie beweisen. Die *Ordnung* einer endlichen Gruppe G ist die Anzahl $|G|$ seiner Elemente. Der folgender Satz sagt, dass die Ordnung jeder(!) Untergruppe ein Teiler der Gruppenordnung ist.



Satz von LaGrange:

Ist G eine *endliche* Gruppe und U eine Untergruppe von G , dann gilt $|G| = n \cdot |U|$ wobei n die Anzahl der Nebenklassen von U ist.

Beweis. Seien a_1U, \dots, a_nU alle verschiedenen Nebenklassen von U . Nach Satz 2.29 sind je zwei verschiedene Nebenklassen disjunkt und ihre Vereinigung $a_1U \cup a_2U \cup \dots \cup a_nU$ ist gleich G . Nach der Kürzungsregel¹² gilt $|aU| = |U|$ für jedes $a \in G$, d.h. die Nebenklassen besitzen die gleiche Mächtigkeit. Damit ist

$$|G| = |a_1U \cup a_2U \cup \dots \cup a_nU| = |a_1U| + |a_2U| + \dots + |a_nU| = n \cdot |U|.$$

□

Warum ist dieser Satz auch in der Praxis nützlich? Nur ein Beispiel: Angenommen, wir haben eine Gruppe G und wollen ein “gutes” Element in G finden. Nehmen wir auch an, dass die “schlechten” Elemente eine Untergruppe U von G bilden. Dann wissen wir sofort, dass entweder $U = G$ (keine “guten” Elemente vorhanden sind) oder $|U| \leq |G|/2$ (mindestens die Hälfte der Elemente “gut” sind!). In diesem letzten Fall können wir einfach die Elemente aus G rein zufällig (je mit Wahrscheinlichkeit $1/2$) auswählen. Da die Menge der “guten” Elemente sehr dicht ist, werden wir uns ziemlich schnell auf einen “guten” Element stoßen.

2.6.1 Zyklische Gruppen

Sei (G, \circ) eine Gruppe und $a \in G$, dann ist a^k die Abkürzung für

$$a^k := \overbrace{a \circ a \circ \dots \circ a}^{k\text{-mal}}$$

¹² $a \circ u_1 = a \circ u_2 \implies u_1 = u_2$

Lemma 2.30. Ist (G, \circ) eine endliche Gruppe und $a \in G$, dann ist

$$H_a = \{a, a^2, a^3, \dots\}$$

eine Untergruppe von G .

Beweis. Weil es nur endlich viele Gruppenelemente gibt, muss in der Folge a, a^2, a^3, \dots ein Glied mehrfach vorkommen, z.B. $a^i = a^j$ mit $i < j$. Wegen der Gruppenregeln haben wir sofort $e = a^i \cdot (a^i)^{-1} = a^i \cdot (a^{-1})^i = a^j \cdot (a^{-1})^i = a^{j-i}$. Außerdem gilt: $a^{-1} = a^{j-i-1}$. \square

Die Anzahl $|H_a|$ der Elemente in dieser Untergruppe heißt die *Ordnung* des Elements a . Die Gruppe G heißt *zyklisch*, wenn es ein Element $a \in G$ mit $H_a = G$ gibt; jedes derartige a heißt *erzeugendes Element* (oder Erzeugendes) von G .

Satz 2.31. Jede endliche Gruppe (G, \circ) , deren Ordnung $p = |G|$ eine Primzahl ist, ist zyklisch.

Beweis. Nimm ein beliebiges Element¹³ $a \in G$, $a \neq e$, und betrachte die von a erzeugte zyklische Untergruppe H_a . Nach Satz von Lagrange, muss $|H_a|$ ein Teiler von $p = |G|$ sein. Da aber p prim ist, kann das nur dann sein, wenn $|H_a|$ gleich 1 oder p ist. Aus $a \neq e$ folgt $|H_a| \geq 2$, und damit auch $|H_a| = p$. Da aber $H_a \subseteq G$, folgt die Behauptung: $G = H_a$. \square

Lemma 2.32. Ist (G, \circ) eine endliche Gruppe und $a \in G$, dann gilt: $a^{|G|} = e$.

Beweis. Sei $k = |H_a|$ die Ordnung von $a \in G$. Da H_a eine Untergruppe von G ist, sagt uns der Satz von Lagrange, dass $|G|/k$ eine ganze Zahl sein muss. Dann gilt auch: $a^{|G|} = (a^k)^{|G|/k} = e^{|G|/k} = e$. \square

Sei

$$\mathbb{Z}_n^* := \{a \in \mathbb{Z}_n : a \neq 0 \text{ und } \text{ggT}(a, n) = 1\}.$$

Die Menge \mathbb{Z}_n^* bildet eine multiplikative Gruppe (siehe Satz 2.25) und sein Ordnung $|\mathbb{Z}_n^*|$ ist die berühmte Euler-Funktion $\phi(n)$, d.h.

$$\phi(n) = \text{Anzahl der Zahlen in } 1, 2, \dots, n-1, \text{ die relativ prim zu } n \text{ sind}$$

Korollar 2.32 liefert uns sofort den folgenden Satz von Euler, der in Kryptographie benutzt wird (siehe Abschnitt 2.4.1). Ein Spezialfall dieses Satzes mit $m = p$ eine Primzahl, den kleinen Satz von Fermat, haben wir bereits direkt bewiesen.

Korollar 2.33. (Euler) Für jedes $a \in \mathbb{Z}_n^*$ gilt:

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

Zwei Strukturen (G, \circ) und $(H, *)$ heißen *isomorph*, falls es eine Bijektion $f : G \rightarrow H$ mit $f(a \circ b) = f(a) * f(b)$ für alle $a, b \in G$ gibt.

¹³Warum ist $G \neq \{e\}$?

Satz 2.34. (Cayley) Zwei zyklische Gruppen sind genau dann isomorph, wenn sie den gleichen Ordnung besitzen.

Beweis. “ \Rightarrow ” Aus Anzahlgründen klar.

“ \Leftarrow ”: Es seien (G_1, \circ) und $(G_2, *)$ zwei zyklische Gruppen mit erzeugenden Elementen $a \in G_1$ und $b \in G_2$. Sei $|G_1| = |G_2|$. Wir definieren die Bijektion $f : G_1 \rightarrow G_2$ mit $f(a^k) := b^k$ ($k \in \mathbb{Z}$). Seien $x, y \in G_1$. Dann existieren $j, k \in \mathbb{N}$ mit $x = a^j$ und $y = a^k$. Dann ist

$$\begin{aligned} f(x \circ y) &= f(a^j \circ a^k) = f(a^{j+k}) = b^{j+k} = b^j * b^k \\ &= f(a^j) * f(a^k) = f(x) * f(y). \end{aligned}$$

Also sind G_1 und G_2 isomorph. □



Als Korollar bekommen wir, dass (bis auf Isomorphie) die Gruppen $(\mathbb{Z}_m, +)$ und $(\mathbb{Z}, +)$ die *einzigsten* zyklischen Gruppen sind!

2.7 Ringe und Körper

Eine Menge \mathbb{F} mit zwei binären Operationen $+$ und \cdot heißt *Körper*, wenn

1. $(\mathbb{F}, +)$ und $(\mathbb{F} \setminus \{0\}, \cdot)$ sind kommutative Gruppen und
2. Es gelten die Distributivgesetze: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ und $(a + b) \cdot c = (a \cdot b) + (a \cdot c)$.

Zum Beispiel, $(\mathbb{R}, +, \cdot)$ und $(\mathbb{Q}, +, \cdot)$ sind Körper. Aber beide diese Körper sind *unendlich*. Gibt es Körpern mit *endlich* vielen Elementen? Ja, und sogar sehr viele!

Satz 2.35. Ist p prim, so ist $(\mathbb{Z}_p, +, \cdot)$ ein Körper.

Beweis. $(\mathbb{Z}_p, +)$ ist eine kommutative Gruppe: $0 \in \mathbb{Z}_p$ ist das neutrale Element und die additive Inverse von $a \in \mathbb{Z}_p$ ist $p - a$ und.

$\mathbb{Z}_p \setminus \{0\}$ ist auch eine kommutative Gruppe: $1 \in \mathbb{Z}_p \setminus \{0\}$ ist das neutrale Element und $a^{-1} := a^{p-2}$ ist die multiplikative Inverse von $a \in \mathbb{Z}_p \setminus \{0\}$: $a \cdot a^{-1} := a^{p-1} \equiv 1 \pmod{p}$. Warum? Da laut kleinem Satz von Fermat für jedes $a \in \mathbb{Z}_p$, $a \neq 0$ die Kongruenz $a^{p-1} \equiv 1 \pmod{p}$ gilt. □



Ist $q = p^m$ eine Potenz einer Primzahl p , so gibt es ein Körper mit q Elementen. Dieser Körper heißt *Galois Körper* (engl. Galois^a field) und ist mit $GF(q)$ bezeichnet.

^aEvariste Galois 1811–1832

Sei $(K, +, \cdot)$ ein Körper. Falls es eine natürliche Zahl n gibt, für die

$$\underbrace{1 + 1 + \dots + 1}_{n\text{-mal}} = 0$$

gilt, heißt die kleinste solche Zahl die *Charakteristik* von K und wird mit $\text{char}(K)$ bezeichnet. Gibt es kein solches n , so definiert man $\text{char}(K) = 0$.

Nach dem Satz 2.35 kommt *jede* Primzahl als Charakteristik eines Körpers vor. Inresanterweise gilt auch die Umkehrung:

Satz 2.36. Die Charakteristik eines Körpers ist stets 0 oder eine Primzahl.

Beweis. Angenommen, $\text{char}(K) = pq$ mit $p, q \neq 1$ und $p, q \in \mathbb{N}$. Dann gilt

$$0 = \underbrace{1 + 1 + \cdots + 1}_{pq\text{-Summanden}} = \underbrace{(1 + 1 + \cdots + 1)}_{p\text{-Summanden}} \cdot \underbrace{(1 + 1 + \cdots + 1)}_{q\text{-Summanden}}$$

Wir haben also, dass $ab = 0$ für zwei Elementen $a, b \in \mathbb{F} \setminus \{0\}$ mit

$$a = \underbrace{(1 + 1 + \cdots + 1)}_{p\text{-Summanden}} \quad \text{und} \quad b = \underbrace{(1 + 1 + \cdots + 1)}_{q\text{-Summanden}}$$

gilt. Das bedeutet aber, dass die Menge $\mathbb{F} \setminus \{0\}$ nicht unter der Multiplikation abgeschlossen ist, d.h. $(\mathbb{F} \setminus \{0\}, \cdot)$ keine Gruppe ist. Ein Widerspruch. \square

Es gibt viele Strukturen $(R, +, \cdot)$, die “sehr ähnlich” zu Körpern sind mit einer Ausnahme: man kann da nicht vernünftig *dividieren*, d.h. $(R \setminus \{0\}, \cdot)$ keine Gruppe ist. Ist aber die Multiplikation \cdot mindestens *associativ*, so nennt man dann $(R, +, \cdot)$ *Ring*. Ist \cdot auch kommutativ, so heißt auch der Ring *kommutativ*. Ein Ring, in dem die Regel $a \cdot b = 0 \implies (a = 0 \vee b = 0)$ gilt, heißt *nullteilerfrei*.¹⁴

▷ *Beispiel 2.37:* - $(\mathbb{Z}, +, \cdot)$ ist nullteilerfrei Ring, aber kein Körper.

- Sei $m = ab$ eine natürliche Zahl, die keine Primzahl ist, dann ist $(\mathbb{Z}_m, +, \cdot)$ ein kommutativer Ring mit Eins, aber nicht nullteilerfrei: $ab \equiv 0 \pmod{m}$.
- Sei $(R, +, \cdot)$ ein Ring und R^R die Menge aller Funktionen von R nach R . Durch punktweise Definition der Addition $(f \oplus g)(x) = f(x) + g(x)$ und Multiplikation $(f \odot g)(x) = f(x) \cdot g(x)$ von Funktionen $f, g \in R^R$ erhält (R^R, \oplus, \odot) die Struktur eines Rings, der *Funktorring* genannt wird.

Bemerkung 2.38. Ein kommutativer und nullteilerfreier Ring R ist ein “hübscher Ring”, er kann zu einem Körper (seinem *Quotientenkörper*) erweitert werden (ganz analog wie man den Ring der ganzen Zahlen \mathbb{Z} zum Körper der rationalen Zahlen \mathbb{Q} erweitert): man betrachtet die Menge aller geordneten Paare $(a, b) \in R \times R$ mit $b \neq 0$ und schreibt sie formal als Brüche $\frac{a}{b}$; zwei Brüche $\frac{a}{b}$ und $\frac{c}{d}$ werden als gleich angesehen, wenn $ad = bc$.

2.7.1 Polynomring

Wir betrachten nun *Polynome* in der Variablen x mit rationalen Koeffizienten (oder allgemeiner mit Koeffizienten in einem Körper \mathbb{F}). Ein Polynom $f = f(x)$ über \mathbb{F} ist gegeben durch einen Ausdruck der Gestalt

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

¹⁴Ein kommutativer (bezüglich der Multiplikation) nullteilerfreier Ring mit mindestens zwei Elementen heißt *Integritätsbereich*.

mit $n \in \mathbb{N}$ und $a_i \in \mathbb{F}$, $i = 0, 1, \dots, n$.

Glieder $a_i x^i$ mit dem Koeffizienten $a_i = 0$ dürfen aus der Summe weglassen bzw. zur Summe beliebig hinzufügen werden. Wir betrachten also zwei Polynome als identisch, falls sie dieselben Glieder haben, abgesehen von Summanden mit dem Koeffizient 0.

Der *Grad* $\text{grad}(f)$ eines Polynoms $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + \dots + a_0$ ist die größte Zahl i , so dass $a_i \neq 0$. Das zugehörige a_i bezeichnen wir als den *Anfangskoeffizienten* oder *höchsten Koeffizienten* von f .

Mit Polynomen kann man rechnen wie mit ganzen Zahlen. Zwei polynome $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + \dots + a_0$ und $g(x) = b_m x^m + b_{m-1} x^{m-1} + \dots + b_1 x + \dots + b_0$ lassen sich addieren

$$(f + g)(x) := (a_n + b_n)x^n + \dots + (a_1 + b_1)x + (a_0 + b_0)$$

(o.B.d.A. können wir $m = n$ annehmen) und multiplizieren

$$(fg)(x) := c_{n+m} x^{n+m} + c_{n+m-1} x^{n+m-1} + \dots + c_1 x + \dots + c_0 \quad \text{mit} \quad c_i := \sum_{j+k=i} a_j b_k.$$

Damit bilden die Polynome über \mathbb{F} ein *Polynomring*, der mit $\mathbb{F}[x]$ bezeichnet ist.

Die Polynomringe $\mathbb{F}[x]$ haben viele Gemeinsamkeiten mit dem Ring \mathbb{Z} der ganzen Zahlen. Insbesondere kann man ein Polynom $p(x)$ des Grades n durch ein Polynom $q(x)$ des Grades $m \leq n$ mit Rest dividieren.

▷ *Beispiel 2.39*: Seien $f(x) = 3x^5 - 2x^4 + x^2 - 3x + 5$ und $g(x) = x^2 + 1$. Dann ist:

$$\begin{array}{r|l} 3x^5 - 2x^4 + 0x^3 + x^2 - 3x + 5 & x^2 + 1 \\ -3x^5 & \hline -2x^4 - 3x^3 + x^2 - 3x + 5 & 3x^3 - 2x^2 - 3x + 3 \\ + 2x^4 & \\ \hline -3x^3 + 3x^2 - 3x + 5 & \\ + 3x^3 & \\ \hline 3x^2 + 5 & \\ - 3x^2 - 3 & \\ \hline 2 & \end{array}$$

Damit ist $f(x) = q(x)g(x) + r(x)$ mit $q(x) = 3x^2 - 2x^2 - 3x + 3$ und dem Rest $r(x) = 2$.

Die Polynome kann man also genau wie ganze Zahlen mit dem Rest dividieren.

Satz 2.40. (Division mit Rest) Zu Polynomen $f(x), g(x) \neq 0$ existieren Polynome $q(x), r(x)$, so dass

$$f(x) = q(x)g(x) + r(x), \quad \text{mit } \text{grad}(r) < \text{grad}(g) \text{ oder } r(x) = 0.$$

Beweis. Seien $f, g \neq 0$ Polynome vom Grade n und m , mit Anfangskoeffizienten a_n und b_m . Falls $n \geq m$, können wir g aus f herausdividieren, also das Polynom $f_1(x) = f(x) - p(x)$ mit $p(x) = f(x) - a_n b_m^{-1} x^{n-m} g(x)$ bilden. Offenbar hat f_1 einen kleineren Grad als f , da $p(x)$ den Term $a_n x^n$ enthält. Gilt $\text{grad}(f_1) \geq m$, so kann g aus f_1 ein weiteres Mal herausdividiert werden. Dies läßt sich fortsetzen, bis ein Polynom r vom Grad kleiner m oder aber das Nullpolynom übrigbleibt. \square

Jedes Polynom $p(x)$ bestimmt in natürlicher Weise eine Funktion von \mathbb{F} nach \mathbb{F} , indem für x ein $a \in \mathbb{F}$ eingesetzt und der Ausdruck in \mathbb{F} ausgewertet wird. Das Ergebnis dieser Auswertung wird mit $p(a)$ bezeichnet und *Wert von $p(x)$ an der Stelle a* genannt.



Man beachte, dass zwischen Polynomen und Polynomfunktionen im Allgemeinen streng unterschieden werden muss, da verschiedene Polynome (z.B. x und x^3 über \mathbb{Z}_3) dieselbe Polynomfunktion darstellen können. Betrachtet man jedoch Polynome über *unendlichen* Körpern, dann stellen verschiedene Polynome auch immer verschiedene Polynomfunktionen dar.

Ein $a \in \mathbb{F}$ heißt *Nullstelle* des Polynoms $p(x)$, falls $p(a) = 0$.

Lemma 2.41. Genau dann ist $a \in \mathbb{F}$ eine Nullstelle von $p(x)$, wenn es ein Polynom $q(x)$ mit $p(x) = q(x) \cdot (x - a)$ gibt.

Beweis. Nach Satz 2.40 ist $p(x) = q(x)(x - a) + r(x)$, wobei $r = 0$ oder $\text{grad}(r) < \text{grad}(x - a) = 1$ ist. In jedem Fall ist $r = b$ mit $b \in \mathbb{F}$. Einsetzen $x := a$ liefert $r(a) = p(a) - q(a)(a - a) = 0$, d.h. $b = 0$ (der Rest $r(x)$ ist ein Nullpolynom). \square

Korollar 2.42. Jedes Polynom $p(x) \neq 0$ vom Grad n kann höchstens n verschiedene Nullstellen haben.

Der Euklidische Algorithmus läßt sich nun auch auf rationale Polynome anwenden. Da sich der Grad der Restpolynome bei jeder Division verkleinert, bricht er nach endlich vielen Schritten ab, seine Laufzeit ist durch den Grad des Polynoms g beschränkt. Wie für ganze Zahlen können wir also feststellen:

Zwei rationale Polynome $f(x), g(x) \neq 0$ besitzen einen größten gemeinsamen Teiler $d(x)$, und es gibt rationale Polynome $s(x)$ und $t(x)$, so dass

$$d(x) = s(x)f(x) + t(x)g(x).$$

Eine wichtige Aufgabenstellung ist die sogenannte *Interpolation* von Polynomen, d.h. es soll ein Polynom gefunden werden, das an gegebenen Stellen gegebene Werte annimmt. Die wichtigste Anwendung besteht in der Approximation (Annäherung) von Funktionen durch Polynome. Darüber hinaus ist die Interpolation ein wichtiger Bestandteil für schnelle Algorithmen zur Polynommultiplikation.

Satz 2.43. (Interpolationsformel von Lagrange) Seien n verschiedene Stellen $a_1, a_2, \dots, a_n \in \mathbb{F}$ und n Werte b_1, b_2, \dots, b_n gegeben. Dann erfüllt das Polynom

$$p(x) := \sum_{i=1}^n b_i \prod_{\substack{j=1 \\ j \neq i}}^n \frac{x - a_j}{a_i - a_j}$$

die Bedingung $p(a_i) = b_i$ für alle $1 \leq i \leq n$. Außerdem ist $p(x)$ das einzige Polynom vom Grad $\leq n - 1$, das diese Bedingung erfüllt.

Beweis. Die erste Aussage ist trivial. Es bleibt also nur die zweite Aussage (Eindeutigkeit) zu beweisen. Angenommen, es gibt ein Polynom $p'(x)$ vom Grad $\leq n - 1$ mit $p'(a_i) = b_i$ für alle $1 \leq i \leq n$. Dann hat das Polynom $q(x) := p(x) - p'(x)$ mindestens n verschiedenen Nullstellen a_1, \dots, a_n . Da aber $\text{grad}(q) \leq n - 1$, kann das (laut Korollar 2.42) nur dann sein, wenn $q(x) = 0$ ein Nullpolynom ist, d.h. nur wenn die Polynome $p(x)$ und $p'(x)$ gleich sind. \square

2.7.2 Komplexe Zahlen*

Beim Rechnen mit reellen Zahlen stößt man auf das Problem, dass gewisse algebraische Gleichungen – die einfachste solche Gleichung ist $x^2 + 1 = 0$ – in \mathbb{R} keine Lösungen besitzen (das Quadrat einer reellen Zahl kann nicht negativ sein). Andererseits gibt es Gleichungen mit rationalen, ja sogar mit ganzzahligen Koeffizienten, die zwar keine Lösung in \mathbb{Q} , wohl aber eine in \mathbb{R} besitzen, wie zum Beispiel $x^2 - 2 = 0$. Diese in \mathbb{Q} nicht lösbare Gleichung wird also lösbar, indem man \mathbb{Q} zu \mathbb{R} erweitert. Das legt den Gedanken nahe, dass vielleicht eine nochmalige Erweiterung von \mathbb{R} zu einem noch größeren Körper auch eine Lösung von $x^2 + 1 = 0$ (und möglicherweise von weiteren, in \mathbb{R} unlösbaren Gleichungen) führt.

Man kann die Erweiterung von \mathbb{R} bis \mathbb{C} genauso konstruieren, wie die Erweiterung von \mathbb{Z} bis \mathbb{Q} konstruiert ist. Nämlich, man kann \mathbb{Q} als die Menge aller Paare (a, b) mit $a, b \in \mathbb{Z}$ und $b \neq 0$ betrachten mit der Addition

$$(a, b) + (a', b') = (ab' + a'b, bb') \quad \text{oder} \quad \frac{a}{b} + \frac{a'}{b'} = \frac{ab' + a'b}{bb'}$$

und der Multiplikation

$$(a, b) \cdot (a', b') = (aa', bb') \quad \text{oder} \quad \frac{a}{b} \cdot \frac{a'}{b'} = \frac{aa'}{bb'}$$

Analog kann man \mathbb{C} als die Menge $\mathbb{R} \times \mathbb{R}$ mit der Verknüpfungen

$$\begin{aligned} (a, b) + (a', b') &= (a + a', b + b') \\ (a, b) \cdot (a', b') &= (aa' - bb', ab' + a'b) \end{aligned}$$

definieren. Man kann dann zeigen, dass \mathbb{C} ein Körper ist mit

$$\mathbf{1} = (1, 0)$$

als neutralem Element hinsichtlich Multiplikation: $(1, 0) \cdot (a, b) = (a, b)$. Da $(0, 1) \cdot (0, 1) = (-1, 0) = -\mathbf{1}$, ist

$$i = (0, 1)$$

die Lösung von $x^2 = 1$ in \mathbb{C} . D.h. das Paar $i = (0, 1)$ bezeichnet die “Zahl”

$$i = \sqrt{-1}.$$

Komplexe Zahlen sind also *Paare* $z = (a, b)$ von reellen Zahlen. Ein solches Paar schreibt man auch in der Form

$$z = a + bi \quad \text{mit} \quad i = (0, 1)$$

und nennt dies die *Normalform* von z . Jede komplexe Zahl $z = a + bi$ wird also eindeutig durch die beiden reellen Zahlen a und b beschreiben. Man nennt a den *Realteil* von z und b den *Imaginärteil*, kurz $a = \operatorname{Re} z$ und $b = \operatorname{Im} z$.

Der Grund, warum die komplexen Zahlen gut sind, erklärt der folgender Satz.

Satz 2.44. (Fundamentalsatz der Algebra) Sind a_0, a_1, \dots, a_n komplexe Zahlen, $a_n \neq 0$, so ist die Polynomgleichung $a_0 + a_1x + a_2x^2 + \dots + a_nx^n = 0$ in \mathbb{C} stets lösbar.

Ein Beweis für den Fundamentalsatz war schon Gauß bekannt. Der Beweis ist aber nicht einfach, und wir verzichten auf ihm.

2.8 Allgemeine Vektorräume

In Kapitel 5 werden wir sehen, wie man mit Vektoren über einem Körper \mathbb{F} rechnen kann. Es gibt aber auch andere Strukturen, die nicht Teilmengen von \mathbb{F}^n sind und trotzdem dieselben Eigenschaften, wie Vektorräume $V \subseteq \mathbb{R}^n$, haben. Deshalb lohnt es, einen allgemeineren Begriff des “Vektorraums” kennenzulernen.

Sei $\mathbb{F} = (\mathbb{F}, +, \cdot, 0, 1)$ ein Körper, wobei 0 das neutrale Element der Gruppe $(\mathbb{F}, +)$ und 1 das neutrale Element der Gruppe (\mathbb{F}^*, \cdot) ist.

Ein *Vektorraum* über den Körper \mathbb{F} ist eine additive abelsche Gruppe $G = (V, +, \mathbf{0})$, die mit einer Operation $\mathbb{F} \times V \ni (\lambda, v) \rightarrow \lambda v \in V$ (“Multiplikation mit Skalar”) versehen ist, für die folgende Axiome für alle $\lambda, \mu \in \mathbb{F}$ und $\mathbf{u}, \mathbf{v} \in V$ gelten:

- (a) $(\lambda \cdot \mu)\mathbf{v} = \lambda(\mu\mathbf{v})$
- (b) $\lambda(\mathbf{u}+\mathbf{v}) = \lambda\mathbf{u}+\lambda\mathbf{v}$ (beide Additionen in G)
- (c) $(\lambda + \mu)\mathbf{v} = \lambda\mathbf{u}+\lambda\mathbf{v}$ (die erste Addition in \mathbb{F} , die zweite in G)
- (d) $1\mathbf{v} = \mathbf{v}$

Die Elemente von V heißen dann *Vektoren*.

In anderen Wörtern, ein Vektorraum ist eine Menge, deren Elemente sich addieren und mit Skalar multiplizieren lassen, wobei die Summe von Vektoren und das Vielfache eines Vektors wieder Elemente der Menge sind. Die Elemente so eines Vektorraumes sind (heißen) Vektoren.

Hier sind drei Standardbeispiele für Vektorräume.

1. Der üblicher Vektorraum \mathbb{F}^n von (richtigen) Vektoren mit komponentenweise Addition und Skalarmultiplikation, wie oben.

2. Es sei P_n die Menge aller reellen Polynome vom Grade höchstens n . Jedes Polynom p mit

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

wird durch seine Koeffizienten a_n, a_{n-1}, \dots, a_0 repräsentiert, also durch ein $(n + 1)$ -Tupel $(a_n, a_{n-1}, \dots, a_0)$ von reellen Zahlen. Umgekehrt entspricht jedem solchen Tupel von reellen Zahlen genau ein Polynom vom Grade höchstens n . Zwei Polynome aus P_n kann man addieren und mit einer reellen Zahl multiplizieren. Damit bildet P_n einen Vektorraum über dem Körper der reellen Zahlen.

3. Es sei \mathbb{F} ein Körper, X eine (beliebige) Menge und sei \mathbb{F}^X die Menge aller Funktionen $f : X \rightarrow \mathbb{F}$ mit der Addition

$$(f + g)(x) := f(x) + g(x)$$

und der Skalarmultiplikation mit einer reellen Zahl λ gemäß

$$\lambda f(x) := \lambda \cdot f(x) \quad \forall x \in X.$$

Man überzeugt sich sofort, dass \mathbb{F}^X Vektorraum über dem Körper \mathbb{F} ist.

Alle Konzepte – wie lineare Unabhängigkeit, Basis und Dimension, die wir im Kapitel 5 nur für Vektorräume $V \subseteq \mathbb{F}^n$ betrachten werden – kann man auch auf allgemeine Vektorräume direkt übertragen!

2.9 Aufgaben

- 2.1.** Zeige: Für beliebige ganze Zahlen a, b, c gilt $\text{ggT}(a, b - ca) = \text{ggT}(a, b)$.
- 2.2.** Zeige, dass $n^4 + 4$ für $n > 1$ keine Primzahl ist. *Hinweis:* Stelle $n^4 + 4$ als ein Produkt von zwei Summen dar.
- 2.3.** Zeige: Eine ganze Zahl $n \in \mathbb{Z}$ in Dezimaldarstellung $n = \sum_{i=0}^{\infty} a_i 10^{-i}$ ist durch 3 teilbar genau dann, wenn $\sum_{i=0}^{\infty} a_i$ durch 3 teilbar ist.
- 2.4.** Beweise oder widerlege: Ist $s \equiv r \pmod{m}$, so gilt $a^s \equiv a^r \pmod{m}$.
- 2.5.** Gilt Lemma 2.11 auch wenn $\text{ggT}(a, m) > 1$?
- 2.6.** Zeige folgendes: Sind die Zahlen m und n durch d teilbar, so ist auch $\text{ggT}(m, n)$ durch d teilbar. Fazit: der größte gemeinsame Teiler $\text{ggT}(m, n)$ ist nicht nur der "größter" im Bezug seiner Größe; er ist auch der "größter" in dem Sinne, dass jeder gemeinsame Teiler d von m, n muss auch $\text{ggT}(m, n)$ teilen! *Hinweis:* Satz 2.6.
- 2.7.** Zeige, dass auch die Umkehrung von Lemma 2.11 (Kürzungsregel) gilt: $ax \equiv ay \pmod{m} \Rightarrow x \equiv y \pmod{m}$ für alle $x, y \in \mathbb{Z}_m$ gilt, so müssen a und m teilerfremd sein.
- 2.8.** Zeige, dass es in $(\mathbb{Z}_6, +, \cdot)$ (modulo 6) Elemente $a \neq 0$ gibt, die keine multiplikative Inverse haben.
- 2.9.** Zeige, dass $a \in \mathbb{Z}_m$ ein Nullteiler¹⁵ genau dann ist, wenn $\text{ggT}(a, m) > 1$ gilt.
- 2.10.** Fibonacci-Zahlen f_0, f_1, f_2, \dots sind rekursiv wie folgt definiert: $f_0 = 0, f_1 = 1$ und $f_{n+2} = f_{n+1} + f_n$ für $n \geq 0$. Zeige, dass für $n \geq 2$ jede zwei aufeinander folgende Fibonacci-Zahlen f_n und f_{n+1} teilerfremd sind, d.h. $\text{ggT}(f_n, f_{n+1}) = 1$ gilt.
- 2.11.** Leite den Satz von Bézout aus dem Satz 2.6 ab.

¹⁵ $a \in \mathbb{Z}_m$ ist ein Nulteiler, wenn $a \neq 0$ und es ein $x \in \mathbb{Z}_m$ mit $x \neq 0$ und $a \cdot x \equiv 0 \pmod{m}$ gibt.

2.12. Sei $\phi(n)$ die Eulersche Funktion, d.h. $\phi(n) = |\mathbb{Z}_n^*|$ wobei

$$\mathbb{Z}_n^* = \{a \in \mathbb{Z}_n : a \neq 0 \text{ und } \text{ggT}(a, n) = 1\}$$

Zeige:

- (i) Ist p prim, so gilt $\phi(p) = p - 1$.
- (ii) Für jede Primzahl p und alle $k \in \mathbb{N}$ gilt $\phi(p^k) = p^{k-1}(p - 1)$.
- (iii) Sind p, q verschiedene Primzahlen, so gilt $\phi(pq) = \phi(p)\phi(q)$. *Hinweis:* Um $\phi(n)$ zu bestimmen, lohnt es sich oft zuerst die Anzahl der Zahlen in \mathbb{Z}_n , die *nicht* teilerfremd mit n sind, zu bestimmen.

2.13. Sei $p > 2$ eine ungerade Primzahl. Zeige:

$$\sum_{i=1}^{p-1} i^{p-1} \equiv -1 \pmod{p} \quad \text{und} \quad \sum_{i=1}^{p-1} i^p \equiv 0 \pmod{p}.$$

2.14. (Wie alt Daisy von Donald eigentlich ist?) Löse das folgende modulare Gleichungssystem:

$$\begin{aligned} x &\equiv 2 \pmod{3} \\ x &\equiv 2 \pmod{5} \\ x &\equiv 5 \pmod{7} \end{aligned}$$

2.15. Seien $1 \leq a \neq b \leq N$ zwei ganze Zahlen. Wieviele Primzahlen p mit $p \geq M$ mit $a \equiv b \pmod{p}$ kann dann es geben? *Hinweis:* Chinesischer Restsatz.

2.16. Zeige folgendes:

1. In einer Halbgruppe (H, \circ) gibt es höchstens ein neutrales Element e . (Also ist das neutrale Element einer Gruppe eindeutig.)
2. In einem Monoid (M, \circ) hat jedes Element $a \in M$ höchstens ein Inverses. (D.h. in einer Gruppe sind die Inversen eindeutig.)
3. In einer Gruppe (G, \circ) ist für alle $a, b \in G$ das Inverse $(a \circ b)^{-1} = a^{-1} \circ b^{-1}$.
4. In einer Gruppe gilt die Kürzungsregel, d.h. man kann von links und von rechts "kürzen".
5. In einer Halbgruppe (H, \circ) gilt das *allgemeine Assoziativitätsgesetz*: Seien $a_1, a_2, \dots, a_n \in H$. Dann liefert jede sinnvolle Beklammersung von $a_1 \circ a_2 \circ \dots \circ a_n$ denselben Wert. *Hinweis:* Zeige mit Induktion über n , dass die Menge P_n aller Werte der verschiedenen Beklammersungen von $a_1 \circ a_2 \circ \dots \circ a_n$ genau ein Element enthält.

2.17. (Putnam Exam, 1971) Sei M eine Menge und \circ eine binäre Operation auf M mit folgenden zwei Eigenschaften

$$(x \circ y) \circ z = (y \circ z) \circ x \tag{*}$$

und

$$x \circ x = x \tag{**}$$

für alle $x, y, z \in M$. Zeige, dass dann \circ auch kommutativ ist.

2.18. Sei (G, \circ) eine multiplikative Gruppe mit $a^2 = e$ für alle $a \in G$. Zeige, dass dann \circ auch kommutativ sein muss.

2.19. Zwei Gruppen (G, \circ) und $(H, *)$ sind *isomorph*, falls es eine bijektive Abbildung $f : G \rightarrow H$ mit $f(a \circ b) = f(a) * f(b)$ für alle $a, b \in G$ gibt.

Zeige, dass die Gruppen (\mathbb{R}_+, \cdot) und $(\mathbb{R}, +)$ isomorph sind. *Hinweis:* Betrachte die Logarithmus-Funktion $\ln x$.

2.20.

1. Zeige: (\mathbb{N}, \circ) mit $a \circ b = a^b$ ist keine Halbgruppe.
2. Ist die Potenzmenge einer Menge eine Gruppe bezüglich Vereinigung (oder bezüglich Schnitt)?
3. Sei M eine endliche Menge, S_M die Menge aller bijektiven Abbildungen von M auf M und bezeichne \circ die Komposition von Abbildungen. Dann ist (S_M, \circ) eine Gruppe mit der identischen Abbildung id_M als neutralem Element. Diese Gruppe ist genau dann abelsch, wenn M nicht mehr als 2 Elemente hat.

2.21. Zeige, dass in einem Ring $(R, +, \cdot)$ gilt: $0 \cdot a = 0 = a \cdot 0$ und $(-a) \cdot b = -(a \cdot b) = a \cdot (-b)$. Hinweis: $0 + 0 = 0$.

2.22. Zeige, dass jeder Körper nullteilerfrei ist, d.h. $ab = 0 \Rightarrow a = 0 \vee b = 0$.

2.23. Zeige, dass $\{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ ein Körper ist. Hinweis: $a^2 - 2b^2 \neq 0$, da $\sqrt{2}$ irrational ist.

2.24. Sei $(R, +, \cdot)$ ein kommutativer Ring (d.h. $a \cdot b = b \cdot a$ für alle $a, b \in R$) mit $|R| \geq 2$, in dem die Gleichung $a \cdot x = b$ für alle $a, b \in R$, $a \neq 0$ eine Lösung hat. Zeige, dass dann R ein Körper ist.

2.25. Man kann Ringe bis zum Körper erweitern, wie auch Körper bis zum anderen (größeren) Körper erweitern.

1. Erweiterung von \mathbb{Z} bis \mathbb{Q} . Betrachte \mathbb{Q} als die Menge aller Paare (a, b) mit $a, b \in \mathbb{Z}$ und $b \neq 0$ mit der Addition

$$(a, b) + (a', b') = (ab' + a'b, bb') \quad \text{oder} \quad \frac{a}{b} + \frac{a'}{b'} = \frac{ab' + a'b}{bb'}$$

und der Multiplikation

$$(a, b) \cdot (a', b') = (aa', bb') \quad \text{oder} \quad \frac{a}{b} \cdot \frac{a'}{b'} = \frac{aa'}{bb'}$$

Zeige, dass dann \mathbb{Q} ein Körper ist.

2. Erweiterung von \mathbb{R} bis \mathbb{C} . Betrachte \mathbb{C} als die Menge $\mathbb{R} \times \mathbb{R}$ mit der Verknüpfungen

$$\begin{aligned} (a, b) + (a', b') &= (a + a', b + b') \\ (a, b) \cdot (a', b') &= (aa' - bb', ab' + a'b). \end{aligned}$$

Zeige, dass dann \mathbb{C} ein Körper ist mit $\mathbf{1} = (1, 0)$ als neutralem Element hinsichtlich Multiplikation ist. Wie sieht dann die Lösung von $x^2 = 1$ in \mathbb{C} aus?

Kapitel 3

Einschub aus der Analysis

Contents

3.1	Endliche Folgen und Reihen	89
3.2	Unendlicher Folgen	98
3.2.1	Konvergenzkriterien für Folgen	101
3.2.2	Bestimmung des Grenzwertes	104
3.3	Unendliche Reihen	106
3.3.1	Konvergenzkriterien für Reihen	112
3.3.2	Anwendung: Warum Familiennamen aussterben?	117
3.3.3	Umordnungssatz	119
3.4	Grenzwerte bei Funktionen	123
3.5	Differentiation	124
3.6	Mittelwertsätze der Differentialrechnung	126
3.7	Approximation durch Polynome: Taylorentwicklung	129
3.8	Extremalstellen	131
3.9	Die Bachmann-Landau-Notation: klein o und groß O	132
3.10	Rekurrenzen*	140
3.10.1	Das Master Theorem	149
3.11	Aufgaben	151

3.1 Endliche Folgen und Reihen

Eine *Folge* (engl. *progression*) ist eine Funktion $f : \mathbb{N} \rightarrow \mathbb{R}$, deren Definitionsbereich gleich der Menge der natürlichen Zahlen \mathbb{N} ist. D.h. eine Folge ist eine (potentiell unendliche) Folge reeller Zahlen

$$f(0), f(1), f(2), \dots, f(n), \dots$$

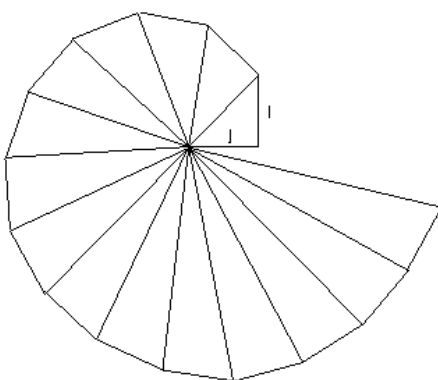
Gewöhnlich wird eine Folge f in der Form

$$\langle a_n \rangle = a_0, a_1, a_2, \dots$$

einfach aufgeschrieben, also als Abfolge der Folgenglieder $a_n = f(n)$. Der Funktionswert a_n heißt in diesem Zusammenhang auch das n -te *Folgenglied* der Folge. Man kann eine bestimmte Folge auf zwei Arten beschreiben:

1. Durch eine explizite Definition: Man gibt eine Formel an, aus der man jedes Glied sofort berechnen kann, z.B. $a_n = n^2$.
 2. Durch eine rekursive Definition: Zuerst gibt man das erste Glied a_0 (oder a_1) der Folge an, dann gibt man zusätzlich eine Formel an, mit der man das Glied a_{n+1} aus dem Glied a_n (oder aus den Gliedern a_0, a_1, \dots, a_n) berechnen kann.
- *Beispiel 3.1* : Das Rad des Theodorus (griechischer Gelehrter, 465 v.Chr.) ist eines der ersten Beispiele einer Rekursion. Die Konstruktion trägt seinen Namen, weil er mit ihrer Hilfe erstmals bewies, dass $\sqrt{3}, \sqrt{5}, \sqrt{7}, \dots$ irrationale Zahlen sind.

Dieses Rad kann durch einen rekursiven Algorithmus gebildet werden:



Die Bildungsregel lautet folgendermaßen:

D_1 Rechtwinkliges Dreieck mit Seitenlänge 1

D_2 Die Hypotenuse von D_1 ist ein Schenkel. Der andere Schenkel besitzt die Länge 1.

D_3 Die Hypotenuse von D_2 ist ein Schenkel. Der andere Schenkel besitzt die Länge 1.

D_4 Die Hypotenuse von D_3 ist ein Schenkel. Der andere Schenkel besitzt die Länge 1.

...

D_n Die Hypotenuse von D_{n-1} ist ein Schenkel. Der andere Schenkel besitzt die Länge 1.

In mathematischer Notation sieht die rekursive Bildungsregel so aus:

$$\begin{aligned} a_1 &= \sqrt{2} \\ a_{n+1} &= \sqrt{a_n^2 + 1} \end{aligned}$$

Gegeben sei eine Folge $\langle a_n \rangle = a_1, a_2, a_3, \dots$. Dann nennt man die Folge $\langle S_n \rangle = S_1, S_2, S_3, \dots$, deren Elemente S_n nach der Vorschrift

$$S_n = a_1 + a_2 + a_3 + \dots + a_n = \sum_{i=1}^n a_i$$

gebildet werden, die *Reihe* $\langle S_n \rangle$ der Folge $\langle a_n \rangle$. Prominente Reihen sind:

Arithmetische Reihe:

$$\sum_{k=1}^n k = 1 + 2 + 3 + \dots + n.$$

Geometrische Reihe:¹

$$\sum_{k=0}^n x^k = 1 + x + x^2 + \dots + x^n.$$

Harmonische Reihe:

$$\sum_{k=1}^n \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}.$$

Wie kann man eine "geschlossene" Form für die Funktion $S_n : \mathbb{N} \rightarrow \mathbb{R}$ mit $S_n = \sum_{k=0}^n a_k$ finden? Es gibt zwar eine sogenannte *Theorie der erzeugenden Funktionen*, die in vielen Fällen (aber nicht immer!) eine solche geschlossene Form für S_n finden lässt, das ist aber ein zeitaufwändiges Thema und wir werden diese Theorie nicht betrachten. Stattdessen werden wir uns auf ein Paar wichtiger Regeln beschränken.

Arithmetische Reihe

Zuerst betrachten wir die arithmetische Reihe $S_n = 1 + 2 + \dots + n$. Eine einfache (aber sehr kluge) Idee² ist die Reihenfolge von Zahlen umzukehren und die beiden Reihen aufzuaddieren:

$$\begin{array}{r} S_n = 1 + 2 + 3 + \dots + n-1 + n \\ + \\ S_n = n + n-1 + n-2 + \dots + 2 + 1 \\ \hline 2S_n = (n+1) + (n+1) + (n+1) + \dots + (n+1) + (n+1) \end{array}$$

Dies ergibt $2 \cdot S_n = n(n+1)$ und wir haben eine nützliche Formel bewiesen:

Satz 3.2. (Arithmetische Reihe)

$$S_n = \sum_{k=1}^n k = \frac{n(n+1)}{2}. \quad (3.1)$$

¹Der Grund, warum die Folge $a_k = x^k$ eine "geometrische Folge" heißt, ist dass der Betrag jedes Folgengliedes (für $n \geq 2$) das geometrische Mittel der beiden Nachbarn ist:

$$\frac{a_{n+1}}{a_n} = \frac{a_n}{a_{n-1}} \quad \Big| \cdot a_n a_{n-1}$$

$$\Rightarrow a_{n+1} a_{n-1} = a_n^2 \Rightarrow |a_n| = \sqrt{a_{n+1} a_{n-1}}.$$

²Dieser Trick hat Gauß als Kind erfunden. Um eine Ruhepause in der Klasse zu bekommen, hat der Lehrer die folgende Aufgabe gestellt: Wie groß ist die Summe $S = 1 + 2 + \dots + 100$ der ersten 100 Zahlen. Leider hat der Lehrer Pech gehabt: Bereits nach wenigen Minuten hat der kleine Gauß die Antwort 5050 gegeben.

Geometrische Reihe

Der folgender Trick ist als *Verschiebungs-Trick* bekannt: Ist $S_n = \sum_{k=0}^n a_k$, dann versuche die rechte Summe (*) in der Gleichung

$$S_n + a_{n+1} = a_0 + \overbrace{\sum_{k=0}^n a_{k+1}}^{(*)}$$

mittels S_n darstellen.

Um diesen Trick zu demonstrieren, betrachten wir die *geometrische Reihe*:

$$S_n = \sum_{k=0}^n x^k$$

Dann ist

$$S_n + x^{n+1} = 1 + \sum_{k=0}^n x^{k+1} = 1 + x \cdot \overbrace{\sum_{k=0}^n x^k}^{S_n} = 1 + x \cdot S_n$$

und wir erhalten:

$$S_n \cdot (x - 1) = x^{n+1} - 1$$

Damit haben wir eine weitere nützliche Formel bewiesen: ³

Satz 3.3. (Geometrische Reihe)

$$\sum_{k=0}^n x^k = \frac{x^{n+1} - 1}{x - 1} = \frac{1 - x^{n+1}}{1 - x} \quad \text{für } x \neq 1 \quad (3.2)$$

Manchmal lohnt es sich zu probieren, das n -te Folgenglied a_n als die Differenz $b_{n+1} - b_n$ von zwei aufeinanderfolgenden Folgenglieder einer anderen Folge b_n darzustellen.

Teleskop-Trick: Gegeben sei $S_n = \sum_{k=s}^t a_k$ mit $s \leq t$. Bestimme eine Folge $\langle b_k \rangle$, so dass für alle k gilt: $a_k = b_{k+1} - b_k$. Dann gilt: $S_n = b_{t+1} - b_s$.

Beweis.

$$\begin{aligned} S_n &= a_s + a_{s+1} + a_{s+2} + \dots + a_{t-1} + a_t \\ &= (b_{s+1} - b_s) + (b_{s+2} - b_{s+1}) + (b_{s+3} - b_{s+2}) + \dots + (b_t - b_{t-1}) \\ &= -b_s + (b_{s+1} - b_{s+1}) + (b_{s+2} - b_{s+2}) + (b_{s+3} - b_{s+3}) + \dots + (b_t - b_t) + b_{t+1} \\ &= b_{t+1} - b_s \end{aligned}$$

□

³Sie sollten sich diese Formel gut merken – sie taucht in vielen Situationen immer wieder auf!

▷ *Beispiel 3.4* : Wir wenden den Teleskop-Trick auf

$$S_n := \sum_{k=1}^n \frac{1}{k(k+1)}$$

an. Da

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1},$$

erhalten wir

$$S_n = \overbrace{\left(\frac{1}{1} - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \left(\frac{1}{3} - \frac{1}{4} \right) + \dots + \left(\frac{1}{n} - \frac{1}{n+1} \right)}^{\text{Teleskopsumme}} = 1 - \frac{1}{n+1}.$$

▷ *Beispiel 3.5* : Wir wenden den Teleskop-Trick auf $\sum_{k=0}^n x^k$ mit $x \neq 0$ an. Gesucht sind b_k mit $x^k = b_{k+1} - b_k$ für alle $k \geq 0$. Ansatz: $b_k = f(k)x^k$. Dann gilt:

$$x^k = f(k+1)x^{k+1} - f(k)x^k \iff 1 = f(k+1)x - f(k)$$

Der Ansatz $f(k) = f(k+1)$ ergibt $f(k) = \frac{1}{x-1}$ und damit $b_k = \frac{x^k}{x-1}$. Also gilt:

$$\sum_{k=0}^n x^k = b_{n+1} - b_0 = \frac{x^{n+1}}{x-1} - \frac{1}{x-1} = \frac{x^{n+1} - 1}{x-1}.$$

▷ *Beispiel 3.6* : (**Ratenzahlung**) Im täglichen Leben hat man meist nicht nur Einmalzahlungen zu leisten, sondern häufig werden gleiche Beiträge R in regelmäßigen Abständen (Zeiteinheiten wie Monat, Vierteljahr, Jahr, usw.) ein- oder ausgezahlt (Raten, Renten, Pensionen, usw.). Das eingezahlte Geld wird wieder mit $x\%$ verzinst.

Bei der *nachschüssigen* Rente erfolgt die Zahlung R am Ende des Zeiteinheits, bei der *vorschüssigen* Rente dagegen zu Beginn der Zeiteinheit. In Zusammenhang mit Renten interessiert man sich vor allem für den Gesamtwert, unter Berücksichtigung von Zinseszins, den eine Rente am Anfang bzw. Ende der Ratenzahlungen hat. Hierbei kommt es darauf an, ob die Rente nach- oder vorschüssig ist. Denn bei nachschüssiger Zahlung einer n -maligen Rente wird die erste Rentenzahlung nur $(n-1)$ -mal verzinst, bei vorschüssiger aber n -mal.

Nehmen wir an, Sie zahlen monatlich *am Anfang* des Monats (vorschüssige Zahlung) den Betrag R auf ein Konto ein und das Geld werde wieder mit $x\%$ verzinst (z.B. ein Festgeldkonto). So haben Sie *am Ende* des ersten Monats

$$K_1 = R \cdot q \quad \text{mit} \quad q := 1 + \frac{x}{100}.$$

Am Ende des zweiten Monats

$$K_2 = \underbrace{R \cdot q}_{\text{Geld von diesem Monat}} + \underbrace{R \cdot q^2}_{\text{Geld vom Vormonat}}$$

und allgemeiner am Ende der n -ten Monats

$$\begin{aligned} K_n &= Rq + Rq^2 + \dots + Rq^n \\ &= R \cdot (q + q^2 + \dots + q^n) \\ &= Rq(1 + q + q^2 + \dots + q^{n-1}) \quad \text{mit } q := 1 + \frac{x}{100} \end{aligned}$$

Bei der nachschüssigen Zahlungen ist

$$\begin{aligned} K_n &= R + Rq + Rq^2 \dots + Rq^{n-1} \\ &= R(1 + q + q^2 + \dots + q^{n-1}) \end{aligned}$$

Falls zu Beginn des Zeitraums auch noch ein Anfangskapital K_0 vorliegt (wie oft bei Ratensparverträgen), so ergeben sich als Endwerte entsprechend

$$\begin{aligned} K_n &= K_0q^n + Rq \frac{q^n - 1}{q - 1} && \text{(vorschüssiger Zahlung)} \\ K_n &= K_0q^n + R \frac{q^n - 1}{q - 1} && \text{(nachschüssiger Zahlung)} \end{aligned}$$

Angenommen, Sie schließen mit Ihrer Bank einen Ratensparvertrag über 10 Jahre und zu einem Zinsfuß von $x = 6\%$ ab. Zu beginn des ersten Jahres zahlen Sie einen Einmalbetrag von 1500 € ein und anschließend jeweils am Ende des Jahres eine Rate von 200 €. Welchen Wert hat das Kapital nach 10 Jahren?

Da es sich um nachschüssige Ratenzahlung handelt, ergibt sich als Endwert

$$K_{10} = 1500 \cdot 1,06^{10} + 200 \cdot \frac{1,06^{10} - 1}{1,06 - 1} \approx 5.322 \text{ €}$$

Das von Ihnen eingezahlte Kapital beträgt $1500 + 10 \cdot 200 = 3.500 \text{ €}$. Der Rest stammt aus den Zinseszinsen.

Löst man die Rentenendwertformeln (ohne Anfangskapital) nach R auf, so kann man bestimmen, welche jährliche Rate zu zahlen ist, um bei einem Zinsfuß von $x\%$ nach n Jahren ein gewünschtes Kapital K_n zu erhalten. Mit dem Aufzinsungsfaktor $q = 1 + \frac{x}{100}$ ergeben sich bei nach- und vorschüssiger Zahlung:

$$R = (K_n - K_0q^n) \cdot \frac{q - 1}{q^n - 1} \quad \text{und} \quad R = \frac{K_n - K_0q^n}{q} \cdot \frac{q - 1}{q^n - 1}$$

- *Beispiel 3.7*: Für ihre Wohnung haben Sie sich endlich den neuen Fernseher mit Videorekorder für $A \text{ €}$ gekauft. Bei einem monatlichen Zinssatz von $x\%$ zahlen Sie für n Monate pro Monat $R \text{ €}$ zurück und haben dann alles getilgt.

Frage: Wieviel müssen Sie monatlich bezahlen, d.h. wie groß ist R ?

Um dieses Problem zu behandeln, machen wir eine Art doppelte Buchführung, nämlich das verzinste Darlehen bei der Bank sowie die Gesamtsumme der gezahlten Beträge. Stichtag ist jeweils der 1. Tag im Monat nach der Einzahlung.

	Darlehen	eingezahlte Beträge
1. Monat	A	R
2. Monat	Aq	$Rq + R$
3. Monat	Aq^2	$Rq^2 + Rq + R$
⋮	⋮	⋮
k -ter Monat	Aq^{k-1}	$Rq^{k-1} + \dots + Rq + R$

wobei wieder $q = 1 + \frac{x}{100}$.

Wenn nach n Monaten alles bezahlt ist, muss

$$Aq^{n-1} = R(q^{n-1} + \dots + Rq + R) = R \frac{q^n - 1}{q - 1}$$

gelten. Also ist

$$R = A \cdot \frac{q^n - q^{n-1}}{q^n - 1}$$

ihr monatlich zu bezahlender Betrag.

Als Beispiel betrachten wir $A = 1000 \text{ €}$ und $n = 24$ bei $x = \frac{6}{12} = 0,5$; also $q = 1 + \frac{x}{100} = 1,005$. Dann müssen Sie monatlich $R = 44,10 \text{ €}$ einzahlen. In dieser Zeit haben Sie also $R \cdot 24 = 1058,- \text{ €}$ gezahlt.

- **Beispiel 3.8 : (Jahresrenten)** Sie gewinnen in einem Lotto 600.000 €. Der Lotteriehhaber macht Ihnen einen Vorschlag: statt diesen Betrag bereits jetzt bar auszuzahlen, können Sie 20 Jahre lang jedes Jahr 50.000 € ausgezahlt bekommen. In 20 Jahren hätten Sie also auf Ihrem Konto 1.000.000 € (eine Million!) Würden Sie einen solchen Vorschlag annehmen, wenn eine Ihnen bekannte Bank⁴ 8% Zinsen für Ihr Kapital jährlich (mit Garantie!) bezahlen bereit ist?

Der Grund, warum der Zinssatz p hier wichtig ist, ist folgender. Wenn wir 10 € heute mit Zinssatz p anlegen, so werden wir im nächsten Jahr $(1 + p) \cdot 10 = 10,80 \text{ €}$ haben, $(1 + p)^2 \cdot 10 \approx 11,66 \text{ €}$ in zwei Jahren, usw. Oder anders gesagt, 10 €, die sie im nächsten Jahr bekommen, sind heute nur $1/(1 + p) \cdot 10 \approx 9,26 \text{ €}$ wert. Der Grund: wenn wir heute 9,26 € anlegen, so haben wir in einem Jahr $(1 + p) \cdot 9,26 = 10 \text{ €}$.

Die Auszahlung $m = 50.000 \text{ €}$ am Anfang des ersten Jahres ist natürlich auch m Euro wert. Aber die Anzahlung am Anfang des nächsten Jahres ist nur $m/(1 + p)$ wert, die am Anfang des dritten Jahres ist nur $m/(1 + p)^2$ wert, usw. und die am Anfang des n -Jahres ist nur $m/(1 + p)^{n-1}$ wert. Die gesamte Auszahlung ist also in Wirklichkeit (betrachtet von heute) nur

$$V(n) = \sum_{i=1}^n \frac{m}{(1 + p)^{i-1}} \quad (3.3)$$

wert mit $p = 0,08$. Setzen wir $x := 1/(1 + p)$, so erhalten wir

$$\begin{aligned} V(n) &= m \cdot \sum_{j=0}^{n-1} x^j \\ &= m \cdot \frac{1 - x^n}{1 - x} \quad (\text{nach (3.2)}) \\ &= m \cdot \frac{1 - \left(\frac{1}{1+p}\right)^n}{1 - \frac{1}{1+p}} \\ &= m \cdot \frac{1 + p - \left(\frac{1}{1+p}\right)^{n-1}}{p}. \end{aligned}$$

Für $m = 50.000 \text{ €}$, $n = 20$ und $p = 0,08$ ergibt sich $V \approx 530.180,- \text{ €}$.

⁴Um der Realität nahe zu kommen, sollte man einen Zinssatz p mit $3\% \leq p \leq 5\%$ annehmen. Und mit der Garantie ist auch so eine Sache ...

Harmonische Reihe

Im Allgemeinen sind die Summen $S_n = \sum_{k=0}^n a_k$ sehr schwer *exakt* zu bestimmen. Andererseits reicht es in vielen Fällen insbesondere in der Laufzeitanalyse von Algorithmen, nur eine vernünftige *Abschätzung* für S_n zu finden.

Als Beispiel betrachten wir die Teilsummen der *harmonischen Reihe*:

$$H_n := 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} = \sum_{k=1}^n \frac{1}{k}. \quad (3.4)$$

Es ist für die Summe H_n keine geschlossene Form bekannt, die sie vereinfacht. Wir schätzen nun H_n gegen den Logarithmus ab. Dafür teilen wir die Summanden in Päckchen auf und zwar setzen wir

$$P_k := \left\{ \frac{1}{2^{k-1}}, \frac{1}{2^{k-1}+1}, \frac{1}{2^{k-1}+2}, \dots, \frac{1}{2^k-1} \right\}$$

D.h.

$$P_1 = \{1\}, P_2 = \left\{ \frac{1}{2}, \frac{1}{3} \right\}, P_3 = \left\{ \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7} \right\}, \dots$$

Ist K die Anzahl der vollen Päckchen, so muss $2^{K-1} \leq n < 2^K - 1$ gelten, woraus

$$\log_2(n+1) < K \leq \log_2 n + 1$$

folgt. Ein volles Päckchen P_k enthält $|P_k| = 2^k - 2^{k-1} = 2^{k-1}$ Zahlen. Die größte Zahl in P_k ist $\frac{1}{2^{k-1}}$, die kleinste ist $\frac{1}{2^k-1}$ und $|P_k| = 2^{k-1}$. Hieraus schließen wir (für jedes volle Päckchen)

$$\frac{1}{2} = |P_k| \cdot \frac{1}{2^k} < \sum_{x \in P_k} x \leq |P_k| \cdot \frac{1}{2^{k-1}} = 1$$

Aufsummiert erhalten wir

$$\frac{1}{2} \cdot \log_2(n+1) = \sum_{k=1}^K \frac{1}{2} < H_n \leq \sum_{k=1}^{K+1} 1 = \log_2 n + 2.$$

Genauer kann man sogar zeigen, dass

$$\ln n < H_n \leq \ln n + 1.$$

In gewissem Sinne ist diese Abschätzung nicht wesentlich schärfer als die oben angegebene. Der natürliche Logarithmus ist ein konstantes Vielfaches des Zweierlogarithmus und beide Abschätzungen sagen aus, dass die Summen H_n "asymptotisch" wie ein Logarithmus wachsen:

Satz 3.9. (Harmonische Reihe)

$$\ln n < H_n = \sum_{k=1}^n \frac{1}{k} \leq \ln n + 1 \quad (3.5)$$

► **Beispiel 3.10: (Laufzeit vom Quicksort Algorithmus)** Die (erwartete) Laufzeit von Quicksort Algorithmus kann man durch

$$T(n) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{j-i+1}$$

darstellen. Nun wollen wir diesen Ausdruck vereinfachen.

$$\begin{aligned} T(n) &= \sum_{i=1}^n \sum_{j=i+1}^n \frac{2}{j-i+1} = 2 \cdot \sum_{i=1}^n \sum_{j=i+1}^n \frac{1}{j-i+1} = 2 \cdot \sum_{i=1}^n \sum_{j=2}^{n-i+1} \frac{1}{j} \\ &\leq 2 \cdot \sum_{i=1}^n H_n \quad (H_n \text{ die harmonische Reihe}) \\ &= 2n \cdot H_n \leq 3(n \ln n) \end{aligned}$$

Der folgender Satz ermöglicht es, endliche Summen durch Integrale abzuschätzen. Eine Funktion $F(x)$ heißt *Stammfunktion* der Funktion $f(x)$, wenn $F'(x) = f(x)$ für alle x gilt. Der Hauptsatz der Differential- und Integralrechnung besagt, dass dann $\int_a^b f(t)dt = F(b) - F(a)$ gilt.

Satz 3.11. (Integral-Kriterium) Sei $F(x)$ eine Stammfunktion von $f : \mathbb{R} \rightarrow \mathbb{R}$.

(a) Wenn f monoton wachsend ist, dann gilt

$$F(n) - F(m-1) \leq \sum_{i=m}^n f(i) \leq F(n+1) - F(m).$$

(b) Wenn f monoton fallend ist, dann gilt

$$F(n+1) - F(m) \leq \sum_{i=m}^n f(i) \leq F(n) - F(m-1).$$

Beweis. Wir verifizieren nur Teil (a). Teil (b) folgt mit analogem Argument. Da f monoton wachsend ist, gilt $\int_i^{i+1} f(x) dx \leq f(i+1)$, denn die Fläche unter der Kurve $f(x)$, zwischen i und $i+1$, ist nach oben beschränkt durch die Fläche des Rechtecks zwischen i und $i+1$ mit der Höhe $f(i+1)$. Die Fläche dieses Rechtecks ist $f(i+1) \cdot 1 = f(i+1)$.

Es ist also

$$F(n) - F(m-1) = \int_{m-1}^n f(x) dx = \sum_{i=m-1}^{n-1} \int_i^{i+1} f(x) dx \leq \sum_{i=m-1}^{n-1} f(i+1) = \sum_{i=m}^n f(i).$$

Analog erhalten wir

$$f(i) \leq \int_i^{i+1} f(x) dx,$$

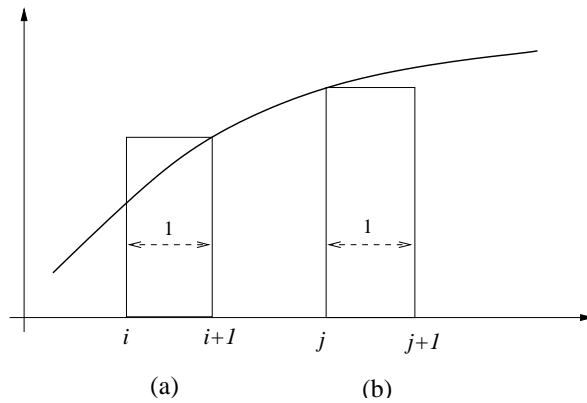


Abbildung 3.1: Fall (a) $\int_i^{i+1} f(x) dx \leq f(i+1)$ und Fall (b) $f(j) \leq \int_j^{j+1} f(x) dx$

denn f ist monoton wachsend und damit

$$\sum_{i=m}^n f(i) \leq \sum_{i=m}^n \int_i^{i+1} f(x) dx = \int_m^{n+1} f(x) dx = F(n+1) - F(m).$$

□

► *Beispiel 3.12*: Wir betrachten die Reihe $S_n := \sum_{k=1}^n k^a$ für $a \neq -1$. (Für $a = -1$ das ist die harmonische Reihe und wir haben bereits sie durch das Logarithmus abgeschätzt.) Sei $F(x) = x^{a+1}/(a+1)$. Da $F(x)' = x^a$, ist F eine Stammfunktion von x^a und wir erhalten als Konsequenz:

$$n^{a+1} \leq S_n \cdot (a+1) \leq (n+1)^{a+1}$$

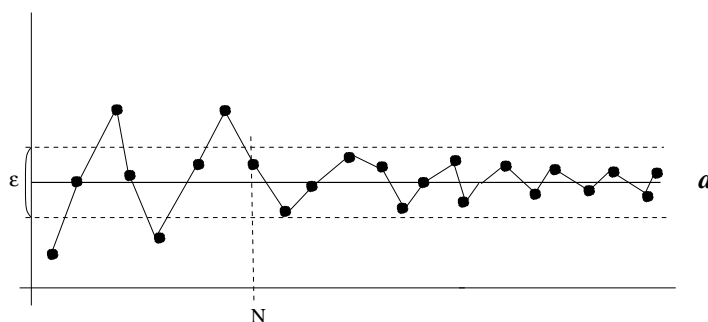
3.2 Unendlicher Folgen

Wir wollen das Verhalten einer *unendlichen* Folge

$$\langle a_n \rangle = a_0, a_1, a_2, \dots$$

betrachten. Insbesondere wollen wir wissen, ob sich die Folge irgendwann stabilisiert, d.h. ob es eine Zahl $a \in \mathbb{R}$ gibt, so dass für jede (beliebig kleine) Zahl $\epsilon > 0$ “fast alle” Glieder a_n um höchstens $\pm\epsilon$ von a entfernt sind. “Fast alle Glieder” heißt hier “alle außer vielleicht endlich vielen ersten Gliedern”. Gilt dies, so sagt man, dass die Folge gegen a *konvergiert* (oder *strebt*), und nennt diese Zahl den *Grenzwert* von $\langle a_n \rangle$. D.h. die Folge $\langle a_n \rangle$ strebt gegen a , wenn die folgende Aussage gilt:

$$\forall \epsilon > 0 \exists N \in \mathbb{N} \forall n \geq N : |a_n - a| < \epsilon \quad (3.6)$$



Um die sprachlichen Formulierungen zu vereinfachen, vereinbaren wir nun, dass die Aussage
für *fast alle* n gilt $P(n)$

bedeutet, dass

Eigenschaft $P(n)$ für *alle, bis auf endlich viele* n gilt

Also konvergiert eine Folge $\langle a_n \rangle$ gegen a genau dann, wenn fast alle Folgenglieder in der ϵ -Umgebung $U_\epsilon(a) := \{x \in \mathbb{R} : |x - a| < \epsilon\}$ von a liegen und $\epsilon > 0$ kann beliebig klein gemacht werden. Für die Konvergenz einer Folge sind also nur die “hinteren Glieder” verantwortlich, was am Anfang passiert ist egal.

Wiederum anders formuliert: Eine Folge konvergiert gegen einen Grenzwert, falls es

zu jeder *Toleranzgrenze* $\epsilon > 0$ einen *Schwellenwert* $N \in \mathbb{N}$ gibt, ab dem die Folge die vorgegebene *Genauigkeit* $|a_n - a| < \epsilon$ erzielt.

Die Negation der obigen Aussage (3.6) ist

$$\exists \epsilon > 0 \forall N \in \mathbb{N} \exists n \geq N : |a_n - a| \geq \epsilon$$

D.h. die Folge $\langle a_n \rangle$ strebt *nicht* gegen a , wenn es eine Umgebung $U_\epsilon(a)$ gibt, so dass *unendlich viele* Folgenglieder *außerhalb* dieser Umgebung liegen.

Diese Begriffsbestimmung (Konvergenz) ist die Grundlage der Analysis! Seine Bedeutung beruht darauf, dass viele Größen nicht durch einen in endlich vielen Schritten *exakt* berechenbaren Ausdruck gegeben, sondern nur mit beliebiger Genauigkeit *approximiert* werden können.

Eine wichtige Eigenschaft konvergierender Folgen ist, dass der Grenzwert *eindeutig* bestimmt ist.

Behauptung 3.13. Jeder konvergente Folge $\langle a_n \rangle$ hat *genau einen* Grenzwert a . Dieser Grenzwert wird mit $a = \lim_{n \rightarrow \infty} a_n$ bezeichnet.

Beweis. Sind a und a' Grenzwerte ein und derselben Folge $\langle a_n \rangle$ und $a \neq a'$, dann wählen wir $\epsilon := \frac{|a - a'|}{2}$. Ab einem Schwellenwert $N \in \mathbb{N}$ liegen alle Folgenglieder a_n mit $n > N$ in der ϵ -Umgebung von a . Es gibt auch einen Schwellenwert N' , ab dem alle Folgenglieder in der ϵ -Umgebung von a' liegen. Ab dem größeren der beiden Schwellenwerte müssen die Folgenglieder also im Durchschnitt der beiden ϵ -Umgebungen liegen. Der Durchschnitt ist aber leer: Gebe es ein x in beiden Umgebungen, so käme man zu der unsinnigen Abschätzung:

$$|a - a'| = |a - x + x - a'| \leq |x - a| + |x - a'| < \epsilon + \epsilon = |a - a'|.$$

□

Im Umgang mit dem Begriff der Konvergenz benutzt man ein wichtiges Prinzip, das als *Archimedisches Prinzip* bekannt ist:

Zu jeder reellen Zahl $x \in \mathbb{R}$ gibt es ein $n \in \mathbb{N}$ mit $x < n$.

Insbesondere erlaubt dieses Prinzip zu zeigen, dass eine Folge $\langle a_n \rangle$ eine *Nullfolge* ist, d.h. gegen 0 konvergiert. Der Grund, warum Nullfolgen wichtig sind, ist offensichtlich: $\lim_{n \rightarrow \infty} a_n = a$ genau dann, wenn $\langle a_n - a \rangle$ eine Nullfolge ist.

▷ *Beispiel 3.14*: (Das fundamentale Beispiel) Die Folge $a_n = 1/n$, $n \geq 1$, ist eine Nullfolge:

$$\lim_{n \rightarrow \infty} (1/n) = 0.$$

Beweis: Sei $\epsilon > 0$. Dann gibt es eine natürliche Zahl N mit $N > 1/\epsilon$. Wir wählen ein solches N und erhalten für alle $n \geq N$ ebenfalls $n > 1/\epsilon$ und daher $|1/n - 0| = 1/n < \epsilon$.

▷ *Beispiel 3.15*: Die Folge $a_n = x^n$ mit $x \in \mathbb{R}$ und $|x| < 1$ ist eine Nullfolge, d.h. $\lim_{n \rightarrow \infty} x^n = 0$ für $|x| < 1$.

Beweis: Sei $\epsilon > 0$ und $q = 1/|x|$. Da $q > 1$, gibt es eine natürliche Zahl N mit $q^N > \frac{1}{\epsilon}$. Wir wählen eine solche Zahl N und erhalten für alle $n \geq N$ ebenfalls $q^n > 1/\epsilon$ und daher $|1/q^n - 0| = 1/q^n < \epsilon$.

Eine divergente Folge $\langle a_n \rangle$ ist “böartig”, da sie keinen Grenzwert besitzt, d.h. die Folge keinen einzelnen “Häufungspunkt” hat. Soll man sofort solche Folgen wegwerfen? Nicht unbedingt – es kann sein, dass die Folge trotzdem “stabil genug” ist, wenn sie nur wenige verschiedene “Häufungspunkte” hat.

Ist $\langle a_n \rangle$ eine Folge und $0 \leq n_0 < n_1 < n_2 \dots$ eine *unendliche* Folge von natürlichen Zahlen, dann wird $(a_{n_k}) = a_{n_0}, a_{n_1}, a_{n_2}, \dots$ eine *Teilfolge* von $(a_n)_{n \in \mathbb{N}}$ genannt.

Zahl $a \in \mathbb{R}$ heißt *Häufungspunkt* von $\langle a_n \rangle$, wenn es eine Teilfolge $\langle a_{n_k} \rangle$ gibt mit $\lim_{k \rightarrow \infty} a_{n_k} = a$.

In anderen Worten, a ist ein Häufungspunkt von $\langle a_n \rangle$, wenn in jeder Umgebung von a *unendlich viele* Folgenglieder liegen.⁵ Der größte Häufungspunkt heißt *Limes-Superior* und wird als

$$\limsup a_n \quad \text{oder} \quad \overline{\lim} a_n$$

bezeichnet. Der kleinste Häufungspunkt heißt *Limes-Inferior* und wird mit

$$\liminf a_n \quad \text{oder} \quad \underline{\lim} a_n$$

bezeichnet. Bei konvergenten Folgen stimmen beide überein: $\limsup a_n = \liminf a_n$.

▷ *Beispiel 3.16*: Betrachte die Folge $\langle a_n \rangle$ mit $a_n = (-1)^n + \frac{1}{n}$. Dann strebt die “gerade” Teilfolge $\langle a_{2k} \rangle$ gegen 1, während die “ungerade” Teilfolge $\langle a_{2k+1} \rangle$ gegen -1 strebt. Die Häufungspunkte sind also $\limsup a_n = 1$ und $\liminf a_n = -1$.

Satz 3.17. (Bolzano–Weierstraß) Jede beschränkte Folge reeller Zahlen hat mindestens einen Häufungspunkt.

⁵Zur Erinnerung: a ist ein Grenzwert von $\langle a_n \rangle$, wenn in jeder Umgebung von a *fast alle* (d.h. alle, bis auf endlich viele) Folgenglieder liegen.

Beweis. Ist die Folge $\langle a_n \rangle$ beschränkt, so gibt es zwei Zahlen $a_0 \leq b_0$ mit der Eigenschaft, dass fast alle Folgenglieder in dem Intervall $[a_0, b_0]$ liegen. Wir halbieren das Intervall $[a_0, b_0]$. In einer der beiden Hälften sind unendlich viele Folgenglieder, die eine Teilfolge von $\langle a_n \rangle$ bilden. Diese Hälfte sei $[a_1, b_1]$. Setzen wir das fort, so erhalten wir eine Intervallschachtelung mit gemeinsamen Punkt h . Dies h ist Häufungspunkt. Ist nämlich $\epsilon > 0$ gegeben, so gibt es ein N mit $|a_N - b_N| < \epsilon$ und $h \in [a_N, b_N]$. Nach Konstruktion sind in $[a_N, b_N] \subset (h - \epsilon, h + \epsilon)$ unendlich viele Folgenglieder. \square

3.2.1 Konvergenzkriterien für Folgen

Natürlich möchten wir den Grenzwert einer Folge bestimmen. Aber zuerst ist zu klären, ob die Folge überhaupt konvergiert und hier helfen uns sogenannte *Konvergenzkriterien*.

Die Folge $\langle a_n \rangle$ heißt *beschränkt*, falls es ein $L \in \mathbb{R}$ mit $|a_n| \leq L$ für alle n gibt.

Es ist klar, dass unbeschränkte Folgen keinen (endlichen) Grenzwert haben können. Also sind solche Folgen besonders "böartig" und man kann solche Folgen sofort weg werfen. Was aber wenn eine Folge beschränkt ist? Dann muss sie auch nicht unbedingt einen Grenzwert haben: Nimm zum Beispiel $a_n = (-1)^n$.

Ist die Folge *monoton fallend* ($a_n \geq a_{n+1}$ für alle n) oder *monoton wachsend* ($a_n \leq a_{n+1}$ für alle n), so können die Folgenglieder nicht hin und hier springen – sie können sich nur in einer Richtung (nur nach oben oder nur nach unten) bewegen. Natürlich, reicht die Monotonie alleine noch nicht: z.B. ist die Folge $a_n = n$ (streng) monoton wachsend aber $\lim a_n = \infty$. Der Grund hier ist, dass diese Folge *unbeschränkt* wächst.

Beide Eigenschaften – Monotonie und Beschränkung – haben allein keinen bzw. nur einen geringen Bezug zur Konvergenz, ihre Kombination ist aber überraschenderweise sehr mächtig und liefert ein oft benutztes Konvergenzkriterium.

Satz 3.18. (Monotonie-Kriterium) Ist die Folge $\langle a_n \rangle$ monoton, so gilt:

$$\langle a_n \rangle \text{ ist konvergent} \iff \langle a_n \rangle \text{ ist beschränkt}$$

Beweis. (\Leftarrow): Sei $a_0 \leq a_1 \leq a_2 \leq \dots$ eine monoton wachsende und beschränkte Folge (der Fall einer monoton fallender Folge ist analog). Dann gibt es eine obere Schranke $L \in \mathbb{R}$ mit $a_n \leq L$ für alle n und sogar eine *kleinste* obere Schranke a mit der Eigenschaft: ⁶ für jedes $\epsilon > 0$ gibt es ein $N = N(\epsilon)$ mit $a - a_N < \epsilon$. Damit gilt für alle $n \geq N$:

$$|a - a_n| = a - a_n = \underbrace{(a - a_N)}_{< \epsilon} - \underbrace{(a_n - a_N)}_{\geq 0} < \epsilon.$$

Da dies für beliebig kleines $\epsilon > 0$ gilt, strebt $\langle a_n - a \rangle$ gegen 0 und deshalb muss a_n gegen a streben.

(\Rightarrow) Dieser Richtung gilt für beliebige (nicht nur für monotone) Folgen: Ist die Folge $\langle a_n \rangle$ konvergent, so ist sie beschränkt. Sei $\lim a_n = a$. Dann gibt es N mit $|a_n - a| < 1$, d.h. $a - 1 < a_n < a + 1$ für alle $n \geq N$. Setzen wir $b = \max\{a - 1, a + 1\}$, so ist $|a_n| \leq b$ für alle $n \geq N$. Damit ist aber auch $|a_n| \leq L$ für alle n , wenn wir L als $L = b + |a_0| + \dots + |a_N|$ nehmen. \square

⁶D.h. die Folgenglieder beliebig nahe zu a kommen.

Es gibt auch ein paar anderen nützlichen Konvergenzkriterien.

Satz 3.19.

1. **Majorantenkriterium für Nullfolgen:** Ist $\lim_{n \rightarrow \infty} a_n = 0$ und $|b_n| \leq L \cdot |a_n|$ für $L > 0$ und fast alle n , dann gilt $\lim_{n \rightarrow \infty} b_n = 0$. Die Folge $\langle a_n \rangle$ heißt dann die *Majorante* für die Folge $\langle b_n \rangle$.
2. Ist $\langle a_n \rangle$ beschränkt und $\lim_{n \rightarrow \infty} b_n = 0$, dann gilt $\lim_{n \rightarrow \infty} a_n \cdot b_n = 0$.
3. **Vergleichskriterium:** Seien $\langle b_n \rangle$ und $\langle c_n \rangle$ zwei Folgen mit demselben Grenzwert b . Gilt $b_n \leq a_n \leq c_n$ für fast alle n , so gilt $\lim_{n \rightarrow \infty} a_n = b$.

Beweis. Zu 1: Sei $\epsilon > 0$ beliebig. Da $\langle a_n \rangle$ eine Nullfolge ist, gibt es einen Schwellenwert N , so dass $|a_n| < \epsilon/L$ für alle $n \geq N$. Dann gilt auch für alle $n \geq N$: $|b_n| \leq L \cdot |a_n| < L \cdot (\epsilon/L) = \epsilon$.

Zu 2: Da $\langle a_n \rangle$ beschränkt ist, gibt es ein $L > 0$, so dass $|a_n| \leq L$ gilt. Sei nun $\epsilon > 0$ beliebig. Da $\langle b_n \rangle$ eine Nullfolge ist, gibt es einen Schwellenwert N , so dass $|b_n| < \epsilon/L$ für alle $n \geq N$ gilt. Dann gilt auch für alle $n \geq N$: $|a_n \cdot b_n| = |a_n| \cdot |b_n| < L \cdot (\epsilon/L) = \epsilon$.

Den Beweis des Vergleichskriteriums lassen wir als Übungsaufgabe. □

▷ *Beispiel 3.20:* Zu zeigen:

$$\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1. \quad (3.7)$$

Dazu betrachten wir die Folge $b_n = \sqrt[n]{n} - 1$ mit $n \geq 1$. Nach dem binomischen Lehrsatz (siehe Satz 1.41) gilt für jedes $n \geq 2$:

$$n = (1 + b_n)^n = \sum_{i=0}^n \binom{n}{i} b_n^i > \binom{n}{2} b_n^2 = \frac{n(n-1)}{2} \cdot b_n^2$$

und daraus durch Umstellung

$$b_n^2 < \frac{2}{n-1} \leq 4 \cdot \frac{1}{n}$$

Da $1/n$ eine Nullfolge ist, ist auch $\lim b_n = 0$, woraus $\lim \sqrt[n]{n} = \lim(1 + b_n) = 1$ folgt.

Genauso kann man

$$\lim_{n \rightarrow \infty} \sqrt[n]{n^c} = 1 \quad (3.8)$$

für jedes $c \in \mathbb{R}$, $c \geq 0$ zeigen.

▷ *Beispiel 3.21:* Sei $c \geq 1$ eine ganze Zahl und betrachten die Folge $a_n = \sqrt[n]{c}$. Dann gilt $1 \leq \sqrt[n]{c} \leq \sqrt[n]{n}$, wenn $n \geq c$ ist. Der Grenzwert von $\sqrt[n]{n}$ ist 1 (siehe das vorige Beispiel). Auf der linken Seite haben wir eine 1, also ist der Grenzwert auch 1. Nach dem Vergleichskriterium gilt: $\lim_{n \rightarrow \infty} \sqrt[n]{c} = 1$.

▷ *Beispiel 3.22:* Dieses Beispiel soll zeigen, dass man unter Umständen die Glieder einer Nullfolge mit dem Gliedern einer *unbeschränkten* Folge multiplizieren darf, ohne die Nullfolgeneigenschaft zu verlieren.

Wir betrachten die beiden Zahlenfolgen $\langle x^n \rangle = 1, x, x^2, x^3, \dots$ und $\langle n \rangle = 0, 1, 2, 3, \dots$ mit $0 < |x| < 1$, und bilden daraus die Folge $\langle a_n \rangle$ mit $a_n = nx^n$. Indem wir $|x|$ als $\frac{1}{1+q}$ für ein $q \in \mathbb{R}$ schreiben, folgt für $n \geq 2$:

$$\begin{aligned} |nx^n| &= \frac{n}{(1+q)^n} = \frac{n}{1 + \binom{n}{1}q + \binom{n}{2}q^2 + \dots + q^n} \\ &< \frac{n}{\binom{n}{2}q^2} = \frac{2}{(n-1)q^2} = \underbrace{\frac{2}{q^2}}_L \cdot \frac{1}{n-1}. \end{aligned}$$

Damit haben wir gezeigt, dass die Nullfolge $(\frac{1}{n-1})$ eine Majorante für die Folge $a_n = nx^n$ darstellt, und das Majorantenkriterium für Nullfolgen sagt uns, dass auch $\langle a_n \rangle$ eine Nullfolge ist. Die Folge $\langle \frac{1}{n-1} \rangle$ ist erst recht eine Majorante für die Folge $\langle x^n \rangle$, was uns das Nebenergebnis liefert, dass $\langle x^n \rangle$ für $|x| < 1$ eine Nullfolge darstellt.

Wenn man nicht zeigen kann, dass eine Folge konvergent ist, wie kann man dann zeigen, dass diese Folge *divergiert*? Dazu kann man das folgende Kriterium benutzen.

Satz 3.23. (Teilfolgenkriterium) Wenn eine Folge gegen a konvergiert, dann konvergiert *jede* ihre Teilfolge gegen a .

Beweis. Sei $(a_n)_{n \in \mathbb{N}}$ eine Folge und $(a_{n_k})_{k \in \mathbb{N}}$ ihre Teilfolge. Sei $a_n \rightarrow a$ und $\epsilon > 0$ beliebig klein. Dann gibt es ein m_0 , so dass $|a_n - a| < \epsilon$ für alle $n \geq m_0$. Wähle k_0 so, dass $n_{k_0} \geq m_0$. Dann ist $n_k \geq n_{k_0} \geq m_0$ für alle $k \geq k_0$ und damit auch $|a_{n_k} - a| < \epsilon$. \square

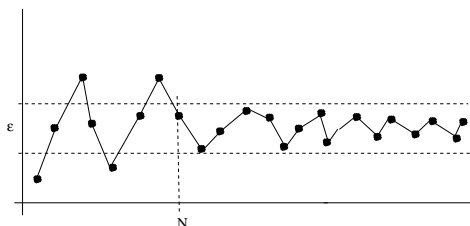
Wie kann man dieses Kriterium benutzen, um Divergenz einer Folge $\langle a_n \rangle$ zu zeigen? In vielen Fällen reicht es eine *monotone* und *unbeschränkte* Teilfolge von $\langle a_n \rangle$ zu finden. Dann muss auch nach Monotonie-Kriterium auch die Folge $\langle a_n \rangle$ selbst divergieren!

Konvergente Folgen nähern sich ihrem Grenzwert beliebig nahe an. Das hat Konsequenzen für die Beziehung der Folgenglieder untereinander: Auch sie müssen beliebig nahe aneinander rücken. Diese Eigenschaft nennt man auch das Cauchy-Kriterium.

Eine Folge $\langle a_n \rangle$ reeller Zahlen heißt *Cauchy-Folge*, wenn

$$\forall \epsilon > 0 \exists N \forall n, m > N : |a_n - a_m| < \epsilon.$$

Man nennt diese Eigenschaft "Konvergenz in sich". Informell: Egal wie klein die (von ϵ bestimmte) Umgebung ist, kommt die Folge irgendwann mal hinein und bleibt für immer da.





Cauchy-Kriterium

Eine Folge reeller Zahlen ist genau dann konvergent, wenn sie eine Cauchy-Folge ist.

Die Behauptung “konvergiert \Rightarrow Cauchy-Folge” ist trivial. Weniger trivial ist die andere Richtung “Cauchy-Folge \Rightarrow konvergiert”: Es ist nicht sofort klar, warum sich eine beliebige, hin und hier springende Cauchy-Folge irgendwann mal doch stabilisieren muss.

Beweis. (\Rightarrow): Sei $\langle a_n \rangle$ konvergent mit Grenzwert a . Dann gibt es zu $\epsilon > 0$ ein $N = N(\epsilon)$ mit $|a_n - a| < \epsilon$ für alle $n > N$. Es folgt

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| < 2\epsilon$$

für alle $n, m > N$ (Dreiecksungleichung). Das ist “Cauchy” mit 2ϵ für ϵ .

(\Leftarrow): Sei $\langle a_n \rangle$ eine Cauchy-Folge. Dann gibt es $N \in \mathbb{N}$ mit $|a_n - a_m| < 1$ für alle $n, m \geq N$. Damit ist für alle $n \geq N$

$$|a_n| = |a_n - a_N + a_N| \leq |a_n - a_N| + |a_N| < 1 + |a_N|$$

Setze $s = \max\{|a_0|, |a_1|, \dots, |a_{N-1}|, 1 + |a_N|\}$. Dann gilt $|a_k| \leq s$ für alle $k \in \mathbb{N}$, d.h. die Folge $\langle a_n \rangle$ ist beschränkt. Damit hat die Folge $\langle a_n \rangle$ einen Häufungspunkt b (nach dem Satz von Bolzano-Weierstraß). Wir werden zeigen, dass b der Grenzwert von $\langle a_n \rangle$ ist.

Sei $\epsilon > 0$ gegeben. Es gibt $N \in \mathbb{N}$ mit $|a_n - a_m| < \epsilon/2$ für alle $n, m \geq N$ (Cauchy-Folge). Andererseits (da b ein Häufungspunkt ist) gibt es für jedes $n \geq N$ ein $m \geq n$ mit $|a_m - b| < \epsilon/2$. Damit gilt für jedes $n \geq N$

$$|a_n - b| = |a_n - a_m + a_m - b| \leq |a_n - a_m| + |a_m - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

Vorsicht mit dem Cauchy-Kriterium: Aus der schwächeren Eigenschaft $|a_{n+1} - a_n| < \epsilon$ für alle $n > N(\epsilon)$ kann *nicht* auf die Konvergenz von $\langle a_n \rangle$ geschlossen werden! So gilt etwa für $H_n = \sum_{k=1}^n \frac{1}{k}$ wegen $H_{n+1} - H_n = \frac{1}{n+1}$ sicherlich $\lim_{n \rightarrow \infty} (H_{n+1} - H_n) = 0$. Aber wir haben bereits gezeigt (siehe 3.5), dass $H_n \geq \ln n$ gilt. D.h. die Folge H_n ist nicht beschränkt und somit (nach Monotonie-Kriterium) ist sie divergent.

3.2.2 Bestimmung des Grenzwertes

Soll der Grenzwert einer Folge $\langle a_n \rangle$ bestimmt werden, so muss zunächst die Konvergenz der Folge gezeigt werden. Dazu benutzt man eins der oben genannten Konvergenzkriterien. Angenommen, $a = \lim_{n \rightarrow \infty} a_n$ existiert. Man muss nun a bestimmen. Wir beschreiben eine allgemeine Vorgehensweise.

Definition: Eine reelle Funktion f heißt *stetig* in $a \in \mathbb{R}$, falls für jede Folge $\langle a_n \rangle$ mit $\lim_{n \rightarrow \infty} a_n = a$ gilt: $\lim_{n \rightarrow \infty} f(a_n) = f(a)$, d.h. $\lim_{n \rightarrow \infty} f(a_n) = f(\lim_{n \rightarrow \infty} a_n)$.

Intuitiv bedeutet die Stetigkeit von f , dass die Funktion keinen Sprung in a macht. So ist zum Beispiel die Funktion $f : [0, 1] \rightarrow \{0, 1\}$ mit

$$f(x) = \begin{cases} 0 & \text{falls } x \in [0, 1) \\ 1 & \text{falls } x = 1 \end{cases}$$

nicht stetig in $x = 1$. Die Funktion $f(x) = \frac{1}{x^2}$ ist überall stetig, außer im Punkt $x = 0$. Zum Beispiel stetig überall sind:

Polynome, e^x , \sqrt{x} , x^a mit $a > 0$, $n^x \sin(x)$, $\cos(x)$, usw.

Satz 3.24. Sei die Folge $\langle a_n \rangle$ rekursiv als $a_{n+1} = f(a_n)$ gegeben, wobei $f : \mathbb{R} \rightarrow \mathbb{R}$ eine stetige Funktion ist. Existiert der Grenzwert $a = \lim_{n \rightarrow \infty} a_n$, so gilt

$$f(a) = a.$$

D.h. der Grenzwert a ein Fixpunkt von $f(x)$ ist.

Beweis.

$$a = \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} f(a_n) \stackrel{\text{stetigkeit}}{=} f\left(\lim_{n \rightarrow \infty} a_n\right) = f(a).$$

□

► *Beispiel 3.25:* Sei $\langle a_n \rangle$ durch $a_0 = 2$ und $a_{n+1} = \sqrt{a_n}$ gegeben. Die Folge ist beschränkt: $1 \leq a_n \leq 2$ (Induktion). Da $a_n \geq 1$, gilt $a_{n+1} = \sqrt{a_n} \leq a_n$; damit ist die Folge auch monoton fallend. Mit dem Monotoniekriterium existiert $a = \lim_{n \rightarrow \infty} a_n$. Dann ist

$$a = \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} \sqrt{a_n} \stackrel{\text{stetigkeit}}{=} \sqrt{\lim_{n \rightarrow \infty} a_n} = \sqrt{a},$$

also muss der Grenzwert entweder 0 oder 1 sein, da dies die einzigen Lösungen der Gleichung $a = \sqrt{a}$ sind. Da alle Folgenglieder ≥ 1 sind, kann 0 nicht Grenzwert sein, also gilt $\lim_{n \rightarrow \infty} a_n = 1$.

► *Beispiel 3.26: (Existenz der Wurzel)* Sei $b \in \mathbb{R}$, $b > 0$ fest. Wir betrachten die Folge $\langle a_n \rangle$ mit $a_0 > 0$ und

$$a_{n+1} = \frac{1}{2} \left(a_n + \frac{b}{a_n} \right)$$

Zu zeigen: $\lim_{n \rightarrow \infty} a_n = \sqrt{b}$. Die Folge ist monoton fallend⁷ $a_n \geq a_{n+1}$ und beschränkt: Aus $a_0 > 0$ folgt $a_n > 0$ für alle n . Nach dem Monotoniekriterium existiert $a = \lim_{n \rightarrow \infty} a_n$. Dann ist mit

⁷Sei $f(x) = \frac{1}{2} \left(x + \frac{b}{x} \right)$. Dann gilt $f(x)^2 \geq b \iff x^4 + 2x^2b + b^2 \geq 4x^2b \iff x^4 - 2x^2b + b^2 \geq 0 \iff (x^2 - b)^2 \geq 0$. Also ist $f(x)^2 \geq b$ für alle x und somit auch $a_n^2 \geq b$ für alle $n \geq 1$. Deshalb gilt:

$$a_{n+1} - a_n = \frac{1}{2} \left(a_n + \frac{b}{a_n} \right) - a_n = \frac{\overbrace{b - a_n^2}^{\leq 0}}{2a_n} \leq 0$$

$$f(x) := \frac{1}{2} \left(x + \frac{b}{x} \right)$$

$$\begin{aligned} a &= \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} f(a_n) \\ &\stackrel{\text{stetigkeit}}{=} f\left(\lim_{n \rightarrow \infty} a_n\right) = f(a) \\ &= \frac{1}{2} \left(a + \frac{b}{a} \right) \end{aligned}$$

und damit gilt $a^2 = b$.

Häufig benutzt man bei der Bestimmung der Grenzwerte von Folgen nicht direkt die Definition, sondern führt die Konvergenz nach gewissen Regeln auf schon bekannte Folgen zurück. Dazu dienen die folgenden Regeln (deren Ableitung Regel ist als Übungsaufgabe gestellt).


Grenzwertregeln:

Seien $\langle a_n \rangle$ und $\langle b_n \rangle$ konvergente Folgen mit $\lim_{n \rightarrow \infty} a_n = a$ und $\lim_{n \rightarrow \infty} b_n = b$, dann gilt

1. $\lim_{n \rightarrow \infty} (a_n + b_n) = a + b$
2. $\lim_{n \rightarrow \infty} (a_n \cdot b_n) = a \cdot b$ Spezialfall: $\lim_{n \rightarrow \infty} (c \cdot a_n) = c \cdot a$
3. $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{a}{b}$ (falls $b \neq 0$ und $b_n \neq 0$ für $n > n_0$)
4. $\lim_{n \rightarrow \infty} |a_n| = |a|$
5. $\lim_{n \rightarrow \infty} \sqrt{|a_n|} = \sqrt{|a|}$

3.3 Unendliche Reihen

Die zu einer unendlichen Folge $\langle a_n \rangle = a_0, a_1, a_2, \dots$ gehörende *Reihe* ist die Folge $\langle S_n \rangle$ mit $S_n = \sum_{i=0}^n a_i$. Den Grenzwert der Teilsummen-Folge $\langle S_n \rangle$ bezeichnet man oft als $\sum_{n=0}^{\infty} a_n$.

 Diese (am häufigsten benutzte) Bezeichnung “ $\sum_{n=0}^{\infty} a_n$ ” ist nur eine Verabredung – unendliche Summen als solche sind nicht definiert! Hat aber die Folge $\langle S_n \rangle$ einen Grenzwert S , so kann man bereits $\sum_{n=0}^{\infty} a_n = S$ schreiben. Damit versteht man dann, dass “die Folge $\langle S_n \rangle$ mit $S_n = \sum_{i=0}^n a_i$ gegen S strebt”. Also ist

$$\sum_{n=0}^{\infty} a_n = \text{die Bezeichnung für } \lim_{n \rightarrow \infty} \sum_{n=0}^n a_n.$$

▷ *Beispiel 3.27*: Betrachtet man die (alternierende) Folge $a_n = (-1)^n$, so kann man nicht ohne weiteres die entsprechende Reihe als “unendliche Summe” hinschreiben: $\sum_{n=0}^{\infty} a_n = 1 - 1 + 1 - 1 + 1 - 1 + \dots$, da man sonst “zeigen” könnte, dass die Reihe gegen 0 konvergiert

$$\underbrace{(1 - 1)}_0 + \underbrace{(1 - 1)}_0 + \underbrace{(1 - 1)}_0 + \dots = 0$$

wie auch gegen 1

$$1 + \underbrace{(-1 + 1)}_0 + \underbrace{(-1 + 1)}_0 + \underbrace{(-1 + 1)}_0 + \dots = 1,$$

was natürlich ein Schwachsinn ist. Deshalb muss man die entsprechende Reihe als die Folge (S_n) mit $S_n = \sum_{n=0}^n (-1)^n$ betrachten. Dann haben wir: $S_0 = +1$, $S_1 = 0$, $S_2 = +1$, $S_3 = 0$, $S_4 = +1$, usw. – eine Folge, die offensichtlich divergiert: sie springt immer zwischen 1 und 0.

Aus dem Monotonie-Kriterium für Folgen (Satz 3.18) ergibt sich das folgende Konvergenzkriterium für Reihen.

Satz 3.28. Majorantenkriterium für Reihen: Gilt $0 \leq a_k \leq b_k$ für fast alle k und konvergiert die Reihe $\sum_{k=0}^{\infty} b_k$, so konvergiert auch die Reihe $\sum_{k=0}^{\infty} a_k$.

Beweis. Nach dem Monotonie-Kriterium für Folgen (Satz 3.18) ist die konvergente Folge $B_n = \sum_{k=0}^n b_k$ beschränkt ($b_k \geq 0$). Folglich (da $a_k \leq b_k$) ist auch die Folge $A_n = \sum_{k=0}^n a_k$ beschränkt und damit muss sie auch konvergieren (widerum nach dem Monotonie-Kriterium für Folgen). \square

Geometrische Reihe

Wir wissen bereits, dass

$$\sum_{k=0}^n x^k = \frac{1 - x^{n+1}}{1 - x} = \frac{1}{1 - x} - \frac{x^{n+1}}{1 - x}$$

für beliebiges $x \neq 1$ gilt. Da für $|x| < 1$ die Folge $a_n = x^n$ eine Nullfolge ist (siehe Beispiel 3.15), erhalten wir (mit der Benutzung der Grenzwertregeln):

Unendliche geometrische Reihe

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x} \quad \text{falls } |x| < 1 \quad (3.9)$$

► *Beispiel 3.29 :*

$$0,999999999\dots = 0,9 \sum_{k=0}^{\infty} (1/10)^k = 0,9 \cdot \frac{1}{1 - (1/10)} = 0,9 \cdot \frac{10}{9} = 1.$$

► *Beispiel 3.30 :*

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = \sum_{k=0}^{\infty} (1/2)^k = \frac{1}{1 - (1/2)} = 2.$$

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \dots = \sum_{k=0}^{\infty} (-1/2)^k = \frac{1}{1 - (-1/2)} = 2/3.$$

Wir betrachten die folgende Reihe:

$$\sum_{i=1}^n i \cdot x^i = x + 2x^2 + 3x^3 + \dots + nx^n.$$

Das ist *keine* geometrische Reihe, da der Quotient a_{i+1}/a_i nicht konstant ist. In solchen Fällen kann uns eine unserer alten Freundinnen – die Differenzierung – helfen. In unserem Fall können wir zum Beispiel beide Seiten der uns schon bekannten Gleichung (3.2)

$$\sum_{i=0}^n x^i = \frac{1 - x^{n+1}}{1 - x}$$

differenzieren und erhalten

$$\left(\sum_{i=0}^n x^i \right)' = \left(\frac{1 - x^{n+1}}{1 - x} \right)'$$

Also folgt ⁸

$$\begin{aligned} \sum_{i=1}^n ix^{i-1} &= \frac{-(n+1)x^n(1-x) - (-1)(1-x^{n+1})}{(1-x)^2} \\ &= \frac{-(n+1)x^n + (n+1)x^{n+1} + 1 - x^{n+1}}{(1-x)^2} \\ &= \frac{1 - (n+1)x^n + nx^{n+1}}{(1-x)^2}. \end{aligned}$$

Die Differenzierung verändert die Exponenten von x in jedem Term. In unserem Fall haben wir eine geschlossene Form nicht für $\sum_{i=1}^n ix^i$ sondern für $\sum_{i=1}^n ix^{i-1}$ ausgerechnet (der i -te Term in der letzten Summe ist x^{i-1} , nicht x^i). Nichtsdestotrotz ist die Lösung (mindestens in diesem Fall) ziemlich einfach: multipliziere beide Seiten mit x . Dies ergibt:

$$\sum_{i=1}^n ix^i = \frac{x - (n+1)x^{n+1} + nx^{n+2}}{(1-x)^2}. \quad (3.10)$$

Obwohl dieser Ausdruck auf dem ersten Blick nicht viel einfacher aussieht, hat er aber eine schöne Eigenschaft: für beliebiges (aber festes) $x \in \mathbb{R}$ mit $|x| < 1$ “strebt” die rechte Seite gegen $x/(1-x)^2$, wenn $n \rightarrow \infty$ (da für $|x| < 1$ gilt $\lim_{n \rightarrow \infty} x^n = 0$). Damit haben wir noch eine nützliche Gleichung erhalten:

Verallgemeinerte geometrische Reihe

$$\sum_{k=0}^{\infty} k x^k = \frac{x}{(1-x)^2}, \quad \text{falls } |x| < 1 \quad (3.11)$$

⁸Zur Erinnerung: $(x^n)' = nx^{n-1}$ und $\left(\frac{f(x)}{g(x)} \right)' = \frac{f(x)' \cdot g(x) - f(x) \cdot g(x)'}{g(x)^2}$.

► **Beispiel 3.31 : (Unendliche Jahresrenten)** Ein junger Familienvater hat einen Jackpot von 10.000.000 € gewonnen. Das ist schon eine Menge Geld, man kann damit was anfangen. Wir nehmen wieder an, dass der Zinssatz bei $p = 8\%$ liegt (und zwar garantiert). Der Mann ist aber ein guter Familienvater und will, dass von diesem Geld nicht nur er, sondern seine Kinder, Enkelkinder usw. profitieren könnten. Der Lottoinhaber weiß das und macht einen Vorschlag: statt diesen Betrag bereits jetzt bar auszuzahlen, will er (und später seine Nachfolger) für alle $k = 1, 2, \dots$ (bis zu unendlich!) am Ende k -tes Jahres einen Betrag von $k \cdot m$ Euro mit $m = 50.000$ € zahlen. D.h. im 1. Jahr bekommt der Vater 50.000 €, im 2. Jahr bereits 100.000 €, im 3. Jahr 150.000 € usw. Ab dem 10. Jahr soll der Familienvater bereits eine halbe Million und weiter noch mehr (jährlich!) bekommen, ab dem 20. Jahr bereits eine runde Million und mehr (wiederum jährlich) bekommen usw. Falls dem Vater etwas passieren sollte, werden die Kinder jedes Jahr mehr als eine Million Euro pro Jahr bekommen, und spätere Enkelkinder sollten mehrere Millionen jährlich ausbezahlt bekommen. So rechnet der Mann: schon die Enkelkinder sollten viel mehr jährlich bekommen, als er von heutigem Gewinn für Zukunft verteilen könnte. Klingt sehr attraktiv: der Betrag kann *unbeschränkt* wachsen – ein Paradies für seine Nachfolger! Es ist doch schwer zu glauben, dass der Familienvater besser die heutigen 10.000.000 € + Zinseszins wählen sollte. Ist das so?

Die Antwort ist: Der Familienvater sollte das Angebot besser ablehnen. In unserem Fall hat die gesamte Auszahlung in Wirklichkeit (betrachtet von heute) nur den Wert (siehe (3.3)):

$$\begin{aligned} V &= \sum_{k=1}^{\infty} \frac{k \cdot m}{(1+p)^k} \\ &= m \cdot \frac{\frac{1}{1+p}}{\left(1 - \frac{1}{1+p}\right)^2} \quad (\text{nach (3.11) mit } x = \frac{1}{1+p}) \\ &= m \cdot \frac{1+p}{p^2} \quad (\text{multipliziere mit } (1+p)^2). \end{aligned}$$

Für $m = 50.000$ € und $p = 8\% = 0,08$ ist $V = 8.437.500$ €. Und eine (intuitive) Erklärung ist sehr einfach: Obwohl die Auszahlungen jedes Jahr wachsen, ist das Wachstum mit der Zeit nur *additiv*, während in Zukunft bezahlte Euros *exponentiell* schnell ihren Wert (vom heutigen Standpunkt) verlieren. Die Auszahlungen in der weiteren Zukunft sind (im wesentlichen) wertlos.

Allgemeine harmonische Reihen

Wir betrachten nun die Reihen von der Form

$$\sum_{k=1}^{\infty} \frac{1}{f(k)}$$

mit $f(k) \geq k$ für alle k . Für $f(k) = k$ ist das genau die harmonische Reihe.

► **Beispiel 3.32 : Harmonische Reihe**

$$H_n = \sum_{k=1}^n \frac{1}{k} \quad \text{ist divergent,}$$

da (wie wir bereits gezeigt haben) $H_n \geq \ln n$ gilt und daher ist die Folge H_n nicht beschränkt. Divergenz von H_n kann man auch direkt zeigen:

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \overbrace{\frac{1}{3} + \frac{1}{4}}^{> 2 \cdot \frac{1}{4} = \frac{1}{2}} + \overbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}^{> 4 \cdot \frac{1}{8} = \frac{1}{2}} + \overbrace{\frac{1}{9} + \frac{1}{10} + \dots + \frac{1}{16}}^{> 8 \cdot \frac{1}{16} = \frac{1}{2}} + \dots$$

Da unendlich oft mehr als $1/2$ hinzuaddiert wird, ist $\lim_{n \rightarrow \infty} H_n = \infty$.

► *Beispiel 3.33* : Interessanterweise ist die “ähnliche” Reihe

$$S_n = \sum_{k=1}^n \frac{1}{k^2}$$

bereits konvergent! Da die Folge $\langle S_n \rangle$ monoton wachsend ist, reicht es zu zeigen, dass sie auch beschränkt ist.

Behauptung: Für alle $n \geq 1$ gilt $S_n \leq 2 - \frac{1}{n}$.

Beweis: Induction nach n . Für $n = 1$ ist die Aussage offensichtlich richtig. Induktionsschritt: $n \mapsto n + 1$

$$S_{n+1} = S_n + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n+1}.$$

Mit sogenanntem “Verdichtungskriterium” kann man zeigen, dass die Reihe

$$\sum_{k=1}^{\infty} \frac{1}{f(k)}$$

bereits konvergiert, falls $f(k)$ um mehr als einem logarithmischen Faktor schneller als k wächst.

Satz 3.34. (Verdichtungssatz) Sei $\langle a_k \rangle$ eine monoton fallende Nullfolge mit $a_k > 0$ für alle k .

Dann konvergiert die Reihe $\sum_{k=1}^{\infty} a_k$ gegen einen Grenzwert S genau dann, wenn die “verdichtete”

Reihe $\sum_{k=0}^{\infty} 2^k \cdot a_{2^k}$ gegen einen Grenzwert S' konvergiert. Außerdem gilt $\frac{1}{2}S' \leq S \leq S'$.

Die Beweisidee ist einfach: Teile die Ausgangsreihe in Abschnitte der Länge 2^k für $k = 0, 1, 2, \dots$ auf und betrachte die Summe von der Folgenglieder in jedem Abschnitt als den entsprechenden Folgenglied der verdichteten Reihe.

Beweis. Wir setzen $S_n := \sum_{k=1}^n a_k$ und $S'_n := \sum_{k=1}^n 2^k a_{2^k}$. Damit ist für $n < 2^{k+1}$

$$\begin{aligned} S_n &\leq a_1 + (a_2 + a_3) + (a_4 + \dots + a_7) + \dots + (a_{2^k} + \dots + a_{2^{k+1}-1}) \\ &\leq a_1 + 2a_2 + 4a_4 + \dots + 2^k a_{2^k} = S'_n \end{aligned}$$

und für $n \geq 2^{k+1}$

$$\begin{aligned} S_n &\geq a_1 + a_2 + (a_3 + a_4) + (a_5 + \dots + a_8) + \dots + (a_{2^{k+1}} + \dots + a_{2^{k+1}}) \\ &\geq a_1 + a_2 + 2a_4 + 4a_8 + \dots + 2^k a_{2^{k+1}} \geq \frac{1}{2} S'_n \end{aligned}$$

Ist nun die verdichtete Reihe konvergent, d.h. $\lim_{n \rightarrow \infty} S'_n = S'$ existiert, so ist auch die Ausgangsreihe konvergent (Monotonie-Kriterium) mit dem Grenzwert $\frac{1}{2} S' \leq \lim_{n \rightarrow \infty} S_n \leq S'$. Ist dagegen die verdichtete Reihe divergent, so folgt aus der für $n \geq 2^{k+1}$ gültigen Beziehung $S_n \geq \frac{1}{2} S'_n$ auch die Divergenz der Ausgangsreihe. \square

Satz 3.35. Die Reihe $\sum_{k=1}^{\infty} \frac{1}{f(k)}$ konvergiert, falls

$$f(k) \geq k \log_2^r k$$

für ein festes $r > 1$ und alle k gilt.

Beweis. Zuerst betrachten wir die Reihe $\sum_{k=1}^{\infty} k^{-r}$ mit $r > 1$. Anwendung des Verdichtungssatzes führt auf

$$2^k a_{2^k} = 2^k (2^{-kr}) = x^k \quad \text{mit} \quad x := 2^{1-r}.$$

Da $r > 1$, ist die verdichtete Reihe eine geometrische Reihe $\sum_{k=1}^{\infty} x^k$ mit $|x| < 1$, welche konvergent ist mit dem Grenzwert

$$\sum_{k=1}^{\infty} x^k = \frac{1}{1-x} = \frac{1}{1-2^{1-r}} = \frac{2^r}{2^r-2} = 1 + \frac{1}{2^{r-1}-1}.$$

Wenn wir nun die verdichtete Reihe von $\sum_{k=1}^{\infty} \frac{1}{f(k)}$ mit $f(k) \geq k \log_2^r k$ betrachten, dann erhalten wir

$$\frac{2^k}{f(2^k)} \leq \frac{2^k}{2^k (\log_2(2^k))^r} = \frac{1}{k^r}.$$

Da (wie wir gerade gezeigt haben) die Reihe $\sum_{k=1}^{\infty} k^{-r}$ konvergiert, muss nach dem Majorantenkriterium für Reihen auch die Reihe $\sum_{k=1}^{\infty} \frac{2^k}{f(2^k)}$ und damit (nach dem Verdichtungssatz) auch die Reihe $\sum_{k=1}^{\infty} \frac{1}{f(k)}$ konvergieren. \square

Verallgemeinerte harmonische Reihe: Für jedes festes $r > 1$ gilt:

$$\sum_{k=1}^{\infty} \frac{1}{k^r} = a \quad \text{mit} \quad 1 < a \leq 1 + \frac{1}{2^{r-1}-1}. \quad (3.12)$$

3.3.1 Konvergenzkriterien für Reihen

Aus der Konvergenz der geometrischen Reihe ergeben sich die folgenden zwei wichtige spezifische Konvergenzkriterien für Reihen.

Satz 3.36. Die Reihe $\sum_{k=0}^{\infty} a_k$ mit $a_k \geq 0$ ist konvergent, wenn eine der folgenden zwei Bedingungen erfüllt sind:

1. **Wurzelkriterium:** Es gibt eine reelle Zahl $0 < \theta < 1$, so dass für fast alle k gilt: $\sqrt[k]{a_k} < \theta$.
2. **Quotientenkriterium:** Es gibt eine reelle Zahl $0 < \theta < 1$, so dass für fast alle k sowohl $a_k \neq 0$ wie auch $\frac{a_{k+1}}{a_k} \leq \theta$ gilt.

Beweis. Um die erste Aussage zu beweisen, wende das Majorantenkriterium für Reihen auf $\sum_{k=0}^{\infty} \theta^k$ an.

Um die zweite Aussage zu beweisen, vergleiche mit der geometrischen Reihe $\sum_{k=0}^{\infty} \theta^k$, $0 \leq \theta < 1$. Sei $\frac{a_{k+1}}{a_k} \leq \theta$ für alle $k \geq n$ ($n \in \mathbb{N}$). Dann gilt: $a_{n+k} \leq \theta^k a_n$ für alle $k \in \mathbb{N}$. Also ist $\sum_{k=0}^{\infty} \theta^k a_n$ Majorante von $\sum_{k=n}^{\infty} a_k$ und die Reihe $\sum_{k=0}^{\infty} a_k$ konvergiert nach dem Majorantenkriterium. \square



Man beachte, dass die Bedingung im Quotientenkriterium *nicht* lautet

$$\text{für fast alle } k \text{ gilt: } \frac{a_{k+1}}{a_k} < 1, \quad (*)$$

sondern “für fast alle k gilt $\frac{a_{k+1}}{a_k} \leq \theta$ ” mit einem von n unabhängigen $\theta < 1$. (Dasselbe gilt auch für das Wurzelkriterium.) Die Quotienten $\frac{a_{k+1}}{a_k}$ dürfen also nicht beliebig nahe an 1 herankommen. Das die Bedingung (*) nicht ausreicht, zeigt das Beispiel der divergenten harmonischen Reihe $\sum_{k=1}^{\infty} \frac{1}{k}$. Mit $a_k := 1/k$ gilt zwar

$$\frac{a_{k+1}}{a_k} = \frac{k}{k+1} < 1 \quad \text{für fast alle } k \geq 1,$$

wegen $\lim_{n \rightarrow \infty} \frac{n}{n+1} = 1$ gibt es jedoch *kein* $\theta < 1$ mit

$$\frac{a_{k+1}}{a_k} \leq \theta \quad \text{für alle } k \geq 1,$$

Eine wichtige Klasse von Funktionen sind Polynome

$$P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n.$$

Sogenannte “Potenzreihen” sind Verallgemeinerungen der Polynome, wenn man unendlich viele Terme zulässt.

Definition: Für eine gegebene Folge $\langle a_n \rangle$ heißt die Reihe

$$P(x) = \sum_{n=0}^{\infty} a_n x^n$$

Potenzreihe mit Koeffizienten a_n .

Man betrachtet auch allgemeinere Potenzreihen um ein Punkt $x_0 \in \mathbb{R}$

$$P(x, x_0) = \sum_{n=0}^{\infty} a_n (x - x_0)^n.$$

Der Punkt x_0 wird auch *Entwicklungspunkt* der Potenzreihe genannt. Wir werden aber nur die Potenzreihen $P(x)$ mit dem Entwicklungspunkt $x_0 = 0$ betrachten (die Eigenschaften der allgemeinen Potenzreihen sind analog).

Einige Potenzreihen haben wir bereits gesehen: Die geometrische Reihe $\sum_{n=0}^{\infty} x^n$ und die verallgemeinerte geometrische Reihe $\sum_{n=0}^{\infty} n \cdot x^n$. Das sind die "einfachsten" Potenzreihen mit Koeffizienten $a_n = 1$ bzw. $a_n = n$. Es gibt aber auch viele andere wichtige Potenzreihen. So kann man zum Beispiel viele analytischen Funktionen $f(x)$ (wie Exponent- oder Logarithmusfunktion) als entsprechende Potenzreihen darstellen.⁹

▷ *Beispiel 3.37*: Die Potenzreihen für e^x und $\ln x$ sind:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots$$

$$\ln(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

Die entsprechenden Koeffizienten sind $a_n = \frac{1}{n!}$ und $a_n = \frac{(-1)^n}{n}$.

Es ist klar, dass nicht alle Potenzreihen konvergieren. Wie kann man erkennen, ob eine gegebene Potenzreihe konvergiert oder nicht? Und wenn konvergiert, dann für welche Werte von x tut sie das? Alle diese Fragen kann man mit dem Begriff vom "Konvergenzradius" leicht beantworten.

Der *Konvergenzradius* R der Potenzreihe ist gegeben durch $R = \frac{1}{r}$, wobei r ist als

$$r = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|} \quad (3.13)$$

definiert. Sind die Koeffizienten a_n einer Potenzreihe für fast alle n von Null verschieden, dann lässt sich r auch berechnen durch

$$r = \limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|. \quad (3.14)$$

Satz 3.38. Sei $R = \frac{1}{r}$, wobei r ist entweder durch (3.13) oder durch (3.14) gegeben. Die Potenzreihe $P(x) = \sum_{n=0}^{\infty} a_n x^n$ ist

- (a) absolut konvergent, falls $|x| < R$
- (b) divergent, falls $|x| > R$

Beweis. Teil (a) folgt unmittelbar aus den Wurzel- und Quotientenkriterien.

⁹Wie man die Potenzreihen für Funktionen bekommt wird im Abschnitt 3.7 skizziert.

Teil (b): Ist $|x| > R = \frac{1}{r}$ mit r durch (3.13) gegeben, so gilt

$$\limsup \sqrt[n]{|a_n x^n|} = |x| \cdot \limsup \sqrt[n]{|a_n|} = |x| \cdot \frac{1}{r} \geq 1 + \delta \quad \text{mit} \quad \delta > 0$$

und damit ist die Folge $|a_n x^n| \geq (1 + \delta)^n$ unbeschränkt. Das gleiche gilt auch wenn r als (3.14) definiert ist. \square

► *Beispiel 3.39*: Wir betrachten die Potenzreihe $e^x = \sum_{n=0}^{\infty} a_n x^n$ mit $a_n = \frac{1}{n!}$.

Aus der Stirling's Formel für $n!$ $\Rightarrow a_n \leq \left(\frac{n}{e}\right)^{-n}$

$$\Rightarrow \lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} \sim \lim_{n \rightarrow \infty} \sqrt[n]{\left(\frac{e}{n}\right)^n} = \lim_{n \rightarrow \infty} \frac{e}{n} = 0$$

\Rightarrow Radius $R = \infty \Rightarrow$ überall konvergiert.

Bemerkung 3.40. Der Satz sagt nichts über die Punkte x mit $|x| = R$ aus. Diese Punkte müssen für jede Reihe neu untersucht werden.

Bemerkung 3.41. Eine wichtige eigenschaft der Potenzreihen ist, dass man sie gliederweise differenzieren darf. Der Konvergenzradius ändert sich dabei nicht.

► *Beispiel 3.42*: (**Dezimaldarstellung**) Jede Zahl in Intervall $[0, 1]$ lässt sich als eine (potenziell unendliche) Folge $0, x_1 x_2 x_3 \dots$ mit $x_k \in \{0, 1, 2, \dots, 9\}$ darstellen. Man kann auch eine umgekehrte Frage stellen:

Ist jede Reihe $\sum_{k=1}^{\infty} x_k 10^{-k}$ mit $x_k \in \{0, 1, 2, \dots, 9\}$ auch konvergent? Dazu wenden wir das

Majorantenkriterium an. Für alle k gilt: $\overbrace{x_k}^{a_k} \cdot 10^{-k} \leq \overbrace{9}^{b_k} \cdot 10^{-k}$, und $\sum_{k=1}^{\infty} b_k$ ist konvergent:¹⁰

$$\sum_{k=1}^{\infty} 9 \cdot 10^{-k} = \sum_{k=0}^{\infty} 0,9 \cdot 10^{-k} = 0,9 \cdot \frac{1}{1 - \frac{1}{10}} = 1.$$

► *Beispiel 3.43*: (**Die Euler-Reihe**) In der Beschreibung von *Wachstumsverhalten* (in Physik, Biologie und Wirtschaft) taucht die Folge

$$c_n = \left(1 + \frac{1}{n}\right)^n$$

auf. Zum Beispiel c_n ist der Vergrößerungsfaktor eines Kapitals in einem Jahr bei Zinssatz 100% und n -maliger Aufzinsung:

$n = 1$	jährliche Aufzinsung	$c_1 = \left(1 + \frac{1}{1}\right)^1 = 1 + 1 = 2$
$n = 12$	monatliche Aufzinsung	$c_{12} = \left(1 + \frac{1}{12}\right)^{12} = 2,61303529$
$n = 365$	tägliche Aufzinsung	$c_{365} = \left(1 + \frac{1}{365}\right)^{365} = 2,714567455$
$n = 525600$	Aufzinsung je Minute	$c_n = 2,718279 \dots$

¹⁰ $\sum_{k=0}^{\infty} 10^{-k}$ ist eine geometrische Reihe.



Die Euler-Zahl e ist als der Grenzwert

$$e := \lim_{n \rightarrow \infty} e_n \quad \text{von} \quad e_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{n!}$$

definiert.

Existenz von e folgt aus dem Quotientenkriterium, da $e = \sum_{n=0}^{\infty} a_n$ mit $a_n = 1/n!$ (und $0! = 1$) gilt und $\langle a_n/a_{n-1} \rangle = \langle 1/n \rangle$ eine Nullfolge ist.

Die Folgen e_n und c_n sehen als vollkom verschieden aus. Nichtdestotrotz, gilt folgendes:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \right)^n = e$$

Das kann man relativ leicht zeigen. Zuerst zeigt man, dass die Folge $\langle c_n \rangle$ monoton wachsend und durch e nach oben beschränkt ist. Damit gilt $\lim_{n \rightarrow \infty} c_n \leq e$. Es reicht dann zu zeigen, dass $\lim_{n \rightarrow \infty} c_n \geq e_N$ für jedes festes $N \geq 1$ gilt. Die Gleichheit $\lim_{n \rightarrow \infty} c_n = e$ folgt dann aus dem Vergleichskriterium. Wir lassen die Details als Übungsaufgabe (Aufgabe 15).

Die Darstellung von e als Grenzwert der Folge e_n ist gut, da die Folge e_n sehr rasch gegen $e = 2,7182818\dots$ konvergiert, während die Folge c_n nur relativ langsam gegen denselben Grenzwert e konvergiert. Deshalb ist e_n besser für die numerische Berechnung von e geeignet.

► *Beispiel 3.44*: Sei $x \in \mathbb{R}$, $0 \leq x < 1$. Wir haben bereits gezeigt, dass die Reihe $\sum_{k=0}^{\infty} k x^k$ gegen $x/(1-x)^2$ konvergiert. Wir zeigen nun, dass auch die allgemeinere Reihe $\sum_{k=1}^{\infty} k^c x^k$ für jedes $c \in \mathbb{R}$ konvergiert.

Wurzelkriterium: Da $\lim_{k \rightarrow \infty} \sqrt[k]{k^c} = 1$ (siehe (3.8)), erhalten wir

$$\lim_{k \rightarrow \infty} \sqrt[k]{k^c x^k} = \lim_{k \rightarrow \infty} \sqrt[k]{x^k} = x < 1.$$

► *Beispiel 3.45*: Die Reihe $\sum_{k=1}^{\infty} \frac{c^k}{k!}$ ist konvergent für jedes $c \geq 0$. Sei $c > 0$ (für $c = 0$ ist nichts zu beweisen). Dann gilt $\lim_{k \rightarrow \infty} a_{k+1}/a_k = \lim_{k \rightarrow \infty} c/(k+1) = 0$ und wir können das Quotientenkriterium anwenden.

Aus dem Cauchy-Kriterium für Folgen unmittelbar, dass eine Reihe $\langle S_n \rangle = \sum_{n=0}^{\infty} a_n$ genau dann konvergiert, wenn es zu jedem $\epsilon > 0$ ein n_0 mit

$$|S_n - S_m| = \left| \sum_{k=m+1}^n a_k \right| < \epsilon \quad \text{für alle } n > m \geq n_0 \quad (3.15)$$

gilt.

Es gibt auch einige spezielle Konvergenzkriterien für Reihen.

Satz 3.46.

1. Ist $\langle a_n \rangle$ keine Nullfolge, so ist die Reihe $\sum_{k=0}^{\infty} a_k$ divergent.
2. **Dirichlet-Kriterium:** Ist $\langle a_k \rangle$ eine monoton fallende Nullfolge und $\langle b_k \rangle$ eine Folge mit beschränkten Partialsummen (d.h. $|\sum_{k=0}^n b_k| \leq B$ für alle n), dann konvergiert $\sum_{k=0}^{\infty} a_k b_k$.
3. **Leibniz-Kriterium:** Es sei $\langle a_k \rangle$ eine monoton fallende Nullfolge. Dann konvergiert die alternierende Reihe $S_n = \sum_{k=0}^n (-1)^k a_k$ gegen einen Wert, der zwischen S_{2n+1} und S_{2n} für alle n liegt.

Beweis. Die erste Aussage folgt aus (3.15): Ist $\langle a_n \rangle$ keine Nullfolge, dann gibt es ein $\epsilon > 0$, so dass $|a_n| \geq \epsilon$ für unendlich viele n . Dann gilt auch $|S_n - S_{n-1}| = |a_n| \geq \epsilon$ für unendlich viele n und die Folge $\langle S_n \rangle$ muss nach (3.15) divergieren.

Um die zweite Aussage zu beweisen, setze $B_k = b_0 + b_1 + \dots + b_k$. Dann gilt

$$\begin{aligned} \left| \sum_{k=m+1}^n a_k b_k \right| &= \left| \underbrace{a_n}_{\geq 0} (B_n - B_m) + \sum_{k=m+1}^{n-1} \underbrace{(a_k - a_{k+1})}_{\geq 0} (B_k - B_m) \right| \\ &\leq |(B_n - B_m)| a_n + \sum_{k=m+1}^{n-1} |B_k - B_m| (a_k - a_{k+1}) \\ &\leq 2B \left(a_n + \sum_{k=m+1}^{n-1} (a_k - a_{k+1}) \right) \\ &= 2B(a_n + a_{m+1} - a_n) \\ &= 2Ba_{m+1}. \end{aligned}$$

Sei nun $\epsilon > 0$ beliebig klein (aber fest). Man wähle n_0 so, dass $a_{m+1} < \frac{\epsilon}{2B}$ für alle $m \geq n_0$ gilt (a_n ist monoton fallend). Dann ist

$$\left| \sum_{k=m+1}^n a_k b_k \right| < 2B \cdot \frac{\epsilon}{2B} = \epsilon$$

für alle $n > m \geq n_0$ und die Reihe $\sum_{k=0}^{\infty} a_k b_k$ konvergent nach Cauchy-Kriterium.

Um die dritte Aussage zu beweisen, setze $b_k := (-1)^k$ und $B_k := \sum_{j=0}^k b_j$. Es ist $B_k = 0$ oder $B_k = 1$, d.h. B_k ist beschränkt und damit konvergiert $\sum_{k=0}^{\infty} (-1)^k a_k$ nach Dirichlet-Kriterium gegen einen Grenzwert S . Es gilt auch: $S_{2n+2} - S_{2n} = (-1)^{2n+2} a_{2n+2} + (-1)^{2n+1} a_{2n+1} = a_{2n+2} - a_{2n+1} \leq 0$ (da a_k monoton fallend ist) und damit $S_0 \geq S_2 \geq \dots \geq S_{2n} \geq S_{2n+2} \geq \dots$ woraus $S \leq S_{2n}$ für alle n folgt. Entsprechend ist wegen $S_{2n+3} - S_{2n+1} = -a_{2n+3} + a_{2n+2} \geq 0$ $S_1 \leq S_3 \leq \dots \leq S_{2n+1} \leq S_{2n+3} \leq \dots$ woraus $S_{2n+1} \leq S$ für alle n folgt. Außerdem gilt wegen $S_{2n+1} - S_{2n} = -a_{2n+1} \leq 0$ $S_{2n+1} \leq S_{2n}$ für alle $n = 0, 1, 2, \dots$. Also gilt $S_{2n+1} \leq S \leq S_{2n}$ für alle $n = 0, 1, 2, \dots$, wie gewünscht. \square

Bemerkung 3.47. Die Umkehrung der ersten Aussage im Satz 3.46 gilt nicht! Zum Beispiel ist die harmonische Folge $a_k = 1/k$ eine Nullfolge, aber die harmonische Reihe $H_n = \sum_{k=1}^n \frac{1}{k}$ divergiert:

$$\sum_{k=m+1}^{2m} \frac{1}{k} = \frac{1}{m+1} + \frac{1}{m+2} + \dots + \frac{1}{2m} > m \cdot \frac{1}{2m} = \frac{1}{2}.$$

Da es für $\epsilon < 1/2$ kein n_0 mehr gibt \implies Divergenz nach Cauchy-Kriterium.

3.3.2 Anwendung: Warum Familiennamen aussterben?

Es war schon seit langem aufgefallen, dass Familiennamen ehrwürdiger Adelsgeschlechter über Jahrhunderte hinweg mit der Zeit aussterben. Dies inspirierte den englischen Naturforscher Franzis Galton 1873 das folgende Problem zu stellen:

Es seien p_0, p_1, p_2, \dots die Wahrscheinlichkeiten dafür, dass ein Mann 0, 1, 2, ... Söhne hat. Jeder Sohn habe die gleichen Wahrscheinlichkeiten für eigene Söhne usw. Wie groß ist die Wahrscheinlichkeit, dass die männliche Linie nach n Generationen ausgestorben ist? D.h. mit welcher Wahrscheinlichkeit stirbt der Name des Mannes aus?

Das ist eine typische Fragestellung aus der sogenannten *Theorie der Verzweigungsprozesse*, die heutzutage zahlreiche Anwendungen in der Physik, Chemie, Biologie, Sozialologie und Technik hat. Statt dem Mann hätten wir zum Beispiel ein Neutron betrachten können, das einen Urkern spaltet, wobei wieder Neutronen (Söhne) freigesetzt werden usw.

Der Mann stelle die nullte Generation dar, dessen Söhne die erste usw. Für $n \geq 0$ wollen wir mit a_n die Wahrscheinlichkeit bezeichnen, dass die männliche Linie nach n Generationen ausgestorben ist. Offenbar gilt $a_0 = p_0$.

Angenommen wir würden a_n kennen. Dann wissen wir auch, dass jede Verzweigungslinie der ersten Generation mit Wahrscheinlichkeit a_n nach n weiteren Generationen ausgestorben ist. Damit ist die Wahrscheinlichkeit, dass der Mann i Söhne hat und alle Nachkommen dieser i Söhne nach n weiteren Generationen ausgestorben sind, gleich

$$p_i \cdot a_n^i.$$

Insgesamt können wir demnach a_{n+1} durch

$$a_{n+1} = p_0 + p_1 a_n + p_2 a_n^2 + p_3 a_n^3 + \dots$$

berechnen, weil der Mann entweder gar keinen Sohn hat oder einen Sohn oder zwei Söhne oder ... hat.

Unmittelbar aus der Definition von a_n folgt, dass $a_n \leq a_{n+1}$, und somit ist die Folge $\langle a_n \rangle$ monoton wachsend. Berücksichtigt man noch, dass die a_n als Wahrscheinlichkeiten durch 1 nach oben beschränkt sind, so kann man nach **Monotonie-Kriterium** unmittelbar auf die Existenz von

$$a = \lim_{n \rightarrow \infty} a_n \leq 1$$

schließen. Die Zahl a beschreibt gerade die Wahrscheinlichkeit, ob die männliche Linie ausstirbt. Diese Zahl hängt natürlich von der Wahrscheinlichkeitsverteilung p_0, p_1, p_2, \dots ab. Als nächstes wollen wir diese Abhängigkeit genauer untersuchen.

Wenn wir die Potenzreihe

$$f(x) = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + \dots$$

einführen, so haben wir die Rekursionsvorschrift

$$a_{n+1} = f(a_n) \tag{3.16}$$

gefunden.

Zuerst beobachten wir, dass die Reihe $f(x)$ für alle $x \in [0, 1]$ konvergiert. Für $x = 0$ ist das klar, da dann $f(0) = p_0$ gilt. Für $x = 1$ gilt $f(1) = p_0 + p_1 + p_2 + p_3 + \dots = 1$, da p_0, p_1, p_2, \dots

eine Wahrscheinlichkeitsverteilung ist. Für $0 < x < 1$ können wir Dirichlet-Kriterium anwenden: (x^k) ist eine monoton fallende Nullfolge und (p_k) ist eine Folge mit beschränkten Partialsummen (da $\sum_{k=0}^n p_k \leq 1$ für alle $n \geq 0$ gilt). Deshalb muss nach **Dirichlet-Kriterium** auch die Reihe $f(x) = \sum_{i=0}^{\infty} p_i x^i$ konvergieren. Ausserdem, kann man zeigen, dass die Funktion f auf dem Intervall $[0, 1]$ konvex¹¹ und stetig ist.

Die Zahl $a = \lim_{n \rightarrow \infty} a_n$ beschreibt gerade die Wahrscheinlichkeit, ob die männliche Linie ausstirbt. Geht man nun in (3.16) zum Limes über, so ergibt sich unter Beachtung der Stetigkeit von f die Beziehung

$$a = f(a),$$

das heißt, a ist ein Fixpunkt von f . Dieser Fixpunkt muss auf der Geraden $y = x$ liegen. Da $f(0) = p_0$, $f(1) = \sum_{i=0}^{\infty} p_i = 1$ und die Funktion $f(x)$ monoton wachsend im Intervall $[0, 1]$ ist, muss der erste Fixpunkt die kleinste Zahl zwischen p_0 und 1 sein. Da $f(1) = 1$, ist $a = 1$ ein Fixpunkt von f . Gibt es weitere Fixpunkte?

Um diese Frage zu beantworten, betrachten wir die Ableitung¹²

$$f'(x) = p_1 + 2p_2x + 3p_3x^2 + \dots = \sum_{i=1}^{\infty} ip_i x^{i-1}.$$

Das ist der Anstieg der Tangente an den Graphen von f im Punkt x . Da f auf $[0, 1]$ konvex ist, liegen alle Werte $f(x)$ mit $x \in [0, 1]$ oberhalb dieser Tangente.

Wenn also $f'(1) \leq 1$ gilt, so liegen auch alle Werte $f(x)$ mit $x \in [0, 1]$ oberhalb der Geraden $y = x$, die den Anstieg 1 besitzt (siehe Abb. 3.2(A)). Im Falle $f'(1) \leq 1$ kann es also keinen weiteren Fixpunkt als 1 geben.

Anders im Fall $f'(1) > 1$. Hier muss der Graph der konvexen Funktion f die Gerade $y = x$ für ein $x < 1$ schneiden (da $f'(0) = p_1 \leq 1$), das heißt, es muss ein weiterer Fixpunkt existieren (siehe Abb. 3.2(B)).

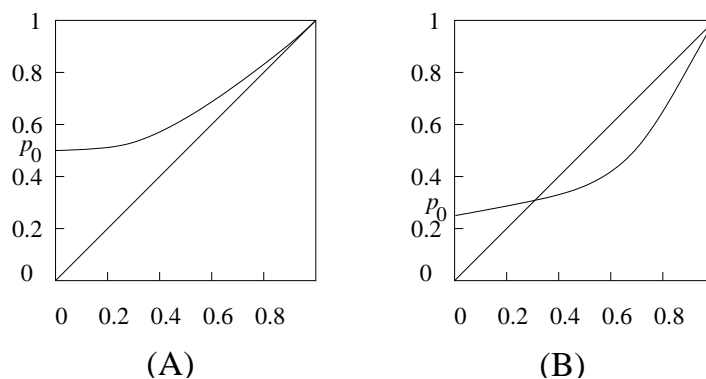


Abbildung 3.2:

Zusammengefasst: Im Falle

$$\sum_{i=0}^{\infty} ip_i \leq 1$$

¹¹Da die zweite Ableitung $f''(x) = \sum_{i=2}^{\infty} i(i-1)x^{i-2}$ nicht negativ ist.

¹²Die Summe $\sum_{i=0}^{\infty} ip_i$ ist die *erwartete Anzahl* der Söhne, die ein Mann haben kann, unter der Wahrscheinlichkeitsverteilung p_0, p_1, p_2, \dots . Wir werden bald dieses Konzept genauer kennenlernen.

stirbt demnach die männliche Linie mit Wahrscheinlichkeit 1 aus, während es im Fall

$$\sum_{i=0}^{\infty} ip_i > 1$$

mit positiver Wahrscheinlichkeit immerfort Namensträger gibt. Das bedeutet also, dass der Familienname langfristig nicht sicher ausstirbt, wenn im Durchschnitt mehr als ein Sohn da ist.

In der Biologie hat dieser Prozess Anwendung auf das Aussterben bzw. Überleben vorteilhafter Gene. Hierbei setzt man meist noch

$$p_i := e^{-\lambda} \frac{\lambda^i}{i!} \quad (\text{sogenannte Poisson-Verteilung, siehe Abschnitt 4.10})$$


so dass

$$f(x) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i x^i}{i!} = e^{\lambda(x-1)}.$$

In diesem Fall gilt $\lambda = p_1 + 2p_2 + 3p_3 + \dots = f'(1)$.

3.3.3 Umordnungssatz

Sei $\langle a_k \rangle$ eine Folge und $\langle S_n \rangle$ die entsprechende Reihe der Partialsummen $S_n = \sum_{k=0}^n a_k$. Hat die Reihe $\langle S_n \rangle$ den Grenzwert S , so schreibt man $\sum_{k=0}^{\infty} a_k = S$.

 Wir haben bereits erwähnt, dass das *nur eine Schreibweise ist* – unendliche Summen als solche existieren nicht! Insbesondere können wir i.A. nicht die Terme in $\sum_{k=0}^{\infty} a_k$ umordnen (obwohl für endlichen Summen das immer erlaubt ist, da + eine kommutative Operation ist, d.h. $x + y = y + x$ gilt).

▷ *Beispiel 3.48*: Nach dem Leibniz-Kriterium konvergiert die alternierende harmonische Reihe $\sum_{k=1}^{\infty} \frac{(-1)^k}{k}$ gegen einen Grenzwert s . Wir vergleichen die Reihen, die s und $\frac{3}{2}s$ darstellen:

$$\begin{aligned} s &= -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \dots \\ s + \frac{1}{2}s &= -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \dots \\ &\quad - \frac{1}{2} + \frac{1}{4} - \frac{1}{6} + \frac{1}{8} - \dots \\ &= -1 - \frac{1}{3} + \frac{1}{2} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \dots \end{aligned}$$

Damit ist:

$$s = -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \frac{1}{9} + \frac{1}{10} \pm \dots$$

und

$$s + \frac{1}{2}s = -1 - \frac{1}{3} + \frac{1}{2} - \frac{1}{5} - \frac{1}{7} + \frac{1}{4} - \frac{1}{9} - \frac{1}{11} + \frac{1}{6} - \frac{1}{13} - \frac{1}{15} \pm \dots$$

Was auffällt: In der Reihe für $s + \frac{1}{2}s$ tauchen dieselben Summanden auf, nur in einer anderen Reihenfolge. Wenn man in konvergenten Reihen die Reihenfolge der Summation ändert, bekommt man eventuell einen anderen Grenzwert! Der Grund für dieses Phänomen liegt, so stellt sich heraus, in der Tatsache, dass die Summe der Absolutbeträge $\sum_{k=1}^{\infty} \frac{1}{k}$ divergiert (harmonische Reihe).

Die Reihe $\sum_{k=0}^{\infty} a_k$ heißt *absolut konvergent*, wenn die Reihe $\sum_{k=0}^{\infty} |a_k|$ konvergiert. Sie heißt *bedingt konvergent*, wenn sie zwar konvergiert, die Reihe $\sum_{k=0}^{\infty} |a_k|$ aber divergiert. Eine solche ist zum Beispiel die Reihe $\sum_{k=1}^{\infty} \frac{(-1)^k}{k}$.

Bedingt konvergente Reihen sollte man tunlichst meiden. Absolut konvergente Reihen aber sind harmlos. Mit ihnen kann man alles das machen, was man sich so naiverweise vorstellt! Es gilt nämlich folgendes:

Satz 3.49. (Umordnungssatz) Sei $\sum_{k=0}^{\infty} a_k$ absolut konvergent (gegen a) und sei $\varphi : \mathbb{N} \rightarrow \mathbb{N}$ eine beliebige bijektive Abbildung (Umordnung). Dann konvergiert auch die Reihe $\sum_{k=0}^{\infty} a_{\varphi(k)}$ absolut gegen a .

Beachte, dass die Folgen $S_n = \sum_{k=0}^n a_k$ und $S'_n = \sum_{k=0}^n a_{\varphi(k)}$ total verschieden sein können! Deshalb ist der Satz nicht trivial: konvergiert $\sum_{k=0}^{\infty} |a_k|$, so haben die beiden Folgen $\langle S_n \rangle$ und $\langle S'_n \rangle$ den selben Grenzwert!

Beweis. Angenommen, konvergiert die Reihe $\sum_{k=0}^{\infty} a_k$ absolut gegen einen Grenzwert A . Sei $\epsilon > 0$ beliebig klein. Da die Reihe absolut konvergent ist, gibt es nach dem Cauchy-Kriterium für Reihen ein $n_0 \in \mathbb{N}$ mit

$$\sum_{k=n_0+1}^{\infty} |a_k| < \frac{\epsilon}{2}$$

Daraus folgt

$$\left| \sum_{k=0}^{n_0} a_k - A \right| = \left| \sum_{k=n_0+1}^{\infty} a_k \right| \leq \sum_{k=n_0+1}^{\infty} |a_k| < \frac{\epsilon}{2} \quad (3.17)$$

Da φ bijektiv ist, gibt es ein $N > n_0$ mit

$$\varphi(\{0, 1, \dots, N\}) \supseteq \{0, 1, \dots, n_0\} \quad (3.18)$$

Dann gilt für $n \geq N$:

$$\begin{aligned} \left| \sum_{k=0}^n a_{\varphi(k)} - A \right| &= \left| \sum_{k=0}^n a_{\varphi(k)} - \sum_{k=0}^{n_0} a_k + \sum_{k=0}^{n_0} a_k - A \right| \\ &\leq \left| \sum_{k=0}^n a_{\varphi(k)} - \sum_{k=0}^{n_0} a_k \right| + \left| \sum_{k=0}^{n_0} a_k - A \right| \\ &\stackrel{(3.17)}{<} \left| \sum_{k=0}^n a_{\varphi(k)} - \sum_{k=0}^{n_0} a_k \right| + \frac{\epsilon}{2} \\ &\stackrel{(3.18)}{\leq} \sum_{k=n_0+1}^{\infty} |a_k| + \frac{\epsilon}{2} < \epsilon, \end{aligned}$$

die umgeordnete Reihe konvergiert also gegen denselben Grenzwert wie die Ausgangsreihe. Das die umgeordnete Reihe auch absolut konvergiert, folgt aus der Anwendung des gerade Bewiesenen auf die Reihe $\sum_{k=0}^{\infty} |a_k|$. \square

Satz 3.50. (Großer Umordnungssatz) Sei $I = \bigcup_j I_j$ (endlich oder unendlich) mit $I_j \cap I_k = \emptyset$ für alle $j \neq k$. Dann gilt

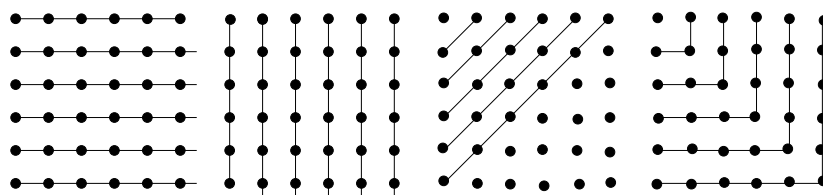
$$\sum_{i \in I} a_i = \sum_j \sum_{i \in I_j} a_j$$

sofern eine der beiden Seiten existiert, mit a_j durch $|a_j|$ ersetzt.

Specialfall: $\mathbb{N}_+^2 = \{1, 2, \dots\} \times \{1, 2, \dots\}$ (*Doppelreihensatz* oder *doppeltes Zählen*)

$$\begin{aligned} \sum_{(n,m) \in \mathbb{N}_+^2} a_{n,m} &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{n,m} \quad \text{Zeilensummen} \\ &= \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} a_{n,m} \quad \text{Spaltensummen} \\ &= \sum_{N=1}^{\infty} \sum_{\substack{(n,m) \\ n+m=N}} a_{n,m} \quad \text{Diagonalsummen} \\ &= \sum_{N=1}^{\infty} \sum_{\max\{n,m\}=N} a_{n,m} \quad \text{Quadratsummen} \end{aligned}$$

falls eine der 5 Summen existiert, mit $a_{n,m}$ durch $|a_{n,m}|$ ersetzt.



Mit diesem Fakt kann man einige ‘Monster-Reihen’ erledigen.

► *Beispiel 3.51* : Die Reihe

$$A = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{n^2 + m^2}$$

divergiert. Um das zu zeigen, betrachten wir die Quadratsummen

$$S_N = \sum_{\substack{(n,m) \\ \max\{n,m\}=N}} \frac{1}{n^2 + m^2}$$

Da $n^2 + m^2 \leq 2N^2$, haben wir

$$S_N \geq \frac{\text{Anzahl der Summanden}}{2N^2} = \frac{2N}{2N^2} = \frac{1}{N}.$$

Damit ist $\langle S_N \rangle$ eine Majorante der harmonischen Folge $\langle 1/N \rangle$, woraus die Divergenz der Reihe $A = \sum_{N=1}^{\infty} S_N$ folgt.

► *Beispiel 3.52* : Interessant ist, dass die “ähnliche” Reihe $\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{n^3 + m^3}$ bereits konvergiert! Um das zu zeigen, betrachten wir wieder die Quadratsummen

$$S_N = \sum_{\substack{(n,m) \\ \max\{n,m\}=N}} \frac{1}{n^3 + m^3}.$$

Dann ist

$$S_N \leq \frac{\text{Anzahl der Summanden}}{N^3} = \frac{2N}{N^3} = \frac{2}{N^2},$$

da $n^3 + m^3 \geq N^3$ gilt. Damit gilt auch (nach Lemma ??)

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{n^3 + m^3} = \sum_{N=1}^{\infty} S_N < 2 \cdot \sum_{N=1}^{\infty} \frac{1}{N^2} < 4.$$

► *Beispiel 3.53* : $\sum_{n=2}^{\infty} \sum_{k=2}^{\infty} \frac{1}{n^k} \leq 2$.

$$\begin{aligned} \sum_{n=2}^{\infty} \sum_{k=2}^{\infty} \left(\frac{1}{n}\right)^k &= \sum_{n=2}^{\infty} \sum_{k=0}^{\infty} \left(\frac{1}{n}\right)^{k+2} \\ &= \sum_{n=2}^{\infty} S_n \quad \text{mit } S_n := \sum_{k=0}^{\infty} a^{k+2} \text{ und } a := \frac{1}{n} \\ &= \sum_{n=2}^{\infty} \left(\frac{1}{n^2} \cdot \frac{n}{n-1}\right) \quad \text{da } S_n = a^2 \cdot \sum_{k=0}^{\infty} a^k = a^2 \cdot \frac{1}{1-a} \text{ (geom. Reihe)} \\ &= \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \leq 2 \text{ (verallgemeinerte harmonische Reihe, Lemma ??)} \end{aligned}$$

► *Beispiel 3.54* :

$$\begin{aligned} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \binom{n}{m} 2^{-n-m} &= \sum_{n=0}^{\infty} \left(\sum_{m=0}^n \binom{n}{m} 2^{-m} \right) \cdot 2^{-n} \\ &= \sum_{n=0}^{\infty} \left(1 + \frac{1}{2}\right)^n \cdot 2^{-n} \quad \text{(binomischer Lehrsatz)} \\ &= \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n = \frac{1}{1-3/4} = 4. \end{aligned}$$

3.4 Grenzwerte bei Funktionen

Folgen sind Funktionen von \mathbb{N} nach \mathbb{R} . Den Begriff des Grenzwertes kann man auch auf allgemeine Funktionen $f : \mathbb{R} \rightarrow \mathbb{R}$ erweitern.

Sei $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \rightarrow \mathbb{R}$ eine Funktion. Sei $a \in I$ (evtl. ist f in a nicht definiert). Eine α -Umgebung einer Zahl y ist die Menge U_α aller Zahlen, die sich von y um höchstens $\pm\alpha$ unterscheiden.

Definition: Eine Zahl A heißt *Grenzwert* von $f(x)$ in a , geschrieben

$$\lim_{x \rightarrow a} f(x) = A$$

falls

$$\forall \epsilon > 0 \exists \delta > 0 \forall x \neq a : x \in U_\delta(a) \Rightarrow f(x) \in U_\epsilon(f(a)).$$

Kurz: Es gibt eine δ -Umgebung $U_\delta(a)$ von a , so dass in dieser Umgebung die Funktion $f(x)$ sich um höchstens $\pm\epsilon$ vom A abweicht. Ausserdem kann $\epsilon > 0$ beliebig klein gemacht werden.

Bemerkung 3.55. Ist $f : \mathbb{N} \rightarrow \mathbb{R}$ eine Folge, so kann man die Mengen $\{n_0, n_0 + 1, n_0 + 2, \dots\}$ mit $n_0 \in \mathbb{N}$ als “ δ -Umgebungen” der Unendlichkeit ∞ betrachten. Damit ist der Begriff des Limes für Folgen ein Spezialfall dieser mehr allgemeiner Definition.

Wir erinnern an die folgenden Regeln zur Grenzwertberechnung, die direkt aus der Definition des Limes folgen. Aus $\lim_{x \rightarrow a} f(x) = c$ und $\lim_{x \rightarrow a} g(x) = d$ folgt

1. $\lim_{x \rightarrow a} (f(x) \pm g(x)) = c \pm d$
2. $\lim_{x \rightarrow a} (f(x) \cdot g(x)) = c \cdot d$ Spezialfall: $\lim_{x \rightarrow a} (b \cdot f(x)) = b \cdot d$ für $b \in \mathbb{R}$
1. $\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{c}{d}$ (falls $d \neq 0$)

► *Beispiel 3.56:* Umformung von Termen

$$\begin{aligned} \lim_{x \rightarrow 1} \frac{x^n - 1}{x - 1} &= \lim_{x \rightarrow 1} \frac{(x - 1) \cdot (x^{n-1} + x^{n-2} + \dots + x + 1)}{x - 1} \\ &= \lim_{x \rightarrow 1} x^{n-1} + \lim_{x \rightarrow 1} x^{n-2} + \dots + \lim_{x \rightarrow 1} x + \lim_{x \rightarrow 1} 1 \\ &= n \end{aligned}$$

► *Beispiel 3.57* :

$$\begin{aligned}
 \lim_{x \rightarrow 0} \frac{\sqrt{x+1} - 1}{x} &= \lim_{x \rightarrow 0} \frac{(\sqrt{x+1} - 1) \cdot (\sqrt{x+1} + 1)}{x \cdot (\sqrt{x+1} + 1)} \\
 &= \lim_{x \rightarrow 0} \frac{x + 1 - 1}{x \cdot (\sqrt{x+1} + 1)} \\
 &= \lim_{x \rightarrow 0} \frac{x}{x \cdot (\sqrt{x+1} + 1)} \\
 &= \lim_{x \rightarrow 0} \frac{1}{\sqrt{x+1} + 1} \\
 &= \frac{1}{2}.
 \end{aligned}$$

Auch wenn der Grenzwert $\lim_{x \rightarrow a} f(x) = A$, muss $A = f(a)$ nicht unbedingt gelten! Dafür muss die Funktion $f(x)$ stetig in a sein. Zur Erinnerung: Eine reelle Funktion f heißt *stetig* in $a \in \mathbb{R}$, falls für jede Folge $\langle a_n \rangle$ mit $\lim_{n \rightarrow \infty} a_n = a$ gilt: $\lim_{n \rightarrow \infty} f(a_n) = f(a)$.

Satz 3.58. $f(x)$ ist stetig in $a \iff \lim_{x \rightarrow a} f(x) = f(a)$ gilt

Beweis. (\Leftarrow): Sei $\lim_{x \rightarrow a} f(x) = f(a)$ und sei $\epsilon > 0$ beliebig klein. Dann gibt es ein $\delta > 0$, so dass $|x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon$. Sei nun $\langle a_n \rangle$ eine beliebige Folge mit $\lim_{n \rightarrow \infty} a_n = a$. Dann gibt es ein n_0 , so dass $|a_n - a| < \delta$ für alle $n \geq n_0$. Also muss auch $|f(a_n) - f(a)| < \epsilon$ für alle $n \geq n_0$ gelten, d.h. $\lim_{n \rightarrow \infty} f(a_n) = f(a)$.

(\Rightarrow): Kontraposition: Nehmen wir an, dass $\lim_{x \rightarrow a} f(x) = f(a)$ *nicht* gilt. Dann gibt es ein $\epsilon > 0$, so dass $\forall \delta > 0$ es ein $\exists x_\delta$ mit $|x_\delta - a| < \delta$ und $|f(x_\delta) - f(a)| \geq \epsilon$ gibt. Betrachte die Folge $\langle a_n \rangle$ mit $a_1 := x_1, a_2 := x_{1/2}, \dots, a_n := x_{1/n} \dots$. Diese Folge konvergiert mit $\lim_{n \rightarrow \infty} a_n = a$ (da $|a_n - a| < 1/n$ gilt), aber $|f(a_n) - f(a)| \geq \epsilon$ für alle $n \in \mathbb{N}$ und damit auch $\lim_{n \rightarrow \infty} f(a_n) \neq f(a)$. Somit ist die Funktion $f(x)$ nicht stetig in a . \square

3.5 Differentiation

Das Grundproblem der Differentialrechnung ist die Berechnung der Steigung einer Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ in einem beliebigen Punkt x .

Definition: Die *Ableitung* einer Funktion f im Punkt x ist

$$f'(x) := \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

Hier sind die Ableitungen einiger häufig benutzten Funktionen.

Satz 3.59. Seien $c \in \mathbb{R}$ und $n \in \mathbb{N}$ beliebige Konstanten.

$$f(x) = c \Rightarrow f'(x) = 0$$

$$f(x) = x \Rightarrow f'(x) = 1$$

$$f(x) = x^n \Rightarrow f'(x) = nx^{n-1}$$

$$f(x) = \ln g(x) \Rightarrow f'(x) = \frac{1}{g(x)} g'(x) \quad \text{Spezialfall: } (\ln x)' = \frac{1}{x}$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2}$$

$$f(x) = c^{g(x)} \Rightarrow f'(x) = (\ln c) e^{g(x)} g'(x) \quad \text{Spezialfall: } (c^x)' = (\ln c) e^x$$

$$f(x) = e^{c g(x)} \Rightarrow f'(x) = c e^{c g(x)} g'(x) \quad \text{Spezialfall: } (e^x)' = e^x$$

Beweis. Wir beweisen nur einige ausgewählte Fälle, um den allgemeinen Zugang zu demonstrieren.

1. Sei $f(x) = x^n$. Dann ist $f'(x) = n \cdot x^{n-1}$:

$$(x+h)^n = x^n + n \cdot x^{n-1} \cdot h + \dots + h^n,$$

$$f'(x) = \lim_{h \rightarrow 0} \frac{1}{h} (n \cdot x^{n-1} \cdot h + \dots + h^n) = n \cdot x^{n-1}.$$

2. Sei $f(x) = \frac{1}{x}$. Dann ist $f'(x) = -\frac{1}{x^2}$:

$$f'(x) = \lim_{h \rightarrow 0} \frac{1}{h} \left(\frac{1}{x+h} - \frac{1}{x} \right) = \lim_{h \rightarrow 0} \frac{1}{h} \left(\frac{x - x - h}{x(x+h)} \right) = -\frac{1}{x^2}.$$

3. Sei $f(x) = e^x$. Dann ist $f'(x) = e^x$:

$$f'(x) = \lim_{h \rightarrow 0} \frac{1}{h} (e^{x+h} - e^x) = e^x \cdot \lim_{h \rightarrow 0} \frac{e^h - 1}{h}$$

Es reicht also zu zeigen, dass $\lim_{h \rightarrow 0} \frac{e^h - 1}{h} = 1$. Wir haben $e = \lim_{y \rightarrow 0} (1+y)^{\frac{1}{y}}$ (Eulerische Zahl) und somit

$$\lim_{y \rightarrow 0} \frac{\ln(1+y)}{y} = \ln e = 1 \quad (\text{da } \ln a^{1/b} = (\ln a)/b) \quad (3.19)$$

Setzt man $y := e^h - 1$, dann gilt: $h = \ln(1+y)$. Somit folgt:

$$\lim_{h \rightarrow 0} \frac{e^h - 1}{h} = \lim_{y \rightarrow 0} \frac{y}{\ln(1+y)} = \lim_{y \rightarrow 0} \frac{1}{\frac{\ln(1+y)}{y}} = \frac{1}{\ln e} = 1.$$

4. Sei $f(x) = \ln x$, $0 < x < \infty$. Dann ist $f'(x) = \frac{1}{x}$

$$\frac{f(x+h) - f(x)}{h} = \frac{\ln(x+h) - \ln x}{h} = \frac{1}{x} \cdot \frac{\ln(1 + \frac{h}{x})}{\frac{h}{x}}$$

Mit (3.19) folgt:

$$f'(x) = \frac{1}{x} \lim_{h \rightarrow 0} \frac{\ln(1 + \frac{h}{x})}{\frac{h}{x}} = \frac{1}{x}$$

□

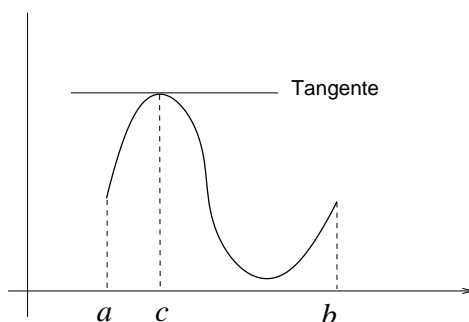
3.6 Mittelwertsätze der Differentialrechnung

Wozu sind Ableitungen gut? Ableitungen erlauben uns die *Entwicklung* der Funktion zu analysieren. Zum Beispiel sie helfen bei:

- Approximation von $f(x)$ durch Polynome (Taylorentwicklung).
- Bestimmung lokaler Extremalstellen von $f(x)$.
- Bestimmung der Grenzwerte für Quotienten $f(x)/g(x)$ (Regeln von de l'Hôpital).

Als Basis für alle diese Anwendungen sind sogenannte "Mittelwertsätze der Differentialrechnung".

Geometrisch besagt der folgende Satz von Rolle (Michel Rolle, 18. Jahrhundert), dass eine differenzierbare Funktion, die mit dem selben Wert startet und endet, irgendwo dazwischen eine waagrechte Tangente haben muss:



Satz 3.60. (Satz von Rolle) Sei $f : [a, b] \rightarrow \mathbb{R}$ stetig und im offenen Intervall (a, b) differenzierbar. Sei ferner $f(a) = f(b)$. Dann gibt es ein $c \in (a, b)$ mit $f'(c) = 0$.

Beweis. Ist f konstant, so ist $f'(x) = 0$ für alle x . Sei nun f nicht konstant. Dann hat f in $[a, b]$ ein Minimum an einer Stelle x_{min} und ein Maximum an einer Stelle x_{max} (hier benutzen wir die Stetigkeit von f). Da f nicht konstant ist, ist $f(x_{min}) < f(x_{max})$. Wegen $f(a) = f(b)$ liegt einer der beiden Punkte x_{min} oder x_{max} echt zwischen a und b . Angenommen, das ist x_{max} (der Fall von x_{min} ist analog). Setze $c := x_{max}$ und betrachte die Steigung

$$S(x) := \frac{f(x) - f(c)}{x - c}$$

Dann ist $f'(c)$ der Grenzwert von $S(x)$ wenn x von links ($x < c$) oder von rechts ($x > c$) gegen c strebt. Im ersten Fall ist $f'(c) \geq 0$ während im zweiten Fall ist $f'(c) \leq 0$. Zusammengefasst ergibt dies $f'(c) = 0$. \square

Eine differenzierbare Funktion wird lokal durch ihre Tangente gut approximiert. Lokal bedeutet hier in einer Umgebung von x_0 , und gut, dass der Fehler in Verhältnis zu $(x - x_0)$ klein ist, d.h. $\frac{\text{Fehler}}{x - x_0} \rightarrow 0$ für $x \rightarrow x_0$. Wir haben also

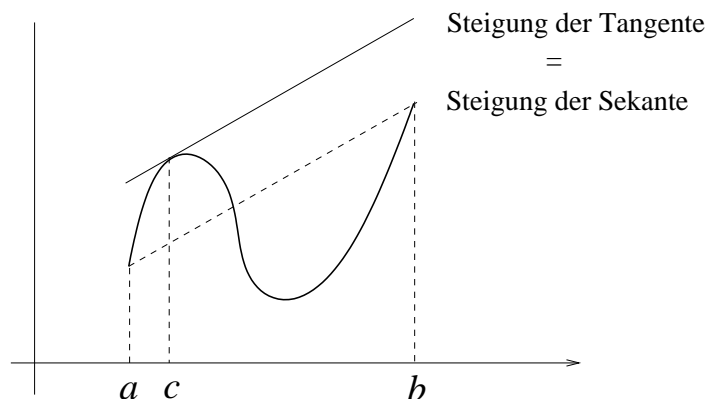
$$f(x) \approx f(x_0) + (x - x_0)f'(x_0)$$

Genauer gilt

Satz 3.61. (1. Mittelwertsatz von Lagrange 1736) Ist $f : [a, b] \rightarrow \mathbb{R}$ stetig und auf (a, b) differenzierbar, so gibt es ein $c \in (a, b)$ mit (mittlerer Ableitung)

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

Hinter dem Mittelwertsatz steht die folgende Überlegung: Es gibt ein c zwischen a und b , wo die Tangentensteigung $f'(c)$ gleich der Sekantensteigung $\frac{f(b)-f(a)}{b-a}$ ist, d.h. wo Tangente und Sekante parallel sind:



Beweis. Wir wenden den Satz von Rolle auf $g(x) := f(x) - \frac{f(b)-f(a)}{b-a}(x-a)$ an. Wir können dies tun, da $g(a) = f(a) = g(b)$. Es gibt danach ein $c \in (a, b)$ mit¹³ $0 = g'(c) = f'(c) - \frac{f(b)-f(a)}{b-a}$. \square

Mit dem 1. Mittelwertsatz kann man die folgenden, in vielen Situationen sehr nützlichen Abschätzungen beweisen.

Lemma 3.62. Für jedes $x \in (-1, 1)$ gilt

$$e^{x/(1+x)} \leq 1 + x \leq e^x.$$

Beweis. Wir betrachten die Funktion $f(z) = \ln z$. Wir wählen zwei Punkte $a = 1$ und $b = 1 + x$. Nach dem 1. Mittelwertsatz (Satz 3.61) gibt es dann $c \in (a, b)$ mit

$$f'(c) = \frac{f(b) - f(a)}{b - a} = \frac{\ln(1 + x)}{x}.$$

Wegen $f''(z) = -\frac{1}{z^2} \leq 0$ ist die Ableitung $f'(z)$ monoton fallend, woraus $f'(a) \geq f'(c) \geq f'(b)$ folgt. Da $f'(a) = f'(1) = 1$ und $f'(b) = f'(1 + x) = 1/(1 + x)$, ergibt dies die Ungleichungen

$$1 \geq \frac{\ln(1 + x)}{x} \geq \frac{1}{1 + x}$$

oder äquivalent

$$\frac{x}{1 + x} \leq \ln(1 + x) \leq x.$$

Die Behauptung folgt durch Eponentieren. \square

¹³Da $c' = 0$ und $x' = 1$.

Eine reellwertige Funktion $f(x)$ heißt *konvex*, falls für alle $\lambda \in (0, 1)$ gilt:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y),$$

Die Funktion f heißt *konkav*, wenn $-f$ konvex ist.

Kovexität von Funktionen muss man sehr oft bestimmen, wenn man zum Beispiel die Jensen's Ungleichung ¹⁴ (siehe Satz 1.23) für nicht-triviale Funktionen $f(x)$ anwenden will. Glücklicherweise gibt es dafür ein einfaches Kriterium: Es reicht, dass die zweite Ableitung von f nicht negativ ist.

Lemma 3.63. (Konvexität) Gilt $f''(x) \geq 0$ für alle x , so ist $f(x)$ konvex.

Beweis. Wegen $f''(x) \geq 0$ ist die Ableitung f' monoton wachsend. Seien nun x, y und $\lambda \in (0, 1)$. Wir setzen $z := \lambda x + (1 - \lambda)y$. Nach dem 1. Mittelwertsatz (Satz 3.61) gibt es dann $c \in (x, z)$ und $d \in (z, y)$ mit

$$\frac{f(z) - f(x)}{z - x} = f'(c) \overset{c < d}{\leq} f'(d) = \frac{f(y) - f(z)}{y - z}$$

Wegen

$$\begin{aligned} z - x &= \lambda x + (1 - \lambda)y - x = (1 - \lambda)(y - x) \\ y - z &= y - \lambda x + (1 - \lambda)y = \lambda(y - x) \end{aligned}$$

ergibt sich somit

$$\frac{f(z) - f(x)}{1 - \lambda} \leq \frac{f(y) - f(z)}{\lambda}$$

bzw.

$$f(z) = \lambda f(z) + (1 - \lambda)f(z) \leq \lambda f(x) + (1 - \lambda)f(y).$$

Die Funktion ist also konvex. □

Eine allgemeinere (als der 1. Mittelwertsatz) Aussage liefert der folgender Satz.¹⁵

Satz 3.64. (2. Mittelwertsatz von Cauchy 1823) Sei $f : [a, b] \rightarrow \mathbb{R}$ stetig und auf (a, b) differenzierbar. Es habe g' keine Nullstelle. Dann ist $g(a) \neq g(b)$ und es gibt eine Zwischenstelle $c \in (a, b)$ mit

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

Beweis. Da $g'(x) \neq 0$ für alle x , ist g streng monoton (wachsend oder fallend) und deshalb gilt $g(a) \neq g(b)$. Wir wenden nun den Satz von Rolle ¹⁶ auf

$$h(x) := f(x) - \left(f(a) + \frac{f(b) - f(a)}{g(b) - g(a)} (g(x) - g(a)) \right)$$

¹⁴Ist f konvex, so gilt:

$$f\left(\sum_{i=1}^r \lambda_i x_i\right) \leq \sum_{i=1}^r \lambda_i f(x_i)$$

für alle $0 \leq \lambda_i \leq 1$ mit $\sum_{i=1}^r \lambda_i = 1$.

¹⁵Satz von Lagrange entspricht dem Fall $g(x) = x$.

¹⁶Wir können das tun, da $h(a) = h(b) = 0$.

an und erhalten eine Zwischenstelle c mit

$$0 = h'(c) = f'(c) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot g'(c).$$

□

3.7 Approximation durch Polynome: Taylorentwicklung

Durch die erste Ableitung könnten wir eine Funktion f schreiben als

$$f(x+h) = \underbrace{f(x) + f'(x) \cdot h}_{\text{Geradengleichung}} + R(x, h)$$

mit $\lim_{h \rightarrow 0} R(x, h) = 0$. In vielen Fällen reicht diese "lokale" Annäherung durch eine Gerade nicht aus, wir benötigen Annäherungen durch Parabeln, kubische Parabeln, ..., kurz: durch Polynome.

Sei $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n$ ein Polynom n -ten Grades. Dann gilt:

$$\begin{aligned} p'(x) &= a_1 + 2a_2x + 3 \cdot a_3x^2 + \dots + n \cdot a_nx^{n-1} \\ p''(x) &= 2a_2 + 2 \cdot 3 \cdot a_3x + \dots + n \cdot (n-1)a_nx^{n-2} \\ &\vdots \\ p^{(n)}(x) &= 1 \cdot 2 \cdot \dots \cdot n \cdot a_n \end{aligned}$$

Damit kann man die Koeffizienten wie folgt bestimmen:¹⁷

$$a_0 = p(0), \quad a_1 = \frac{p'(0)}{1!}, \quad a_2 = \frac{p''(0)}{2!}, \quad \dots, \quad a_n = \frac{p^{(n)}(0)}{n!}$$

und somit auch

$$p(x) = p(0) + \frac{p'(0)}{1!}x + \frac{p''(0)}{2!}x^2 + \dots + \frac{p^{(n)}(0)}{n!}x^n. \quad (3.20)$$

Interessanterweise kann man auch Nicht-Polynome $f(x)$ in der ähnlichen Form darstellen mit einem *Restglied*

$$R_n(x, c) := \frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1}.$$

¹⁷Mit $f^{(k)}(a)$ bezeichnet man die k -te Ableitung von $f(x)$ im Punkt a ist, d.h. $f^{(0)}(x) := f(x)$ und $f^{(k+1)}(x) := (f^{(k)}(x))'$.

**Taylorformel:^a**

Sei $f : \mathbb{R} \rightarrow \mathbb{R}$ beliebig oft differenzierbar. Dann gibt es für jede $x \in \mathbb{R}$ ein c zwischen 0 und x mit

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(0)}{k!} \cdot x^k + R_n(x, c). \quad (3.21)$$

Gilt $\lim_{n \rightarrow \infty} R_n(x, c) = 0$ für alle c zwischen 0 und x , so gilt auch

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k. \quad (3.22)$$

^aBrook Taylor, 1685-1731

Beweis. Wir werden nur den ersten Teil (3.21) beweisen. Sei

$$T(x) = \sum_{k=0}^n \frac{f^{(k)}(0)}{k!} x^k = f(0) + f'(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \dots + \frac{f^{(n)}(0)}{n!}x^n$$

das Taylorpolynom. Weiterhin sei $h(x) := f(x) - T(x)$ und $g(x) := x^{n+1}$. Beachte, dass für alle $k \in \{1, 2, \dots, n\}$ folgendes gilt:¹⁸

$$T^{(k)}(0) = \underbrace{0 + \dots + 0}_{\text{da } \text{const}' = 0} + f^{(k)}(0) \frac{k!}{k!} + \underbrace{0 + \dots + 0}_{\text{da } x = 0} = f^{(k)}(0)$$

und


$$g^{(k)}(x) = (n+1)n(n-1) \cdots (n+1-k)x^{n+1-k}.$$

Deshalb gilt $h^{(k)}(0) = g^{(k)}(0) = 0$ für alle $k = 1, 2, \dots, n$.

Wir wenden nun den 2. Mittelwertsatz auf die Funktionen $f^{(k)}$ und $g^{(k)}$ an und erhalten

$$\begin{aligned} \frac{h(x)}{g(x)} &= \frac{h(x) - h(0)}{g(x) - g(0)} = \frac{h'(c_1)}{g'(c_1)} \\ &= \frac{h'(x_1) - h'(0)}{g'(x_1) - g'(0)} = \frac{h''(c_2)}{g''(c_2)} \\ &= \dots \\ &= \frac{h^{(n)}(x_n) - h^{(n)}(0)}{g^{(n)}(x_n) - g^{(n)}(0)} = \frac{h^{(n+1)}(c_{n+1})}{(n+1)!} \end{aligned}$$

mit $0 \leq c_{n+1} \leq c_n \leq \dots \leq c_1 \leq x$. Multiplikation mit $g(x) = x^{n+1}$ liefert die Behauptung (mit $c = c_{n+1}$). \square

 Beachte, dass die zweite Formel (3.22) nur dann gilt, wenn das Restglied $R_n(x, c)$ gegen 0 für alle $0 \leq c \leq x$ konvergiert (wenn $n \rightarrow \infty$). Das Problem ist, dass i.A. die Potenzreihe $\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$ kann entweder überhaupt divergent sein oder kann nicht notwendig gegen $f(x)$ konvergieren. Nimmt man zum Beispiel die Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ mit $f(x) = e^{-1/x^2}$ für $x \neq 0$ und $f(0) = 0$, so kann man zeigen, dass $f^{(k)}(0) = 0$ für alle $k = 0, 1, \dots$ gilt. Damit ist die Taylor-Reihe $T(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$ von f identisch 0, während $f(x) \neq 0$ für alle $x \neq 0$ gilt.

¹⁸Da die k -te Ableitung von x^n gleich $n(n-1) \cdots (n-k+1)x^{n-k}$ ist.

► *Beispiel 3.65*: Sei $f(x) = e^x$. Dann ist $f(x)$ beliebig oft differenzierbar und es gilt: $f^{(n)}(x) = f(x)$ für jedes $n \in \mathbb{N}$. Damit ist das Taylorpolynom n -ten Grades gleich

$$T_n(x) = \sum_{k=0}^n \frac{x^k}{k!}$$

Das Restglied ist

$$R_n(x, c) = \frac{e^c x^{n+1}}{(n+1)!}$$

für ein c zwischen 0 und x . Es gilt:

$$|R_n(x, c)| = \left| \frac{e^c x^{n+1}}{(n+1)!} \right| \leq \frac{e^{|x|} |x|^{n+1}}{(n+1)!}.$$

Da $\lim_{n \rightarrow \infty} x^n/n! = 0$, haben wir

$$\lim_{n \rightarrow \infty} \frac{e^{|x|} |x|^{n+1}}{(n+1)!} = 0.$$

Also $R_n(x) \rightarrow 0$ für $n \rightarrow \infty$. Nach der Taylorformel gilt:

$$e^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots = \sum_{i=0}^{\infty} \frac{x^i}{i!}. \quad (3.23)$$

Für $x = 1$ gibt dies die Formel zur Berechnung von e :

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots + \frac{1}{n!} + \dots = 2,7182818\dots$$

Ähnlich bekommt man

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - \dots = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i} \quad (3.24)$$

Dazu reicht es zu beobachten, dass für $f(x) = \ln(1+x)$ nach der Quotientenregel für Ableitungen gilt (nachrechnen!): $f^{(k)}(0) = (-1)^{k+1}(k-1)!$.

3.8 Extremalstellen

Sei $f : \mathbb{R} \rightarrow \mathbb{R}$ eine Funktion und $a \in \mathbb{R}$. Dann heißt a eine *lokale Maximalstelle* von f , falls es ein $\epsilon > 0$ gibt, so dass $f(a) \geq f(x)$ für alle $x \in U_\epsilon(a)$ gilt. Die *lokale Minimalstelle* ist analog definiert.

Lemma 3.66.

1. Extremalstellen-Test: Gilt $f'(a) = 0$ und hat $f'(x)$ an der Stelle a ein Vorzeichenwechsel, dann liegt in a ein Extremum vor.

Dabei ist a ein lokales Maximum, wenn $f'(x)$ von $+$ nach $-$ wechselt, sonst ist a ein lokales Minimum.

2. Extremalstellen-Test Gilt $f'(a) = 0$ und $f''(a) > 0$ (bzw. $f'(a) = 0$ und $f''(a) < 0$), so ist a lokale Minimalstelle (bzw. Maximalstelle) von f .

Beweis. Teil (1) folgt unmittelbar aus der Definition von $f'(x)$.

Teil (2): Sei $f''(a) > 0$. Zu $\epsilon = f''(a)/2$ gibt es (nach der Definition der Ableitung) ein $\delta > 0$ mit

$$-f''(a)/2 < f''(x) - f''(a) < f''(a)/2$$

also $f''(x) > f''(a)/2 > 0$ für alle x mit $|x - a| < \delta$. Für solche x gibt es (laut Taylorformel) ein c zwischen x und a (also insbesondere $|c - a| < \delta$) mit

$$f(x) = f(a) + \underbrace{f'(a)}_{=0}(x-a) + \underbrace{f''(c)}_{>0}(x-c)^2/2 = f(a) + \underbrace{f''(c)}_{>0}(x-c)^2/2 > f(a).$$

$f(a)$ ist also ein lokales Minimum. Ist $f''(a) < 0$, so betrachte $g = -f$. □

3.9 Die Bachmann-Landau-Notation: klein o und groß O

In diesem Abschnitt werden wir die asymptotische Notation einführen, die Sie während Ihres weiteren Studiums ständig begleiten wird.

Seien A und B zwei Algorithmen mit Laufzeiten $T_A(n)$ und $T_B(n)$. Um diese Algorithmen (bezüglich ihrer Laufzeit) zu vergleichen, fragt man

- Ist A im “wesentlichen” genau so schnell wie B ?
- Ist A “viel” schneller als B ?

Um solche (und ähnliche) Fragen mathematisch zu präzisieren, hat sich die folgende asymptotische Notation als sehr hilfreich erwiesen. Die Notation $O(x)$ war erstmal von Bachmann (1894) in seinem Buch (über Zahlentheorie) und Landau (wie er selbst in 1909 schrieb) hat diese Notation erst aus diesem Buch kennengelernt. Die Notation $o(x)$ hat Landau selbst eingeführt.

Um unnötigen “Feinheiten” zu vermeiden, werden wir in diesem Abschnitt hauptsächlich nur die Funktionen $f : \mathbb{N} \rightarrow \mathbb{R}$ betrachten.

Definition:

1. $f = O(g) \Leftrightarrow$ Es gibt eine Konstante $c \geq 0$ und eine Zahl $n_0 \in \mathbb{N}$, so dass $f(n) \leq c \cdot g(n)$ für alle $n \geq n_0$ gilt.
2. $f = o(g) \Leftrightarrow \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$ (man schreibt auch $f \ll g$).

Davon abgeleitete Notationen:

3. $f = \Omega(g) \Leftrightarrow g = O(f)$.
4. $f = \omega(g) \Leftrightarrow \lim_{n \rightarrow \infty} \frac{g(n)}{f(n)} = 0$ (man schreibt auch $f \gg g$)
5. $f = \Theta(g) \Leftrightarrow f = O(g)$ und $g = O(f)$ (man schreibt auch $f \asymp g$)
6. $f \sim g \Leftrightarrow f = (1 + o(1))g$, d.h. wenn $\lim_{n \rightarrow \infty} \frac{g(n)}{f(n)} = 1$

Was besagen die einzelnen Notationen? $f = O(g)$ drückt aus, dass f asymptotisch nicht stärker als g wächst (obwohl $f(n)$ durchaus stets größer als $g(n)$ sein kann). $f = \Omega(g)$ besagt, dass f zumindest

so stark wie g wächst. $f = \Theta(g)$ impliziert, dass f und g gleichstark wachsen. $f = o(g)$ drückt aus, dass f echt schwächer als g wächst.

Bemerkung 3.67. Um den Unterschied zwischen klein- o und groß- O besser zu verstehen, lohnt es sich ihre Definitionen formal zu vergleichen.

$$\begin{aligned} f = O(g) &\iff \exists c > 0 \exists n_0 \forall n \geq n_0 : f(n) \leq c \cdot g(n) \\ f = o(g) &\iff \forall c > 0 \exists n_0 \forall n \geq n_0 : f(n) \leq c \cdot g(n) \end{aligned}$$

Der einzige Unterschied ist also im ersten Quantor! Für $f = O(g)$ reicht es, dass es *mindestens eine* (auch wenn sehr große!) Konstante $c > 0$ gibt, so dass ab bestimmten Schwellwert n_0 die Funktion $f(n)$ durch $cg(n)$ nach oben beschränkt ist. Im Gegensatz dazu sagt $f = o(g)$, dass es für *jede* (auch wenn sehr kleine!) Konstante $c > 0$ wird $f(n)$ ab einem bestimmten Schwellwert n_0 nicht mehr den Wert $cg(n)$ überschreiten.

Die asymptotischen Notationen erwecken zunächst den Anschein, als würden einfach Informationen *weggeworfen*. Dieser Eindruck ist auch sicher nicht falsch. Man mache sich jedoch klar, dass wir, wenn wir eine Aussage wie etwa

$$4n^3 + \frac{2}{3}n - \frac{1}{n^2} = \Theta(n^3)$$

machen, nur untergeordnete Summanden und konstante Faktoren weglassen. Wir werfen also sehr gezielt die Teile weg, die für hinreichend hohe Werte von n nicht ins Gewicht fallen (nachrangige Terme), und wir verzichten auf die führende Konstante (hier 4), da die konkreten Konstanten bei realen Anwendungen ohnehin keine *absolute* konstanten sind. Man kann also sagen: „*Wir reduzieren einen Ausdruck auf das asymptotisch Wesentliche.*“

Alternativ formuliert, die asymptotischen Notationen erlauben es uns, Funktionen in verschiedene *Wachstumsklassen* zusammenzufassen. Obiger Ausdruck hat die Bedeutung: „*Der Ausdruck links gehört in die Klasse der Funktionen, die nicht schneller als n^3 wachsen.*“

Der Einwand, dass bei so einer Zielsetzung das Elementzeichen, also

$$4n^3 + \frac{2}{3}n - \frac{1}{n^2} \in \Theta(n^3),$$

wesentlich intuitiver wäre, ist gerechtfertigt und nur ungenügend mit dem Verweis, das Gleichheitszeichen sei Konvention, zu entkräften.



Die Benutzung vom Gleichheitssymbol in der Bezeichnung “ $f = O(g)$ ” ist zwar fast überall in der Literatur verbreitet, man muss aber immer erinnern, dass das mit der *Gleichheit* zweier Funktionen nichts zu tun hat! Wäre nämlich $f = O(g)$ als eine Gleichheit verstanden, so könnte man auch $O(g) = f$ schreiben. Aber dann kommt man schnell zum Unsinn: es ist $2n = O(n)$ also ist auch $O(n) = 2n$ und, da $n = O(n)$, sollte auch auch $n = O(n) = 2n$ “gelten”. Wir werden deshalb nie “ $O(g) = f$ ” anstatt “ $f = O(g)$ ” schreiben. Man muss sich immer erinnern, dass unter “ $O(g)$ ” eine *Klasse* von Funktionen versteckt ist. Deswegen schreibt man manchmal $f \in O(g)$ anstatt von $f = O(g)$



Aus $f = o(g)$ folgt zwar $g = \Omega(f)$, aber

$$f = o(g) \implies g = O(f)$$

gilt *nicht!* In der Tat, wäre es $g = O(f)$, so sollte $g(n) \leq cf(n)$ für eine positive (da $g \neq 0$) Konstante $c > 0$ und alle $n \geq n_0$ gelten. Aber dann hätten wir $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} \leq \lim_{n \rightarrow \infty} \frac{f(n)}{cf(n)} = \frac{1}{c} \neq 0$, ein Widerspruch mit $f = o(g)$.



Vorsicht mit Exponenten! Zum Beispiel ist $4^n \neq O(2^n)$: $2^n/4^n = 1/2^n$ und somit $\lim_{n \rightarrow \infty} 2^n/4^n = 0$. D.h. $2^n = o(4^n)$ woraus $4^n \neq O(2^n)$ folgt (siehe vorige Bemerkung).



Vorsicht mit Logarithmen! Für das Wachstum der Logarithmus-Funktion gilt:

$$f = O(g) \implies \log_a f = O(\log_a g)$$

aber

$$f = o(g) \implies \log_a f = o(\log_a g)$$

gilt im Allgemeinen *nicht!* Gegenbeispiel: $f(n) = \sqrt{n}$ und $g(n) = n$. Dann ist $\log_a f(n) = \frac{1}{2} \log_a n = \Theta(\log_a g(n))$.

Die Symbolen O, o und Ω, ω haben unterschiedliche Bedeutung: Die ersten zwei geben eine *obere* während die letzten zwei eine *untere* Schranke an.

obere Schranke	untere Schranke
$f = O(g)$	$f = \Omega(g)$
$f = o(g)$	$f = \omega(g)$

Gilt zum Beispiel $f(n) = O(n^2)$, dann bedeutet dies nur, dass $f(n)$ *nicht schneller* als n^2 wächst – es kann gut sein, dass (in Wirklichkeit) $f(n)$ nur linear oder sogar noch langsamer wächst. Zeigt man aber, dass auch $f(n) = \Omega(n^2)$ gilt (eine *untere* Schranke), so kann man bereits sagen, dass (bis zu multiplikativen Faktoren) die Funktion $f(n)$ quadratisch wächst.

Behauptung 3.68. Der Grenzwert $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = L$ möge existieren.

1. Wenn $L = 0$, dann ist $f = o(g)$.
2. Wenn $L < \infty$, dann ist $f = O(g)$.
3. Wenn $0 < L < \infty$, dann ist $f = \Theta(g)$.

Beweis. Übungsaufgabe. □

Diese Behauptung erlaubt es also, eine asymptotische Relation durch Berechnung des entsprechenden Grenzwertes zu verifizieren. Diese Methode versagt nur dann, wenn der Grenzwert nicht existiert.

► *Beispiel 3.69*: Sei $f(n) = \frac{1}{6}n^3 + \frac{1}{2}n^2 + \frac{1}{3}n - 1 = \Theta(n^3)$. Nach der Grenzwertregeln gilt:

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{f(n)}{n^3} &= \lim_{n \rightarrow \infty} \frac{\frac{1}{6}n^3}{n^3} + \lim_{n \rightarrow \infty} \frac{\frac{1}{2}n^2}{n^3} + \lim_{n \rightarrow \infty} \frac{\frac{1}{3}n}{n^3} - \lim_{n \rightarrow \infty} \frac{1}{n^3} \\ &= \frac{1}{6} + \lim_{n \rightarrow \infty} \frac{1}{2n} + \lim_{n \rightarrow \infty} \frac{1}{3n^2} + \lim_{n \rightarrow \infty} \frac{1}{n^3} \\ &= \frac{1}{6}.\end{aligned}$$

Damit ist $0 < \lim_{n \rightarrow \infty} f(n)/n^3 = 1/6 < \infty$ und somit auch $f(n) = \Theta(n^3)$.

Genauso kann man zu Beispiel zeigen, dass die Funktion

$$\pi^2 3^{n-7} + \frac{(2,7n^{113} + n^9 - 86)^4}{\sqrt{n}} - 1,08^{3n}$$

nichts anderes als $\Theta(3^n)$ ist.

► *Beispiel 3.70*: Wir definieren $f(n) = \begin{cases} 1 & n \text{ gerade} \\ 0 & \text{sonst} \end{cases}$

Es sei $g(n) = 1$ für alle $n \in \mathbb{N}$. Dann existiert der Grenzwert von $\frac{f(n)}{g(n)}$ nicht, da die Werte 0 und 1 unendlich oft auftauchen. Es ist aber

$$f = O(g)$$

(mit $c = 1$ und $n_0 = 0$). Man verifiziere, dass mit Ausnahme der Beziehung $g = \Omega(f)$ alle weiteren asymptotischen Relationen zwischen f und g nicht gelten. (Warum ist zum Beispiel die Beziehung $g = \Theta(f)$ falsch?)

Um zu bestimmen, zu welcher Θ -Klasse eine *endliche* Summe $f(n) = g_1(n) + \dots + g_k(n)$ (wenn k ist eine Konstante und damit hängt von n nicht ab) gehört, reicht es die Θ -Klassen der Funktionen $g_i(n)$ zu bestimmen und die größte davon zu nehmen.



Wenn es zwei (oder mehrere) gleichgroße Terme sind, aber einige davon *negativ* sind, dann muss man aufpassen, ob der größte Term nach der Umformung da überhaupt bleibt. So ist zum Beispiel

$$3n + 5n - 6n = 2n = \Theta(n) \text{ aber}$$

$$3(n+2) + 5(n+3) - 8n = \Theta(1),$$

$$\sum_{i=1}^n i - \frac{n^2}{2} = \frac{n+1}{2} = \Theta(n).$$



Gefährlich ist auch dann, wenn die Funktion *unendlich oft* negative Werte annimmt:

$$2 + \sin n = \Theta(1) \text{ aber}$$

$$\sin n \neq \Theta(1) \text{ (unendlich oft negativ)}$$

$$1 + \sin n \neq \Theta(1) \text{ (unendlich oft 0 vorkommt)}$$

$$-2n \neq \Theta(n) \text{ (ist negativ)}$$



Vorsicht mit Exponenten! Die Aussage

$$f = \Theta(g) \implies F(f) = \Theta(F(g))$$

gilt nur, wenn $F(x)$ durch ein Polynom nach oben beschränkt ist. So gilt zum Beispiel

$$f = \Theta(g) \implies f^2 = \Theta(g^2);$$

$$f = \Theta(g) \implies \log f = \Theta(\log g);$$

$$f = \Theta(g) \implies f^{100} = \Theta(g^{100})$$

$$\text{aber } f = \Theta(g) \implies 2^f = \Theta(2^g) \text{ gilt nicht!}$$

Um $f = o(g)$ zu zeigen, muss man den Grenzwert $\lim_{n \rightarrow \infty} f(n)/g(n)$ der Quotientenfunktion $f(n)/g(n)$ betrachten. Das ist aber oft nicht so einfach. Streben zum Beispiel die Funktionen $f(n)$ und $g(n)$ beide gegen 0 oder beide gegen ∞ , so bekommen wir unbestimmte Ausdrücke $0/0$ oder ∞/∞ . Was dann? Bedeutet nun das, dass $f(n)/g(n)$ keinen Grenzwert hat? Nicht unbedingt! Um solche unbestimmte Ausdrücke zu behandeln, gibt es einige Regeln.

Diese Regeln stammen aus dem Buch *Analyse des infiniment petits* (1696) von de l'Hospital.¹⁹



Regeln von de l'Hôpital: Sei $a < b \leq \infty$ und $-\infty \leq L \leq +\infty$. Ferner seien $f, g : [a, b) \rightarrow \mathbb{R}$ differenzierbare Funktionen mit $g'(x) \neq 0$ für alle $x \in [a, b)$, und es gelte $\lim_{x \rightarrow b} f(x) = \lim_{x \rightarrow b} g(x) = 0$ oder $\lim_{x \rightarrow b} g(x) = \pm\infty$. Gilt dann $\lim_{x \rightarrow b} \frac{f'(x)}{g'(x)} = L$, so ist $\lim_{x \rightarrow b} \frac{f(x)}{g(x)} = L$.
Dasselbe gilt auch für Grenzübergang $x \rightarrow a$ anstatt $x \rightarrow b$.

Beweis. Wir betrachten nur den Fall $\lim_{x \rightarrow b} f(x) = \lim_{x \rightarrow b} g(x) = 0$ (der Fall $\lim_{x \rightarrow b} g(x) = \pm\infty$ ist analog). Wir können o.B.d.A. annehmen, dass $g(x) \neq 0$ in $[a, b)$ (sonst kleineres Intervall wählen).²⁰ Wir definieren zwei Hilfsfunktionen:

$$F(x) = \begin{cases} f(x) & \text{falls } x \neq b \\ 0 & \text{falls } x = b \end{cases} \quad \text{und} \quad G(x) = \begin{cases} g(x) & \text{falls } x \neq b \\ 0 & \text{falls } x = b \end{cases}$$

und erhalten mit dem zweiten Mittelwertsatz (Satz 3.64)

$$\frac{f(x)}{g(x)} = \frac{F(x) - F(b)}{G(x) - G(b)} = \frac{F'(c)}{G'(c)} \quad \text{mit } x < c < b$$

Durch Grenzübergang $x \rightarrow b$ (und damit auch $c \rightarrow b$) folgt die Behauptung. □

Bemerkung 3.71. Entsprechen $f(n)$ und $g(n)$ den Laufzeiten irgendwelchen Algorithmen, so sind f, g Abbildungen von \mathbb{N} (und nicht von \mathbb{R}) nach \mathbb{R} , da n die Eingabelänge ist, die üblicher Weise ganzzahlig ist. Deshalb wissen wir nicht, was die Ableitungen $f'(n)$ und $g'(n)$ sein sollten. Trotzdem gibt es einen einfachen Trick, wie man dieses Problem umgehen kann: Man erweitert den Definitionsbereich von \mathbb{N} auf \mathbb{R} , indem man anstatt einer natürlicher Zahl n in den Formeln für $f(n)$ und $g(n)$

¹⁹Guilame Francois Antoine Marquis de l'Hôpital (1661-1704) war der Schuler von Johann Bernoulli (1667-1748). Sein Buch war das erste Buch in Analysis überhaupt. Der Satz selbst war eigentlich von Bernoulli bewiesen worden, trägt aber die Name des Buchauthors. Der Name wird entweder als "l'Hôpital" (alt) oder als "l'Hospital" (neu) geschrieben. Beide spricht man als "Lopital" aus.

²⁰Da $g'(x)$ keine Nullstelle in (a, b) hat, kann $g(x)$ nach dem Satz von Rolle höchstens *einmal* den Wert 0 annehmen.

eine reelle Zahl betrachtet, und zeigt, dass $\frac{f(x)}{g(x)}$ den Grenzwert 0 für $x \rightarrow \infty$ hat. Nach der Definition des Grenzwertes reellen Funktionen, muss dann auch die Folge $\frac{f(n)}{g(n)}$ den Grenzwert 0 haben.

▷ *Beispiel 3.72*: Der Grenzfalle $\frac{\infty}{\infty}$: Für $n \in \mathbb{N}$ erhält man durch n -malige Anwendung der l'Hospitalschen Regeln:

$$\lim_{x \rightarrow \infty} \frac{x^n}{e^x} = \lim_{x \rightarrow \infty} \frac{nx^{n-1}}{e^x} = \dots = \lim_{x \rightarrow \infty} \frac{n!}{e^x} = 0.$$

▷ *Beispiel 3.73*: Der Grenzfalle $\frac{0}{0}$: Auf dem Intervall $I = (0, 1)$ gilt:

$$\lim_{x \rightarrow 1} \frac{\ln x}{x-1} = \lim_{x \rightarrow 1} \frac{x^{-1}}{1} = 1$$

▷ *Beispiel 3.74*: Der Grenzfalle 1^∞ : $\lim_{x \rightarrow 1} x^{1/(x-1)} = ?$

Logarithmieren und Anwenden der l'Hospitalschen Regeln ergibt:

$$\ln x^{1/(x-1)} = \frac{\ln x}{x-1}$$

und somit $\lim_{x \rightarrow 1} x^{1/(x-1)} = e^1 = e$.

▷ *Beispiel 3.75*: Der Grenzfalle $\frac{\infty}{\infty}$: Für $n \in \mathbb{N}$ erhält man durch n -malige Anwendung der l'Hospitalschen Regeln:

$$\lim_{x \rightarrow \infty} \frac{x^n}{e^x} = \lim_{x \rightarrow \infty} \frac{nx^{n-1}}{e^x} = \dots = \lim_{x \rightarrow \infty} \frac{n!}{e^x} = 0.$$

▷ *Beispiel 3.76*: Der Grenzfalle $\frac{0}{0}$: Seien $a, b > 0$. Dann gilt

$$\lim_{x \rightarrow 0} \frac{a^x - b^x}{x} = \ln \frac{a}{b}$$

Beweis: Setze $f(x) := a^x - b^x$, $g(x) := x \Rightarrow f'(x) = a^x \ln a - b^x \ln b$ und $g'(x) = 1 \Rightarrow$

$$\lim_{x \rightarrow 0} \frac{a^x - b^x}{x} = \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} = \lim_{x \rightarrow 0} a^x \ln a - \lim_{x \rightarrow 0} b^x \ln b = \ln a - \ln b = \ln \frac{a}{b}$$

Bemerkung 3.77. Neben der oben diskutierten Grenzübergängen in Quotienten treten auch irreguläre Produktausdrücke der folgender Art auf:

$$\lim_{x \rightarrow a} f(x) = 0, \lim_{x \rightarrow a} g(x) = \infty \quad \Longrightarrow \quad \lim_{x \rightarrow a} f(x) \cdot g(x) = ?$$

Die kann man häufig in der Form

$$\lim_{x \rightarrow a} f(x) \cdot g(x) = \lim_{x \rightarrow a} \frac{f(x)}{g(x)^{-1}}$$

mit den obigen l'Hospitalschen Regeln behandelt werden.

Manchmal sind auch irreguläre Exponentialausdrücke der Form

$$\lim_{x \rightarrow a} f(x)^{g(x)} = ?$$

zu untersuchen, was zu Grenzfällen der Art 0^0 , ∞^0 und 0^∞ führen kann. In diesem Fall wird zunächst logarithmiert,

$$\lim_{x \rightarrow a} g(x) \ln(f(x)) = ?$$

was zu obinem Fall führt. Der Grenzwert des gegebenen Ausdrucks ist dann wegen der Stetigkeit der Exponentialfunktion gegeben durch

$$\lim_{x \rightarrow a} f(x)^{g(x)} = \exp\left(\lim_{x \rightarrow a} g(x) \ln(f(x))\right).$$

Mittels der Transformation

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

wird der Grenzfall $\infty - \infty$ in den Fall $\frac{0}{0}$ überführt.

▷ *Beispiel 3.78* : $\lim_{x \rightarrow 0} x^x = ?$. Logarithmieren und die l'Hospital Regel ergibt:

$$\begin{aligned} \lim_{x \rightarrow 0} x \cdot \ln x &= \lim_{x \rightarrow 0} \frac{\ln x}{x^{-1}} \quad (\rightarrow \frac{\infty}{\infty} \Rightarrow \text{kann l'Hospital anwenden}) \\ &= - \lim_{x \rightarrow 0} \frac{1}{x^2} \quad (\text{l'Hospital}) \\ &= 0 \end{aligned}$$

und somit

$$\lim_{x \rightarrow 0} x^x = e^0 = 1.$$



Manchmal funktionieren die Regeln von l'Hospital nicht, da die Ausdrücke hin-und-hier oszillieren. Zum Beispiel

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{x}{(x^2 + 1)^{1/2}} &= \lim_{x \rightarrow \infty} \frac{1}{x(x^2 + 1)^{-1/2}} = \lim_{x \rightarrow \infty} \frac{(x^2 + 1)^{1/2}}{x} \\ &= \lim_{x \rightarrow \infty} \frac{x(x^2 + 1)^{-1/2}}{1} = \lim_{x \rightarrow \infty} \frac{x}{(x^2 + 1)^{1/2}}. \end{aligned}$$

Wir betrachten als nächstes eine Wachstums-Hierarchie wichtiger Laufzeitfunktionen.

Lemma 3.79. (Wachstum von Standardfunktionen) Seien $a, b \in \mathbb{R}_+$, dann gilt:

1. Wenn $a < b$, dann gilt $x^a = o(x^b)$.
2. Es gilt (auch wenn b sehr groß und a sehr klein): $(\ln x)^b = o(x^a)$. D.h. logarithmisches Wachstum ist unwesentlich gegenüber dem Wachstum von Polynomen.
3. Es gilt (auch wenn b sehr groß und a sehr klein): $x^b = o(2^{a \cdot x})$. D.h. polynomiales Wachstum ist unwesentlich gegenüber dem Wachstum von Potenzen.

Beweis.

$$\lim_{x \rightarrow \infty} x^a/x^b = \lim_{x \rightarrow \infty} \frac{1}{x^{b-a}} = 0 \quad (\text{da } b - a > 0).$$

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{(\ln x)^b}{x^a} &= \lim_{x \rightarrow \infty} \frac{\ln x}{x^{a/b}} = \lim_{x \rightarrow \infty} \frac{\frac{1}{x}}{\frac{a}{b} x^{\frac{a}{b}-1}} \quad (\text{Regel von de l'Hôpital}) \\ &= \frac{b}{a} \cdot \lim_{x \rightarrow \infty} \frac{1}{x^{a/b}} = 0 \end{aligned}$$

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{x^b}{2^{a \cdot x}} &= \lim_{x \rightarrow \infty} \frac{x}{2^{a \cdot x/b}} = \lim_{x \rightarrow \infty} \frac{x}{e^{c \cdot x}} \quad \text{mit } c = (a \ln 2)/b > 0 \\ &= \lim_{x \rightarrow \infty} \frac{1}{c e^{c \cdot x}} = 0 \quad (\text{Regel von de l'Hôpital}) \end{aligned}$$

□

Als Folgerung haben wir die folgenden Wachstumsstufen.²¹ Für $a > 1, k > 1$ und $b > 1$ gilt:

$$\log_a n \ll n \ll n \log_a n \ll n^k \ll b^n \ll n!$$

► *Beispiel 3.80* : Gegeben sind 4 Algorithmen A_1, A_2, A_3, A_4 mit entsprechenden Laufzeiten

$$\begin{aligned} T_1(n) &= 1000n, \\ T_2(n) &= 200n \log_2 n, \\ T_3(n) &= 10n^2, \\ T_4(n) &= 2^n. \end{aligned}$$

$$\text{Schnellster Algorithmus} : = \begin{cases} A_4 & \text{für } 1 \leq n \leq 9 \\ A_3 & \text{für } 10 \leq n \leq 100 \\ A_1 & \text{für } n \geq 101. \end{cases}$$

► *Beispiel 3.81* : Bei einer Millisekunde (10^{-3}) pro Operation und einer Stunde zur Verfügung ist die größte mögliche Problemstellung (d.h. die größte mögliche Input-Länge):

$$\overline{\max n : T(n) \leq 3,6 \cdot 10^3} \quad \left\| \begin{array}{c|c|c|c} A_1 & A_2 & A_3 & A_4 \\ \hline 3600 & 1600 & 600 & 21 \end{array} \right.$$

Verzehnfachung der Maschinengeschwindigkeit:

$$\overline{\max n : T(n) \leq 3,6 \cdot 10^3} \quad \left\| \begin{array}{c|c|c|c} A_1 & A_2 & A_3 & A_4 \\ \hline 36000 & 13500 & 1900 & 25 \end{array} \right.$$

²¹Zur Erinnerung: $f \ll g$ ist einfach eine andere Schreibweise für $f = o(g)$.

► *Beispiel 3.82* : In einem Wettbewerb sollen 1.000.000 Zahlen sortiert werden. Teilnehmer sind ein schneller Rechner und ein PC, auf deren aber *verschieden schnelle* Sortier-Algorithmen laufen.

	Op. pro sek.	Anzahl d. Operationen
schneller Rechner	100 Mio.	$2n^2$
PC	1 Mio	$50n \log n$

Rechenzeit:

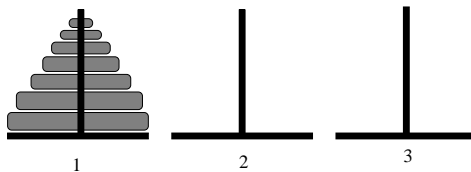
schneller Rechner	$\frac{2 \cdot (10^6)^2 \text{ Oper.}}{10^8 \text{ Oper./sec}}$	= 20.000 sec	≈ 5.56 h
PC	$\frac{50 \cdot 10^6 \cdot \log 10^6 \text{ Oper.}}{10^6 \text{ Oper./sec}}$	≈ 1.000 sec	≈ 16,67 min

Fazit: Es lohnt sich, sowohl eine größenordnungsbezogene wie auch eine asymptotische Analyse durchzuführen.

3.10 Rekurrenzen*

Im Spiel „Türme von Hanoi“ (siehe Abschnitt 1.4.1) haben wir drei Stäbe 1, 2 und 3. Ursprünglich besitzt Stab 1 n Ringe, wobei die Ringe in absteigender Größe auf dem Stab aufgereiht sind (mit dem größten Ring als unterstem Ring). Die Stäbe 2 und 3 sind leer.

Ein Zug besteht darin, einen zuoberst-liegenden Ring von einem Stab zu einem anderen zu bewegen. Der Zug ist aber nur dann erlaubt, wenn der Ring auf einen größeren Ring gelegt wird oder wenn der Stab leer ist. Das Spiel ist erfolgreich beendet, wenn alle Ringe von Stab 1 nach Stab 2 bewegt wurden.



Wir haben dieses Spiel bereits in Abschnitt 1.4.1 betrachtet und haben einen folgenden rekursiven Algorithmus entworfen:

Algorithmus Hanoi(n ; 1, 2, 3)

While $n > 0$ do

Rufe Hanoi($n - 1$; 1, 3, 2) auf [die obersten $n - 1$ Scheiben von Stapel 1 zum Stapel 2]

Verlege die zuoberst liegende Scheibe auf 1 nach 3

Rufe $\text{Hanoi}(n-1; 2, 1, 3)$ auf [verlege $n-1$ Scheiben vom Hilfsstapel 2 nach 3]

Sei a_n die minimale Anzahl der Züge, die ausreichen sind, alle Ringe von Stab 1 nach Stab 2 zu bewegen. Aus dem oben angegebenen Algorithmus folgt: $a_n \leq 2 \cdot a_{n-1} + 1$. Man kann auch zeigen (probiere das!), dass $a_n \geq 2 \cdot a_{n-1} + 1$ gilt. Also haben wir die folgende Rekursionsgleichung:

$$\begin{aligned} a_0 &= 0 \\ a_n &= 2 \cdot a_{n-1} + 1 \quad \text{für } n > 0 \end{aligned}$$

Wenn wir nun die Rekursionsgleichung entwickeln wollen, bekommen wir in jedem Schritt einen "störenden" Term $+1$. Stattdessen wenden wir einen (of sehr hilfreichen) Trick an: konstruiere eine neue einfachere Rekursionsgleichung. In unserem Fall reicht es eine Eins auf beiden Seiten zu addieren:

$$\begin{aligned} a_n + 1 &= 1 \\ a_n + 1 &= 2 \cdot a_{n-1} + 2 \quad \text{für } n > 0 \end{aligned}$$

Nun setzen wir $b_n := a_n + 1$ und erhalten

$$\begin{aligned} b_0 &= 1 \\ b_n &= 2 \cdot b_{n-1} \quad \text{für } n > 0 \end{aligned}$$

und diese letzte Rekursionsgleichung ist leicht zu lösen:

$$b_n = 2b_{n-1} = 2^2b_{n-2} = 2^3b_{n-3} = \dots = 2^i b_{n-i} = \dots = 2^n b_0 = 2^n.$$

Also ist $a_n = b_n - 1 = 2^n - 1$.



(Das Pizzaproblem, Jacob Steiner 1826) Wieviele Pizzascheiben kann man bekommen, wenn man die Pizza mit n geraden Schnitten aufteilt? Oder mehr mathematisch: Was ist die maximale Anzahl x_n der Flächen in der Ebene, die man mit n Linien bekommen kann?

Bestimmt kann man denken, dass $x_n \approx 2^n$ gelten sollte: jede neue Gerade verdoppelt die Anzahl der Pizzascheiben! Diese erste Reaktion ist aber total daneben! Die Funktion x_n wächst viel langsamer, nämlich $x_n = \Theta(n^2)$.

1. Die n -te Gerade ergibt k neuen Flächen genau dann, wenn sie k alten Flächen schneidet, und sie kann so viele alten Flächen schneiden genau dann, wenn sie die alten Geraden in genau $k-1$ Punkte trifft.
2. Zwei Geraden können sich in höchstens einem Punkt treffen.
3. Aus (2) folgt, dass die neue Gerade kann die $n-1$ alten Geraden in höchstens $n-1$ Punkt treffen.
4. Aus (1) und (3) folgt, dass $k \leq n$.

Damit erhalten wir die Rekursionsungleichung

$$x_n \leq x_{n-1} + n \quad \text{für } n > 0$$

Man kann zeigen, dass hier auch Gleichheit gilt: es reicht die neue Gerade so zu wählen, dass sie parallel zu keiner der alten ist und keinen der alten Treffpunkte berührt. Damit haben wir die Rekursionsgleichung:

$$\begin{aligned}x_0 &= 1 \\x_n &= x_{n-1} + n \quad \text{für } n > 0\end{aligned}$$

Damit ist $x_n = \sum_{k=1}^n k$ die (bereits und bekannte) arithmetische Reihe, und wir wissen die Antwort (siehe (3.1):

$$x_n = \frac{n(n+1)}{2} = \Theta(n^2).$$

Rekurrenzen vom Grad 2

Die Beiden oben betrachteten Folgen waren “linear” und hatten die Form $x_n = Ax_{n-1} + B$. In Anwendungen kommen aber kompliziertere Folgen vor, wo jeder Folgeglied x_n durch eine lineare Funktion von $d \geq 2$ letzten Folgeglieder $x_{n-1}, x_{n-2}, \dots, x_{n-d}$ bestimmt ist. Solche Rekurrenzen nennt man *Rekurrenzen vom Grad d* . Für solchen Folgen gibt es auch einige Tricks, um die geschlossene Form für x_n zu finden. In diesem Abschnitt betrachten wir den Falls $d = 2$. (Rekurrenzen vom höheren Grad werden wir im nächsten Abschnitt betrachten.)

Die Folge $\langle x_n \rangle = x_0, x_1, x_2, \dots$ sei durch die ersten zwei Zahlen $x_0 = a_0, x_1 = a_1$ und eine Rekurrenz

$$x_n = Ax_{n-1} + Bx_{n-2} \tag{3.25}$$

oder äquivalent durch die Gleichung

$$x_n - Ax_{n-1} - Bx_{n-2} = 0$$

gegeben. Wir wollen eine Funktion $f(n)$ mit $x_n = f(n)$ für alle n bestimmen. Dazu gibt es ein allgemeines Verfahren. Dazu betrachtet man die Nullstellen des *charakteristischen Polynoms* $z^2 - Az - B$, d.h. man betrachtet die Lösungen des quadratischen Gleichungs

$$z^2 - Az - B = 0.$$

Satz 3.83. Seien r und s die Lösungen von $z^2 - Az - B = 0$. Dann hat die Lösung der Rekursionsgleichung (3.25) die Form

$$x_n = \begin{cases} ar^n + bs^n & \text{falls } r \neq s \\ ar^n + bnr^n & \text{falls } r = s \end{cases}$$

wobei die Zahlen a und b (eindeutig) durch die Randbedingungen $a + b = a_0, ra + sb = a_1$ bzw. $a = a_0, ra + sb = a_1$ bestimmt sind.

Beweis. Zuerst beobachten wir, dass $x_n = ar^n$ für jede reelle Zahl $a \neq 0$ und jedes r mit $r^2 - Ar - B = 0$ eine Lösung der Rekursionsgleichung (3.25) ist:

$$ar^n = Aar^{n-1} + Bar^{n-2} \iff r^2 = Ar + B \iff r^2 - Ar - B = 0.$$

Es reicht also zu zeigen (Übungsafgabe!), dass die Summe $u_n = s_n + t_n$ je zwei Lösungen s_n und t_n von (3.25) auch eine Lösung von (3.25) ist. Um die Zahlen a und b zu bestimmen, reicht es die ersten zwei Folgenglieder $x_0 = a_0$ und $x_1 = a_1$ zu betrachten:

$$\begin{aligned} a_0 &= ar^0 + bs^0 = a + b \\ a_1 &= ar^1 + bs^1 = ar + bs. \end{aligned}$$

□

► *Beispiel 3.84*: Die berühmte Folge der *Fibonacci-Zahlen* $(x_n) = x_1, x_2, x_3, \dots$ ist durch $x_1 = x_2 = 1$ und die Rekursion $x_n = x_{n-1} + x_{n-2}$ gegeben. Die Nullstellen des charakteristischen Polynoms $z^2 - z - 1$ sind

$$r = \frac{1 + \sqrt{5}}{2} \quad \text{und} \quad s = \frac{1 - \sqrt{5}}{2}.$$

Die Randbedingungen $a + b = 1$ und $ar + bs = 1$ ergeben: $a = 1/\sqrt{5}$ und $b = -1/\sqrt{5}$.

Damit gilt:

$$x_n = \frac{1}{\sqrt{5}} \left(\frac{1 + \sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1 - \sqrt{5}}{2} \right)^n.$$

► *Beispiel 3.85*: Sei die Folge (x_n) durch die Randbedingungen $x_0 = 2$, $x_1 = 7$ und die Rekursionsgleichung $x_n = x_{n-1} + 2x_{n-2}$ gegeben. Das charakteristische Polynom in diesem Fall ist $z^2 - z - 2 = (z - 2)(z + 1)$. Seine Nullstellen sind also $r = 2$ und $s = -1$. Damit hat x_n die Form $x_n = ar^n + bs^n = a2^n + b(-1)^n$, wobei a und b sind durch das Gleichungssystem

$$\begin{aligned} a + b &= x_0 = 2 \\ 2a - b &= x_1 = 7 \end{aligned}$$

bestimmt, woraus $a = 3$ und $b = -1$ folgt. Damit ist $x_n = 3 \cdot 2^n - (-1)^n$ die gesuchte Lösung.

In der Stochastik treten endliche Folgen x_0, x_1, \dots, x_N , die durch die Rekursionsgleichungen von der Form $x_n = px_{n+1} + (1-p)x_{n-1}$ mit $x_0 = 0$, $x_N = 1$ und $0 < p \leq 1/2$ auf.

► *Beispiel 3.86*: (**Gambler's Ruin**) Ein Spieler namens Theo Retiker nimmt in einem Casino an einem Spiel mit Gewinnwahrscheinlichkeit $0 < p \leq 1/2$ teil.²² Zum Beispiel wirft man eine (nicht unbedingt faire) Münze, dessen Seiten mit rot und blau gefärbt sind, und wir gewinnen, falls rot kommt.

Wir nehmen an, dass Theo in jedem Schritt (oder Spielrunde) nur 1 € einsetzen kann. Geht die Runde zu Theos Gunsten aus, erhält er den Einsatz zurück und zusätzlich denselben Betrag aus der Bank (Gewinn = 1 €). Endet die Runde ungünstig, verfällt der Einsatz (Gewinn = -1 €).

Theo kommt ins Casino mit n Euro (Anfangskapital) und sein Ziel ist m Euro zu gewinnen (dann will er aufhören); in diesem Fall sagen wir, dass Theo gewinnt. Theo spielt bis er m Euro gewinnt oder bis er alle seine mitgenommenen n Euro verliert.

Sei $N = n + m$ fest, und sei x_n die Wahrscheinlichkeit, dass Theo gewinnt, wenn sein Anfangskapital n ist. Also $x_0 = 0$ und $x_N = 1$. Nehmen wir nun an, dass Theo mit Anfangskapital n

²²Natürlich, wird kein Casino eine Gewinnwahrscheinlichkeit $p > 1/2$ zulassen.

($0 < n < N$) beginnt. Nach der ersten Runde wird Theo mit Wahrscheinlichkeit p gewinnen und $n + 1$ Euro haben; danach wird er Gewinner mit Wahrscheinlichkeit x_{n+1} . Andererseits kann Theo die erste Wette mit Wahrscheinlichkeit $q = 1 - p$ verlieren; dann wird er nur $n - 1$ Euro haben und kann nur mit Wahrscheinlichkeit x_{n-1} das ganze Spiel gewinnen. Insgesamt ist Theo der Gewinner mit Wahrscheinlichkeit $x_n = px_{n+1} + qx_{n-1}$.

Wir können die Gleichung $x_n = px_{n+1} + qx_{n-1}$ als

$$px_{n+1} - x_n + qx_{n-1} = 0 \quad (3.26)$$

umschreiben. Wir wiederum raten die Lösung in der Form $x_n = z^n$ mit $z > 0$. Dann erhalten wir aus (3.26) die Gleichung $pz^{n+1} - z^n + z^{n-1}q = 0$. Dividieren wir dann beide Seiten durch z^{n-1} und erhalten eine quadratische Gleichung:

$$pz^2 - z + (1 - p) = 0$$

Lösen wir diese Gleichung, so bekommen wir:

$$\begin{aligned} z_{1,2} &= \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2p} \\ &= \frac{1 \pm (1 - 2p)}{2p} \\ &= \frac{1-p}{p} \text{ oder } 1 \end{aligned}$$

Die Lösungen sind verschieden genau dann, wenn $p \neq 1/2$. Damit ergeben sich zwei Fälle $p < 1/2$ und $p = 1/2$.

Fall 1: $p < 1/2$. Im diesem Fall haben wir zwei *verschiedene* Lösungen $r = (1-p)/p$ und $s = 1$. Wir können deshalb entweder $x_n = r^n$ oder $x_n = s^n = 1$ nehmen, und die Gleichung (3.26) wird erfüllt. Da die linke Seite von (3.26) für $x_n = r^n$ und für $x_n = 1$ gleich Null ist, wird auch

$$x_n := a \cdot r^n + b \cdot 1$$

für beliebige a und b die Gleichung (3.26) erfüllen. Es bleibt also die Parameter a und b zu bestimmen. Hier benutzen wir die Randbedingungen:

$$\begin{aligned} 0 = a_0 &= a + b \\ 1 = x_N &= a \cdot r^N + b. \end{aligned}$$

Wir lösen wir dieses Gleichungssystem und erhalten:

$$b = -a, \quad a = \frac{1}{r^N - 1}$$

und deshalb

$$\begin{aligned} x_n &= a \cdot r^n + b = \frac{1}{r^N - 1} \cdot r^n - \frac{1}{r^N - 1} = \frac{r^n - 1}{r^N - 1} \\ &\leq \frac{r^n}{r^N} = \frac{r^n}{r^{n+m}} = r^{-m} \\ &= \left(\frac{p}{1-p} \right)^m \leq e^{-m/(1-p)} \\ &< e^{-m} \end{aligned}$$

Fall 2: $p = 1/2$. In diesem Fall hat die Gleichung $pz^2 - z + q = 0$ nur eine Lösung $r = 1$, und die Lösung $x_n = r^n = 1$ für (3.26) sagt uns nichts. Aber in diesem Fall (wenn $p = 1/2$) ist $x_n = an + b$ eine offensichtliche Lösung für (3.26). Aus $x_0 = 0$ und $x_N = 1$ folgt, dass $b = 0$ und $a = 1/N$. Also auch in diesem Fall haben wir die gleiche Lösung

$$x_n = \frac{n}{N} = \frac{n}{n+m}.$$

Damit haben wir die folgende interessante Aussage über die Gewinnchancen bewiesen.

Satz 3.87. Die Gewinnwahrscheinlichkeit in jeder Spielrunde sei $0 < p \leq 1/2$. Die Gewinnwahrscheinlichkeiten mit Anfangskapital n Euro und dem Wunsch, m Euro zu gewinnen, sind

1. genau $n/(n+m)$, falls $p = 1/2$.
2. kleiner als e^{-m} , falls $p < 1/2$.

► **Beispiel 3.88: (Gambler's Ruin - Vortsetzung)** Schauen wir nun an, was dieser Satz für Theo bedeutet.

Fall 1: Faire Münze ($p = 1/2$). In diesem Fall hängt für jedes (feste) m die Wahrscheinlichkeit m Euro zu gewinnen vom Anfangskapital n ab: Je größer es ist, desto größer ist die Wahrscheinlichkeit. Wenn zum Beispiel Theo mit $n = 500$ € startet und $m = 100$ € gewinnen will, dann ist seine Gewinnwahrscheinlichkeit

$$\frac{n}{n+m} = \frac{500}{500+100} = \frac{5}{6}$$

Nicht schlecht – mit einer fairen Münze kann man wohl spielen! Wenn Theo mit $n = 1.000.000$ € startet und $m = 100$ € gewinnen will, dann ist seine Gewinnwahrscheinlichkeit

$$\frac{n}{n+m} > 0,9999 \dots$$

In diesem Fall gewinnt Theo (ein Millionär) seine 100 € fast sicher.

Fall 2: Unfaire Münze ($p < 1/2$) (amerikanisches Casino). Nehmen wir nun an, dass Theo in ein Casino in U.S.A. geht und immer auf "rot" wettet. Das Rouletterad in Amerika hat 18 schwarze Nummern, 18 rote Nummern und 2 grüne Nummern. Man kann aber nur auf eine rote oder schwarze Nummer setzen, aber nicht auf grünen. Also ist in diesem Fall die Gewinnwahrscheinlichkeit in jeder Runde gleich $p = 18/38 \approx 0,47$. Die 0,03 Gewinnchance hat das Casino für sich selbst bestimmt aber die Gewinnchance für Theo sieht trotzdem "fast" fair aus. Natürlich ist dann die Gewinnwahrscheinlichkeit ein bisschen kleiner als $5/6$, aber Theo hofft, dass das nicht so dramatisch ist: Wie kann dieser kleine $-0,03$ Unterschied in Gewinnchance irgendwas essenziell verändern? Leider, leider ... Nach dem Satz 3.87 ist in diesem Fall die Gewinnwahrscheinlichkeit sogar kleiner als $1/37.000$, und zwar – egal mit welchem Anfangskapital n Theo sein Spiel beginnt!

Beachte, dass die obere Schranke e^{-m} in diesem Fall (wenn $p < 1/2$) *nicht mehr* vom Anfangskapital n abhängt! Und die Konsequenzen sind erstaunlich: Angenommen, Theo startet mit $n = 500$ € und will $m = 100$ € gewinnen. Dann ist

$$\Pr \{ \text{Theo gewinnt} \} < \left(\frac{18/38}{20/38} \right)^{100} = \left(\frac{9}{10} \right)^{100} < \frac{1}{37.648}.$$

Das ist ein dramatischer Unterschied zu dem fairen Spiel, wo (wie wir bereits wissen) die Gewinnchance sogar $5/6$ war. Wir haben auch gesehen, dass im fairen Spiel mit Anfangskapital $n = 1.000.000 \text{ €}$ Theo die $m = 100 \text{ €}$ fast sicher gewinnen kann (mit Wahrscheinlichkeit $0,9999$). Aber im amerikanischen Casino wird Theo auch mit so großem Anfangskapital und so kleinen Ambitionen (nur 100 € zu gewinnen) *fast sicher sein Million verlieren!*

Fazit: Wenn überhaupt, dann nicht in Amerika spielen!

Rekurrenzen vom höheren Grad

Im allgemeinen haben (lineare) Rekurrenzen die Form für $n \geq d$

$$x_n = A_1 x_{n-1} + A_2 x_{n-2} + \dots + A_d x_{n-d} + f(n) \quad (3.27)$$

oder äquivalent

$$x_n - A_1 x_{n-1} - A_2 x_{n-2} - \dots - A_d x_{n-d} = -f(n). \quad (3.28)$$

Die Rekurrenz ist *homogen*, falls $f(n) = 0$ für alle n gilt, und für solchen Rekurrenzen kann man denselben Trick wie im Fall $d = 2$ anwenden. Man betrachtet wiederum das charakteristische Polynom

$$p(z) = z^d - A_1 z^{d-1} - A_2 z^{d-2} - \dots - A_d.$$

Eine Nullstelle r von $p(z)$ hat *Vielfachheit* k , falls $p(z)$ ist durch $(z - r)^k$ teilbar. Der folgender Satz ist eine direkte Verallgemeinerung des Satzes 3.83.

Satz 3.89. Sei r eine Nullstelle des charakteristischen Polynoms $p(z)$ von (3.27) von Vielfachheit k . Dann erfüllen alle k Folgen

$$r^n, nr^n, n^2 r^n, \dots, n^{k-1} r^n$$

die Rekursionsgleichung (3.27). Außerdem, läßt sich *jede* Lösung von (3.27) als eine Linearkombination von solchen Lösungen über alle Nullstellen von $p(z)$ darstellen.

▷ *Beispiel 3.90*: Betrachte die Rekursionsgleichung

$$x_n = 3x_{n-1} - 4x_{n-3}$$

mit Randbedingungen $x_0 = 0$, $x_1 = 1$ und $x_2 = 13$. Wenn man beginnt die ersten Folgeglieder zu berechnen, bekommt man $39, 113, 287, 705, 1663, \dots$. Es scheint keine vernünftige Formel für x_n zu existieren. Trotzdem, wir können eine solche Formel ziemlich leicht finden. Das charakteristische Polynom für diese Rekurrenz hat die Form

$$p(z) = z^3 - 3z^2 + 4 = (z + 1)(z - 2)^2.$$

Das Polynom hat also drei Nullstellen: -1 mit Vielfachheit 1 und 2 mit Vielfachheit 2. Nach Satz 3.89 ist jede Lösung unserer Rekursionsgleichung eine Linearkombination von $(-1)^n$, 2^n und $n2^n$. D.h.

$$x_n = A(-1)^n + B2^n + Cn2^n$$

für irgendwelche Konstanten A, B und C . Um diese Konstanten zu bestimmen, benutzen wir die Randbedingungen:

$$\begin{aligned}x_0 = 0 &\implies A + B = 0 \\x_1 = 1 &\implies -A + 2B + 2C = 1 \\x_2 = 13 &\implies A + 4B + 8C = 13.\end{aligned}$$

Löst man dieses Gleichungssystem, so bekommt man $A = 1$, $B = -1$ und $C = 2$, und wir sind fertig:

$$x_n = (-1)^n - 2^n + 2n2^n = (2n - 1)2^n + (-1)^n.$$

Die Situation mit *inhomogenen* Rekurrenzen, d.h. Rekurrenzen (3.27) mit $f(n) \neq 0$, ist komplizierter. In solchen Fällen kann man versuchen, die Rekurrenz auf eine homogene Rekurrenz zu reduzieren.

▷ *Beispiel 3.91* : Betrachte die Rekurrenz

$$a_{n+2} = a_{n+1} + a_n + 2^n. \quad (3.29)$$

Diese Rekurrenz ist inhomogen wegen dem Term $f(n) = 2^n$. Man kann aber sie “homogenisieren”, indem man eine neue “verschobene” Rekurrenz betrachtet und ihre Vielfache aus der Original-Rekurrenz abzieht. In unserem Beispiel können wir die Rekurrenz nach links verschieben

$$\begin{aligned}a_{n+2} &= a_{n+1} + a_n + 2^n \\a_{n+1} &= a_n + a_{n-1} + 2^{n-1}.\end{aligned}$$

Nun kann man die zweifache der zweiten Gleichung von der ersten abziehen, um den Term 2^n zu eliminieren:

$$a_{n+2} - 2a_{n+1} + a_{n+1} + a_n - 2a_n - 2a_{n-1}$$

oder äquivalent

$$a_{n+2} = 3a_{n+1} - a_n - 2a_{n-1}.$$

Das ist bereits eine homogene Rekurrenz und wir können sie mit Hilfe von Satz 3.89 lösen.

Die Allgemeine Methode

Wir betrachten die folgende zwei Operatoren, die eine Folge $\langle a_n \rangle$ in eine andere Folge $\langle b_n \rangle$ überführen:

$$\mathbf{E}\langle a_n \rangle := \langle a_{n+1} \rangle \quad \text{und} \quad c\langle a_n \rangle := \langle ca_n \rangle.$$

D.h. der Operator \mathbf{E} überführt die Folge a_0, a_1, a_2, \dots in die Folge a_1, a_2, a_3, \dots indem er einfach das erste Element eliminiert. Der Operator überführt die Folge a_0, a_1, a_2, \dots in die Folge ca_0, ca_1, ca_2, \dots . Addition und Multiplikation von Operatoren A und B sind definiert durch:

$$\begin{aligned}(A + B)\langle a_n \rangle &:= A\langle a_n \rangle + B\langle a_n \rangle \\(A \cdot B)\langle a_n \rangle &:= A(B\langle a_n \rangle).\end{aligned}$$

Zum Beispiel

$$\begin{aligned}(2 + \mathbf{E})\langle a_n \rangle &= \langle 2a_n + a_{n+1} \rangle \\ \mathbf{E}^2 \langle a_n \rangle &= \langle a_{n+2} \rangle.\end{aligned}$$

Beobachtung: Ist $\langle a_n \rangle$ durch eine homogene Rekursionsgleichung definiert und ist $p(z)$ das charakteristische Polynom dieser Gleichung, so gilt

$$p(\mathbf{E})\langle a_n \rangle = \langle 0 \rangle = 0, 0, 0, \dots$$

In diesem Fall sagt man auch, dass $p(\mathbf{E})$ ein *Annihilator* für die Folge $\langle a_n \rangle$ ist. Zum Beispiel ist $\mathbf{E} - 2$ ein Annihilator für die Folge $\langle 2^n \rangle$, da

$$(\mathbf{E} - 2)\langle 2^n \rangle = \langle 2^{n+1} - 2 \cdot 2^n \rangle = \langle 0 \rangle.$$

Annihilatoren für einige wichtige Folgen sind:

Folge	Annihilator
$\langle c \rangle$	$\mathbf{E} - 1$
$\langle \text{Polynom in } n \text{ vom Grad } k \rangle$	$(\mathbf{E} - 1)^{k-1}$
$\langle c^n \rangle$	$(\mathbf{E} - c)$
$\langle c^n \text{ mal Polynom in } n \text{ vom Grad } k \rangle$	$(\mathbf{E} - c)^{k-1}$

So ist zum Beispiel $(\mathbf{E} - 2)^2$ der Annihilator für die Folge $\langle n2^n \rangle$.

Eine nützliche Eigenschaft der Annihilatoren ist folgende:

Ist A ein Annihilator für $\langle a_n \rangle$ und B der Annihilator für $\langle b_n \rangle$, so ist $A \cdot B$ der Annihilator für $\langle a_n + b_n \rangle$.

Zum Beispiel ist $(\mathbf{E} - 3)^2(\mathbf{E} - 1)$ ein Annihilator für $\langle n2^n + 1 \rangle$.

Die allgemeine Vorgehensweise zur Lösung von Rekursionsgleichungen in der Form (3.28) ist folgende:

1. Wende den Annihilator für die rechte Seite von (3.28) auf *beide* Seiten an.
2. Löse die resultierende *homogene* Rekursionsgleichung.

▷ *Beispiel 3.92* : Wir betrachten die Rekursionsgleichung

$$a_n - 5a_{n-1} + 6a_{n-2} = 4$$

mit $a_0 = 5$ und $a_1 = 7$. In unserer neuen Notation hat diese Rekursionsgleichung die Form

$$(\mathbf{E}^2 - 5\mathbf{E} + 6)\langle a_n \rangle = \langle 4 \rangle.$$

Wir wenden den Operator $\mathbf{E} - 1$ an, um 4 zu annihilieren:

$$(\mathbf{E} - 1)(\mathbf{E}^2 - 5\mathbf{E} + 6)\langle a_n \rangle = \langle 0 \rangle.$$

Das charakteristische Polynom hat also die Form $p(z) = (z-1)(z^2-5z+6) = (z-1)(z-2)(z-3)$, und seine Nullstellen sind 1, 2 und 3. Das ergibt die Lösung in der Form

$$a_n = A + B2^n + C3^n.$$

Wir wissen, dass $a_0 = 5, a_1 = 7$ und $a_2 = 5a_1 + 6a_0 = 65$. Das gibt uns die Gleichungssystem

$$\begin{aligned} A + B + C &= a_0 = 5 \\ A + 2B + 3C &= a_1 = 7 \\ A + 4B + 9C &= a_2 = 9 \end{aligned}$$

woraus $A = 2, B = 4$ und $C = -1$ folgt. Die Lösung der Rekursionsgleichung also ist

$$a_n = 2 + 4 \cdot 2^n - 3^n.$$

► *Beispiel 3.93* : Wir betrachten die Rekursionsgleichung

$$a_n - 2a_{n-1} = 2^n - 1$$

mit $a_0 = 0$. In unserer neuen Notation hat diese Rekursionsgleichung die Form

$$(\mathbf{E} - 2)\langle a_n \rangle = \langle 2^{n+1} - 1 \rangle.$$

Wir wenden den Operator $(\mathbf{E} - 2)(\mathbf{E} - 1)$ an, um $2^{n+1} - 1$ zu annihilieren:

$$(\mathbf{E} - 1)(\mathbf{E} - 2)^2 \langle a_n \rangle = \langle 0 \rangle.$$

Das charakteristische Polynom hat also die Form $p(z) = (z-1)(z-2)^2$, und seine Nullstellen sind 1 und 2 (mit Vielfachheit 2). Das ergibt die Lösung in der Form

$$a_n = (A + Bn)2^n + C.$$

Wir wissen, dass $a_0 = 0, a_1 = 2 \cdot 0 + 2^1 - 1 = 1$ und $a_2 = 2a_1 + 2^2 - 1 = 5$. Das gibt uns die Gleichungssystem

$$\begin{aligned} A + 0 + C &= a_0 = 0 \\ 2A + 2B + C &= a_1 = 1 \\ 4A + 8B + C &= a_2 = 5 \end{aligned}$$

woraus $A = -1, B = 1$ und $C = 1$ folgt. Die Lösung der Rekursionsgleichung also ist

$$a_n = (n-1)2^n + 1.$$

3.10.1 Das Master Theorem

Viele Algorithmen arbeiten *rekursiv*, d.h. sie bestehen aus Schritten, in denen der Algorithmus auf einigen *kleineren* Eingaben aufgerufen wird. Ihre Laufzeit-Funktion $T(n)$ (= maximale Laufzeit der Eingaben der Länge n) ist dann auch rekursiv definiert. Damit bekommt man oft eine Rekursionsgleichung der Form

$$T(n) = a \cdot T(n/b) + f(n) \tag{3.30}$$

Eine solche Rekursionsgleichung beschreibt zum Beispiel die Laufzeit eines Algorithmus, der die Eingabe der Länge n in a Teilprobleme der Länge n/b zerlegt; diese durch a rekursive Aufrufe löst, und aus den erhaltenen Teillösungen die Gesamtlösung zusammensetzt. Hierbei ist $f(n)$ der Aufwand, der für das Zerlegen in Teilprobleme und für das Zusammensetzen der Gesamtlösung benötigt wird.

Wie findet man eine geschlossene Form für solchen Rekursionsgleichungen? Es gibt ein Satz, der solchen Rekursionsgleichungen sehr einfach auslösen lässt. Da der Satz alle solchen Gleichungen mit einem Schuß “meistert”, nennt man ihm das “Master Theorem”.

Satz 3.94. (Master Theorem) Gegeben sei eine Rekursionsgleichung der Form $T(0) = 0$ und

$$T(n) = a \cdot T(n/b) + n^k$$

wobei $a \geq 1$, $b > 1$. Dann kann $T(n)$ asymptotisch wie folgt abgeschätzt werden:

$$T(n) = \begin{cases} \Theta(n^k) & \text{falls } a < b^k \\ \Theta(n^k \log n) & \text{falls } a = b^k \\ \Theta(n^{\log_b a}) & \text{falls } a > b^k \end{cases}$$

Beweis. Sei $f(n) := \Theta(n^k)$. Wir entwickeln die Rekurrenz und bekommen

$$T(n) = f(n) + af(n/b) + a^2 f(n/b^2) + \cdots + a^i f(n/b^i) + \cdots + a^L f(n/b^L)$$

wobei $n/b^L = 1$, d.h.

$$L = \log_b n.$$

Da $f(n) = \Theta(n^k)$, erhalten wir

$$T(n) = \sum_{i=0}^L a^i f(n/b^i) = \sum_{i=0}^L a^i \cdot (n/b^i)^k = n^k \cdot \underbrace{\sum_{i=0}^L x^i}_{S(x)} \quad (3.31)$$

mit $x := a/b^k$. Die Summe $S(x) = \sum_{i=0}^L x^i$ ist eine geometrische Reihe und wir wissen bereits, dass

$$S(x) = \frac{x^{L+1} - 1}{x - 1} = \frac{1 - x^{L+1}}{1 - x}$$

für $x \neq 1$ gilt. Da $x = a/b^k$ eine Konstante ist, gilt

$$S(x) = \begin{cases} \Theta(1) & \text{falls } x < 1 \\ L & \text{falls } x = 1 \\ \Theta(x^L) & \text{falls } x > 1 \end{cases} \quad (3.32)$$

Da $L = \log_b n$ und

$$x^L = \left(\frac{a}{b^k}\right)^{\log_b n} = \frac{a^{\log_b n}}{b^{k \cdot \log_b n}} = \frac{n^{\log_b a}}{n^k}$$

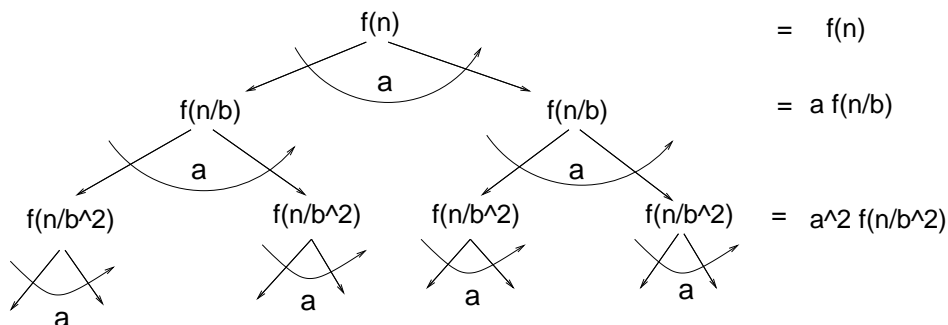
liefern uns (3.31) und (3.32) folgendes:

- Ist $a < b^k$, so ist $x < 1$ und damit auch $T(n) = n^k \cdot S(x) = \Theta(n^k)$.

- Ist $a = b^k$, so ist $x = 1$ und damit auch $T(n) = n^k \cdot S(x) = n^k \cdot L = n^k \cdot \log_b n$.
- Ist $a > b^k$, so ist $x > 1$ und damit auch $T(n) = n^k \cdot S(x) = \Theta(n^k \cdot x^L) = \Theta(n^{\log_b a})$.

□

Bemerkung 3.95. Intuitiv ist Satz 3.94 ziemlich einfach. Betrachte den Rekursionsbaum für $T(n) = aT(n/b) + f(n)$:



Der Baum hat die Tiefe $L = \log_b n$. Die i -te Ebene summiert sich zu $a^i \cdot f(n/b^i)$. Der Wert $T(n)$ selbst ist die Summe über alle Ebenen. Ist $f(n)$ groß, so kann man den Rest ignorieren: $T(n) = \Theta(f(n))$. Ist $f(n)$ klein, so trägt jeder innere Knoten nur sehr wenig bei und die ganze Summe $T(n)$ ist im wesentlichen auf den Blätter konzentriert; da jedes Blatt die selbe Konstante beiträgt und wir insgesamt $n^{\log_b a}$ Blätter haben, ist in diesem Fall $T(n) = \Theta(n^{\log_b a})$.

► *Beispiel 3.96* : Die drei Rekursionsgleichungen

$$\begin{aligned} T(n) &= 8 \cdot T(n/3) + n^2 \\ T(n) &= 9 \cdot T(n/3) + n^2 \\ T(n) &= 10 \cdot T(n/3) + n^2 \end{aligned}$$

haben der Reihe nach die Lösungen:

$$\begin{aligned} T(n) &= \Theta(n^2) \\ T(n) &= \Theta(n^2 \log n) \\ T(n) &= \Theta(n^{\log_3 10}) = \Theta(n^{2.09}) \end{aligned}$$

3.11 Aufgaben

3.1. Richtig oder falsch: Die Folge $\langle a_n \rangle$ konvergiert gegen a , wenn

- (a) ... sie a immer näher kommt.
- (b) ... sie a beliebig nahe kommt.
- (c) .. sie a beliebig nahe kommt, es aber nie erreicht.

3.2. Benutze die Formel $\lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n = e^x$ zur Untersuchung der Folge $a_n = (1 - \frac{1}{n+1})^n$ auf Konvergenz. Bestimme gegebenenfalls den Grenzwert $\lim_{n \rightarrow \infty} a_n$.

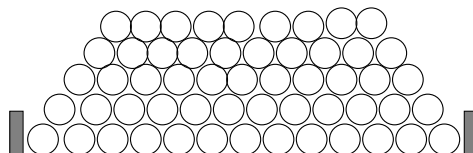
3.3. Es sei $c_n(x) = \frac{x^n n!}{n^n}$, wobei $x > 0$. Für welche $x > 0$ ist $\lim_{n \rightarrow \infty} \frac{c_{n+1}(x)}{c_n(x)} < 1$? Untersuche die Reihe $\sum_{n=1}^{\infty} c_n(x)$ auf Konvergenz in den Fällen $x = 2$ und $x = 4$.

3.4. Ein Turm wird aus Würfeln gebaut. Der erste Würfel hat eine Kantenlänge von $l = 1\text{m}$, der zweite $l = 0,5\text{m}$. Jeder weitere hat die halbe Kantenlänge der darunter liegenden Würfels. Welche Höhe nimmt der Turm an, wenn unendlich viele Würfel aufeinander gesetzt werden?

3.5. Sei $\langle a_n \rangle$ eine arithmetische Folge mit $a_n \neq 0$ für alle n . Zeige, dass für alle $n \geq 2$ gilt:

$$\frac{1}{a_1 \cdot a_2} + \frac{1}{a_2 \cdot a_3} + \cdots + \frac{1}{a_{n-1} \cdot a_n} = \frac{n-1}{a_1 \cdot a_n}.$$

3.6. Ein Ästhet will seinen Weinkeller verschönern. Dazu will er die a vorhandene Weinflaschen wie folgt auslegen:



Dabei will er, dass: (i) mindestens zwei Reihen entstehen und (ii) die oberste Reihe vollständig gefüllt ist.

Gebe zuerst eine mathematische Formulierung des Problems an.

Angenommen, der Ästhet hat $a = pm$ Weinflaschen, wobei $p \leq 2m + 1$ und p ungerade ist.

Zeige, dass dann das Problem lösbar ist. *Hinweis:*

$$\dots + (m-2) + (m-1) + m + (m+1) + (m+2) + \dots$$

3.7. Ein Frosch springt über die Straße. Beim ersten Sprung springt er 1 m. Dabei ermüdet er, so dass er bei jedem folgenden Sprung nur noch $2/3$ des vorigen Sprungs erreicht.

Welche Weglänge wird der Frosch nach n Sprüngen zurücklegen?

Die Straße ist 3 m breit und nach 6 Sprüngen wird an dieser Stelle ein Auto vorbei kommen. Wird dann der Frosch überleben, d.h. nach diesen 6 Sprüngen die Straße überquert haben?

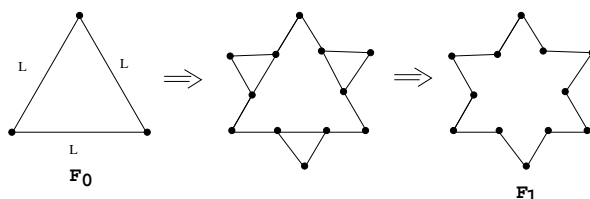
3.8. Wir betrachten ein Tenisturnier mit N Spielern. Die Spieler spielen paarweise und nach jedem Spiel fliegt der Verlorene raus aus dem Turnier. Wieviele Spiele insgesamt müssen gespielt werden, bis nur ein Gewinner bleibt? *Hinweis:* Ist N eine Zweierpotenz, so kann man die Anzahl S der Spiele leicht bestimmen; Die Anzahl der Spiele halbiert sich nach jeder Runde. Was aber wenn N keine Zweierpotenz ist? Probiere eine Bijektion zwischen Spielen und bestimmten Spielern zu finden.

3.9. Sie haben K Euro geerbt und legen diesen Betrag auf einem Konto an. Der Zinssatz beträgt $p\%$. Sie wollen n Jahre lang einen festen Betrag von x Euro jeweils am Ende jedes Jahres aus dem Konto herausnehmen, so dass nach n Jahren das Konto leer wird.

Wie groß ist der Betrag x ?

Hinweis: Sei $F_i(x)$ der Kontostand am Ende des i -ten Jahres. Probiere zuerst $F_i(x)$ für die ersten i 's zu bestimmen, danach eine allgemeine Vermutung für allgemeines i zu finden und diese Vermutung mittels Induktion zu beweisen. Die geometrische Reihe wird bestimmt in Spiel kommen.

3.10. Wir erzeugen rekursiv bestimmte geometrische Figuren F_0, F_1, F_2, \dots wie folgt. Wir beginnen mit einem gleichseitigen Dreieck F_0 mit Seitenlänge L . Dann teilen wir jede Kante in drei Teile auf, und erweitern jedes mittlere Teil zu einem gleichseitigen Dreieck



und lassen dann im Inneren verlaufenden Kanten weg (siehe die Skizze). So erhalten wir die Figur F_1 . Dann teilen wir wieder jede Kante von F_1 in drei Teile auf, erweitern jedes mittlere Teil zu einem gleichseitigen Dreieck und lassen dann im Inneren verlaufenden Kanten weg. So erhalten wir die Figur F_2 , usw. Die n -te Figur besteht also aus $3 \cdot 4^n$ Kanten.

Sei a_n der Umfang (d.h. die Gesamtlänge der Kanten in) der n -ten Figur und sei b_n die Fläche der n -ten Figur, $n = 0, 1, 2, \dots$

Untersuche die Folgen $\langle a_n \rangle$ und $\langle b_n \rangle$ auf Konvergenz.

Hinweis: Der Umfang der n -ten Figur ist gleich die Anzahl der Kanten mal die Länge einer Kante. Zum Beispiel $a_0 = 3 \cdot L$, $a_1 = (3 \cdot 4) \cdot (L/3)$, $a_2 = (3 \cdot 4 \cdot 4) \cdot (L/3^2)$, usw. Hat ein gleichseitiges Dreieck die Seitenlänge ℓ , so hat es nach dem Satz von Pythagoras²³ die Fläche $\frac{\sqrt{3}}{2} \cdot \ell^2$.

3.11. Die Folge $\langle a_k \rangle$ sei monoton wachsend und möge den Grenzwert a haben. Zeige, dass dann auch die Folge $\langle b_n \rangle$ mit

$$b_n = \frac{a_0 + a_1 + \dots + a_n}{n + 1}$$

gegen a konvergiert.

3.12. Sei $b \in \mathbb{Z}$ und $x \in \mathbb{R}$ eine reelle Zahl mit $|x| < 1$. Zeige, dass dann $\lim_{n \rightarrow \infty} n^b x^n = 0$ gilt.

3.13. Seien $a, b \in \mathbb{R}$, $a > 0$ und $|b| > 1$. Zeige, dass

(a) $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$

(b) $\lim_{n \rightarrow \infty} \sqrt[n]{n^a} = 1$

3.14. (Multiplikation statt Division) Sei $a > 0$ eine reelle Zahl. Wir wollen $1/a$ berechnen, ohne dabei irgendwelchen Zahlen zu dividieren. Wir suchen also eine Lösung x für die Gleichung $ax = 1$. Diese Gleichung läßt sich äquivalent als $x = 2x - ax^2$ umschreiben; das gesuchte x ist dann die von Null verschiedene Lösung dieser Gleichung. Setzen wir $f(x) := 2x - ax^2 = x(2 - ax)$, so erhalten wir die Rekursionsgleichung: $x_{n+1} = x_n(2 - ax_n)$.

Zeige, dass die Folge (x_n) mit $0 < x_0 < 2/a$ und $x_{n+1} = x_n(2 - ax_n)$ gegen $1/a$ strebt.

Hinweis: Monotonie-Kriterium.

3.15. Zeige, dass $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e$ gilt. *Hinweis:* Zuerst zeige, dass die Folge $c_n = \left(1 + \frac{1}{n}\right)^n$ monoton wachsend und durch e nach oben beschränkt ist. Dann zeige, dass $\lim_{n \rightarrow \infty} c_n \geq e_N$ für jedes festes $N \geq 1$ gilt. Um $(1 + 1/n)^n$ als eine Summe darzustellen, benutze den binomischen Lehrsatz.

3.16. Zeige die absolute Konvergenz folgenden Reihen: $\sum_{k=1}^{\infty} \frac{x^k}{k!}$ und $\sum_{k=1}^{\infty} \frac{\sin(kx)}{k^2}$.

3.17. Seien $x, q \in \mathbb{R}$, $x \geq 0$ und $0 \leq q < 1$. Zeige, dass die binomische ²⁴ Reihe $\sum_{k=0}^{\infty} \binom{x}{k} q^k$ ist absolut

²³Im rechtwinkligen Dreieck ist die Summe der Kathetenquadrate gleich dem Hypotenusenquadrat.

²⁴Wir haben die Binomialkoeffizienten $\binom{n}{k}$ im Abschnitt 1.6.2 nur für natürliche Zahlen n und k definiert. Man kann aber $\binom{x}{k}$ auch für $x \in \mathbb{R}$ und $k \in \mathbb{N}$ als Produkt definieren:

$$\binom{x}{k} := \prod_{i=1}^k \frac{x - i + 1}{i}$$

konvergent.

3.18. Bestimme den Grenzwert $\lim_{x \rightarrow \infty} x^{1/x}$.

3.19. Bestimme den Grenzwert $\lim_{x \rightarrow 0} \frac{e^x - 1 - x}{x^2}$.

3.20. Wie sehen die Grenzwerte $\lim_{x \rightarrow 0} \frac{x}{e^x}$ und $\lim_{x \rightarrow 0} \frac{1}{e^x}$ aus?

3.21. Bestimme den Grenzwert

$$\lim_{x \rightarrow 1} \frac{\ln x}{x - 1}$$

Hinweis: de l'Hospital

3.22. Seien $a, b \in \mathbb{R}$ nicht negativ. Bestimme die folgenden Grenzwerte:

(a) $\lim_{x \rightarrow \infty} \left(1 + \frac{a}{x}\right)^{bx}$

(b) $\lim_{x \rightarrow 1} \left(\frac{1}{\ln x} - \frac{1}{x-1}\right)$

Hinweis: Exponentialausdrücke der Form $f(x)^{g(x)}$ werden zunächst logarithmiert. Ausdrücke der Form $f(x) - g(x)$, die zum Grenzfäll der Art $\infty - \infty$ führen können, kann man mit der Transformation

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

zu dem Fall $\frac{0}{0}$ überführen.

3.23. Seien $f, g : \mathbb{N} \rightarrow \mathbb{R}$ zwei Funktionen mit $f = O(g)$. Seien

$$F(n) := \sum_{i=1}^n f(i) \quad \text{und} \quad G(n) := \sum_{i=1}^n g(i).$$

Zeige oder wiederlege: $F = O(G)$.

3.24. Seien $f_1(n), \dots, f_N(n)$ und $g_1(n), \dots, g_N(n)$ Funktionen mit $f_k(n) = O(g_k(n))$ für alle $k = 1, 2, \dots, N$. Zeige oder wiederlege: Dann gilt $\sum_{k=1}^N f_k(n) = O\left(\sum_{k=1}^N g_k(n)\right)$.

3.25. Sei $g : \mathbb{R} \rightarrow \mathbb{R}$ mit $\lim_{x \rightarrow \infty} g(x) = \infty$. Zeige oder wiederlege:

$$f(x) = o(g(x)) \Rightarrow e^{f(x)} = o\left(e^{g(x)}\right)$$

3.26. Für eine Funktion $f(n)$ definieren wir drei Funktionen

$$g_1(n) := f(2n), \quad g_2(n) := f(n/2), \quad g_3(n) := f(n+2).$$

In welcher asymptotischen Beziehung stehen die in der Tabelle angegebenen Funktionen $f(n)$ mit den Funktionen g_1, g_2, g_3 ?

Man sollte die *stärkstmögliche* Beziehung aus den fünf möglichen

$$f = o(g), \quad f = O(g), \quad f = \Theta(g), \quad f = \Omega(g), \quad f = \omega(g)$$

angeben. Zum Beispiel $\sqrt{n} = O(n)$ ist zwar richtig aber $\sqrt{n} = o(n)$ ist stärker. Oder $\binom{n}{2} = O(n^2)$ ist zwar richtig aber $\binom{n}{2} = \Theta(n^2)$ ist stärker.

$f(n)$	$\log_2 n$	\sqrt{n}	n	$n \log_2 n$	n^{10}	$n^{\log_2 n}$	$2^{n/100}$
$g_1(n)$					Θ		
$g_2(n)$							
$g_3(n)$							

Begründung des Beispielintrags: Da $f(n) = n^{10}$ folgt

$$g_1(n) = (2n)^{10} = 1024 \cdot n^{10} = \Theta(n^{10}) = \Theta(f(n)).$$

3.27. Gebe die best mögliche asymptotische Beziehungen zwischen Funktionen an:

- (a) $f(x) = e^{(\ln \ln x)^2}$ und $g(x) = \sqrt{x}$
 (b) $f(x) = x \log_4 x$ und $g(x) = \sqrt{x}(\log_2 x)^3$
 (c) $f(x) = (\log_4 x)^{1/2}$ und $g(x) = (\log_2 x)^{1/3}$

3.28. Zeige, dass für beliebige zwei Zahlen $a, b > 1$ und für beliebige Funktion $f : \mathbb{N} \rightarrow \mathbb{N}$ die Beziehung $\log_a f(n) = \Theta(\log_b f(n))$ gilt. Fazit: Für den Wachstum von Logarithmen ist die Basis unwesentlich!

3.29. Zeige, dass es zwei nicht fallende Funktionen $f, g : \mathbb{R} \rightarrow \mathbb{R}$ gibt (d.h. $x \leq y \Rightarrow f(x) \leq f(y)$ und $g(x) \leq g(y)$) gibt, so dass

$$f = O(g) \text{ aber weder } f = o(g) \text{ noch } f = \Theta(g) \text{ gilt.}$$

Hinweis: Für $f \neq o(g)$ reicht es, dass $f(x) = g(x)$ für *unendlich* viele x gilt. Für $f \neq \Theta(g)$ reicht es, dass die Differenz $|g(x) - f(x)|$ nicht von oben beschränkt ist.

3.30. Lokalisiere so genau wie möglich den Fehler im folgenden „Induktionsbeweis“.

Behauptung:

$$\sum_{i=1}^n (2i + 1) = O(n)$$

Beweis durch Induktion nach n . Induktionsbasis $n = 1$: In diesem Fall ist

$$\sum_{i=1}^1 (2i + 1) = 3 = O(1).$$

Induktionsschritt $n \mapsto n + 1$: Wir beginnen mit der Induktionsannahme

$$\sum_{i=1}^n (2i + 1) = O(n).$$

Wir addieren auf beiden Seiten $2(n + 1) + 1$ und erhalten

$$\sum_{i=1}^n (2i + 1) + 2(n + 1) + 1 = O(n) + 2(n + 1) + 1.$$

Nach einer Vereinfachung folgt

$$\sum_{i=1}^{n+1} (2i + 1) = O(n) + 2n + 3.$$

Aber $(2n + 3)$ ist sicher in $O(n)$ und somit

$$\sum_{i=1}^n (2i + 1) = O(n) + O(n) = O(n).$$

Kapitel 4

Diskrete Stochastik

Contents

4.1 Intuition und Grundbegriffe	158
4.2 Drei Modellierungsschritte	161
4.2.1 Das Geburtstagsproblem	162
4.3 Stochastische Unabhängigkeit	163
4.4 Bedingte Wahrscheinlichkeit	165
4.4.1 Multiplikationssatz für Wahrscheinlichkeiten	168
4.4.2 Satz von der totalen Wahrscheinlichkeit	169
4.4.3 Satz von Bayes	171
4.5 Stochastische Entscheidungsprozesse	174
4.5.1 Das „Monty Hall Problem“	177
4.5.2 Stichproben	178
4.5.3 Das “Sekretärinnen-Problem” an der Börse	179
4.6 Zufallsvariablen	184
4.7 Erwartungswert und Varianz	186
4.8 Analytische Berechnung von $E[X]$ und $\text{Var}[X]$	190
4.9 Eigenschaften von $E[X]$ und $\text{Var}[X]$	191
4.10 Verteilungen diskreter Zufallsvariablen	199
4.11 Abweichung vom Erwartungswert	205
4.11.1 Markov-Ungleichung	206
4.11.2 Tschebyschev-Ungleichung	208
4.11.3 Chernoff-Ungleichungen	212
4.12 Das Urnenmodell – Hashing*	218
4.13 Bedingter Erwartungswert*	224
4.14 Summen von zufälliger Länge – Wald’s Theorem	228
4.15 Irrfahrten und Markov-Ketten	232
4.16 Statistisches Schätzen: Die Maximum-Likelihood-Methode *	240
4.17 Die probabilistische Methode*	243
4.18 Aufgaben	247

4.1 Intuition und Grundbegriffe

Die Stochastik bedient sich gerne Beispielen aus der Welt des Glücksspiels, sie ist deswegen aber noch lange keine “Würfelbudenmathematik”. Ihr geht es darum, die Vorstellung einer Zufallsentscheidung so allgemein zu fassen, dass sie auch in ganz anderen Bereichen – von der Genetik bis zur Börse – zum Tragen kommen kann.

Der Begriff *Zufallsexperiment* steht für jeden realen Vorgang, der vom Zufall beeinflusst wird. Typischerweise liefert ein Zufallsexperiment ein *Ergebnis*, das “zufällig” (zumindest teilweise) ist.

Beispiele für Zufallsexperimente:

- Glücksspiele (z.B. Münzwurf, Würfeln, Lotto)
- 0-1-Experimente (Bernoulli-Experimente), wobei z.B. “1” für *Erfolg* und “0” für *Misserfolg* steht (z.B. Therapie, Platzierung, Schießen). Das betrachtete Zufallsexperiment kann die einmalige Durchführung eines 0-1-Experiments sein oder auch eine mehrmalige unabhängige Durchführung eines 0-1-Experiments.
- Zufällige Anzahlen (z.B. Anzahl von Kunden oder Jobs, Anzahl verkaufter Zeitungen, Anzahl von Verkehrsunfällen).
- Lebensdauer (z.B. von technischen Bauteilen, von Lebewesen).

Die **mathematische Modellierung** eines Zufallsexperiments erfolgt durch

- die Festlegung einer Menge Ω , die alle möglichen Ergebnisse des Zufallsexperiments enthält (das ist i.d.R. eine leichte Aufgabe); die Elemente von Ω heißen *Elementarereignisse*
- die Festlegung einer passenden *Wahrscheinlichkeitsverteilung* auf Ω (das ist i.d.R. die schwierige Aufgabe); das bedeutet vereinfacht gesagt: Für jedes mögliche Elementarereignis $\omega \in \Omega$ ist die Wahrscheinlichkeit $\Pr\{\omega\}$ seines Eintretens festzulegen.

Definition: Ein *diskreter Wahrscheinlichkeitsraum* besteht aus einer endlichen oder abzählbaren Menge Ω von *Elementarereignissen* und einer Funktion (einer Wahrscheinlichkeitsverteilung) $\Pr : \Omega \rightarrow [0, 1]$ mit der Eigenschaft, dass

$$\sum_{\omega \in \Omega} \Pr\{\omega\} = 1$$

gilt. Eine Teilmenge $A \subseteq \Omega$ heißt *Ereignis*. Seine Wahrscheinlichkeit ist durch

$$\Pr\{A\} = \sum_{\omega \in A} \Pr\{\omega\}$$

definiert.

Die Menge¹ Ω ist die Menge aller möglichen Ergebnissen eines Zufallsexperiments und $\Pr\{\omega\}$ ist die Wahrscheinlichkeit, dass der Zufall das Ergebnis ω liefern wird.

Die Funktion \Pr selbst heißt *Wahrscheinlichkeitsmaß* oder *Wahrscheinlichkeitsverteilung*. Zum Beispiel, *Gleichverteilung* (auch als *Laplace-Verteilung* bekannt) ist ein Wahrscheinlichkeitsmaß $\Pr : \Omega \rightarrow [0, 1]$ mit $\Pr\{\omega\} = \frac{1}{|\Omega|}$ für alle $\omega \in \Omega$. Damit ist

$$\Pr\{A\} = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der günstigen Elementarereignisse}}{\text{Anzahl aller Elementarereignisse}}$$

¹Auf englisch heißt Ω *sample space*.

Diese Verteilung entspricht unserer gängigen Vorstellung: Ein Ereignis wird umso wahrscheinlicher, je mehr Elementarereignisse an ihm beteiligt sind. Bei einer Gleichverteilung wird kein Element von Ω bevorzugt, man spricht daher auch von einer *rein zufälligen Wahl* eines Elements aus Ω .

► *Beispiel 4.1* : Zufallsexperiment: Einmaliges Werfen eines Spielwürfels. Mit welcher Wahrscheinlichkeit kommt eine gerade Zahl?

1. *Wahrscheinlichkeitsraum* Ω ist die Menge aller möglichen Ergebnissen des Experiments, d.h. Augenzahlen $1, 2, \dots, 6$, je mit Wahrscheinlichkeit $1/6$:

$$\left\{ \begin{array}{c} \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \end{array} & \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \end{array} \\ \hline 1/6 & \dots & 1/6 \end{array} \right\}$$

2. Ereignisse sind Teilmengen von $\{1, 2, 3, 4, 5, 6\}$. Z.B. Ereignis “Würfeln einer geraden Zahl” ist die Teilmenge $E = \{2, 4, 6\}$, und seine Wahrscheinlichkeit ist $\Pr\{E\} = 3 \cdot (1/6) = 1/2$.

In dieser Vorlesung werden wir nur Wahrscheinlichkeitsräume Ω betrachten, die entweder *endlich* oder *abzählbar* sind – deshalb das Wort “diskrete” vor der “Stochastik”. Da die Informatik sich hauptsächlich mit diskreten Strukturen beschäftigt, reicht uns diese (einfachere) Teil der Stochastik völlig aus.

Ist Ω überzählbar, so kann man nicht ohne weiteres die Wahrscheinlichkeiten $\Pr\{A\}$ für die Teilmengen $A \subseteq \Omega$ (die Ereignisse) einfach als die Summe von $\Pr\{\omega\}$ über die Elementarereignisse $\omega \in A$ definieren. Dazu braucht man den Begriff der sogenannten σ -Algebra, den wir hier nicht betrachten werden. Wir beschränken uns auf einem Beispiel.

► *Beispiel 4.2* : Romeo und Juliet haben eine Verabredung am bestimmten Zeitpunkt (sei es Zeitpunkt 0) und jeder kann mit einer Verzögerung von 0 bis auf 1 Stunde kommen. Die Verzögerungszeiten sind unabhängig und gleichwahrscheinlich. Derjenige, der als erste kommt, wird nur 15 Minuten warten, und dann wird weg gehen. Was ist die Wahrscheinlichkeit dafür, dass Romeo und Juliet sich treffen?

Wir können unseren Wahrscheinlichkeitsraum als das Quadrat $\Omega = [0, 1] \times [0, 1]$ darstellen, dessen Elemente (x, y) (Elementarereignisse) alle mögliche Ankunftszeiten von Romeo (x) und Julia (y) sind. Es gibt überzählbar viele solche Elementarereignisse und wir können nicht zu jedem seine Wahrscheinlichkeit zuweisen. Warum? Dann sollten wir für fast alle (x, y) (für alle außer abzählbar vielen Paaren) $\Pr\{x, y\} = 0$ setzen. In einer solchen Situation geht man anders rum. Zuerst schaut man, welches Ereignis $A \subseteq \Omega$ für uns interessant ist. In unserem Fall ist das die Menge

$$A = \{(x, y) : |x - y| \leq 1/4, 0 \leq x, y \leq 1\}$$

d.h. der schattierte Bereich im Abbildung 4.1. Man definiert dann die Wahrscheinlichkeit von A als

$$\Pr\{A\} = \frac{\text{Fläche von } A}{\text{Gesamtfläche}}$$

In unserem Beispiel ist die Fläche von A genau 1 minus die Fläche $(3/4) \cdot (3/4) = 9/16$ von zwei ungeschattierten Dreiecken. Da die Gesamtfläche gleich 1 ist, gilt somit $\Pr\{A\} = 1 - 9/16 = 7/16$.

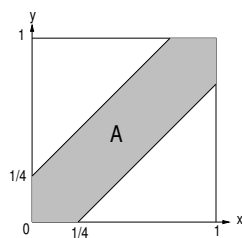


Abbildung 4.1: Das Ereignis, dass Romeo und Juliet sich treffen.

Für (diskrete) Wahrscheinlichkeitsmaße gelten die folgenden Rechenregeln. Für ein Ereignis $A \subseteq \Omega$ ist $\bar{A} = \Omega \setminus A$ das komplementäre Ereignis zu A .

Satz 4.3. Sei (Ω, \Pr) ein endlicher Wahrscheinlichkeitsraum und A, B Ereignisse. Es gilt:

- (a) $\Pr\{\Omega\} = 1$, $\Pr\{\emptyset\} = 0$ und $\Pr\{A\} \geq 0$ für alle $A \subseteq \Omega$.
- (b) $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}$.
- (c) $A \cap B = \emptyset \implies \Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$ (disjunkte Ereignisse).
- (d) $\Pr\{\bar{A}\} = 1 - \Pr\{A\}$ (komplementäre Ereignisse).
- (e) $\Pr\{A \cap B\} \geq \Pr\{A\} - \Pr\{\bar{B}\}$.
- (f) $\Pr\{A \setminus B\} = \Pr\{A\} - \Pr\{A \cap B\}$.
- (g) Ist $A \subseteq B$, so gilt $\Pr\{A\} \leq \Pr\{B\}$ (Monotonie).

Beweis. (a) gilt nach der Definition von \Pr . Zu (b):

$$\begin{aligned}
 \Pr\{A \cup B\} &= \sum_{\omega \in A \cup B} \Pr\{\omega\} \\
 &= \sum_{\omega \in A} \Pr\{\omega\} + \sum_{\omega \in B} \Pr\{\omega\} - \sum_{\omega \in A \cap B} \Pr\{\omega\} \\
 &= \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}
 \end{aligned} \tag{4.1}$$

da für $\omega \in A \cap B$, $\Pr\{\omega\}$ in (4.1) zweimal gezählt wird. (c) folgt aus (b). (d) folgt aus (c) und (a). (e) folgt aus (b), da $\Pr\{A \cup B\} \leq 1$ ist. (f) folgt aus (c). \square

► *Beispiel 4.4:* Wir wollen einen Schaltkreis mit n Verbindungen konstruieren. Aus früheren Erfahrungen wissen wir, dass jede Verbindung mit Wahrscheinlichkeit p falsch sein kann. D.h. für $1 \leq i \leq n$ ist

$$\Pr\{i\text{-te Verbindung ist falsch}\} = p.$$

Was kann man über die Wahrscheinlichkeit, dass das Schaltkreis *keine* falschen Verbindungen haben wird, sagen?

Sei A_i das Ereignis, dass die i -te Verbindung korrekt ist; also ist $\Pr\{\bar{A}_i\} = p$. Dann ist

$$\Pr\{\text{alle Verbindungen sind richtig}\} = \Pr\left\{\bigcap_{i=1}^n A_i\right\}$$

Obwohl es schwer ist, diese Wahrscheinlichkeit exakt auszurechnen, man kann vernünftige Abschätzungen finden. Einerseits, ist laut der Monotonie-Eigenschaft (g)

$$\Pr \left\{ \bigcap_{i=1}^n A_i \right\} = \Pr \left\{ A_1 \cap \left(\bigcap_{i=2}^n A_i \right) \right\} \leq \Pr \{A_1\} = 1 - p.$$

Andererseits, ist laut der Eigenschaften (d) und (b)

$$\Pr \left\{ \bigcap_{i=1}^n A_i \right\} = 1 - \Pr \left\{ \overline{\bigcap_{i=1}^n A_i} \right\} = 1 - \Pr \left\{ \bigcup_{i=1}^n \overline{A_i} \right\} \geq 1 - \sum_{i=1}^n \Pr \{ \overline{A_i} \} = 1 - np.$$

Ist zum Beispiel $n = 10$ und $p = 0,01$, so gilt

$$0,9 = 1 - 10 \cdot 0,01 \leq \Pr \{ \text{alle Verbindungen sind richtig} \} \leq 1 - 0,01 = 0,99.$$

4.2 Drei Modellierungsschritte

Keiner weiß so genau, was der Zufall eigentlich ist, aber eine *intuitive* Vorstellung darüber hat fast jeder! Und genau da steckt die Gefahr – genauso wie mit der Unendlichkeit, versagt oft unsere Intuition wenn man mit dem Zufall als einem “halb-definierten” Objekt jongliert. Deshalb werden wir uns in dieser Vorlesung nur auf die (oben gegebene) *mathematische* Definition der Wahrscheinlichkeit verlassen und unsere Intuition nur zur Interpretation der Resultate benutzen.

Will man ein Zufallsexperiment analysieren, so sind in vielen Fällen die folgende “Drei-Schritt-Methode” sehr hilfreich.

1. **Finde den Wahrscheinlichkeitsraum:** Bestimme alle möglichen Ergebnisse des Experiments und ihre Wahrscheinlichkeiten, d.h. bestimme die Menge Ω und die Wahrscheinlichkeiten $\Pr \{ \omega \}$ der Elementarereignisse $\omega \in \Omega$.
2. **Bestimme die Ereignisse E :** bestimme welche von den Ergebnissen $E \subseteq \Omega$ “interessant” sind.
3. **Bestimme die Wahrscheinlichkeit des Ereignis E :** kombiniere die Wahrscheinlichkeiten der Elementarereignisse in E um $\Pr \{E\}$ zu bestimmen, $\Pr \{E\} = \sum_{\omega \in E} \Pr \{ \omega \}$.

► *Beispiel 4.5* : Zufallsexperiment: Würfle zwei Spielwürfel.

1. **Wahrscheinlichkeitsraum Ω :** $6^2 = 36$ möglichen Ausgänge des Experiments, je mit Wahrscheinlichkeit $\frac{1}{36}$.
2. **Ereignisse:** Mögliche Ereignisse: $E_1 =$ “die Summe der Augenzahlen ist > 10 ” oder $E_2 =$ “die zweite Zahl ist größer als die erste”. D.h. $E_1 = \{(5, 6), (6, 5), (6, 6)\}$ und $E_2 = \{(i, j) : 1 \leq i < j \leq 6\}$.
3. **Wahrscheinlichkeiten:** $\Pr \{E_1\} = \frac{|E_1|}{36} = \frac{1}{12}$ und $\Pr \{E_2\} = \frac{|E_2|}{36} = \frac{15}{36} = \frac{5}{12}$.

► *Beispiel 4.6* : In einem Dorf lebt die *Hälfte* aller Menschen alleine, die andere Hälfte mit genau einem Partner.

Wenn ich zufällig jemanden auf dem Marktplatz anspreche, mit welcher Wahrscheinlichkeit lebt derjenige allein? Antwort: $1/2$. Warum? In diesem Fall besteht der W’raum Ω aus allen

0/1-Strings (a_1, \dots, a_n) mit $a_i = 1$ genau dann, wenn der i -ter Mensch alleine lebt. Dann ist $\Pr\{a_i = 1\} = \Pr\{a_i = 0\} = 1/2$.

Wenn ich nun zufällig an eine Wohnungstür klopfe und frage, mit welcher Wahrscheinlichkeit lebt dort jemand allein? Antwort: $2/3$. Warum? In diesem Fall besteht der Wahrscheinlichkeitsraum Ω aus allen 0/1-Strings (b_1, \dots, b_m) mit $b_i = 1$ genau dann, wenn das i -te Haus ein Familienhaus ist. Da genau die Hälfte der Menschen alleine leben, in genau $1/3$ der Häuser Familien leben. Also ist in diesem Fall $\Pr\{\text{im Haus lebt jemand allein}\} = 2/3$.



Immer den *richtigen* Wahrscheinlichkeitsraum wählen!

4.2.1 Das Geburtstagsproblem

Um einen schnellen Zugriff auf Daten zu haben, kann man sie in Listen aufteilen. Beim Abspeichern von Daten in Computern kommt diese Idee in der Technik des *Hashings* zur Anwendung. Nur bei kurzen Listen sind auch die Suchzeiten kurz, daher stellt sich die Frage, mit welcher Wahrscheinlichkeit es zu "Kollisionen" kommt, zu Listen, die mehr als einen Eintrag enthalten. Wir betrachten diese Wahrscheinlichkeit für n Listen und m Daten unter der Annahme, dass alle möglichen Belegungen der Listen mit den Daten gleich wahrscheinlich sind. Wir werden sehen, dass mit Kollisionen schon dann zu rechnen ist, wenn m von der Größenordnung \sqrt{n} ist.

Diese Fragestellung ist in der Stochastik unter dem Namen *Geburtstagsproblem* bekannt. Gefragt ist nach der Wahrscheinlichkeit, dass in einer Klasse mit m Schülern alle verschiedene Geburtstage haben.

1. Finde den Wahrscheinlichkeitsraum: Wir lassen uns von der Vorstellung leiten, dass das Tupel $\omega = (x_1, \dots, x_m)$ der m Geburtstage ein rein zufälliges Element aus

$$\Omega := \left\{ (x_1, \dots, x_m) : x_i \in \{1, \dots, n\} \right\}$$

ist, mit $n = 365$.

2. Bestimme das Ereignis: Uns interessiert das Ereignis E = "alle Geburtstage x_1, \dots, x_m sind verschieden":

$$E = \left\{ (x_1, \dots, x_m) \in \Omega : x_i \neq x_j \text{ für alle } i \neq j \right\}.$$

3. Bestimme die Wahrscheinlichkeit des Ereignis: Es gilt $|E| = n(n-1) \cdots (n-m+1)$. Nehmen wir also an, dass es sich um eine rein zufällige Wahl der Geburtstage aus Ω handelt, so ist die gesuchte

Wahrscheinlichkeit²

$$\begin{aligned}
 \Pr\{E\} &= \frac{|E|}{|\Omega|} = \frac{n(n-1)\cdots(n-m+1)}{n^m} = \prod_{i=1}^{m-1} \left(1 - \frac{i}{n}\right) \\
 &\stackrel{(*)}{\leq} \prod_{i=1}^{m-1} e^{-\frac{i}{n}} = \exp\left(-\sum_{i=1}^{m-1} \frac{i}{n}\right) \\
 &\stackrel{(**)}{=} \exp\left(-\frac{m(m-1)}{2n}\right). \tag{4.2}
 \end{aligned}$$

Für $m = 1 + \sqrt{2n}$ ist diese Wahrscheinlichkeit durch e^{-1} nach oben beschränkt und fällt dann für wachsendes m rapide gegen Null. Diese Abschätzung drückt das *Geburtstag-Phänomen* aus:

In einer Gruppe von $m = 1 + \sqrt{2 \cdot 365} \leq 28$ Leuten haben zwei denselben Geburtstag mit Wahrscheinlichkeit $\geq 1 - e^{-1}$.

Oder in der Perspektive von Hashing mit Verkettung: Anfänglich erhalten wir nur Einerlisten. Wenn aber m in den Bereich von $\Omega(\sqrt{n})$ kommt, dann entwickeln sich erste Zweierlisten.


4.3 Stochastische Unabhängigkeit

Definition: Zwei Ereignisse A und B sind (stochastisch) *unabhängig*, falls

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

gilt.

Das ist die *Definition* der Unabhängigkeit. Aussagen wie “zwei Ereignisse sind unabhängig, falls diese Ereignisse einander nicht beeinflussen” sind *keine* Definitionen!

 Erst richtig falsch ist zu sagen, dass je zwei disjunkte Ereignisse unabhängig sind. Unabhängigkeit von Ereignissen hat nichts mit ihrer Disjunktheit zu tun! Zum Beispiel, sind $\Pr\{A\} > 0$, $\Pr\{B\} > 0$ und $A \cap B = \emptyset$, dann sind A und B abhängig, da dann $\Pr\{A \cap B\} = \Pr\{\emptyset\} = 0$ und $\Pr\{A\} \cdot \Pr\{B\} > 0$ gilt.

Um Unabhängigkeit von Ereignissen zu zeigen, ist der folgender einfacher Fakt oft nützlich.

Behauptung 4.7. Sind A und B zwei unabhängige Ereignisse, so sind auch die Ereignisse A und \bar{B} wie auch \bar{A} und \bar{B} unabhängig.

Beweis.

$$\begin{aligned}
 \Pr\{A \cap \bar{B}\} &= \Pr\{A\} - \Pr\{A \cap B\} \\
 &= \Pr\{A\} - \Pr\{A\} \cdot \Pr\{B\} \\
 &= \Pr\{A\} (1 - \Pr\{B\}) = \Pr\{A\} \cdot \Pr\{\bar{B}\}.
 \end{aligned}$$

²Hier haben wir in (*) die Ungleichung $1 + x \leq e^x$ (gültig für alle $x \in \mathbb{R}$) und in (**) die Gleichung $\sum_{i=1}^{m-1} i = \frac{m(m-1)}{2}$ (arithmetische Reihe) ausgenutzt.

□

► *Beispiel 4.8*: Wir werfen zweimal eine faire Münze und betrachten die Ereignisse:

A = “erster Wurf ergibt Wappen”

B = “beide Ausgänge sind gleich”

A = “beide Ausgänge sind Wappen”

Obwohl die Ereignisse A und B sich gegenwärtig zu “beeinflußen” scheinen, sind sie in Wirklichkeit unabhängig:

$$\Pr\{A \cap B\} = \Pr\{WW\} = \frac{1}{4}$$

$$\Pr\{A\} \cdot \Pr\{B\} = \Pr\{WW, WK\} \cdot \Pr\{WW, KK\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Die Ereignisse A und C sind aber bereits abhängig!

$$\Pr\{A \cap C\} = \Pr\{WW\} = \frac{1}{4}$$

$$\Pr\{A\} \cdot \Pr\{B\} = \Pr\{WW, WK\} \cdot \Pr\{WW\} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}.$$

► *Beispiel 4.9*: Wir werfen dreimal eine faire Münze. Der Wahrscheinlichkeitsraum besteht aus 8 Elementarereignissen

$$\Omega = \{KKK, KKW, KWK, KWW, WKK, WKW, WWK, WWW\}$$

und jedes davon kann mit Wahrscheinlichkeit $1/8$ eintreten. Wir betrachten die Ereignisse:

A = “es gibt mehr W’s als K’s”

B = “die ersten zwei Ausgänge sind gleich”

Dann ist

$$A = \{KWW, WKW, WWK, WWW\} \Rightarrow \Pr\{A\} = 1/2$$

$$B = \{KKK, KKW, WWK, WWW\} \Rightarrow \Pr\{B\} = 1/2$$

$$A \cap B = \{WWW, WWK\} \Rightarrow \Pr\{A \cap B\} = 1/4$$

Also sind A und B unabhängig. Aber beide Ereignisse A und B beinhalten die Ausgänge der ersten zwei Würfe, und es ist schwer zu argumentieren, warum denn ein der Ereignisse keinen Einfluss auf den anderen hat (oder haben soll). Wenn wir zum Beispiel das Ereignis

C = “Wappen im dritten Schritt”

betrachten, dann haben wir

$$C = \{WWW, WKW, KWW, KKW\} \Rightarrow \Pr\{C\} = 1/2$$

$$A \cap C = \{WWW, WKW, KWW\} \Rightarrow \Pr\{A \cap C\} = 3/8$$

und somit sind A und C schon abhängig! Obwohl, wie es leicht zu sehen ist, B und C immer noch unabhängig sind.

Wir haben gesehen, dass es kann drei Ereignisse A, B, C geben, so dass A und B wie auch B und C unabhängig sind, aber die Ereignisse A und C sind *nicht* unabhängig. Waren nun *alle* drei Paare unabhängig, könnte dann man daraus schließen, dass $\Pr\{A \cap B \cap C\} = \Pr\{A\} \cdot \Pr\{B\} \cdot \Pr\{C\}$ gelten soll? Leider, so einfach ist die Sache nicht ...

▷ *Beispiel 4.10*: Wir werfen dreimal eine Münze und betrachten die Ereignisse

A = "die ersten zwei Ausgänge sind gleich"

B = "der erste und der dritte Ausgang sind gleich"

C = "die letzten zwei Ausgänge sind gleich"

Dann gilt $\Pr\{A\} = \Pr\{B\} = \Pr\{C\} = 1/2$, und alle Ereignisse $A \cap B$, $B \cap C$, $A \cap C$ und $A \cap B \cap C$ sind gleich dem Ereignis $\{WWW, KKK\}$, das mit Wahrscheinlichkeit $1/4$ eintritt. Damit sind alle drei Paare unabhängig, aber

$$\begin{aligned}\Pr\{A \cap B \cap C\} &= 1/4 \\ \Pr\{A\} \cdot \Pr\{B\} \cdot \Pr\{C\} &= 1/8\end{aligned}$$

Also sind die Ereignisse A, B, C *nicht* "total" unabhängig.

Die allgemeine Definition ist folgende. Die Ereignisse A_1, \dots, A_n sind (heißen) *total unabhängig*, falls für alle $1 \leq i_1 < i_2 < \dots < i_k \leq n$ mit $k \geq 1$ die Ereignisse

$$A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_{k-1}} \quad \text{und} \quad A_{i_k}$$

unabhängig sind.

Behauptung 4.11. Sind die Ereignisse A_1, \dots, A_n total unabhängig, so gilt:

$$\Pr\{A_1 \cap A_2 \cap \dots \cap A_n\} = \Pr\{A_1\} \cdot \Pr\{A_2\} \cdots \Pr\{A_n\}$$

Beweis. Induktion über n . □

4.4 Bedingte Wahrscheinlichkeit

Alice und Bob gehen zum Abendessen. Um zu entscheiden, wer jetzt bezahlen soll, werfen sie dreimal eine faire Münze. Falls es mehr mals Wappen (W) als Kopf (K) rauskommt, bezahlt Alice, sonst bezahlt Bob. Es ist klar, dass die Chancen gleich sind. Der Wahrscheinlichkeitsraum sieht folgendermaßen aus

$$\Omega = \{KKK, KKW, KWK, KWW, WKK, WKW, WWK, WWW\}$$

und die Ereignisse "bezahlt Alice" und "bezahlt Bob" sind entsprechend

$$A = \{KWW, WKW, WWK, WWW\}$$

$$B = \{KKK, KKW, KWK, WKK\}$$

Sie werfen die Münze einmal und das Resultat ist “Wappen”; bezeichne dieses Ereignis durch E . Wie sollte man jetzt (nachdem das Ereignis E bereits eingetreten ist) die Chancen berechnen? Es gilt

$$E = \{WWW, WWK, WKW, WKK\}$$

Da wir bereits wissen, dass E eingetreten ist, hat sich unser Wahrscheinlichkeitsraum von Ω auf E verkleinert, da die Ereignisse, die nicht in E liegen, nicht mehr möglich sind. In diesem neuen Experiment sehen die Ausgänge “bezahlt Alice” und “bezahlt Bob” folgendermaßen aus:

$$A \cap E = \{WKW, WWK, WWW\}$$

$$B \cap E = \{WKK\}$$

Die neuen Wahrscheinlichkeiten, wer nun bezahlen soll, sind jetzt $3/4$ für Alice und nur $1/4$ für Bob.

Die allgemeine Situation ist folgende: Ist ein Ereignis E bereits eingetreten, wie sieht dann die Wahrscheinlichkeit, dass ein anderes Ereignis A eintreten wird? Im Allgemeinen können wir nicht mehr einfach die Wahrscheinlichkeiten der Elementarereignisse $\omega \in A$ aufsummieren, denn (nachdem E eingetreten ist) werden sich auch die Wahrscheinlichkeiten der Elementarereignisse ändern. Die allgemeine Definition ist folgende:

Definition: Für zwei Ereignisse A und B mit $\Pr\{B\} \neq 0$ die *bedingte Wahrscheinlichkeit* $\Pr\{A | B\}$ für das Ereignis A unter der Bedingung B ist definiert durch

$$\Pr\{A | B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}. \quad (4.3)$$

Die Wahrscheinlichkeit $\Pr\{A | B\}$ bezeichnet man als *a-posteriori-Wahrscheinlichkeit* von A (bezüglich B).

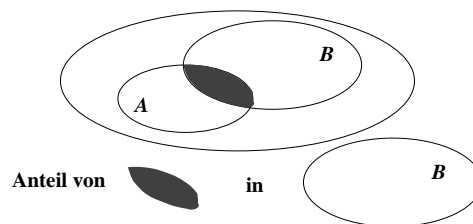


Abbildung 4.2: Bedingte Wahrscheinlichkeit bei der Gleichverteilung.

Für den Beispiel oben (mit Alice und Bob) gilt

$$\Pr\{A | E\} = \frac{\Pr\{A \cap E\}}{\Pr\{E\}} = \frac{3/8}{1/2} = \frac{3}{4},$$

$$\Pr\{B | E\} = \frac{\Pr\{B \cap E\}}{\Pr\{E\}} = \frac{1/8}{1/2} = \frac{1}{4}$$

Bemerkung 4.12. Mit Hilfe der bedingten Wahrscheinlichkeit kann man eine äquivalente Definition der stochastischer Unabhängigkeit zweier Ereignisse A und B angeben:

$$A \text{ und } B \text{ sind unabhängig} \iff \Pr\{A \cap B\} = \Pr\{A\} \Pr\{B\} \iff \Pr\{A | B\} = \Pr\{A\}$$

Bemerkung 4.13. Die bedingte Wahrscheinlichkeit $\Pr\{A|B\}$ kann als Wahrscheinlichkeit für das Eintreten des Ereignisses A interpretiert werden, wenn das Ereignis B bereits eingetreten ist.³ Ist B eine Gleichverteilung, dann ist die angegebene Definition von $\Pr\{A|B\}$ intuitiv klar: Ist das Ereignis B eingetreten, dann sind jene Elementarereignisse aus B für das Ereignis A günstig, die zu A gehören, und dies sind die Elementarereignisse aus $A \cap B$; damit gilt

$$\Pr\{A|B\} = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|}{|\Omega|} \cdot \frac{|\Omega|}{|B|} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}.$$

Insbesondere sind die Ereignisse A und B genau dann unabhängig, wenn der Anteil $|A|/|\Omega|$ des Ereignisses A im ganzen Wahrscheinlichkeitsraum Ω gleich dem Anteil $|A \cap B|/|B|$ des Teilereignisses $A \cap B$ von A in dem Ereignis B ist (siehe Abb. 4.2).

▷ *Beispiel 4.14:* Die Polizei stoppt Autos an der Miquell-Allee rein zufällig. Aus der Erfahrung folgendes ist bekannt: Mit $3/100$ Wahrscheinlichkeit wird das (angehaltene) Auto gelb sein und mit $1/50$ Wahrscheinlichkeit⁴ wird das Auto gelb *und* der Fahrer blond sein.

Nun kommen wir eines Tages die Allee entlang und sehen, dass die Polizei gerade ein gelbes Auto angehalten hat. Wie groß ist die Wahrscheinlichkeit, dass der Fahrer blond ist?

Wir haben also zwei Ereignisse $G =$ “das Auto ist gelb” und $B =$ “der Fahrer ist blond”, und wissen, dass $\Pr\{G\} = 3/100$ und $\Pr\{B \cap G\} = 1/50$. Also gilt

$$\Pr\{B|G\} = \frac{\Pr\{B \cap G\}}{\Pr\{G\}} = \frac{0,02}{0,03} = 0,667 > \frac{1}{2}.$$

▷ *Beispiel 4.15:* In einem großen Haus wohnen mehrere Familien, je mit zwei Kindern. Wir wissen auch, dass es immer die Tür von einem Jungen geöffnet wird, falls mindestens ein Junge in Familie ist (Jungs sind schneller).

Wir klingeln an einer Wohnungstür und ein Junge, der Peter, hat uns gerade die Tür geöffnet. Nun biete ich Ihnen eine Wette an. Wenn das *andere* Kind ebenfalls ein Junge ist, kriegen Sie 5,- €, wenn es ein Mädchen ist, kriege ich 5,- €.

Ist dies eine faire Wette? Natürlich nicht, denn meine Gewinnchancen stehen 2 : 1 dabei!

Nehmen wir der Einfachheit halber an, ein Neugeborenes sei mit Wahrscheinlichkeit $1/2$ ein Mädchen (M) bzw. ein Junge (J), unabhängig vom Geschlecht früher oder später geborener Geschwister. Dann gibt es unter Berücksichtigung der Reihenfolge der Geburten bei 2 Kindern die vier gleichwahrscheinliche Fälle: $\Omega = \{MM, MJ, JM, JJ\}$. Durch die Beobachtung von Peter (J) scheidet der Fall MM aus. Bei den verbleibenden Fällen $A = \{MJ, JM, JJ\}$ ist ein M doppelt so wahrscheinlich wie ein zweites J .

Warum wäre die Wette fair, wenn uns der Peter gesagt hätte, dass er das ältere Kind ist? Da dann hätten wir anstatt $A = \{MJ, JM, JJ\}$ das Ereignis $\{MJ, JJ\}$.

▷ *Beispiel 4.16:* Beim gleichzeitigen Werfen zweier Spielwürfel ist die Menge der Elementarereignisse $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, also $|\Omega| = 36$. Ist A das Ereignis, dass die gewürfelte Augensumme mindestens 10 beträgt, so sind die für A günstigen Elementarereignisse gegeben durch:

³Aber das ist nur eine *Interpretation*, die nicht 100% richtig ist: Es gibt keinen Grund, warum das Ereignis B vor dem Ereignis A eintreten *sollte*! Die Situation ist ähnlich wie mit der logischer Implikation $B \rightarrow A$. Diese Implikation sagt nicht, dass die Aussage B bereits als wahr bewiesen würde; sie sagt nur, dass A richtig wäre, *falls* B gelten würde.

⁴Gelbe Autos sind also selten, aber die blondhaarige wählen ein gelbes Auto noch seltener.

$(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)$. Damit erhält man: $\Pr\{A\} = \frac{6}{36} = \frac{1}{6}$. Sei nun B das Ereignis, dass die gewürfelte Augensumme gerade ist. Von der 36 möglichen Elementarereignissen sind die 18 Ereignisse für B günstig und man erhält: $\Pr\{B\} = \frac{18}{36} = \frac{1}{2}$. Für das Ereignis $A \cap B$ sind nur die Ereignisse $(4, 6), (5, 5), (6, 4), (6, 6)$ günstig, so dass $\Pr\{A \cap B\} = \frac{4}{36} = \frac{1}{9}$ gilt.

Weiß man nun, dass Ereignis B bereits eingetreten ist, so findet man für die a-posteriori-Wahrscheinlichkeit von A : $\Pr\{A | B\} = \frac{(1/9)}{(1/2)} = \frac{2}{9} < \frac{1}{3}$.

Weiß man, dass Ereignis A bereits eingetreten ist, so findet man für die a-posteriori-Wahrscheinlichkeit von B : $\Pr\{B | A\} = \frac{(1/9)}{(1/6)} = \frac{2}{3}$.

4.4.1 Multiplikationssatz für Wahrscheinlichkeiten

Ist $\Pr\{B\} \neq 0$ und kennt man die Wahrscheinlichkeit $\Pr\{A | B\}$, dann kann man auch $\Pr\{A \cap B\}$ berechnen:

$$\Pr\{A \cap B\} = \Pr\{B\} \cdot \Pr\{A | B\}. \quad (4.4)$$

Man nennt diese Reformulierung von (4.3) *Multiplikationssatz für Wahrscheinlichkeiten*.

Korollar 4.17. Seien A und B zwei Ereignisse mit $\Pr\{B\} \neq 0$. Dann sind A und B unabhängig genau dann, wenn $\Pr\{A | B\} = \Pr\{A\}$ gilt.

Der Multiplikationssatz für Wahrscheinlichkeiten gilt auch für mehrere Ereignisse:

Satz 4.18. (Multiplikationssatz für Wahrscheinlichkeiten) Ist $\Pr\{A_1 \cap A_2 \cap \dots \cap A_n\} > 0$, so gilt:

$$\Pr\left\{\bigcap_{i=1}^n A_i\right\} = \Pr\{A_1\} \cdot \Pr\{A_2 | A_1\} \cdot \Pr\{A_3 | A_1 \cap A_2\} \cdot \dots \cdot \Pr\{A_n | A_1 \cap \dots \cap A_{n-1}\}$$

Beweis. $\Pr\left\{\bigcap_{i=1}^n A_i\right\}$ ist gleich

$$\Pr\{A_1\} \cdot \frac{\Pr\{A_2 \cap A_1\}}{\Pr\{A_1\}} \cdot \frac{\Pr\{A_3 \cap A_2 \cap A_1\}}{\Pr\{A_2 \cap A_1\}} \cdot \dots \cdot \frac{\Pr\{A_n \cap \dots \cap A_1\}}{\Pr\{A_{n-1} \cap \dots \cap A_1\}}$$

□

▷ *Beispiel 4.19*: Es ist aufgrund einer Umfrage bei der Studierenden der Informatik bekannt, dass die Wahrscheinlichkeit, die Klausur zur Grundvorlesung “Theoretische Informatik 2” beim ersten Versuch zu bestehen, gleich $p_1 \neq 1$ ist. Die Wahrscheinlichkeit dafür, dass ein Studierender die Nachklausur besteht, wenn er beim erstenmal durchgefallen ist, beträgt p_2 . Mit welcher Wahrscheinlichkeit sind mehr als zwei Versuche notwendig, die Klausur zu bestehen?

Seien A_i die Ereignisse, dass der i -te Versuch erfolglos war, $i = 1, 2$. Unser Ziel ist $\Pr\{A_1 \cap A_2\}$ zu berechnen. Wir wissen, dass $\Pr\{A_1\} = 1 - p_1 \neq 0$ und $\Pr\{A_2 | A_1\} = 1 - \Pr\{\overline{A_2} | A_1\} = 1 - p_2$ gilt. Deshalb ist die gesuchte Wahrscheinlichkeit $\Pr\{A_1 \cap A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2 | A_1\} = (1 - p_1)(1 - p_2)$.

- *Beispiel 4.20*: Gegeben sind 32 Spielkarten, darunter 4 Buben. Zwei Spieler, jeder erhält 3 Karten. Ereignis $A = \text{“Jeder der beiden Spieler erhält genau 1 Buben”}$. Frage: $\Pr\{A\} = ?$ Lösung: $A = A_1 \cap A_2$, wobei $A_i = \text{“Spieler } i \text{ erhält genau 1 Buben”}$. Dann ist

$$\Pr\{A_1 \cap A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2 | A_1\} = \frac{\binom{4}{1} \binom{28}{2}}{\binom{32}{3}} \cdot \frac{\binom{3}{1} \binom{26}{2}}{\binom{29}{3}} \approx 0,081.$$

Erklärung: Wieviele Möglichkeiten gibt es, 3 Spielkarten aus 32 auszuwählen, so dass unter ihnen genau 1 Buben sein wird? Wir haben $\binom{4}{1}$ Möglichkeiten, 1 Buben aus 4 auszuwählen, und $\binom{28}{2}$ Möglichkeiten, noch 2 Spielkarten aus den restlichen 28 Spielkarten (ohne Buben) auszuwählen. Das ergibt nach dem Produktregel (siehe Satz 1.34(3)) genau $\binom{4}{1} \binom{28}{2}$ Möglichkeiten. Somit ist $\Pr\{A_1\} = \frac{\binom{4}{1} \binom{28}{2}}{\binom{32}{3}}$. Nachdem die drei Spielkarten entfernt sind, bleibt es nur noch 29 Spielkarten mit nur 3 Buben darunter. Deshalb ist auch $\Pr\{A_2 | A_1\} = \frac{\binom{3}{1} \binom{26}{2}}{\binom{29}{3}}$.

Es gibt ein Paar Regeln, die den Umgang mit bedingter Wahrscheinlichkeit erleichtern. Zuerst beachten wir, dass $\Pr_B(A) := \Pr\{A | B\}$ eine Wahrscheinlichkeitsverteilung \Pr_B auf Ω definiert, denn:

$$\begin{aligned} \sum_{\omega \in \Omega} \Pr_B(\omega) &= \sum_{\omega \in \Omega} \Pr\{\omega | \omega \in B\} \\ &= \sum_{\omega \in \Omega} \frac{\Pr\{\omega \in B\}}{\Pr\{B\}} = \frac{1}{\Pr\{B\}} \sum_{\omega \in \Omega} \Pr\{\omega \in B\} = \frac{1}{\Pr\{B\}} \cdot \Pr\{B\} = 1. \end{aligned}$$

Es gelten also alle im Satz 4.3 angegebene Eigenschaften auch für $\Pr_B(A)$. Man kann diese Regeln auch direkt aus den Regeln für $\Pr\{A\}$ ableiten. Zum Beispiel:

$$\begin{aligned} \Pr_B(\bar{A}) &= \Pr\{\bar{A} | B\} = \frac{\Pr\{\bar{A} \cap B\}}{\Pr\{B\}} = \frac{\Pr\{B \setminus A\}}{\Pr\{B\}} = \frac{\Pr\{B \setminus (A \cap B)\}}{\Pr\{B\}} \\ &= \frac{\Pr\{B\} - \Pr\{A \cap B\}}{\Pr\{B\}} = 1 - \Pr\{A | B\} \\ &= 1 - \Pr_B(A), \end{aligned}$$

andere Eigenschaften analog.

4.4.2 Satz von der totalen Wahrscheinlichkeit

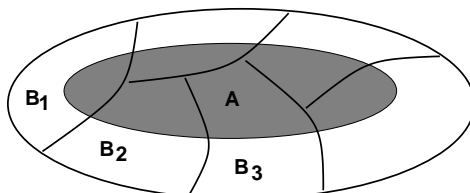
Manchmal wissen wir die Wahrscheinlichkeit von einem Ereignis A im voraus nicht, wissen aber wie die Wahrscheinlichkeit von A aussehen würde, falls irgendwelche andere Ereignisse bereits eingetreten würde.

- *Beispiel 4.21*: (**Eisverkäufer**) Ein Eisverkäufer muss sich entscheiden, ob er mehr Eis für den Feiertag bestellen sollte. Aus der Erfahrung weiss er, dass seine Chance alles zu verkaufen sehr stark vom den Wetter abhängt: Es gibt 90% Chance, falls der Tag sonnig ist, 60% Chance, falls es wolkig ist, und nur 20% Chance, falls es regnet. Nach der Wettervorhersage wird es am den Tag mit 30% Wahrscheinlichkeit sonnig, mit 45% Wahrscheinlichkeit wolkig und mit 25% Wahrscheinlichkeit regnen. Mit welcher Wahrscheinlichkeit wird der Eisverkäufer alles verkaufen?

Diese Frage lässt sich mit dem folgenden Satz zu beantworten. Eine *Zerlegung* des Wahrscheinlichkeitsraums Ω ist eine Menge der Ereignisse B_1, \dots, B_n , so dass $B_1 \cup B_2 \cup \dots \cup B_n = \Omega$ und $B_i \cap B_j = \emptyset$ für alle $i \neq j$ gilt.

Satz 4.22. (Satz von der totalen Wahrscheinlichkeit) Ist B_1, \dots, B_n eine Zerlegung des Wahrscheinlichkeitsraums mit $\Pr\{B_i\} \neq 0$ für alle i , so gilt:

$$\Pr\{A\} = \sum_{i=1}^n \Pr\{A \cap B_i\} = \sum_{i=1}^n \Pr\{B_i\} \cdot \Pr\{A | B_i\}.$$



Diesen Satz bezeichnet man auch als *Adam's Satz*, da er so oft von verschiedenen Mathematikern wiedererfunden wurde.

Beweis.

$$\begin{aligned} \Pr\{A\} &= \Pr\{\Omega \cap A\} = \Pr\left\{\bigcup_{i=1}^n B_i \cap A\right\} = \sum_{i=1}^n \Pr\{B_i \cap A\} \\ &= \sum_{i=1}^n \Pr\{B_i\} \cdot \Pr\{A | B_i\}. \end{aligned}$$

□

▷ *Beispiel 4.23* : (Eisverkäufer – Fortsetzung) Wir betrachten die Ereignisse $B_1 =$ “es ist sonnig”, $B_2 =$ “es ist bevölkert” und $B_3 =$ “es regnet”. Diese Ereignisse zerlegen den Wahrscheinlichkeitsraum, und die entsprechende Wahrscheinlichkeiten sind

$$\Pr\{B_1\} = 0,3 \quad \Pr\{B_2\} = 0,45 \quad \Pr\{B_3\} = 0,25$$

Sei A das Ereignis, dass der Verkäufer alles verkauft. Wir haben die folgenden partiellen Informationen:

$$\Pr\{A | B_1\} = 0,9 \quad \Pr\{A | B_2\} = 0,6 \quad \Pr\{A | B_3\} = 0,2$$

Nach dem Satz von der totalen Wahrscheinlichkeit gilt:

$$\Pr\{A\} = (0,3 \cdot 0,9) + (0,45 \cdot 0,6) + (0,25 \cdot 0,2) = 0,59$$

Da A, \bar{A} für jedes Ereignis A eine Zerlegung des Wahrscheinlichkeitsraums bilden, liefert Satz 4.22 den folgenden:

Korollar 4.24. Seien A und B Ereignisse mit $0 < \Pr\{B\} < 1$. Dann gilt:

$$\Pr\{A\} = \Pr\{B\} \cdot \Pr\{A|B\} + \Pr\{\bar{B}\} \cdot \Pr\{A|\bar{B}\}$$

► *Beispiel 4.25 : (Aus dem Hut gezaubert)* In einem Hut befinden sich drei Karten. Eine ist auf beiden Seiten mit einem Pik versehen $\spadesuit\spadesuit$, eine mit einem Pik und einem Karo $\spadesuit\heartsuit$, und eine mit Karo auf beiden Seiten $\heartsuit\heartsuit$. Wir ziehen rein zufällig eine Karte aus dem Hut. Sei p die Wahrscheinlichkeit dafür, dass die gezogene Karte auf beiden Seiten zwei verschiedene Symbole trägt.

Da wir nur eine solche Karte (aus drei) in dem Hut haben, ist diese Wahrscheinlichkeit offensichtlich $1/3$. Wir wollen aber “zeigen”, dass $p = 1/2$ gilt.

Wir denken uns eine Karte so herausgezogen, dass nur eine Seite sichtbar ist.

Fall 1: Wir sehen ein Pik. Dann kann die gezogene Karte nicht $\heartsuit\heartsuit$ sein. Beide anderen Karten $\spadesuit\heartsuit$ und $\spadesuit\spadesuit$ hatten die gleiche Chance gezogen zu werden. Deshalb ist in diesem Fall $p = 1/2$.

Fall 2: Wir sehen ein Karo. Dieser Fall ist analog.

Wir können diesen “Beweis” auch formalisieren. Sei P das Ereignis “Ich sehe einen Pik” und V sei das Ereignis “die gezogene Karte trägt auf beiden Seiten zwei verschiedene Symbole”. Da $\Pr\{P\} = \Pr\{\bar{P}\} = 1/2$ und $\Pr\{V|P\} = \Pr\{V|\bar{P}\} = 1/2$, liefert uns der Satz von der totalen Wahrscheinlichkeit

$$\Pr\{V\} = \Pr\{P\} \cdot \Pr\{V|P\} + \Pr\{\bar{P}\} \cdot \Pr\{V|\bar{P}\} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Wo ist der Fehler? Der Fehler liegt im falsch gewählten Wahrscheinlichkeitsraum! Da wir nur eine Seite der gezogenen Karte sehen, müssen wir auch unterscheiden, welche von zwei Seiten das ist. Deshalb sind die Elementarereignisse in diesem Experiment nicht die 3 mögliche Kartentypen

$$\{\spadesuit, \spadesuit\}, \{\spadesuit, \heartsuit\}, \{\heartsuit, \heartsuit\}$$

sondern 6 mögliche Paare (Kartentyp, Seite die ich sehe):

$$(\spadesuit\spadesuit, 1), (\spadesuit\spadesuit, 2), (\spadesuit\heartsuit, 1), (\spadesuit\heartsuit, 2), (\heartsuit\heartsuit, 1), (\heartsuit\heartsuit, 2)$$

Deshalb ist $\Pr\{V|P\}$ nicht

$$\Pr\{V|P\} = \frac{|\{\spadesuit, \heartsuit\}|}{|\{\spadesuit\spadesuit, \spadesuit\heartsuit\}|} = \frac{1}{2}$$

sondern

$$\Pr\{V|P\} = \frac{|\{(\spadesuit\heartsuit, 1)\}|}{|\{(\spadesuit\spadesuit, 1), (\spadesuit\spadesuit, 2), (\spadesuit\heartsuit, 1)\}|} = \frac{1}{3}$$

4.4.3 Satz von Bayes

Es gibt einen großen Unterschied zwischen $\Pr\{A|B\}$ und $\Pr\{B|A\}$.

Die Wissenschaftler wollen einen Test für eine Erbkrankheit entwickeln. Natürlich, gibt es keinen perfekten Test: Es werden einige gesunde als kranke eingestuft und umgekehrt. Sei zum Beispiel A das Ereignis “die Testperson ist krank” und B das Ereignis “der Test ist positiv”. Für die Wissenschaftlern

wichtig ist, mit welcher Wahrscheinlichkeit der Testergebnis falsch wird, d.h. für sie sind die Wahrscheinlichkeiten $\Pr\{B|\bar{A}\}$ und $\Pr\{\bar{B}|A\}$ von der Bedeutung. Für die Testperson sind dagegen die Wahrscheinlichkeiten $\Pr\{A|B\}$ und $\Pr\{\bar{A}|\bar{B}\}$ von großer Bedeutung. Ich bin als positiv getestet, mit welcher Wahrscheinlichkeit bin ich wirklich krank? Ich bin als negativ getestet, wie sicher kann ich sein, dass ich tatsächlich gesund bin?

Die bedingten Wahrscheinlichkeiten $\Pr\{A|B\}$ und $\Pr\{B|A\}$ sind durch folgenden Satz verbunden:

Satz 4.26. (Satz von Bayes) Seien A und B Ereignisse mit $\Pr\{A\} \neq 0$ und $\Pr\{B\} \neq 0$. Dann gilt:

$$\Pr\{A|B\} = \frac{\Pr\{A\}}{\Pr\{B\}} \cdot \Pr\{B|A\}$$

Beweis.

$$\Pr\{B\} \cdot \Pr\{A|B\} = \Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B|A\}$$

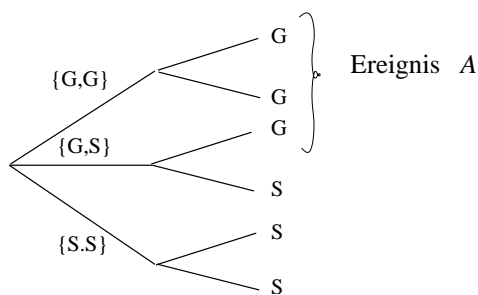
□

Sind $0 < \Pr\{A\} < 1$ und $\Pr\{B\} \neq 0$, so liefert uns Korollar 4.24 die folgende Gleichung:

$$\Pr\{A|B\} = \frac{\Pr\{A\} \cdot \Pr\{B|A\}}{\Pr\{A\} \cdot \Pr\{B|A\} + \Pr\{\bar{A}\} \cdot \Pr\{B|\bar{A}\}}$$

Oft ist der Satz von Bayes in dieser Form formuliert.

► *Beispiel 4.27*: Es gibt drei Beutel, die erste enthält zwei goldene Münzen, die zweite enthält eine goldene und eine silberne Münze, und die dritte enthält zwei silberne Münzen. Ich wähle rein zufällig eine aus der drei Beutel und anschließend ziehe aus dieser Beutel (auch rein zufällig) eine der zwei Münzen. Angenommen, ich habe eine goldene Münze gezogen. Wie groß ist die Wahrscheinlichkeit, dass ich die Beutel mit zwei goldenen Münzen gewählt habe? Sei A das Ereignis "ich habe eine goldene Münze gezogen" und B_1 das Ereignis "ich habe die 1. Beutel (mit zwei goldenen Münzen) gewählt". Gesucht ist also $\Pr\{B_1|A\}$. Der entsprechende Entscheidungsbaum sieht folgendermaßen aus:



Aus dieser Diagramm sehen wir, dass $\Pr\{A\} = 1/2$, $\Pr\{B_1\} = \Pr\{GG\} = 1/3$ und $\Pr\{A|B_1\} = 1$. Die Formel von Bayes liefert uns das Ergebnis:

$$\Pr\{B_1|A\} = \frac{\Pr\{B_1\}}{\Pr\{A\}} \Pr\{A|B_1\} = \frac{(1/3) \cdot 1}{(1/2)} = \frac{2}{3}$$

- **Beispiel 4.28 : (AIDS-Test)** 0,1% aller untersuchten Personen sind HIV-positiv. Der Test ist nicht perfekt: 0,2% der HIV-positiven werden als negative eingestuft; 0,3% der HIV-negativen werden als positive eingestuft.

Frage: $\Pr \{ \text{HIV-positiv} \mid \text{positives Testergebnis} \} = ?$

Lösung: $A =$ "HIV-positives Testergebnis", $B_1 =$ "HIV-positiv", $B_2 =$ "HIV-negativ". Nach der Formel von der totalen Wahrscheinlichkeit:

$$\Pr \{ A \} = \Pr \{ B_1 \} \Pr \{ A \mid B_1 \} + \Pr \{ B_2 \} \Pr \{ A \mid B_2 \} = 0,001 \cdot 0,998 + 0,999 \cdot 0,003 = 0,003996.$$

Nach der Formel von Bayes:

$$\Pr \{ B_1 \mid A \} = \frac{\Pr \{ B_1 \} \Pr \{ A \mid B_1 \}}{\Pr \{ A \}} = \frac{0,001 \cdot 0,998}{0,003996} \approx 0,25.$$

Obwohl der Testfehler so klein ist, wird mit Wahrscheinlichkeit 3/4 eine als HIV-krank eingestufte Person tatsächlich gesund sein! Intuition hier ist klar: obwohl der Fehler wirklich sehr klein sind, ist die (abgeschätzte) Anteil der HIV-positiven Personen noch kleiner.

- **Beispiel 4.29 :** Wir haben eine faire Münze, deren Wurf mit gleicher Wahrscheinlichkeit Kopf oder Zahl ergibt, und eine unfaire Münze, deren Wurf *immer* Kopf ergibt. Wir wählen eine der beiden Münzen zufällig aus und werfen sie zweimal. Angenommen, *beide* Würfe ergeben Kopf. Wie groß ist dann die Wahrscheinlichkeit, dass die unfaire Münze ausgewählt wurde? Dazu betrachten wir zwei Ereignisse:

$A =$ es wurde unfaire Münze gewählt
 $B =$ beide Würfe der Münze ergeben Kopf

Es ist also $\Pr \{ A \mid B \}$ zu bestimmen. Wir haben:

$$\begin{aligned} \Pr \{ A \} &= \Pr \{ \bar{A} \} = 1/2 \\ \Pr \{ B \mid A \} &= 1 \\ \Pr \{ B \mid \bar{A} \} &= \frac{1}{4} \end{aligned}$$

Aus der Formel von der totalen Wahrscheinlichkeit folgt:

$$\begin{aligned} \Pr \{ B \} &= \Pr \{ A \} \Pr \{ B \mid A \} + \Pr \{ \bar{A} \} \Pr \{ B \mid \bar{A} \} \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{8}. \end{aligned}$$

Nun können wir die Formel von Bayes anwenden und erhalten schließlich

$$\Pr \{ A \mid B \} = \frac{\Pr \{ A \} \Pr \{ B \mid A \}}{\Pr \{ B \}} = \frac{\frac{1}{2} \cdot 1}{\frac{5}{8}} = \frac{4}{5}$$

- **Beispiel 4.30 : (Ausschüsse in Lieferungen)** Ein Hersteller von Computern bezieht ein bestimmtes Bauteil von drei Zulieferern: 1, 2 und 3. Die folgenden Anteilswerte an der Gesamtlieferung

sowie die jeweiligen Ausschussanteile innerhalb der drei Lieferungen sind auf Grund längerer Erfahrung bekannt:

	1	2	3
Anteil an Gesamtlieferung	60%	25%	15%
Ausschussanteil in Lieferung	8%	12%	4%

Wir fassen die relativen Häufigkeiten als Wahrscheinlichkeiten auf: Betrachte die Menge Ω aller gelieferten Bauteile mit Gleichverteilung auf Ω . Dann haben wir für die Teilmengen von Ω

$$\begin{aligned} A &= \text{die Ausschuss-Bauteile} \\ B_i &= \text{die von Zulieferer } i \text{ stammenden Bauteile, } i = 1, 2, 3 \end{aligned}$$

die folgenden Wahrscheinlichkeiten und bedingten Wahrscheinlichkeiten:

$$\begin{aligned} \Pr\{B_1\} &= 0,6 & \Pr\{B_2\} &= 0,25 & \Pr\{B_3\} &= 0,15 \\ \Pr\{A | B_1\} &= 0,08 & \Pr\{A | B_2\} &= 0,12 & \Pr\{A | B_3\} &= 0,04 \end{aligned}$$

Frage 1: Wie groß ist der Ausschussanteil in der Gesamtlieferung? Mit der Formel von der totalen Wahrscheinlichkeit:

$$\Pr\{A\} = \sum_{i=1}^3 \Pr\{B_i\} \Pr\{A | B_i\} = 0,08 \cdot 0,6 + 0,12 \cdot 0,25 + 0,04 \cdot 0,15 = 0,084.$$

Frage 2: Welche Anteile am Gesamtausschuss haben die einzelnen Zulieferer? Mit der Formel von Bayes:

$$\Pr\{B_1 | A\} = \frac{\Pr\{B_1\} \Pr\{A | B_1\}}{\Pr\{A\}} = \frac{0,08 \cdot 0,6}{0,084} \approx 0,57;$$

ein Ausschussanteil stammt also mit 57%-iger Wahrscheinlichkeit von Zulieferer Nr. 1. Analog erhalten wir

$$\begin{aligned} \Pr\{B_2 | A\} &= \frac{0,12 \cdot 0,25}{0,084} \approx 0,36 \\ \Pr\{B_3 | A\} &= \frac{0,04 \cdot 0,15}{0,084} \approx 0,07. \end{aligned}$$

Die Anteile der Zulieferer 2 und 3 am Gesamtausschuss betragen also 36% bzw. 7%.

4.5 Stochastische Entscheidungsprozesse

Bei Modellbildung von Experimenten, die in mehreren Schritten ablaufen, ist es oft hilfreich, den ersten Modellierungsschritt (Bestimmung des Wahrscheinlichkeitsraums) als ein *Entscheidungsbaum* darzustellen:

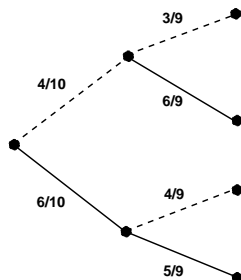
- Verteilung p_1, \dots, p_n des Schrittes 1 sei bekannt;
- Verteilung des Schrittes k sei bekannt *unter der Bedingung, dass Ergebnisse der Schritte 1, \dots, k-1 vorliegen*.

Dies liefert Wahrscheinlichkeiten für spezifische Ergebnisfolgen. Im Allgemeinen gilt für Wahrscheinlichkeitsräume, die als ein Entscheidungsbaum dargestellt sind, folgendes:

- Jede Kante bekommt ihren Gewicht (eine Zahl in $[0,1]$), so dass die Summe der Gewichte der aus einem Knoten ausgehenden Kanten gleich 1 ist. Der Gewicht von $e = (u,v)$ kann man als die *bedingte Wahrscheinlichkeit* interpretieren, entlang dieser Kante weiter zu spazieren, falls man bereits an Knoten u angekommen ist.
- Elementarereignisse ω sind die Wege von der Wurzel zu einem Blatt.
- Das Gewicht $\Pr\{\omega\}$ eines Weges ω ist das *Produkt* der Gewichte seiner Kanten. Das folgt aus dem Multiplikationssatz für Wahrscheinlichkeiten (Satz 4.18).
- Ereignisse E sind Teilmengen der Wege.
- $\Pr\{E\}$ ist die *Summe* der Gewichte der zu E gehörenden Wege.

► *Beispiel 4.31* : Box enthalte 4 weiße und 6 schwarze Kugeln. Ziehe zweimal ohne Zurücklegen.

1. *Wahrscheinlichkeitsraum* $\Omega = \{ww, ws, sw, ss\}$. Hier “ww” entspricht “erste Kugel weiß und zweite Kugel weiß”, “ws” entspricht “erste Kugel weiß und zweite Kugel schwarz” usw. Das kann man als Entscheidungsbaum darstellen (das Ereignis “Kugel ist weiß” ist mit Kante - - - - markiert, und das Ereignis “Kugel ist schwarz” ist mit ——— markiert):



Der Baum besteht aus zwei Ebenen. Die erste Ebene entspricht der Ziehung der ersten Kugel. Die zweite Ebene entspricht der Ziehung der zweiten Kugel *unter der Bedingung, dass die erste Kugel bereits gezogen ist!* So wird z.B. im ersten Schritt die weiße Kugel mit Wahrscheinlichkeit $4/10$ gezogen. Aber die Wahrscheinlichkeit, dass die zweite Kugel auch weiß wird, nachdem die erste gezogene Kugel bereits weiß war, ist gleich $3/9$: Nach dem ersten Schritt bleiben noch 9 Kugeln und nur 3 davon sind weiß. Deshalb ist $\Pr\{ww\} = (4/10) \cdot (3/9) = 2/15$.

2. *Ereignisse*. Für $i = 1, 2$ betrachten wir die Ereignisse $W_i =$ “ i -te Kugel weiß” und $S_i =$ “ i -te Kugel schwarz”. Nach dem Multiplikationssatz für Wahrscheinlichkeiten gilt dann zum Beispiel folgendes:

$$\Pr\{W_1 \cap W_2\} = \Pr\{W_1\} \cdot \Pr\{W_2 | W_1\} = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15}$$

und

$$\begin{aligned}
 \Pr\{W_2\} &= \Pr\{W_2 \cap W_1\} + \Pr\{W_2 \cap S_1\} \\
 &= \Pr\{W_1\} \cdot \Pr\{W_2 | W_1\} + \Pr\{S_1\} \cdot \Pr\{W_2 | S_1\} \\
 &= \frac{4}{10} \cdot \frac{3}{9} + \frac{6}{10} \cdot \frac{4}{9} = \frac{36}{90} = \frac{4}{10} \\
 &= \Pr\{W_1\}.
 \end{aligned}$$

- *Beispiel 4.32*: Wir spielen ein Spiel gegen einen Gegner. Er hat n Zahlen $1, \dots, n$. Der Gegner wählt sich zwei Zahlen $y < z$ aus und schreibt sie für uns nicht sichtbar auf je einem Zettel. Wir wählen *zufällig* einen Zettel und lesen die darauf stehende Zahl $r \in \{y, z\}$, sei $s = \{y, z\} \setminus \{r\}$ die verbleibende Zahl (diese Zahl sehen wir nicht). Wir müssen nun entscheiden, welche der beiden Zahlen r (die uns bekannte Zahl) und s (die uns unbekanntes Zahl) größer ist.

Annahme: Um unsere Entscheidung zu treffen, können wir einen beliebigen Zufallsgenerator benutzen.

Können wir unsere Gewinnchancen größer als 50% machen?

Die Antwort ist: Ja, wir können

$$\Pr\{\text{Gewinn}\} \geq \frac{1}{2} + \frac{1}{2n}$$

erreichen. Für $n = 10$ ist das sogar 55%.

Gewinnstrategie:

1. Rate zufällig eine Zahl \mathbf{x} aus

$$\left\{ 1 - \frac{1}{2}, 2 - \frac{1}{2}, \dots, n - \frac{1}{2} \right\}$$

mit Wahrscheinlichkeit $1/n$. (Da wir $-\frac{1}{2}$ nehmen, gehört \mathbf{x} nicht zu $1, \dots, n$.)

2. Wähle einen der beiden Zettel mit Wahrscheinlichkeit je $1/2$. Diese Wahl ist unabhängig von der Wahl von \mathbf{x} . Sei r die Zahl auf diesem Zettel.
3. Hoffe, dass $y < \mathbf{x} < z$ und gebe die Antwort

$$\text{Antwort} = \begin{cases} r > s & \text{falls } r > \mathbf{x} \\ r < s & \text{falls } r < \mathbf{x} \end{cases}$$

Damit ist (siehe Abb. 4.3):

$$\begin{aligned}
 \Pr\{\text{Gewinn}\} &= \frac{y}{2n} + \frac{z-y}{2n} + \frac{z-y}{2n} + \frac{n-z}{2n} \\
 &= \frac{n+z-y}{2n} \\
 &= \frac{1}{2} + \frac{z-y}{2n} \\
 &\geq \frac{1}{2} + \frac{1}{2n}
 \end{aligned}$$

wobei wir in der letzten Ungleichung die Bedingung $y < z$ ausgenutzt haben.

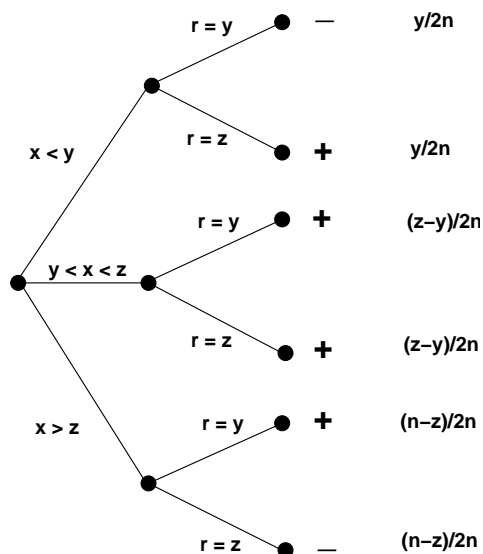


Abbildung 4.3: + = wir gewinnen, - = wir verlieren

4.5.1 Das „Monty Hall Problem“

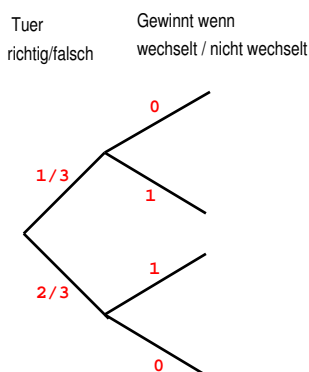
In einer Game Show (z.B. “Gehe auf Ganze”) ist hinter einer von drei Türen ein Hauptpreis (rein zufällig) verborgen.

Ein Zuschauer rät eine der drei Türen und der Showmaster Monty Hall wird daraufhin eine weitere Tür öffnen, hinter der sich aber kein Hauptpreis verbirgt. Der Zuschauer erhält jetzt die Möglichkeit, seine Wahl zu ändern. Sollte er dies tun?

Wir müssen zuerst das Problem genauer beschreiben. Wir nehmen an, dass die folgende drei Bedingungen erfüllt sind:

1. Der Hauptpreis ist mit gleicher Wahrscheinlichkeit $1/3$ hinter jeder der drei Türen verborgen. Der Showmaster weiß, wo der Preis ist, der Zuschauer weiß es natürlich nicht.
2. Egal wo der Hauptpreis ist, wählt der Zuschauer eine der drei Türen mit gleicher Wahrscheinlichkeit $1/3$.
3. Egal wo der Hauptpreis ist, öffnet der Showmaster jede der möglichen Türen (d.h. eine Tür hinter der kein Preis ist) mit gleicher Wahrscheinlichkeit. Also ist diese Wahrscheinlichkeit $1/2$, falls der Zuschauer die Tür mit Hauptpreis gewählt hat, und ist 1 sonst.

Wir betrachten zwei Ereignisse: R = “Zuschauer wählt die *richtige* Tür (die mit dem Preis)” und W = “Zuschauer gewinnt, wenn er die Tür *stets wechselt*”.



Dann gilt:

$$\Pr\{W\} = \Pr\{R\} \cdot \Pr\{W | R\} + \Pr\{\bar{R}\} \cdot \Pr\{W | \bar{R}\} = \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1 = \frac{2}{3}$$

und $\Pr\{\bar{W}\} = 1 - \Pr\{W\} = \frac{1}{3}$. Der Zuschauer sollte seine Wahl stets ändern!

Zu dem selben Ergebnis kann man auch einfach kommen, wenn man die ‘Drei-Schritt-Methode’ anwendet. In unserem Fall besteht Ω aus 9 Elementarereignissen $\omega = (i, j)$ mit $i, j \in \{1, 2, 3\}$. Hier i ist die von dem Zuschauer gewählte Tür und j ist die Tür mit dem Preis. Die Wahrscheinlichkeiten sind $\Pr\{\omega\} = 1/9$ für alle $\omega \in \Omega$. Für uns von Interesse ist das Ereignis $W = \{(i, j) : i \neq j\}$ (Zuschauer gewinnt, wenn er die Tür *stets wechselt*) und $\Pr\{W\} = 6/9 = 2/3$.

4.5.2 Stichproben

Wir haben einen (potentiell unendlichen) Datenstrom x_1, x_2, \dots , wobei alle Elemente x_i verschieden sind. Die Elementen des Datenstroms kommen zu uns ein nacheinanderem, und verschwinden dann für immer. In jedem Zeitpunkt passen in unsem Speicher nur ein der Elemente.

Ein *Representant* des Datenstroms ist ein zufällig gewähltes Element \mathbf{x} mit der Eigenschaft, dass $\Pr\{\mathbf{x} = x_i\} = 1/n$ für jedes $n \geq 1$ und für jedes Element x_i mit $i \in \{1, \dots, n\}$ gilt. D.h. \mathbf{x} muss in jedem Interval x_1, \dots, x_n gleichverteilt sein. Dieses Problem kann man mit dem folgenden Algorithmus (bekannt als *Reservoir Sampling*) lösen.

- Für $n = 1$ setze $\mathbf{x} = x_1$
- Für $n \geq 2$ werfe eine Münze mit Erfolgswahrscheinlichkeit $1/n$. Bei einem Erfolg setze $\mathbf{x} = x_n$. Bei einem Misserfolg wird nichts unternommen.

Satz 4.33. Für jedes $n \geq 1$ und für jedes $1 \leq i \leq n$ gilt: $\Pr\{\mathbf{x} = x_i\} = 1/n$.

Beweis.

$$\begin{aligned} \Pr\{\mathbf{x} = x_i\} &= \Pr\{x_i \text{ ist in } i\text{-ten Schritt gewählt}\} \times \\ &\quad \times \Pr\{\text{keiner von } x_{i+1}, \dots, x_n \text{ ist später gewählt}\} \\ &= \frac{1}{i} \cdot \prod_{j=i+1}^n \left(1 - \frac{1}{j}\right) \\ &= \frac{1}{i} \cdot \frac{i}{i+1} \cdot \frac{i+1}{i+2} \cdot \frac{i+2}{i+3} \cdots \frac{m-2}{m-1} \cdot \frac{m-1}{m} = \frac{1}{n}. \end{aligned}$$

□

Nun wollen wir nicht nur einen Repräsentanten sondern eine Menge der Repräsentanten (eine Stichprobe) $\mathbf{S} \subseteq \{x_1, x_2, \dots\}$ mit $|\mathbf{S}| = s$ Elementen auswählen. Es muss $\Pr\{\mathbf{S} = T\} = \binom{n}{s}^{-1}$ für jedes $n \geq 1$ und für jede s -elementige Teilmenge $T \subseteq \{x_1, \dots, x_n\}$ gelten. Hier nehmen wir an, dass unser Speicher bis zu s Elementen aufnehmen kann. Um das Problem zu lösen, reicht es den Reservoir-Sampling Algorithmus für $s = 1$ nur leicht zu modifizieren.

- Wenn $n \leq s$, dann füge das Element x_n in \mathbf{S} ein.
- Für $n > s$ werfe eine Münze mit Erfolgswahrscheinlichkeit s/n . Bei einem Erfolg bestimme zufällig ein Element aus \mathbf{S} und entferne das Element; das aktuelle Element x_n wird eingefügt. Bei einem Misserfolg wird nichts unternommen.

Satz 4.34. Für jede $n \geq s$ und für jede s -elementige Teilmenge $T \subseteq \{x_1, \dots, x_n\}$ gilt:

$$\Pr\{\mathbf{S} = T\} = \binom{n}{s}^{-1}.$$

Beweis. Sei \mathbf{S}_n die zum Zeitpunkt n vom Algorithmus gewählte s -elementige Stichprobe. Wir wissen, dass $\Pr\{x_n \in \mathbf{S}_n\} = s/n$ für jedes $n > s$ gilt.

Wir führen Induktion nach n . Induktionsbasis $n = s$ gilt, da dann $T = \{x_1, \dots, x_s\}$ die einzige s -elementige Teilmenge ist und deshalb $\Pr\{\mathbf{S}_s = T\} = 1 = \binom{s}{s}^{-1}$ gilt.

Induktionsschritt $n - 1 \rightarrow n$: Sei $T \subseteq \{x_1, \dots, x_n\}$ eine beliebige s -elementige Teilmenge. Wir haben nur zwei Möglichkeiten: entweder T enthält das Element x_n oder nicht.

$$x_n \notin T \Rightarrow \Pr\{\mathbf{S}_n = T\} = \Pr\{x_n \notin \mathbf{S}_n\} \cdot \Pr\{\mathbf{S}_{n-1} = T\} = \left(1 - \frac{s}{n}\right) \binom{n-1}{s}^{-1} = \binom{n}{s}^{-1}$$

$$x_n \in T \Rightarrow \Pr\{\mathbf{S}_n = T\} = \Pr\{x_n \in \mathbf{S}_n\} \cdot \Pr\{\mathbf{S}_{n-1} = T \setminus \{x_n\}\} = \frac{s}{n} \binom{n-1}{s-1}^{-1} = \binom{n}{s}^{-1}.$$

Hier haben wir die Gleichungen $\binom{n}{s} / \binom{n-1}{s-1} = \frac{n}{s}$ und $\binom{n}{s} / \binom{n-1}{s} = \frac{n}{n-s}$ benutzt. □

4.5.3 Das “Sekretärinnen-Problem” an der Börse

Wie wählt man unter 10 Sekretärinnen die beste aus, wenn während des Bewerbungsgesprächs die Zusage erteilt werden soll? Mit diesem “Sekretärinnen-Problem” wird in der Literatur die folgende Aufgabenstellung veranschaulicht: Unter n aufeinanderfolgenden “Gelegenheiten”, für die noch keine Rangfolge bekannt ist, soll die beste ausgewählt werden, indem sie geprüft und sofort zugegriffen wird, andernfalls ist sie für immer verpasst.

In manchen Lehrbüchern findet man dafür die Bezeichnung: “Vermählungs-Problem”. Wie wählt eine Frau am effizientesten unter allen ihren Bekannten einen Mann für das Leben? Dabei entscheidet sie bei jedem Mann, ob er ihr Traummann ist oder nicht; wenn sie ihn abgelehnt hat, kann sie später nicht mehr auf ihn zurückgreifen.

Die allgemeine Strategie lautet: Teste eine gewisse Anzahl von Möglichkeiten und triff deine Wahl aufgrund der Testergebnisse. In unserem Beispiel wird also ein Teil der männlichen Bekannten einer Testprozedur unterzogen und dann trifft die Frau ihre Entscheidung.

Zwei Dinge sind klar:

- Sie sollte nicht den ersten Mann nehmen, denn wer weiß, was noch kommt. Mit anderen Worten: Sehr wahrscheinlich ist der "Erstbeste" nicht der beste.
- Andererseits sollte sie auch nicht zu lange warten, denn dann hat sie mit großer Wahrscheinlichkeit den besten abgelehnt und muss sich also mit einem Mann minderer Qualität begnügen.

Daraus ergibt sich folgende Strategie: Sie testet eine gewisse Anzahl von Männern mit Hilfe eines Verfahrens, über das sie uns nichts zu verraten braucht - und nimmt keinen von diesen! Danach führt sie ihr Testverfahren weiter und nimmt dann den ersten, der besser als alle bisherigen ist.

Die Frage ist nur, wie viele Männer sich ohne Aussicht auf Erfolg dem Testverfahren unterwerfen müssen. Man kann beweisen (und wir werden das bald tun), dass die Frau 37% ihrer in Frage kommenden Bekannten testen soll. Genauer gesagt soll sie einen Bruchteil von $1/e$ testen, wobei $e = 2,718\dots$ die Eulersche Zahl ist.

Interessanterweise ist der Prozentsatz unabhängig von der Zahl der Testmänner: Egal, ob sie 10 oder 1000 Heiratskandidaten ernsthaft in Erwägung zieht, stets ist die beste Strategie, zunächst 37% auszuprobieren und diese zu verwerfen.

Dabei ergeben sich - mindestens aus männlicher Sicht - verschiedene Fragen: Was ist, wenn ich, der Idealmann, unter den ersten 37% und damit von vornherein ausgeschlossen bin? Und was ist, wenn ich, der beste, erst am Ende getestet werde, und also gar nicht zum Zuge komme? Ist das nicht ein total unfaires Verfahren?

Nein! Dieses Verfahren ist ganz gut! Denn mit einer Wahrscheinlichkeit von immerhin $1/e = 37\%$ findet Frau mit dieser Strategie tatsächlich den Mann ihrer Träume! Beachte, dass mit einer trivialen Startegie (wähle aus n bekannten Männer den Traummann rein zufällig einen mit Wahrscheinlichkeit $1/n$) würde bereits bei $n = 4$ Heiratskandidaten die Frau mit viel kleiner Wahrscheinlichkeit glücklich sein.

Aber nicht nur Heiratswillige, sondern auch Aktionäre interessieren sich für die Lösung dieses Problems. Die Lösungsstrategie zum Sekretärinnenproblem wird beim Aktienhandel angewandt, wenn der Kurs einer Aktie ständig schwankt und nicht vorhersagbar ist. Wenn man innerhalb von einem Monat entsprechende Aktien verkaufen möchte, wie kann man mit dieser Strategie den günstigsten Verkaufstag finden?

Um die Lösung dieses Problems zu demonstrieren,⁵ nehmen wir an, dass die Kurse am keinen zwei Tagen gleich sind. Dann gilt der folgende Satz:

⁵Ist der Wahrscheinlichkeitsraum *endlich*, so ist die ganze Stochastik nichts anderes als ein Teil der Kombinatorik. In dieser (endlichen) Form war eigentlich die Stochastik geboren. Das Ziel dieses Abschnitts ist zu zeigen, wie man mit Hilfe von (relativ einfachen) kombinatorischen Überlegungen einige nicht triviale Schlussfolgerungen ziehen kann.

Satz 4.35. Wenn die Anzahl der Handelstage n groß ist, dann sollten die Aktienkurse der ersten $j = n/e$ (knapp 37%) Tage lediglich notiert sein und dann die nächste bessere Gelegenheit ausgewählt werden. Die Wahrscheinlichkeit $P(j)$, dann den günstigsten Verkaufstag zu wählen, beträgt

$$P(j) = \frac{j}{n} \cdot \left(\frac{1}{j} + \dots + \frac{1}{n-1} \right)$$

Für $j = n/e$ ist $P(j) \approx 1/e = 0,367$.

Zum Beispiel für $n = 20$ (Handelstage) ist die optimale Stoppzahl $j = 7$. Die Wahrscheinlichkeit, dann den günstigsten Verkaufstag zu wählen, beträgt $P(7) = 7/20(1/7 + \dots + 1/19) = 0,384\dots$ Für $n = 100$ ist die Stoppzahl $j = 37$ und die Erfolgswahrscheinlichkeit $P(j)$ rund $37/100$ beträgt. Für $n = 1000$ ergibt sich die optimale Stoppzahl $j = 368$ und $P(j) = 368/1000$.

Beweis. Eine Lösungsstrategie zum Börse-Problem (wie auch zum Heirats- oder Sekretärinnen Problem) ist sehr kurz:

j -te Stopppstrategie: Bei $n > 1$ vorgegebenen Tagen verhalte auf folgende Weise: An den ersten j Tagen ($j \in \{1, \dots, n-1\}$) wird lediglich der Kurs beobachtet (und notiert). Sobald der Kurs an einem der nachfolgenden Tage $k > j$ höher ist als das Maximum der j beobachteten Kurse, werden die Aktien verkauft.

Es ist klar, dass die Wahrscheinlichkeit, den bestmöglichen Verkaufstag zu wählen, bei gegebenem n von der Stoppzahl j abhängt. Deswegen bezeichnen wird diese Wahrscheinlichkeit mit $P(j)$ und berechnen sie wie folgt:

Für $k \in \{j+1, \dots, n\}$ betrachten wir das Ereignis

$$A_k = \text{“der } k\text{-te Tag } T_k \text{ ist der beste und } T_k \text{ wird ausgewählt.”}$$

Die Wahrscheinlichkeit, dass der k -te Tag T_k der beste ist, beträgt $1/n$, da wir n Tage haben und jeder davon könnte der beste sein.

Behauptung 4.36.

$$\Pr\{A_k\} = \frac{1}{n} \cdot \frac{j}{k-1} \quad (*)$$

Wir werden diese Behauptung erst später überprüfen (der Beweis ist rein kombinatorisch). An dieser Stelle nehmen wir an, dass die Behauptung 4.36 gilt, und führen den Beweis weiter.

Die j -te Stopppstrategie ist genau dann erfolgreich, wenn das Ereignis $A_{j+1} \cup A_{j+2} \cup \dots \cup A_n$ eintritt. Weil für alle $r, s \in \{j+1, \dots, n\}$ mit $r \neq s$ folgt ⁶ $A_r \cap A_s = \emptyset$, erhalten wir:

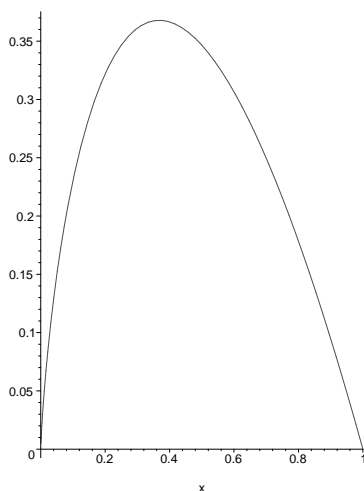
$$\begin{aligned} P(j) &= \Pr\{j\text{-te Stopppstrategie ist erfolgreich}\} = \Pr\{A_{j+1}\} + \Pr\{A_{j+2}\} + \dots + \Pr\{A_n\} \\ &= \frac{j}{n} \cdot \left(\frac{1}{j} + \dots + \frac{1}{n-1} \right) \end{aligned}$$

⁶Da nur ein Tag ausgewählt sein kann.

Um das optimale j zu finden, müssen wir die Funktion $P(j)$ maximieren. Da die harmonische Reihe $H_n = \sum_{i=1}^n \frac{1}{i}$ asymptotisch gleich $\ln n$ ist, erhalten wir:

$$P(j) = \frac{j}{n} (H_{n-1} - H_{j-1}) \sim \frac{j}{n} \ln \frac{n}{j}.$$

Die Funktion $f(x) = x \ln \frac{1}{x}$ sieht folgendermaßen aus:



und erhält ihr Maximum für $x = 1/e$: Die erste Ableitung $f'(x) = \ln(1/x) - 1$ ist in diesem Punkt gleich Null, und die zweite Ableitung $f''(x) = -(1/x)$ ist für $x = 1/e$ negativ.

□

Es bleibt uns, die Behauptung 4.36 zu beweisen. Und hier kommt die Kombinatorik ins Spiel. Wir müssen die Wahrscheinlichkeit $\Pr\{A_k\}$ des Ereignisses A_k bestimmen. Dazu benutzen wir unsere “Drei-Schritte-Methode”.

1. Finde den Wahrscheinlichkeitsraum: Dazu beobachten wir, dass für uns die *tatsächliche* Kurswerte in einem Aktienkursverlauf irrelevant sind – wichtig ist nur die relative Güte dieser Werte: Uns interessiert nur, ob am einem Tag der Kurs besser oder schlechter als am einem anderen ist. Damit können wir⁷ jeden Kursverlauf mit Hilfe eines Urnenmodells betrachten. Wir stellen n Urnen U_1, \dots, U_n in einer Reihe auf. Dann nehmen wir n durchnummerierte Bälle $1, 2, \dots, n$, die den n Handelstagen entsprechen, und verteilen die Bälle (= Tage) in Urnen, so dass jede Urne genau einen Ball enthält. (Eine Verteilung der Bälle ist also nichts anderes als eine Permutation der Handelstage $1, \dots, n$.) Eine solche Verteilung der Bälle entspricht der Anordnung der Handelstage nach ihrem Kurswert. Der Tag in Urne 1 war der schlechteste, der Tag in U_2 war schon besser, der Tag in U_2 war noch besser, usw.

Damit besteht unser Wahrscheinlichkeitsraum Ω aus allen $|\Omega| = n!$ möglichen Verteilungen der n Tage in n Urnen.

2. Bestimme das Ereignis: Welche Elementarereignisse Ω gehören zu dem Ereignis A_k ? Das Ereignis A_k besteht aus zwei Ereignissen:

⁷Dieser Schritt – mathematische Modellierung des Problems – ist sehr wichtig in allen Anwendungen!

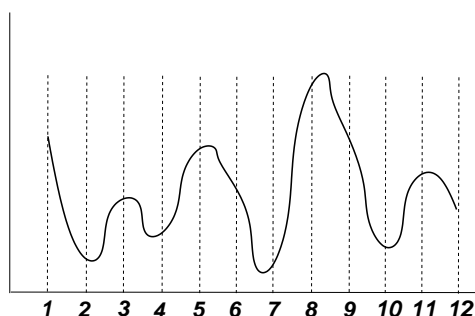


Abbildung 4.4: Aktienkurverlauf in zwei Wochen. Die entsprechende Verteilung der 12 Bälle (= Tage) für diesen Verlauf ist: 7, 2, 10, 4, 12, 3, 6, 11, 5, 1, 9, 8. Tag 8 ist der beste Tag.

1. Zuerst muss der k -te Tag T_k der *beste* sein. Dies bedeutet, dass der Ball k muss in der letzten Urne U_n sitzen.
2. Dann muss noch der Tag T_k auch *ausgewählt* sein. Dies bedeutet folgendes: Wenn wir nur die Urnen anschauen, die die ersten $k - 1$ Bälle $1, \dots, k - 1$ enthalten, dann muss die *letzte* von diesen Urnen einen der ersten j Bälle $s \in \{1, \dots, j\}$ enthalten.⁸

3. Bestimme die Wahrscheinlichkeit des Ereignis: Man kann die zum Ereignis A_k gehörende Verteilungen der Bälle wie folgt erzeugen.

1. Werfe zuerst den Ball k in die letzte Urne $U_n \Rightarrow$ nur eine Möglichkeit.
2. Für jedes $s = 1, 2, \dots, j$ tue folgendes:
 - (a) Werfe die Bälle $k + 1, k + 2, \dots, n$ in die Urnen U_1, \dots, U_{n-1} (die Urne U_n ist ja bereits besetzt) $\Rightarrow P(n - 1, n - k)$ Möglichkeiten, wobei $P(m, r) = m(m - 1) \dots (m - r + 1) = \frac{m!}{(m-r)!}$ die Anzahl der r -Permutationen von $\{1, \dots, m\}$ ist
 - (b) Werfe den Ball s in die *letzte* von $k - 1$ noch freien Urnen (nur eine Möglichkeit), und verteile anschließend die verbleibende $k - 2$ Bälle in $k - 2$ noch verbleibende freie Urnen $\Rightarrow (k - 2)!$ Möglichkeiten.

Nach der Produktregel gilt somit:

$$\begin{aligned}
 |A_k| &= j \cdot P(n - 1, n - k) \cdot (k - 2)! \\
 &= j \cdot \frac{(n - 1)!}{\underbrace{((n - 1) - (n - k))!}_{=(k-1)!}} \cdot (k - 2)! \\
 &= j \cdot \frac{(n - 1)!}{(k - 1)!} \cdot (k - 2)! \\
 &= \frac{j(n - 1)!}{k - 1}.
 \end{aligned}$$

⁸Der beste der ersten $k - 1$ Tage muss ein der ersten j Tage gewesen sein: Nach dem j -ten Tag wählt die j -te Stopppstrategie den *ersten(!)* Tag, der besser als alle Tage $1, \dots, j$ ist.

Da wir insgesamt $|\Omega| = n!$ Verteilungen der Bälle haben, ist die Wahrscheinlichkeit

$$\Pr\{A_k\} = \frac{|A_k|}{|\Omega|} = \frac{j \cdot (n-1)!}{(k-1) \cdot n!} = \frac{1}{n} \cdot \frac{j}{k-1}.$$

Somit ist die Behauptung (*) und damit auch der Satz 4.35 bewiesen. \square

4.6 Zufallsvariablen

“The Holy Roman Empire was neither holy nor Roman, nor an Empire”
–Voltaire

Genauso sind “Zufallsvariablen”: Sie sind weder zufällig noch Variablen:

Eine *Zufallsvariable* ist eine auf dem Wahrscheinlichkeitsraum definierte Funktion.

Die Funktion selbst kann beliebige Werte annehmen, aber normalerweise werden die Zufallsvariablen als Funktionen $X : \Omega \rightarrow S$ mit $S \subseteq \mathbb{R}$ betrachtet.

▷ *Beispiel 4.37*: Wir würfeln zweimal einen Spielwürfel und sind in der Augensumme interessiert. Der Wahrscheinlichkeitsraum ist $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, und die entsprechende Zufallsvariable ist mit $X(i, j) = i + j$ gegeben; der Bildbereich von X ist in diesem Fall $S = \{2, 3, \dots, 12\}$.

Sei $X : \Omega \rightarrow S$ eine Zufallsvariable. Die wichtigste Frage, die wir betrachten werden ist: Für ein gegebenes Element $a \in S$, was ist die Wahrscheinlichkeit, dass X den Wert a annimmt? In anderen Worten was ist die Wahrscheinlichkeit für das Ereignis

$$A = \{\omega \in \Omega : X(\omega) = a\}$$

Weil solche Ereignisse sehr oft auftreten, benutzt man die Abkürzung

$$\Pr\{X = a\}$$

für

$$\Pr\{\{\omega \in \Omega : X(\omega) = a\}\}.$$

Die *Verteilung* einer Zufallsvariablen $X : \Omega \rightarrow S$ ist eine durch $f(a) := \Pr\{X = a\}$ definierte Abbildung $f : S \rightarrow [0, 1]$.

▷ *Beispiel 4.38*: Wir werfen dreimal eine Münze und sei X die Anzahl der Ausgänge “Wappen”. Die mögliche Werte von X sind $S = \{0, 1, 2, 3\}$ und die Verteilung sieht folgendermaßen aus:

a	0	1	2	3
$\Pr\{X = a\}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



Beachte, dass verschiedene Zufallsvariablen dieselbe Verteilung haben können. Wenn wir z.B. den obigen Beispiel betrachten und die Anzahl der Augänge “Kopf” mit Y bezeichnen, dann sind die Zufallsvariablen X und Y verschieden (da $Y = 3 - X$ gilt) aber die Verteilungen von Y und X sind gleich.

Zur Sprechweise: Wir sagen, dass X_1, \dots, X_n *unabhängige Kopieen* einer Zufallsvariable X sind, falls die Zufallsvariablen X_i die gleiche Verteilung wie X haben.

Die einfachsten (und deshalb die wichtigsten) Zufallsvariablen sind *Bernoulli-Variablen*. Jede solche Zufallsvariable X kann nur zwei mögliche Werte 0 und 1 annehmen; $p = \Pr\{X = 1\}$ heißt dann die *Erfolgswahrscheinlichkeit*. Beispiel: Einmaliges werfen einer Münze, wobei der Ausgang “Wappen” mit Wahrscheinlichkeit p kommt. Das entsprechende Zufallsexperiment nennt man *Bernoulli-Experiment*.

Definition: Die *Indikatorvariable* für ein Ereignis $A \subseteq \Omega$ ist die Zufallsvariable $X_A : \Omega \rightarrow \{0, 1\}$ mit

$$X_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{falls } \omega \notin A \end{cases}$$

Beachte, dass eine Indikatorvariable X_A eine Bernoulli-Variable mit der Erfolgswahrscheinlichkeit $\Pr\{X_A = 1\} = \Pr\{A\}$ ist. Somit kann man die Ereignisse als einen Spezialfall der Zufallsvariablen – nämlich als 0-1-wertige Zufallsvariablen – betrachten.

Den Begriff der Unabhängigkeit kann man auch auf Zufallsvariablen übertragen.

Zwei Zufallsvariablen $X : \Omega \rightarrow S$ und $Y : \Omega \rightarrow T$ sind *unabhängig*, falls für jede $s \in S$ und $t \in T$ mit $\Pr\{Y = t\} \neq 0$ gilt

$$\Pr\{X = s | Y = t\} = \Pr\{X = s\}$$

oder äquivalent

$$\Pr\{X = s \wedge Y = t\} = \Pr\{X = s\} \cdot \Pr\{Y = t\}.$$

► *Beispiel 4.39:* Wir würfeln zwei (faire) Spielwürfel. Wir können die entsprechenden Augenzahlen als zwei Zufallsvariablen X_1 und X_2 betrachten. Zum Beispiel ist das Elementarereignis $\omega = (3, 5)$, dann ist $X_1(\omega) = 3$ und $X_2(\omega) = 5$.

Betrachten wir nun $Y = X_1 + X_2$. Dann ist Y auch eine Zufallsvariable, denn Y weist jedem Elementarereignis ω eine reelle Zahl $Y(\omega) = X_1(\omega) + X_2(\omega)$ zu. Sei auch

$$I := \begin{cases} 1 & \text{falls } Y = 7 \\ 0 & \text{falls } Y \neq 7 \end{cases}$$

Dann sind Y und X_1 *abhängig*, da zum Beispiel

$$\Pr\{Y = 2 | X_1 = 3\} = 0 \neq \frac{1}{36} = \Pr\{Y = 2\}.$$

Intuitiv, sollten deshalb auch I und X_1 abhängig sein: Der Wert von I hängt doch davon ab, welchen Wert die Zufallsvariable Y annimmt, und Y hängt doch von X_1 ab.

Überraschenderweise sind I und X_1 *unabhängig*!

Zu zeigen, dass zwei Zufallsvariablen abhängig sind, ist oft viel leichter (es reicht ein Gegenbeispiel, wie oben). Unabhängigkeit braucht mehr Arbeit: Man muss zeigen, dass für *alle* $x, y \in \mathbb{R}$ mit $\Pr\{X_1 = x\} \neq 0$ gilt:

$$\Pr\{I = y | X_1 = x\} = \Pr\{I = y\}$$

Zu betrachten sind also 12 Fälle: $y \in \{0, 1\}$ und $x \in \{1, 2, 3, 4, 5, 6\}$. Zwei Beobachtungen machen uns das Leben einfacher:

- (a) $\Pr\{I = 1\} = 6/36 = 1/6$, da $I(\omega) = 1$ genau dann, wenn eines der 6 unabhängigen Ereignisse $\omega \in \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ eintritt.
- (b) $\Pr\{I = 1 \mid X_1 = x\} = 1/6$ für alle $x \in \{1, 2, 3, 4, 5, 6\}$, da nachdem der erste Würfel x Augen gezeigt hat, dass Ereignis “ $I = 1$ ” nur dann möglich ist, wenn der zweite Würfel $7 - x$ Augen zeigen wird.

Damit haben wir

$$\Pr\{I = 1 \mid X_1 = x\} = \frac{1}{6} = \Pr\{I = 1\} \quad \text{für alle } x = 1, 2, 3, 4, 5, 6$$

Behauptung 4.7 sagt, dass dann auch

$$\Pr\{I = 0 \mid X_1 = x\} = \frac{5}{6} = \Pr\{I = 0\} \quad \text{für alle } x = 1, 2, 3, 4, 5, 6$$

Also sind I und X_1 *unabhängig*!

Die Zufallsvariablen $X_1, \dots, X_k : \Omega \rightarrow S$ mit $S \subseteq \mathbb{R}$ heißen *total unabhängig*, falls für alle $a_1, \dots, a_k \in S$ die Gleichung

$$\Pr\{X_1 = a_1 \text{ und } X_2 = a_2 \text{ und } \dots \text{ und } X_k = a_k\} = \prod_{i=1}^k \Pr\{X_i = a_i\}$$

gilt.

- *Beispiel 4.40*: Seien die Zufallsvariablen $X_1, \dots, X_k : \Omega \rightarrow S$ mit $S \subseteq \mathbb{R}$ total unabhängig. Wie groß kann dann k sein? Um diese Frage zu beantworten, nehmen wir an, dass die Zufallsvariablen nicht trivial sind, d.h. keine der Variablen X_i konstant ist. Dann kann jede der Variablen X_i mindestens zwei verschiedene Werte $a_i \neq b_i \in S$ annehmen (d.h. $\Pr\{X_i = a_i\} > 0$ und $\Pr\{X_i = b_i\} > 0$ gilt). Wir haben mindestens 2^k Möglichkeiten, einen Vektor $c = (c_1, \dots, c_k)$ mit $c_i \in \{a_i, b_i\}$ auszuwählen, und für jede solche Auswahl hat das Ereignis $A_c := \{\omega \in \Omega : X_1 = c_1, \dots, X_k = c_k\}$ die Wahrscheinlichkeit $\Pr\{A_c\} = \prod_{i=1}^k \Pr\{X_i = c_i\} > 0$. Somit sind die Ereignisse A_c nicht leer. Wir haben also mindestens 2^k disjunkten, nicht leeren Teilmengen A_c in Ω gefunden. Somit muss auch die Ungleichung $2^k \leq n := |\Omega|$ gelten.



Fazit: In einem Wahrscheinlichkeitsraum Ω der Größe $n = |\Omega|$ kann es höchstens $\log_2 n$ total unabhängigen nicht trivialen Zufallsvariablen geben.

4.7 Erwartungswert und Varianz

Hat man eine Zufallsvariable $X : \Omega \rightarrow S$, so will man die Wahrscheinlichkeiten $\Pr\{X \in R\}$ für verschiedene Teilmengen $R \subseteq S$ bestimmen (oder mindestens gut abschätzen). Dazu haben sich zwei numerische Charakteristiken der Zufallsvariablen – Erwartungswert und Varianz – als sehr hilfreich erwiesen.

Sei X eine Zufallsvariable, die die Werte a_1, \dots, a_n annimmt.

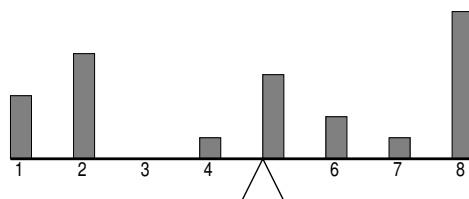
Definition: Der Erwartungswert $E[X]$ von X ist definiert durch:

$$E[X] := \sum_{i=1}^n a_i \cdot \Pr\{X = a_i\}$$

D.h. wir multiplizieren die Werte, die X annehmen kann, mit der entsprechenden Wahrscheinlichkeiten, und summieren die Terme auf. Der Erwartungswert ist also ein "verallgemeinerter Durchschnittswert": Falls X jeden Wert a_i mit gleicher Wahrscheinlichkeit $1/n$ annimmt, so ist

$$E[X] = \frac{a_1 + \dots + a_n}{n}.$$

Man kann den Erwartungswert auch rein mechanisch interpretieren. Wenn wir n Objekte mit Gewichten $p_i = \Pr\{X = a_i\}$ auf der x -Ache in der Positionen a_i ($i = 1, \dots, n$) ablegen, dann wird der Maß-Zentrum genau an der Stelle $E[X]$ sein.



► *Beispiel 4.41*: Wir würfeln zwei (nummerierte) Spielwürfeln und Y sei die Summe der Augenzahlen. Dann ist

$$\begin{aligned} E[Y] &= \sum_{k=2}^{12} k \cdot \Pr\{\text{Summe der beiden Augenzahlen ist } k\} \\ &= \sum_{k=2}^{12} \sum_{i+j=k} k \cdot \Pr\{\text{erster Würfel hat } i \text{ Augen und der zweite hat } j\} \\ &= 2 \cdot \frac{1}{6^2} + 3 \cdot \frac{2}{6^2} + 4 \cdot \frac{3}{6^2} + \dots + 10 \cdot \frac{3}{6^2} + 11 \cdot \frac{2}{6^2} + 12 \cdot \frac{1}{6^2} \\ &= \frac{252}{36} = 7 \end{aligned}$$

Falls die Zufallsvariable *unendlich* viele Werte a_1, a_2, \dots annehmen kann, dann ist der Erwartungswert als

$$E[X] := \lim_{n \rightarrow \infty} \sum_{i=1}^n a_i \Pr\{X = a_i\} = \sum_{i=1}^{\infty} a_i \Pr\{X = a_i\}$$

definiert. Natürlich müssen wir dann uns darum kümmern, ob der Grenzwert überhaupt existiert, d.h. ob die Reihe überhaupt *konvergiert*! Sind alle Zahlen a_i nicht negativ, so ist die Reihe monoton wachsend und (nach Monotonie-Kriterium, Satz 3.46(1)) konvergiert die Reihe genau dann, wenn sie beschränkt ist.⁹

⁹Eine Reihe $\sum_{k=0}^{\infty} a_k$ ist beschränkt, falls es eine Zahl L mit $\sum_{k=0}^n a_k \leq L$ für alle n gilt.

► *Beispiel 4.42* : Sei X eine Zufallsvariable mit der Verteilung $\Pr\{X = 2^i\} = 1/2^i$ für alle $i = 1, 2, \dots$. Das ist eine legale Verteilung, da

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = \sum_{i=0}^{\infty} \frac{1}{2^i} - 1 = \frac{1}{1 - (1/2)} - 1 = 1$$

gilt. Aber der Erwartungswert ist nicht definiert:

$$E[X] = \sum_{i=1}^{\infty} \frac{1}{2^i} 2^i = \sum_{i=1}^{\infty} 1 = \infty.$$

In dieser Vorlesung werden wir meist nur die Fälle treffen, wenn $\sum_{i=1}^{\infty} a_i \Pr\{X = a_i\}$ eine geometrische (und damit auch konvergente) Reihe ist.

Definition: Die *Varianz* $\text{Var}[X]$ einer Zufallsvariable X ist definiert durch:

$$\text{Var}[X] = E[(X - E[X])^2]$$

Was bedeutet eigentlich der Ausdruck $E[(X - E[X])^2]$? Der Ausdruck $X - E[X]$ ist genau die Abweichung der Zufallsvariable X vom seinen Erwartungswert. Dann ist die Zufallsvariable $Y = (X - E[X])^2$ nah an 0, wenn X nah an $E[X]$ ist, und ist eine große Zahl, wenn X weit links oder rechts von $E[X]$ liegt. Die Varianz ist einfach der Durchschnitt dieser Zufallsvariable. Damit ist die Intuition, die dahinter steckt, geklärt:

- Wenn X *immer* nahe an $E[X]$ liegt, dann ist $\text{Var}[X]$ klein.
- Wenn X *oft* weit von $E[X]$ entfernt liegt, dann ist $\text{Var}[X]$ groß.

Bemerkung 4.43. Die Definition der Varianz $E[(X - E[X])^2]$ als ein *Quadrat* von der Abweichung vom Erwartungswert sieht irgendwie künstlich aus. Warum kann man nicht einfach $E[X - E[X]]$ nehmen? Antwort: $E[X - E[X]] = E[X] - E[E[X]] = E[X] - E[X] = 0$. Also hätte dann jede Zufallsvariable die Varianz 0. Nicht sehr nützlich!

Natürlich könnte man die Varianz als $E[|X - E[X]|]$ definieren. Es spricht nichts dagegen. Trotzdem hat die übliche Definition von $\text{Var}[X]$ einige mathematische Eigenschaften, die $E[|X - E[X]|]$ nicht hat.

► *Beispiel 4.44* : Um die Relevanz der Varianz zu demonstrieren, betrachten wir die zwei folgenden Casino-Spiele.

Spiel A: Wir gewinnen 2€ mit Wahrscheinlichkeit 2/3 und verlieren 1€ mit Wahrscheinlichkeit 1/3.

Spiel B: Wir gewinnen 1002€ mit Wahrscheinlichkeit 2/3 und verlieren 2001€ mit Wahrscheinlichkeit 1/3.

Welches Spiel ist für uns günstiger? In beiden Fällen ist unsere Gewinnwahrscheinlichkeit gleich 2/3. Seien A und B die Zufallsvariablen, die dem Gewinn in beiden Spielen entsprechen. Zum

Beispiel ist $A = 2$ mit Wahrscheinlichkeit $2/3$, und -1 mit Wahrscheinlichkeit $1/3$. Dann sind die Erwartungswerte beide gleich:

$$\begin{aligned} E[A] &= 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1 \\ E[B] &= 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1. \end{aligned}$$

Aber das sagt uns nicht die ganze Wahrheit: Intuitiv sieht Spiel B viel gefährlicher aus! Und diesen Unterschied können wir mit Hilfe der Varianz belegen:

$$\begin{aligned} A - E[A] &= \begin{cases} 1 & \text{mit Wahrscheinlichkeit } 2/3 \\ -2 & \text{mit Wahrscheinlichkeit } 1/3 \end{cases} \\ (A - E[A])^2 &= \begin{cases} 1 & \text{mit Wahrscheinlichkeit } 2/3 \\ 4 & \text{mit Wahrscheinlichkeit } 1/3 \end{cases} \\ E[(A - E[A])^2] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\ \text{Var}[A] &= 2 \end{aligned}$$

Andererseits, haben wir für das Spiel B:

$$\begin{aligned} B - E[B] &= \begin{cases} 1001 & \text{mit Wahrscheinlichkeit } 2/3 \\ -2002 & \text{mit Wahrscheinlichkeit } 1/3 \end{cases} \\ (B - E[B])^2 &= \begin{cases} 1.002.001 & \text{mit Wahrscheinlichkeit } 2/3 \\ 4.008.004 & \text{mit Wahrscheinlichkeit } 1/3 \end{cases} \\ E[(B - E[B])^2] &= 1.002.001 \cdot \frac{2}{3} + 4.008.004 \cdot \frac{1}{3} \\ \text{Var}[B] &= 2.004.002 \end{aligned}$$

Damit ist die Varianz im Spiel A nur 2, während sie im Spiel B mehr als zwei Millionen ist! D.h. im Spiel A sollte der Gewinn üblicherweise nah an erwarteten Gewinn von 1 € sein, während er im Spiel B sehr weit von $E[B] = 1$ entfernt liegen kann.

Große Varianz verbindet man oft mit hohem Risiko. So zum Beispiel erwarten wir im Spiel A in 10 Runden einen Gewinn von 10 € zu erzielen, aber auch einen Verlust von 10 € müssen in Kauf nehmen. Im Spiel B können wir auch erwarten, 10 € zu gewinnen, können aber mehr als 20.000 € verlieren!

Die Varianz $E[(X - E[X])^2]$ ist als ein *Quadrat* von der Abweichung vom Erwartungswert definiert. Infolge dieser Quadrierung kann die Varianz sehr weit von der tatsächlichen Abweichung entfernt sein. Zum Beispiel im Spiel B oben ist die Abweichung von $E[B]$ gleich 1001 oder ist gleich -2002 . Aber die Varianz ist ganze kolossale $2.004.002(!)$ Ausserdem ist die Varianz nicht in Euro sondern in "Quadrat-Euro" gemessen. Um diese Effekte zu vermeiden, benutzt man oft anstatt der Varianz die sogenannte *Standardabweichung*, die als

$$\sigma_X := \sqrt{\text{Var}[X]} = \sqrt{E[(X - E[X])^2]}$$

definiert ist. Intuitiv misst σ_X die “erwartete (durchschnittliche) Abweichung” von dem Erwartungswert $E[X]$. Zum Beispiel ist die Standardabweichung für das Spiel B oben gleich

$$\sigma_B = \sqrt{\text{Var}[X]} = \sqrt{2.004.002} \approx 1416.$$

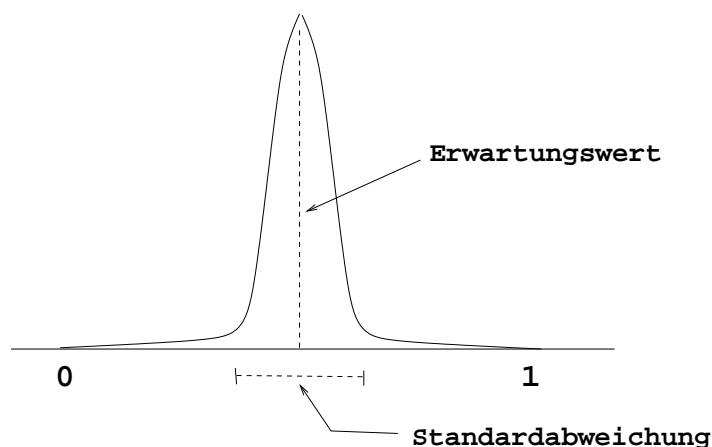


Abbildung 4.5: Die Standardabweichung von der W-Verteilung sagt wie breit ihre “Hauptanteil” ist.

In der Tat weicht die Zufallsvariable B von $E[B]$ entweder um 1001 oder um -2002 ab; also beschreibt σ_B die “durchschnittliche” Abweichung ziemlich gut (siehe Abbildung 4.5).

4.8 Analytische Berechnung von $E[X]$ und $\text{Var}[X]$

Es gibt auch andere allgemeine Methode, die oft hilft, den Erwartungswert $E[X]$ und die Varianz $\text{Var}[X]$ von kompliziert definierten Zufallsvariablen X zu berechnen.

Sei X eine diskrete Zufallsvariable mit der Verteilung $p_k := \Pr\{X = k\}$ für alle k in dem Wertebereich von X . Die *erzeugende Funktion* von X ist definiert durch

$$F_X(x) := \sum p_k x^k$$

wobei die Summe über alle k in dem Wertebereich von X ist. Dann gilt folgendes:

Satz 4.45.

- (a) $E[X] = F'_X(1)$
- (b) $\text{Var}[X] = F''_X(1) + E[X] - E[X]^2$

Beweis. Zu (a): $F'_X(x) = \sum k p_k x^{k-1} \Rightarrow F'_X(1) = \sum k p_k = E[X]$.

Zu (b): $F''_X(x) = \sum k(k-1)p_k x^{k-2} \Rightarrow$

$$F''_X(1) = \sum k(k-1)p_k = \sum k^2 p_k - \sum k p_k = E[X^2] - E[X].$$

Wenn wir also $E[X]$ dazuaddieren und $E[X]^2$ subtrahieren, kommt gerade die Varianz $\text{Var}[X]$ raus. \square

4.9 Eigenschaften von $E[X]$ und $\text{Var}[X]$



Die allerwichtigste Eigenschaft des Erwartungswertes überhaupt ist seine ‘‘Linearitat’’. Diese Eigenschaft des Erwartungswertes ist sehr robust (und deshalb sehr wichtig): sie gilt fur *beliebige* (nicht nur unabhangige) Zufallsvariablen!

Satz 4.46. (Linearitat des Erwartungswertes) Fur beliebigen Zufallsvariablen X, Y gilt:

$$E[X + Y] = E[X] + E[Y].$$

Beweis. Sei a_1, \dots, a_n bzw. b_1, \dots, b_m der Wertebereich von X bzw. Y . Da die Ereignisse $Y = b_j$ fur verschiedene b_j disjunkt sind, gilt nach dem Satz von der totalen Wahrscheinlichkeit

$$\Pr\{X = a_i\} = \sum_{j=1}^m \Pr\{X = a_i, Y = b_j\}$$

und eine analoge Formel fur $\Pr\{Y = b_j\}$. Deshalb gilt:

$$\begin{aligned} E[X + Y] &= \sum_{i=1}^n \sum_{j=1}^m (a_i + b_j) \Pr\{X = a_i, Y = b_j\} \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i \Pr\{X = a_i, Y = b_j\} + \sum_{j=1}^m \sum_{i=1}^n b_j \Pr\{X = a_i, Y = b_j\} \\ &= \sum_{i=1}^n a_i \Pr\{X = a_i\} + \sum_{j=1}^m b_j \Pr\{Y = b_j\} \\ &= E[X] + E[Y]. \end{aligned}$$

\square

Als nachstes schauen wir wie verhalten sich der Erwartungswert und die Varianz, wenn wir die Zufallsvariable durch eine Konstante multiplizieren oder zu einer Zufallsvariable eine Konstante addieren. Dazu bezeichnen wir mit C eine *konstante* Zufallsvariable, die nur einzigen Wert $c \in \mathbb{R}$ annimmt.¹⁰

Lemma 4.47. Sei X eine beliebige Zufallsvariable und C eine konstante Zufallsvariable, die den Wert c annimmt. dann gilt:

- (a) $E[C] = c, \text{Var}[C] = 0$
- (b) $E[cX] = cE[X], \text{Var}[cX] = c^2\text{Var}[X]$
- (c) $E[X + c] = E[X] + c, \text{Var}[X + c] = \text{Var}[X]$

¹⁰Fur diejenigen, die sich unbequem mit dem Begriff ‘‘konstante Variable’’ fuhlen, sei es erinnert, dass eine Zufallsvariable X eigentlich keine ‘‘Variable’’ sondern eine Funktion $X : \Omega \rightarrow \mathbb{R}$ ist.

Beweis. (a) Die Zufallsvariable C nimmt nur einen Wert c mit Wahrscheinlichkeit $\Pr\{C = c\} = 1$. Also gilt $E[C] = c \cdot 1 = c$ und $\text{Var}[C] = E[C^2] - E[C]^2 = c^2 - c^2 = 0$.

Zu (b): Sind a_1, \dots, a_n die Werte von X , so sind ca_1, \dots, ca_n die Werte von cX und

$$\Pr\{cX = ca_i\} = \Pr\{X = a_i\}$$

gilt. Somit erhalten wir

$$E[cX] = \sum_{i=1}^n ca_i \Pr\{X = a_i\} = c \sum_{i=1}^n a_i \Pr\{X = a_i\} = cE[X]$$

und

$$\begin{aligned} \text{Var}[cX] &= E[(cX - E[cX])^2] \\ &= E[c^2X^2 - 2cE[cX] \cdot X + E[cX]^2] \\ &= c^2E[X^2] - 2c^2E[X]^2 + c^2E[X]^2 \\ &= c^2\text{Var}[X]. \end{aligned}$$

Zu (c): Die erste Gleichung $E[X + c] = E[X] + c$ folgt aus der Linearität des Erwartungswertes. Die zweite lassen wir als Übungsaufgabe. \square

Ähnlich kann man die folgende nützliche Formel für die Berechnung der Varianz beweisen.

Satz 4.48. $\text{Var}[X] = E[X^2] - E[X]^2$

Beweis.

$$\begin{aligned} \text{Var}[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2E[X] \cdot X + E[X]^2] \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

\square

► *Beispiel 4.49:* Sei A ein Ereignis und

$$X_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{falls } \omega \notin A \end{cases}$$

sei seine Indikatorvariable. Dann gilt:

$$E[X_A] = \Pr\{A\} \quad \text{und} \quad \text{Var}[X_A] = \Pr\{A\} - \Pr\{A\}^2 = \Pr\{A\} \Pr\{\bar{A}\} \quad (4.5)$$

da $E[X_A] = 1 \cdot \Pr\{A\} + 0 \cdot \Pr\{\bar{A}\}$ und $\Pr\{X_A^2 = 1\} = \Pr\{X_A = 1\}$ gilt.

► **Beispiel 4.50: (Zufällige Teilmengen)** Wir wählen rein zufällig eine m -elementige Teilmenge A aus einer n -elementigen Menge Ω . D.h.

$$\Pr \{ \text{Teilmenge } A \text{ wird ausgewählt} \} = \frac{1}{\binom{n}{m}}.$$

Gegeben sei nun eine Teilmenge $S \subseteq \Omega$. Was kann man über die erwartete Größe $E[|A \cap S|]$ des Schnitts $A \cap S$ sagen?

Jede m -elementige Teilmenge A wird mit Wahrscheinlichkeit $\binom{n}{m}^{-1}$ ausgewählt. Sei $I_{x,A}$ die Indikatorvariable für das Ereignis " $x \in A$ ". Dann wird jedes Element $x \in \Omega$ mit Wahrscheinlichkeit

$$\Pr \{ I_{x,A} = 1 \} = \Pr \{ x \in A \} = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n}$$

in der ausgewählten Teilmenge A liegen.¹¹ Nach der Linearität des Erwartungswertes gilt somit:

$$E[|A \cap S|] = E \left[\sum_{x \in S} I_{x,A} \right] = \sum_{x \in S} E[I_{x,A}] = \sum_{x \in S} \Pr \{ I_{x,A} = 1 \} = \sum_{x \in S} \frac{m}{n} = \frac{m|S|}{n}.$$

⚠ Sind die Zufallsvariablen X und Y nicht unabhängig, so gilt im Allgemeinen die Gleichung $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ nicht! Nimm zum Beispiel $Y = -X$. Dann gilt $\text{Var}[X + Y] = \text{Var}[0] = 0$ aber $\text{Var}[X] + \text{Var}[Y] = \text{Var}[X] + \text{Var}[-X] = 2\text{Var}[X]$. Ein anderes Beispiel: Wir werfen eine Münze und seien X und Y Indikatorvariablen für die Ereignisse "es kommt Kopf" und "es kommt Wappen". Dann ist $\text{Var}[X] = \text{Var}[Y] = 1/4$ aber $\text{Var}[X + Y] = \text{Var}[1] = 0$.

⚠ Im Allgemeinen ist auch die Produktregel $E[X \cdot Y] = E[X] \cdot E[Y]$ falsch! Sei X auf $\{0, 1\}$ gleichwertig verteilt. Dann gilt: $E[X^2] = E[X] = 1/2 \implies E[X]^2 = 1/4 \neq E[X^2]$.

Die Produktregel gilt nur wenn die Zufallsvariablen *unabhängig* sind.

Satz 4.51. Sind X und Y *unabhängige*(!) Zufallsvariablen und a, b beliebige reelle Zahlen, so gilt

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

und

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Beweis. Linearität des Erwartungswertes gibt us: $E[(X + Y)^2] = E[X^2 + 2XY + Y^2] = E[X^2] + 2E[XY] + E[Y^2]$. Wir müssen den Term $E[XY]$ berechnen. Dazu benutzen wir die Unabhängigkeit

¹¹Warum? Da wir genau $\binom{n-1}{m-1}$ m -elementige Teilmengen A mit $x \in A$ haben und es $\binom{n}{m} = \frac{n}{m} \binom{n-1}{m-1}$ gilt.

von X und Y :

$$\begin{aligned}
 E[XY] &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \Pr\{X = a_i, Y = b_j\} \\
 &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \Pr\{X = a_i\} \cdot \Pr\{Y = b_j\} \quad (\text{Unabhängigkeit von } X \text{ und } Y) \\
 &= \left(\sum_{i=1}^n a_i \Pr\{X = a_i\} \right) \cdot \left(\sum_{j=1}^m b_j \Pr\{Y = b_j\} \right) \\
 &= E[X] \cdot E[Y].
 \end{aligned}$$

Somit erhalten wir

$$\begin{aligned}
 \text{Var}[X + Y] &= E[(X + Y)^2] - E[X + Y]^2 \\
 &= (E[X^2] + 2E[XY] + E[Y^2]) - (E[X]^2 + 2E[X] \cdot E[Y] + E[Y]^2) \\
 &= (E[X^2] - E[X]^2) + 2(E[XY] - E[X] \cdot E[Y]) + (E[Y^2] - E[Y]^2) \\
 &= \text{Var}[X] + \text{Var}[Y]
 \end{aligned}$$

□

Ist $X : \Omega \rightarrow S$ mit $S \subset \mathbb{R}$ eine Zufallsvariable und $f : S \rightarrow S$ eine Funktion, so ist $f(X)$ auch eine Zufallsvariable, die jedem Elementarereignis $\omega \in \Omega$ die Zahl $f(X(\omega))$ zuweist. Wie sieht ihr Erwartungswert $E[f(X)]$ aus? Nach der Definition ist

$$E[f(X)] = \sum_{y \in S} y \cdot \Pr\{f(X) = y\}.$$

Man kann aber leicht zeigen, dass auch folgendes gilt.

Lemma 4.52. Ist $X : \Omega \rightarrow S$ eine Zufallsvariable und $f : S \rightarrow S$ eine Funktion, so gilt

$$E[f(X)] = \sum_{x \in S} f(x) \cdot \Pr\{X = x\}$$


Beweis. Sei $y \in S$ ein festes Punkt in dem Wertebereich von f . Dann gilt

$$\begin{aligned}
 \Pr\{f(X) = y\} &= \Pr\{\{\omega \in \Omega : f(X(\omega)) = y\}\} \\
 &= \Pr\{\{\omega \in \Omega : X(\omega) \in f^{-1}(y)\}\} \\
 &= \sum_{x: f(x)=y} \Pr\{\{\omega \in \Omega : X(\omega) = x\}\} \\
 &= \sum_{x: f(x)=y} \Pr\{X = x\}
 \end{aligned}$$

Somit gilt

$$\begin{aligned}
 E[f(X)] &= \sum_y y \cdot \Pr\{f(X) = y\} \\
 &= \sum_y y \cdot \sum_{x:f(x)=y} \Pr\{X = x\} \\
 &= \sum_y \sum_{x:f(x)=y} y \cdot \Pr\{X = x\} \\
 &= \sum_y \sum_{x:f(x)=y} f(x) \cdot \Pr\{X = x\} \\
 &= \sum_x f(x) \cdot \Pr\{X = x\} \quad (\text{da } y \neq y' \Rightarrow f^{-1}(y) \cap f^{-1}(y') = \emptyset)
 \end{aligned}$$

□

 Ist $f(x)$ eine *nicht* lineare Funktion, so gilt $E[f(X)] = f(E[X])$ im Allgemeinen nicht! Das zu “Behaupten” ist ein sehr häufiger Fehler! Ist zum Beispiel $f(x) = x^2$ und X eine Indikatorvariable mit $\Pr\{X = 1\} = p > 0$, dann haben wir $E[f(X)] = E[X] = p$ aber $f(E[X]) = p^2$.

Für allgemeine Funktionen kann man manchmal nur einige Abschätzungen von $E[f(X)]$ mittels $f(E[X])$ finden. So folgt zum Beispiel aus Jensen’s Ungleichung für konvexe Funktionen¹² die folgende Ungleichung:

Lemma 4.53. (Jensen’s Ungleichung) Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable mit endlichem Wertebereich $f(\Omega)$ und $f : \mathbb{R} \rightarrow \mathbb{R}$ sei eine konvexe Funktion. Dann gilt:

$$E[f(X)] \geq f(E[X])$$

Eine Boole’sche Funktionen in n Variablen ist eine Abbildung $f : \{0,1\}^n \rightarrow \{0,1\}$. Eine solche Funktion ist *monoton* (bzw. *anti-monoton*), falls aus¹³ $\mathbf{x} \leq \mathbf{y}$ stets $f(\mathbf{x}) \leq f(\mathbf{y})$ (bzw. $f(\mathbf{x}) \geq f(\mathbf{y})$) folgt.

Lemma 4.54. (Harris’s Ungleichung) Seien $f, g : \{0,1\}^n \rightarrow \{0,1\}$ Boole’sche Funktionen und seien x_1, \dots, x_n unabhängige Bernoulli-Variablen. Sein weiterhin $X = f(x_1, \dots, x_n)$ und $Y = g(x_1, \dots, x_n)$. Sind die Funktionen f und g beide monoton oder beide anti-monoton, so gilt:

$$E[X \cdot Y] \geq E[X] \cdot E[Y].$$

¹²Sind $0 \leq \lambda_i \leq 1$ mit $\sum_{i=1}^r \lambda_i = 1$ und ist f eine konvexe Funktion, so gilt (siehe Satz 1.23):

$$\sum_{i=1}^r \lambda_i f(x_i) \geq f\left(\sum_{i=1}^r \lambda_i x_i\right).$$

¹³ $\mathbf{x} \leq \mathbf{y} \iff x_i \leq y_i$ für alle i .

Beweis. Induktion nach n . Wir werden nur Induktionsbasis $n = 1$ verifizieren. In diesem Fall haben wir $X = f(x)$ und $Y = g(x)$, wobei x eine Bernoulli-Variable mit $\Pr\{x = 1\} = p$ und $\Pr\{x = 0\} = q = 1 - p$ ist. Dann gilt:

$$E[X \cdot Y] - E[X] \cdot E[Y] = f(1)g(1)p + f(0)g(0)q - (f(1)p + f(0)q)(g(1)p + g(0)q)$$

oder äquivalent (nach der Umformung)

$$E[X \cdot Y] - E[X] \cdot E[Y] = pq(f(1) - f(0))(g(1) - g(0)),$$

was nicht negativ sein muss, da $f(1) \geq f(0)$ und $g(1) \geq g(0)$ gilt.

Den Induktionsschritt $n \rightarrow n + 1$ kann man zeigen, indem man die bedingte Erwartungswerte unter der Bedingungen $x_{n+1} = 1$ und $x_{n+1} = 0$ betrachtet. \square

Besteht unser Wahrscheinlichkeitsraum aus allen 0-1 Vektoren der Länge n (d.h. $\Omega = \{0, 1\}^n$), so sagt man, dass ein Ereignis $A \subseteq \Omega$ *monoton* ist, falls aus $\mathbf{x} \in A$ und $\mathbf{x} \leq \mathbf{y}$ stets $\mathbf{y} \in A$ folgt. Und für solche Ereignisse liefert Harris's Ungleichung die folgende interessante untere Schranke.

Korollar 4.55. (Das "Wurzel Trick") Seien A_1, \dots, A_m monotone Ereignisse in $\Omega = \{0, 1\}^n$ mit $\Pr\{A_1\} = \dots = \Pr\{A_m\}$. Dann gilt für jedes $1 \leq i \leq m$:

$$\Pr\{A_i\} \geq 1 - \left(1 - \Pr\left\{\bigcup_{j=1}^m A_j\right\}\right)^{1/m}. \quad (4.6)$$

Beweis.

$$1 - \Pr\left\{\bigcup_{j=1}^m A_j\right\} = \Pr\left\{\bigcap_{j=1}^m \overline{A_j}\right\} \geq \prod_{j=1}^m \Pr\{\overline{A_j}\} = (1 - \Pr\{A_1\})^m,$$

wobei die Ungleichung aus Harris's Ungleichung folgt. \square



Beachte, dass die Monotonität der Ereignisse in Korollar 4.55 sehr wichtig ist. Nehme zum Beispiel eine gleichmäßige Verteilung auf Ω und sei A_1, \dots, A_m eine Zerlegung von Ω in gleichgroße Ereignisse. Dann ist die linke Seite von (4.6) gleich $1/m$, während die rechte Seite gleich 1 ist.

Wenn die Zufallsvariable X nur *natürliche* Zahlen als seine Werte annimmt, gibt es eine alternative (und oft mehr geeignete) Art und Weise den Erwartungswert $E[X]$ zu bestimmen.

Satz 4.56. (Erwartungswert diskreter Zufallsvariablen) Ist $X : \Omega \rightarrow \mathbb{N}$ und $E[X] < \infty$, so gilt

$$E[X] = \sum_{i=0}^{\infty} \Pr\{X > i\}.$$

Beweis. Da X nur Zahlen $0, 1, 2, \dots$ als seine Werte annimmt, gilt

$$\Pr\{X > i\} = \Pr\{X = i + 1\} + \Pr\{X = i + 2\} + \Pr\{X = i + 3\} + \dots$$

und deshalb

$$\begin{aligned} \sum_{i=0}^{\infty} \Pr\{X > i\} &= \Pr\{X > 0\} + \Pr\{X > 1\} + \Pr\{X > 2\} + \dots \\ &= \underbrace{\Pr\{X = 1\}}_{\Pr\{X > 0\}} + \underbrace{\Pr\{X = 2\} + \Pr\{X = 3\} + \dots}_{\Pr\{X > 1\}} \\ &\quad + \underbrace{\Pr\{X = 2\} + \Pr\{X = 3\} + \dots}_{\Pr\{X > 2\}} \\ &= \Pr\{X = 1\} + 2 \cdot \Pr\{X = 2\} + 3 \cdot \Pr\{X = 3\} + \dots \\ &= \sum_{i=1}^{\infty} i \cdot \Pr\{X = i\} \\ &= E[X]. \end{aligned}$$

□

► *Beispiel 4.57:* Wir haben ein Kommunikations-Netz, in dem viele Pakete verschickt werden sollen. Angenommen, der Versand eines Pakets kann sich nur mit Wahrscheinlichkeit $1/k$ um k oder mehr Sekunden verzögern. Das klingt gut; es ist nur 1% Chance, dass der Versand eines Pakets um 100 oder mehr Sekunden verzögert wird. Aber wenn wir die Situation genauer betrachten, ist das Netz gar nicht so gut. Tatsächlich ist die erwartete Verzögerung eines Pakets unendlich! Sei X die Verzögerung eines Pakets. Dann gilt nach Satz 4.56:

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} \Pr\{X > i\} \geq \sum_{i=0}^{\infty} \frac{1}{i+1} \\ &= \infty \text{ (unendliche harmonische Reihe divergiert, siehe Beispiel 3.32).} \end{aligned}$$

In manchen Situationen haben wir mit *unendlichen* Summen (Reihen) der Zufallsvariablen zu tun. Hier können wir nicht mehr ohne weiteres die Linearität des Erwartungswertes benutzen, da diese Eigenschaft nur für *endlichen* Summen gilt. Obwohl gilt die Gleichung

$$E \left[\sum_{i=0}^n X_i \right] = \sum_{i=0}^n E[X_i]$$

für alle n , können wir daraus nicht (ohne weiteres) schliessen, dass auch

$$E \left[\lim_{n \rightarrow \infty} \sum_{i=0}^n X_i \right] = \lim_{n \rightarrow \infty} \sum_{i=0}^n E[X_i]$$

gelten muss. Dazu brauchen wir zusätzlich, dass der Grenzwert $\lim_{n \rightarrow \infty} \sum_{i=0}^n E[|X_i|]$ existiert.

Satz 4.58. (Unendliche Linearität des Erwartungswertes) Seien X_0, X_1, \dots Zufallsvariablen, so dass die Reihe $\sum_{i=0}^{\infty} E[|X_i|]$ konvergiert. Dann gilt

$$E \left[\sum_{i=0}^{\infty} X_i \right] = \sum_{i=0}^{\infty} E[X_i]$$

Beweis. Sei $Y := \sum_{i=0}^{\infty} X_i$. Wir lassen es als Übungsaufgabe, zu verifizieren, dass (wegen der Konvergenz von $\sum_{i=0}^{\infty} E[|X_i|]$) alle Summen in den folgenden Ableitungen absolut konvergent sind, was ihre Vertauschung erlaubt (siehe Satz 3.49).

$$\begin{aligned} \sum_{i=0}^{\infty} E[X_i] &= \sum_{i=0}^{\infty} \sum_{\omega \in \Omega} X_i(\omega) \cdot \Pr\{\omega\} = \sum_{\omega \in \Omega} \sum_{i=0}^{\infty} X_i(\omega) \cdot \Pr\{\omega\} \\ &= \sum_{\omega \in \Omega} Y(\omega) \cdot \Pr\{\omega\} = E[Y] = E \left[\sum_{i=0}^{\infty} X_i \right]. \end{aligned}$$

□

► **Beispiel 4.59: (Casino)** Wir nehmen in einem Casino an einem Spiel mit Gewinnwahrscheinlichkeit $p = 1/2$ teil. Zum Beispiel wirft man eine faire Münze, deren Seiten mit rot und blau gefärbt sind, und wir gewinnen, falls rot kommt. Wir können einen beliebigen Betrag einsetzen. Geht das Spiel zu unseren Gunsten aus, erhalten wir den Einsatz zurück und zusätzlich denselben Betrag aus der Bank. Endet das Spiel ungünstig, verfällt unser Einsatz. Wir betrachten die folgende Strategie:

In jedem Schritt *verdoppeln* wir unseren Einsatz bis erstmals die Farbe rot kommt.

Wir wollen den erwarteten Gewinn dieser Strategie bestimmen. Sei K unser erster Einsatz. Sei X_i das im i -ten Schritt gewonnene Kapital. Dann ist

$$Y = \sum_{i=0}^{\infty} X_i$$

das gewonnene Gesamtkapital. Da in jedem Schritt die Gewinnchance $p = 1/2$ ist, werden wir im i -ten Schritt ($i = 0, 1, 2, \dots$) mit gleicher Wahrscheinlichkeit entweder $K \cdot 2^i$ Euro gewinnen oder denselben Betrag verlieren, d.h. der Gewinn im i -ten Schritt ist entweder positiv ($+K2^i$) oder negativ ($-K2^i$). Also ist der erwartete Gewinn $E[X_i] = 0$ für alle $i = 0, 1, \dots$. Daraus “folgt”:

$$E[Y] = E \left[\sum_{i=1}^{\infty} X_i \right] = \sum_{i=1}^{\infty} E[X_i] = \sum_{i=1}^{\infty} 0 = 0 \quad !?$$

Aber in jedem Schritt ist die Gewinnwahrscheinlichkeit > 0 . Also soll es *mit Sicherheit* irgendwann mal die Münze auf rot landen. D.h. wir sollten mit Wahrscheinlichkeit 1 mindestens K Euro gewinnen. Was war dann in unserer Argumentation falsch? Um diese Frage zu beantworten, müssen wir das Problem genauer betrachten.

Die Wahrscheinlichkeitsraum Ω besteht aus Strings $B^{k-1}R$, $k = 1, 2, \dots$, wobei B für “blau” steht und R für “rot”. Sei X_i das gewonnene Kapital im i -ten Schritt (wir nehmen o.B.d.A. an,

dass $X_i(\omega) = 0$ für Elementarereignisse $\omega = B^k R$ mit $k < i - 1$). Auf jedem Elementarereignis $\omega = B^{k-1} R$ nimmt X_i einen der drei möglichen Werte an:

$$X_i(B^{k-1} R) = \begin{cases} 0 & \text{falls } k < i \\ K \cdot 2^i & \text{falls } k = i \\ -K \cdot 2^i & \text{falls } k > i \end{cases}$$

Unsere Argumentation, dass $E[X_i] = 0$ für alle i , war richtig. Falsch war aber zu sagen, dass

$$E \left[\sum_{i=1}^{\infty} X_i \right] = \sum_{i=1}^{\infty} E[X_i]$$

gilt, da die Reihe $\sum_{i=1}^{\infty} E[|X_i|]$ nicht konvergent ist: $|X_i| = K \cdot 2^i$ mit Wahrscheinlichkeit 2^{-i} und deshalb

$$\sum_{i=1}^{\infty} E[|X_i|] = \sum_{i=1}^{\infty} K \cdot 2^i \cdot 2^{-i} = \sum_{i=1}^{\infty} K = \infty.$$

Die richtige Argumentation ist folgende: Auf *jedem* Elementarereignis $\omega = B^{k-1} R$ ist der Wert von $Y = \sum_{i=1}^{\infty} X_i$ gleich

$$Y(\omega) = K \cdot 2^k - K \cdot \underbrace{\sum_{i=0}^{k-1} 2^i}_{2^k - 1} = K$$

woraus trivialerweise folgt,¹⁴ dass der Erwartungswert von Y auch gleich K sein muss.

4.10 Verteilungen diskreter Zufallsvariablen

In diesem Abschnitt betrachten wir einige wichtige Verteilungen der Zufallsvariablen, die in vielen Anwendungen immer wieder vorkommen, und berechnen ihr Erwartungswert und ihre Varianz.

Bernoulli-Verteilung

Das ist die einfachste Verteilung überhaupt: Jede solche Zufallsvariable X hat nur zwei mögliche Werte 0 und 1; $p = \Pr\{X = 1\}$ heißt dann die Erfolgswahrscheinlichkeit und die Wahrscheinlichkeit eines Misserfolges ist $q = 1 - p$. Beispiel: Einmaliges Werfen einer Münze, wobei der Ausgang ‘‘Wappen’’ mit Wahrscheinlichkeit p kommt. Das entsprechende Zufallsexperiment nennt man *Bernoulli-Experiment*. Den Erwartungswert wie auch die Varianz einer solchen Zufallsvariable sind leicht zu berechnen:

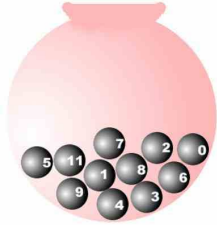
$$\begin{aligned} E[X] &= 1 \cdot \Pr\{X = 1\} + 0 \cdot \Pr\{X = 0\} = p, \\ \text{Var}[X] &= E[X^2] - E[X]^2 = p - p^2 = p(1 - p). \end{aligned}$$

¹⁴Da Y eine konstante Funktion mit $Y(\omega) = K$ für alle ω ist.

Binomialverteilung $B(n, p)$

Eine solche Zufallsvariable S_n gibt uns die *Anzahl* der Erfolge in n unabhängig voneinander ausgeführten Bernoulli-Experimenten X_1, \dots, X_n mit Erfolgswahrscheinlichkeit $\Pr\{X_i = 1\} = p$ für alle $i = 1, \dots, n$. D.h.

$$S_n = X_1 + X_2 + \dots + X_n.$$



Man kann das Experiment auch als ein Urnenmodell vorstellen: Man hat eine Urne mit r roten und s schwarzen Kugeln (also $N = r + s$ Kugeln insgesamt) und zieht n Kugeln rein zufällig eine nacheinander *mit Zurücklegen*. Erfolg ist dann eine rote Kugel und Erfolgswahrscheinlichkeit ist dann $p = r/N$.

Bei einer unabhängigen Wiederholung des Bernoulli-Experiments multiplizieren sich die Wahrscheinlichkeiten, die Wahrscheinlichkeit für *genau* k Erfolge (und $n - k$ Misserfolge) ist also $p^k q^{n-k}$ mit $q = 1 - p$. Da es $\binom{n}{k}$ Möglichkeiten gibt, k Erfolge in einer Versuchsserie der Länge n unterzubringen, ist die Wahrscheinlichkeit, dass X den Wert k annimmt, gerade $\binom{n}{k} p^k q^{n-k}$. Damit gilt:

$$\Pr\{S_n = k\} = \binom{n}{k} p^k q^{n-k}$$

Nach dem binomischen Lehrsatz gilt

$$\sum_{k=0}^n \Pr\{S_n = k\} = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1,$$

wie dies auch sein sollte. Da $E[X_i] = p$, $\text{Var}[X_i] = pq$ und die Zufallsvariablen X_1, \dots, X_n unabhängig sind, kann man den Erwartungswert wie auch die Varianz leicht berechnen:

$$\begin{aligned} E[S_n] &= p + p + \dots + p = np \\ \text{Var}[S_n] &= pq + pq + \dots + pq = npq. \end{aligned}$$

► *Beispiel 4.60*: Wir verteilen m Bonbons an n Kinder. Dazu werfen wir wiederholt ein Bonbon in eine Gruppe aus n Kindern. Der Versuch eines Kindes, das geworfene Bonbon zu fangen, ist ein Bernoulli-Versuch. Jedes der Kinder fängt mit gleicher Wahrscheinlichkeit ein Bonbon. Die Erfolgswahrscheinlichkeit jedes Kindes ist dann $p = \frac{1}{n}$, und die Wahrscheinlichkeit eines Misserfolges ist $q = \frac{n-1}{n} = 1 - \frac{1}{n}$.

Wie groß ist nun die Wahrscheinlichkeit, dass ein bestimmtes Kind von m geworfenen Bonbons genau k fängt? Sei X die Zufallsvariable, deren Wert die Anzahl der von diesem Kind gefangenen Bonbons beschreibt. Dann ist $\Pr\{X = k\}$ durch die Binomial-Verteilung $B(m, p)$ bestimmt. Die erwartete Anzahl gefangener Bonbons ist also $np = \frac{m}{n}$, und die Varianz ist $np(1-p) = \frac{m}{n} \left(1 - \frac{1}{n}\right)$.

Wir wissen bereits, dass für jedes $\alpha \in (0, 1)$

$$\binom{n}{\alpha n} \sim \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot 2^{n \cdot H(\alpha)} \quad (4.7)$$

mit $H(\alpha) = -(\alpha \log_2 \alpha + (1 - \alpha) \log_2(1 - \alpha))$ gilt.

Somit gilt für die Verteilung $\Pr\{S_n = k\}$ einer binomial $B(n, 1/2)$ -verteilten Zufallsvariable

$$\Pr\{S_n = \alpha n\} = \binom{n}{\alpha n} \cdot 2^{-n} \sim \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot 2^{-n(1-H(\alpha))} \quad (4.8)$$

Dann ist für $\alpha = 1/2$

$$\Pr\{S_n = n/2\} \sim \sqrt{\frac{2}{\pi n}}.$$

Die asymptotische Gleichung (4.8) sagt uns, dass eine binomial-verteilte Zufallsvariable tatsächlich nie ihrem Erwartungswert gleich sein wird. Wenn wir zum Beispiel $n = 100$ mal eine faire Münzen werfen würden, dann ist die Wahrscheinlichkeit, dass *genau* 50 mal das Wappen kommt ungefähr 8%. Aber wenn wir die Wahrscheinlichkeit, dass genau 25 mal das Wappen kommt, betrachten (d.h. wenn wir $\alpha = 1/4$ nehmen), dann ist diese Wahrscheinlichkeit nur $\Pr\{S_n = n/4\} \sim 1,913 \cdot 10^{-7}$, was sogar kleiner als 1 zu 5 Millionen ist!

Im Allgemeinen liefert uns die asymptotische Gleichung (4.8) folgendes: Ist $\alpha \neq 1/2$, so ist die Differenz $\epsilon := 1 - H(\alpha)$ positiv, und damit gilt

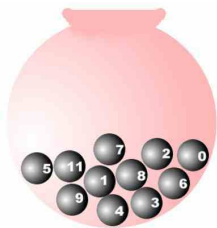
$$\Pr\{S_n = \alpha n\} = O(2^{-\epsilon n}).$$

Somit haben wir die folgende Merkregel:

Für $\alpha \neq 1/2$ ist $\Pr\{S_n = \alpha n\}$ *exponentiell klein* im Vergleich mit n

Geometrische Verteilung $G(p)$

Wir wiederholen das Bernoulli-Experiment X_1, X_2, \dots mit Erfolgswahrscheinlichkeit $p > 0$ oftmals und wollen die Anzahl der Versuche bis zu erstem Erfolg bestimmen.



Man kann auch dieses Experiment auch als ein Urnenmodell vorstellen: Man hat eine Urne mit r roten und s schwarzen Kugeln (also $N = r + s$ Kugeln insgesamt) und zieht n Kugeln rein zufällig eine nacheinander *mit Zurücklegen*. Erfolg ist dann eine rote Kugel und Erfolgswahrscheinlichkeit ist dann $p = r/N$.

Die entsprechende Zufallsvariable

$$X := \min\{i : X_i = 1\}$$

heißt dann *geometrisch verteilt*, da ihre Verteilung eine geometrische Folge ist:

$$\Pr\{X = i\} = q^{i-1}p \quad \text{mit } q = 1 - p$$

Wenn wir diese Wahrscheinlichkeiten aufsummieren, so erhalten wir

$$\sum_{i=1}^{\infty} q^{i-1}p = p \cdot \sum_{i=0}^{\infty} q^i = \frac{p}{1-q} = \frac{p}{1-(1-p)} = 1,$$

wie auch es sein sollte. Um $E[X]$ zu berechnen, können wir Satz 4.56 (über diskretwertige Zufallsvariablen) anwenden. Dazu beobachten wir:

$$\Pr\{X > i\} = \sum_{k=i+1}^{\infty} q^{k-1} p = 1 - p(1 + q + q^2 + \dots + q^{i-1}) = 1 - p \cdot \frac{1 - q^i}{1 - q} = (1 - p)^i.$$

Nach Satz 4.56 gilt:

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} \Pr\{X > i\} \\ &= \sum_{i=0}^{\infty} (1 - p)^i = \frac{1}{1 - (1 - p)} \\ &= \frac{1}{p}. \end{aligned}$$

Wir können $E[X]$ auch mittels der Methode der erzeugenden Funktionen leicht berechnen. Diese Methode ist gut, da sie erlaubt auch "nebenbei" die Varianz $\text{Var}[X]$ zu berechnen.

Die erzeugende Funktion von X ist

$$F(x) = \sum_{i=1}^{\infty} q^{i-1} p x^i = p x \cdot \sum_{i=0}^{\infty} q^i x^i = \frac{p x}{1 - q x}. \quad (\text{geometrische Reihe})$$

Die erste und die zweite Ableitung von $F(x)$ sind

$$\begin{aligned} F'(x) &= \frac{(1 - qx)p + pxq}{(1 - qx)^2} = \frac{p}{(1 - qx)^2} \\ F''(x) &= \frac{2pq}{(1 - qx)^3}. \end{aligned}$$

Setzen wir nun $x = 1$, so erhalten wir (nach Satz 4.45)

$$E[X] = F'(1) = \frac{p}{(1 - q)^2} = \frac{1}{p}$$

und

$$\text{Var}[X] = F''(1) + E[X] - E[X]^2 = \frac{2pq}{p^3} + \frac{1}{p} - \frac{1}{p^2} = \frac{1 - p}{p^2}.$$

► *Beispiel 4.61*: Wir verteilen wiederum Bonbons an n Kinder. Dazu werfen wir wiederholt ein Bonbon in eine Gruppe aus n Kindern. Der Versuch eines Kindes, das geworfene Bonbon zu fangen, ist ein Bernoulli-Versuch. Jedes der Kinder fängt mit gleicher Wahrscheinlichkeit $p = \frac{1}{n}$ ein Bonbon.

Wie viele Bonbons müssen geworfen werden, bis ein bestimmtes Kind ein Bonbon gefangen hat? Sei X die Zufallsvariable, deren Wert die Nummer des Versuchs an, bei dem das Kind erstmals ein Bonbon fängt. Dann ist $\Pr\{X = k\} = (1 - p)^{k-1} \cdot p$ durch die geometrische Verteilung bestimmt. Also ist die erwartete Anzahl der Versuche ein Bonbon zu fangen bis es erstmals klappt genau die Anzahl $E[X] = \frac{1}{p} = n$ der Kinder, die sich auch auf ein Bonbon warten.

Poisson-Verteilung $P(\lambda)$

In vielen Anwendungen (Physik, Biologie, usw.) taucht eine Wahrscheinlichkeitsverteilung – die sogenannte “Poisson-Verteilung” – sehr oft auf. Diese Verteilungen sind nichts anderes als die Grenzwerte der binomialen Verteilungen $B(n, p)$, wenn $n \rightarrow \infty$, $p \rightarrow 0$ und $np = \lambda$ konstant bleibt.

Zum Beispiel wollen wir die Ankunft von Paketen auf einem Internetrouter modellieren. Wir wissen, dass im Durchschnitt der Router λ Pakete pro Sekunde abfertigt. Falls wir diesen Durchschnittswert wissen, wie können wir die tatsächliche Ankunft der Pakete in einer Sekunde modellieren? Eine Möglichkeit ist, die Sekunde in sehr kurze Intervalle der Länge $\delta > 0$ (mit $\delta < \lambda$) aufzuteilen; damit haben wir eine große Anzahl $n = 1/\delta$ der Intervalle. Dann nehmen wir an, dass ein Paket in einem Intervall mit Wahrscheinlichkeit $p = \lambda\delta$ ankommt (dies ergibt die richtige durchschnittliche Anzahl $np = (1/\delta)(\lambda\delta) = \lambda$ der Pakete pro ganze Sekunde). In diesem Model ist die Anzahl X der Intervalle, in denen ein Paket ankommen wird, als binomiale Zufallsvariable verteilt:¹⁵

$$\Pr\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} = \binom{1/\delta}{k} (\lambda\delta)^k (1-\lambda\delta)^{1/\delta-k}.$$

Nur lassen wir δ beliebig klein sein ($\delta \rightarrow 0$), halten aber k fest. Dann gilt:

$$\begin{aligned} \Pr\{X = k\} &= \binom{1/\delta}{k} (\lambda\delta)^k (1-\lambda\delta)^{1/\delta-k} = \binom{1/\delta}{k} (\lambda\delta)^k (1-\lambda\delta)^{(1-\delta k)/\delta} \\ &\approx \frac{(1/\delta)^k}{k!} (\lambda\delta)^k (1-\lambda\delta)^{1/\delta} = \frac{\lambda^k}{k!} (1-\lambda\delta)^{1/\delta} \approx \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

Die resultierende Verteilung

$$\Pr\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

ist als *Poisson-Verteilung* bekannt. Die Tatsache, dass das wirklich eine Wahrscheinlichkeitsverteilung ist, folgt aus der Taylorformel $e = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$ für die Euler-Zahl e :

$$\sum_{k=0}^{\infty} \Pr\{X = k\} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Was den Erwartungswert $E[X]$ betrifft, ist er gleich λ (wie es sein sollte):

$$E[X] = \sum_{k=0}^{\infty} k \cdot \Pr\{X = k\} = e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

► *Beispiel 4.62* : Wir verteilen wiederum m Bonbons an n Kinder. Dazu werfen wir wiederholt einen Bonbon in eine Gruppe aus n Kindern. Der Versuch eines Kindes, das geworfene Bonbon zu fangen, ist ein Bernoulli-Versuch. Jedes der Kinder fängt mit gleicher Wahrscheinlichkeit $p = 1/n$ ein Bonbon.

Mit welcher Wahrscheinlichkeit p_k wird ein bestimmtes Kind *genau* k Bonbon fangen? Da es hier sich um eine binomial verteilte mit Parametern $B(m, 1/n)$ Zufallsvariable handelt, ist

$$p_k = \binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} = \frac{1}{k!} \frac{m(m-1) \cdots (m-k+1)}{n^k} \left(1 - \frac{1}{n}\right)^{m-k}.$$

¹⁵Beachte, dass das nicht genau die Anzahl der Ankünfte der Pakete entspricht, da mehr als ein Paket in einem Zeitintervall ankommen kann. Aber wenn wir die Intervalle wirklich *sehr kurz* wählen werden, wird in einem Intervall nur ein Paket kommen können.

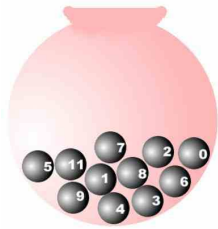
Wenn m und n viel größer als k sind, dann haben wir

$$p_k \approx \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{mit } \lambda = \frac{m}{n}.$$

Also ist die Anzahl der Bonbons, die ein bestimmtes Kind am Ende haben wird, Poisson $P(\lambda)$ mit $\lambda = m/n$ verteilt.

Hypergeometrische Verteilung

Die hypergeometrische Verteilung ist für die Stichprobentheorie von besonderer Bedeutung.¹⁶ Diese Verteilung ergibt die Anzahl der Erfolge bei n -fachem Wiederholen eines 0-1 Experiments *ohne Zurücklegen*. D.h. die Erfolgswahrscheinlichkeit in jedem Schritt kann von früheren Ausgängen des Experiments abhängen.



Man kann das Experiment auch als ein Urnenmodell vorstellen: Man hat eine Urne mit r roten und s schwarzen Kugeln (also $N = r + s$ Kugeln insgesamt) und zieht n Kugeln rein zufällig eine nacheinander *ohne Zurücklegen*. Erfolg ist dann eine rote Kugel. Diesmal hängt aber die Erfolgswahrscheinlichkeit von der Anzahl der bereits gezogenen roten Kugeln ab!

Wir betrachten die Zufallsvariable

$X =$ Anzahl der roten Kugeln in der Stichprobe.

Was ist die Wahrscheinlichkeit, dass X den Wert k annimmt? Dazu müssen k rote und $n - k$ schwarze Kugeln ausgewählt werden, was (ohne Berücksichtigung der Reihenfolge) auf $\binom{r}{k}$ bzw. $\binom{s}{n-k} = \binom{N-r}{n-k}$ Weisen möglich ist. Insgesamt gibt es $\binom{N}{n}$ Stichproben, die gesuchte Wahrscheinlichkeit ist daher

$$\Pr\{X = k\} = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}}. \quad (4.9)$$

Unsere Ableitung der hypergeometrischen Verteilung ergibt als Nebenresultat die kombinatorische Identität (für jedes $0 \leq r \leq N$)

$$\binom{N}{n} = \sum_{k=0}^n \binom{r}{k} \binom{N-r}{n-k} \quad (4.10)$$

denn die Wahrscheinlichkeiten in (4.9) summieren sich zu 1 auf. Wir nutzen sie zur Berechnung des

¹⁶Die hypergeometrische Verteilung kommt zum Beispiel in der Qualitätskontrolle zur Anwendung. Will man die Güte einer Lieferung durch eine Stichprobe überprüfen, so müssen sich die Beteiligten darauf einigen, wieviele fehlerhafte Stücke X die Stichprobe enthalten darf. Der Verkäufer wird darauf achten, dass eine Lieferung mit einem geringen Anteil von Ausschuss mit hoher Wahrscheinlichkeit akzeptiert wird. Der Käufer hat das Interesse, dass eine Lieferung von schlechter Qualität die Kontrolle mit nur geringer Wahrscheinlichkeit passiert. Diese unterschiedlichen Interessen werden sich nur dann vereinbaren lassen, wenn die Stichprobengröße groß genug gewählt ist. Die Wahrscheinlichkeiten werden unter der Annahme bestimmt, dass X hypergeometrisch verteilt ist.

Erwartungswertes einer hypergeometrisch verteilten Zufallsvariable X :¹⁷

$$\sum_{k=0}^n k \cdot \binom{r}{k} \binom{N-r}{n-k} \stackrel{(*)}{=} r \sum_{k=1}^n \binom{r-1}{k-1} \binom{N-r}{n-k} \stackrel{(**)}{=} r \binom{N-1}{n-1} = \frac{nr}{N} \binom{N}{n},$$

also

$$E[X] = \sum_{k=0}^n k \cdot \Pr\{X = k\} = \frac{nr}{N} = np \quad \text{mit } p = \frac{r}{N}.$$

Auf den Erwartungswert hat es also keinen Einfluss, ob man eine Stichprobe mit oder ohne Zurücklegen zieht, er ist in beiden Fällen gleich np . Die Varianz ist (wir verzichten aus dem Beweis)

$$\text{Var}[X] = n \frac{r}{N} \frac{N-r}{N} \frac{N-n}{N-1} = npq \cdot \frac{N-n}{N-1}.$$

D.h. bis auf einem ‘‘Korrekturs-Faktor’’ $\frac{N-n}{N-1}$, der für $N \rightarrow \infty$ und festem n gegen 1 strebt, ist die Varianz dieselbe wie die für binomial $B(n, p)$ -verteilte Zufallsvariable mit $p = r/N$.

Sind r und s groß im Vergleich zu n , so nähert sich die hypergeometrische Verteilung der Binomialverteilung an, denn (mit $l = n - k$)

$$\frac{\binom{r}{k} \binom{s}{l}}{\binom{N}{n}} \approx \frac{\frac{r^k}{k!} \cdot \frac{s^l}{l!}}{\frac{N^n}{n!}} = \binom{n}{k} p^k q^{n-k}, \quad \text{mit } p = \frac{r}{N}.$$

▷ *Beispiel 4.63* : Eine Urne enthält 4 Kugeln: 2 rote und 2 schwarze. Wir ziehen (ohne Zurücklegen) eine Stichprobe aus 2 Kugeln. Sei X die Anzahl der roten Kugeln in der Stichprobe. Hier haben wir also mit einer hypergeometrischen Verteilung mit Parametern $N = 4$ und $r = s = n = 2$ zu tun. Es gilt also für jedes $k = 0, 1, 2$

$$\Pr\{X = k\} = \frac{\binom{2}{k} \binom{2}{2-k}}{\binom{4}{2}} = \frac{1}{6} \binom{2}{k} \binom{2}{2-k}$$

D.h. $\Pr\{X = 0\} = 1/6$, $\Pr\{X = 1\} = 2/3$ und $\Pr\{X = 2\} = 1/6$. Das kann man auch aus der entsprechenden Ziehungs-Diagramm (wie im Beispiel 4.31) entnehmen.

4.11 Abweichung vom Erwartungswert

Bis jetzt haben wir auf den Erwartungswert fokussiert, da er dem ‘‘Durchschnittswert’’ entspricht. Was aber das eigentlich bedeutet? Der Erwartungswert $E[X]$ ist nur eine (speziell definierte) Zahl und

¹⁷Hier benutzt $(*)$ die Identität $\binom{r}{k} = \frac{r}{k} \binom{r-1}{k-1}$ und $(**)$ die Cauchy-Vandermonde Identität:

$$\binom{x+y}{z} = \sum_{i=0}^z \binom{x}{i} \binom{y}{z-i}$$

Beweis. In einer Stadt wohnen x Frauen und y Männer, und die Einwohner haben Lust so viele Clubs wie möglich zu bilden. Die einzige Einschränkung ist, dass jeder Club genau z Teilnehmer haben muss. Dann ist $\binom{x+y}{z}$ die Anzahl aller Clubs, und $\binom{x}{i} \binom{y}{z-i}$ die Anzahl aller Clubs mit i Frauen und $z-i$ Männer. □

als solche sagt uns (bis jetzt) überhaupt nichts. Mehr noch: Die Erwartungswert muss nicht mal im Wertebereich der Zufallsvariable liegen! Zum Beispiel, ist X eine gleichmäßig verteilte Zufallsvariable, die die Werte in $\{0, 1, 9, 10\}$ annimmt, dann ist

$$E[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} + 10 \cdot \frac{1}{4} = 5,$$

die Zahl die zum keinem der Werte 0, 1, 9 oder 10 nah liegt!



Erwarte nicht immer das Erwartete!

Was uns wirklich interessiert ist die Frage, *mit welcher Wahrscheinlichkeit wird die Zufallsvariable nahe an ihrem Erwartungswert liegen?*

Glücklicherweise haben wir ein paar mächtigen Instrumente, um diese Wahrscheinlichkeit zu bestimmen. Dazu gehören die Ungleichungen von Markov, Tschebyschev und Chernoff, die wir jetzt kennenlernen werden.

4.11.1 Markov-Ungleichung



Markov-Ungleichung

Sei $X : \Omega \rightarrow \mathbb{R}_+$ eine nicht-negative Zufallsvariable. Dann gilt für alle $k > 0$:

$$\Pr\{X \geq k\} \leq \frac{E[X]}{k}.$$

Oder äquivalent, für alle $\lambda > 0$ gilt

$$\Pr\{X \geq \lambda \cdot E[X]\} \leq \frac{1}{\lambda}.$$

Beweis.

$$E[X] = \sum_x x \cdot \Pr\{X = x\} \geq \sum_{x \geq k} k \cdot \Pr\{X = x\} = k \cdot \Pr\{X \geq k\}.$$

□



Warum muss die Zufallsvariable X nicht negativ sein? Sei $X \in \{-10, 10\}$ mit $\Pr\{X = -10\} = \Pr\{X = 10\} = 1/2$. Dann ist

$$E[X] = -10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 0.$$

Wir wollen nun die Wahrscheinlichkeit $\Pr\{X \geq 5\}$ ausrechnen. Wenn wir Markov's-Ungleichung "anwenden" kommt

$$\Pr\{X \geq 5\} \leq \frac{E[X]}{5} = \frac{0}{5} = 0$$

raus. Aber das ist doch falsch! Es ist offensichtlich, dass $X \geq 5$ mit Wahrscheinlichkeit $1/2$ gilt (da $X = 10$ mit dieser Wahrscheinlichkeit gilt). Nichtsdestotrotz kann man auch in diesem Fall Markov's-Ungleichung anwenden, aber für eine *modifizierte* Zufallsvariable. Setze nämlich $Y := X + 10$. Das ist bereits eine nicht-negative Zufallsvariable mit $E[Y] = E[X + 10] = E[X] + 10 = 10$, und Markov's-Ungleichung ergibt $\Pr\{Y \geq 15\} \leq 10/15 = 2/3$. Da aber $Y \geq 15 \iff X \geq 5$, haben wir die Abschätzung $\Pr\{X \geq 5\} \leq 2/3$ erhalten.

▷ **Beispiel 4.64 : (Klausuren)** Ich nehme den Stapel Ihrer Klausuren, mische ihn, und verteile wieder die Klausuren an Sie. Jeder bekommt genau eine Klausur und muss sie korrigieren. Sei X die Anzahl von Studenten, die ihre eigene Klausur zurück bekommen. Wie sieht $E[X]$ aus?

Wenn wir direkt die Wahrscheinlichkeiten $\Pr\{X = i\}$ ausrechnen wollten, wäre es nicht so einfach. Wir können aber X als die Summe $X = X_1 + X_2 + \dots + X_n$ von Indikatorvariablen darstellen, wobei $X_i = 1$ wenn der i -te Student seine eigene Klausur bekommt, und $X_i = 0$ sonst. Da jede X_i eine Indikatorvariable ist, gilt $E[X_i] = \Pr\{X_i = 1\}$ (siehe (4.5)). Wie groß ist die Wahrscheinlichkeit $\Pr\{X_i = 1\}$? Jede Verteilung der Klausuren kann man als eine Permutation $f : [n] \rightarrow [n]$ darstellen; der i -te Student seine eigene Klausur bekommt genau dann, wenn $f(i) = i$ gilt. Damit ist für jedes i

$$E[X_i] = \Pr\{X_i = 1\} = \frac{\text{Anzahl der Permutationen } f \text{ mit } f(i) = 1}{\text{Anzahl aller Permutationen } f} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

und die Linearität des Erwartungswertes gibt uns die Antwort:

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = 1.$$

Nun wollen wir die Varianz $\text{Var}[X]$ berechnen. Obwohl X die Summe von Indikatorvariablen ist, können wir *nicht* den Satz 4.51 benutzen, da die Indikatorvariablen X_i nicht *unabhängig* sind: Nach dem Multiplikationssatz für Wahrscheinlichkeiten gilt für $i \neq j$

$$\begin{aligned} \Pr\{X_i = 1, X_j = 1\} &= \Pr\{X_i = 1\} \cdot \Pr\{X_j = 1 \mid X_i = 1\} = \frac{1}{n} \cdot \frac{1}{n-1} \\ &\neq \frac{1}{n} \cdot \frac{1}{n} = \Pr\{X_i = 1\} \cdot \Pr\{X_j = 1\}. \end{aligned}$$

Wir müssen also die Varianz $\text{Var}[X] = E[X^2] - E[X]^2$ direkt ausrechnen. Wir wissen bereits, dass

$$E[X_i \cdot X_j] = \Pr\{X_i = 1, X_j = 1\} = \frac{1}{n(n-1)}$$

gilt. Somit gilt auch

$$\begin{aligned} E[X^2] &= \sum_{i=1}^n E[X_i^2] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E[X_i X_j] \\ &= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} \\ &= 2 \end{aligned}$$

und

$$\text{Var}[X] = E[X^2] - E[X]^2 = 2 - 1 = 1.$$

Die nächste Frage: Wie groß ist die *Wahrscheinlichkeit*, dass mindestens k Studenten ihre eigene Klausur zur Korrektur zurückbekommen werden? Nach Markov's-Ungleichung gilt:

$$\Pr\{X \geq k\} \leq \frac{E[X]}{k} = \frac{1}{k}.$$

Somit gibt es zum Beispiel höchstens 20% Chance, dass 5 Studenten ihre eigene Klausuren bekommen.



Beachte, dass in diesem Beispiel weder der Erwartungswert noch die Varianz von der Anzahl n der Studenten abhängt!

4.11.2 Tschebyschev-Ungleichung

Die Markov-Ungleichung sagt nur, dass mit großer Wahrscheinlichkeit der eigentliche Wert von X *nicht viel größer* als der Erwartungswert $E[X]$ sein wird. Sie sagt aber nicht mit welcher Wahrscheinlichkeit X *nah* an $E[X]$ sein wird – es kann gut sein, dass der eigentliche Wert von X *viel kleiner* als $E[X]$ wird. Es macht deshalb Sinn, die Wahrscheinlichkeit $\Pr\{|X - E[X]| \geq k\}$ zu betrachten.

Da für jede Zufallsvariable Y der Betrag $|Y|$ und damit auch ihre Exponenten $|Y|^r$ nicht negativ sind, können wir Markov-Ungleichung anwenden. Damit gilt für alle $\epsilon > 0$ und alle $r \geq 1$:

$$\Pr\{|Y| \geq k\} = \Pr\{|Y|^r \geq k^r\} \leq \frac{E[|Y|^r]}{k^r}$$

Wenn wir die Zufallsvariable $Y := X - E[X]$ betrachten, ergibt dies (mit $r = 2$)

$$\Pr\{|X - E[X]| \geq k\} \leq \frac{E[(X - E[X])^2]}{k^2}.$$

D.h. die Wahrscheinlichkeit, dass die Zufallsvariable X vom ihrem Erwartungswert $E[X]$ um mehr als $\pm k$ abweicht, kann nicht größer als $1/k^2$ mal die Konstante(!) $E[(X - E[X])^2]$ sein. Diese Konstante haben wir bereits früher kennengelernt und als die *Varianz* $\text{Var}[X]$ von X bezeichnet:

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Damit haben wir die folgende Ungleichung bewiesen.



Tschebyschev-Ungleichung

Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable mit $E[X^2] < \infty$. Dann gilt für alle $k > 0$:

$$\Pr\{|X - E[X]| \geq k\} \leq \frac{\text{Var}[X]}{k^2}.$$

- ▷ **Beispiel 4.65 : (Optimalität der Tschebyschev-Ungleichung)** Dieses Beispiel soll zeigen, dass Tschebyschev-Ungleichung auch *optimal* ist. Sei $a \in \mathbb{R}$, $a \geq 1$ und betrachte die Zufallsvariable X , deren Verteilung folgender Maßen definiert ist:

$$\Pr\{X = -a\} = \frac{1}{2a^2}, \quad \Pr\{X = 0\} = 1 - \frac{1}{a^2} \quad \text{und} \quad \Pr\{X = a\} = \frac{1}{2a^2}$$

Dann gilt

$$E[X] = \frac{-a}{2a^2} + 0 + \frac{a}{2a^2} = 0$$

und

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] = \frac{a^2}{2a^2} + 0 + \frac{a^2}{2a^2} = 1.$$

Setzt man $k = a$, so erhält man unter Beachtung der vorgegebenen Verteilung

$$\Pr\{|X - E[X]| \geq a\} = \Pr\{|X| \geq a\} = \Pr\{X \neq 0\} = \frac{1}{a^2}.$$

Andererseits ist auch der rechter Term $\text{Var}[X]/a^2$ gleich $1/a^2$. D.h. in diesem Fall wird die durch die Tschebyschev'sche Ungleichung gegebene obere Schranke auch tatsächlich angenommen.

- ▷ **Beispiel 4.66 : (Klausuren - Fortsetzung)** Ich nehme den Stapel Ihrer Klausuren, mische ihn, und verteile wieder die Klausuren an Sie. Jeder bekommt genau eine Klausur und muss sie korrigieren. Sei X die Anzahl von Studenten, die ihre eigene Klausur zurück bekommen. Dann gilt $E[X] = \text{Var}[X] = 1$ (siehe Beispiel 4.64). In diesem Beispiel haben wir Markov-Ungleichung benutzt, um

$$\Pr\{X \geq k\} \leq \frac{E[X]}{k} = \frac{1}{k}$$

zu zeigen. Tschebyschev-Ungleichung liefert:

$$\begin{aligned} \Pr\{X \geq k\} &= \Pr\{X - E[X] \geq k - E[X]\} && \text{(ziehe } E[X] \text{ von beiden Seiten ab)} \\ &= \Pr\{X - E[X] \geq k - 1\} && \text{(setze } E[X] = 1 \text{ ein)} \\ &\leq \frac{\text{Var}[X]}{(k-1)^2} = \frac{1}{(k-1)^2} \end{aligned}$$

Und diese obere Schranke ist sogar *quadratisch* besser, als die $\Pr\{X \geq k\} \leq 1/k$, die wir vorher aus der Markov-Ungleichung abgeleitet haben.

- ▷ **Beispiel 4.67 :** Ist $X = X_1 + \dots + X_n$ die Summe von n unabhängigen Bernoulli Variablen je mit Erfolgswahrscheinlichkeit p , so gilt: $E[X] = np$ und $\text{Var}[X] = np(1-p)$ (siehe Abschnitt 4.10). Die Tschebyschev-Ungleichung ergibt dann

$$\Pr\{|X - np| \geq k\} \leq \frac{np(1-p)}{k^2} \leq \frac{n}{4k^2},$$

da $p(1-p) \leq 1/4$ für alle $0 \leq p \leq 1$ gilt: Ist $p = 1/2 + c$ für ein $-1/2 \leq c \leq 1/2$, so gilt $p(1-p) = (1/2 + c)(1/2 - c) = 1/4 - c^2 \leq 1/4$.

Werfen wir zum Beispiel eine faire 0-1 Münze n mal, dann können wir $n/2$ Einsen erwarten. Die Wahrscheinlichkeit, dass die tatsächliche Anzahl der Einsen um mehr als $\lambda \sqrt{n}$ von $n/2$ abweichen wird, ist damit höchstens $1/4\lambda^2$.

▷ **Beispiel 4.68 : (Faire oder unfaire Münze?)** Wir haben zwei Münzen. Wir wissen, dass nur eine der Münzen fair ist. Die andere ist präpariert, so dass die Wahrscheinlichkeit für den Ausgang „Wappen“ gleich $3/4$ ist. Rein äußerlich aber sehen die beiden Münzen völlig gleich aus.

Wir wählen rein zufällig eine der Münzen und werfen sie mehrmals. Wieviel mal müssen wir die Münze werfen, um mit der Wahrscheinlichkeit 0.95 zu bestimmen, welche der Münzen gewählt war?

Sei X die Anzahl der Ausgänge „Wappen“ nach n Wurfen. Um den Typ der Münze festzustellen, wählen wir einen Schwellenwert t zwischen $1/2$ und $3/4$, und schauen, ob die Anteil X/n der Ausgänge „Wappen“ kleiner oder grösser als dieser Schwellenwert t ist. Als natürlichen Schwellenwert wählen wir

$$t = 5/8$$

(genau in der Mitte von $1/2$ und $3/4$). Wir müssen die Münze so lange werfen bis

$$\Pr\{X/n > t\} \leq 0.05 \quad \text{falls die Münze fair ist}$$

$$\Pr\{X/n \leq t\} \leq 0.05 \quad \text{falls die Münze präpariert ist}$$

Dann geben wir die Antwort

$$\text{Antwort} = \begin{cases} \text{fair} & \text{falls } X/n \leq t \\ \text{präpariert} & \text{falls } X/n > t. \end{cases}$$

War die Münze tatsächlich fair, so erhalten wir die richtige Antwort mit Wahrscheinlichkeit

$$\Pr\{\text{Antwort richtig}\} = \Pr\{X/n \leq t\} = 1 - \Pr\{X/n > t\} \geq 1 - 0.05 = 0.95.$$

War die Münze präpariert, so erhalten wir die richtige Antwort mit Wahrscheinlichkeit

$$\Pr\{\text{Antwort richtig}\} = \Pr\{X/n > t\} = 1 - \Pr\{X/n \leq t\} \geq 1 - 0.05 = 0.95.$$

Um die Wahrscheinlichkeit $\Pr\{X/n > 5/8\}$ (und die Wahrscheinlichkeit $\Pr\{X/n < 5/8\}$) abzuschätzen, wenden wir die Tschebyschev-Ungleichung an. Dazu müssen wir das Ereignis „ $X/n > 5/8$ “ in der Form „ $|X - E[X]| \geq k$ “ darstellen.¹⁸ Beachte, dass die Zufallsvariable X binomialverteilt zum Parameter $B(n, p)$ ist, wobei $p = 1/2$ falls die Münze fair ist, und $p = 3/4$ falls die Münze präpariert ist. Wir wissen bereits (siehe Abschnitt 4.10), dass für solcher Zufallsvariablen $E[X] = np$ und $\text{Var}[X] = np(1-p)$ gilt. Also, falls die Münze *fair* ist, dann gilt:¹⁹

$$\begin{aligned} \Pr\left\{\frac{X}{n} > \frac{5}{8}\right\} &= \Pr\left\{\frac{X}{n} - \frac{1}{2} > \frac{5}{8} - \frac{1}{2}\right\} = \Pr\left\{X - \frac{n}{2} > \frac{n}{8}\right\} \\ &= \Pr\left\{X - E[X] > \frac{n}{8}\right\} \leq \Pr\left\{|X - E[X]| > \frac{n}{8}\right\} \\ &\leq \frac{\text{Var}[X]}{(n/8)^2} = \frac{n/4}{n^2/64} = \frac{16}{n} \quad (\text{Tschebyschev-Ungleichung}) \end{aligned}$$

¹⁸Dies ist ein der wichtigsten Schritte, wenn man Tschebyschev-Ungleichung anwenden will!

¹⁹Denn dann $E[X] = n/2$ und $\text{Var}[X] = n/4$.

Ist die Münze präpariert, so gilt: ²⁰

$$\begin{aligned} \Pr \left\{ \frac{X}{n} \leq \frac{5}{8} \right\} &= \Pr \left\{ \frac{3}{4} - \frac{X}{n} \geq \frac{3}{4} - \frac{5}{8} \right\} = \Pr \left\{ \frac{3n}{4} - X \geq \frac{n}{8} \right\} \\ &= \Pr \left\{ E[X] - X \geq \frac{n}{8} \right\} \leq \Pr \left\{ |X - E[X]| \geq \frac{n}{8} \right\} \\ &\leq \frac{\text{Var}[X]}{(n/8)^2} = \frac{3n/16}{n^2/64} = \frac{12}{n} \quad (\text{Tschebyschev-Ungleichung}) \end{aligned}$$

Wir müssen also die Anzahl der Würfe n so wählen, dass die beiden Zahlen $16/n$ und $12/n$ nicht größer als 0,05 sind. Es reicht also $n = 320$ zu nehmen.

Das Gesetz der großen Zahlen besagt, dass sich die relative Häufigkeit eines Zufallsergebnisses immer weiter an die theoretische Wahrscheinlichkeit für dieses Ergebnis (Erwartungswert) annähert, je häufiger das Zufallsexperiment durchgeführt wird.

Wiederholt man ein Zufallsexperiment X mit Erfolgswahrscheinlichkeit p , so stabilisiert sich die relative Häufigkeit $H = X/n$ der Erfolge mit wachsender Versuchszahl n bei p . Allgemeiner gilt, dass das *arithmetische Mittel* von n identisch verteilten, unabhängigen Zufallsvariablen mit wachsendem n gegen den Erwartungswert strebt. In diesem Fall spricht man von einem ‘‘Gesetz der großen Zahlen’’. Eine einfache Version dieses Gesetzes ist das folgende Resultat.

Satz 4.69. (Schwachere Gesetz der großen Zahlen) Sei X eine reellwertige Zufallsvariable mit endlichem Erwartungswert und endlicher Varianz. Seien X_1, \dots, X_n unabhängige Kopien von X . Dann gilt für alle $\epsilon > 0$

$$\Pr \left\{ \left| \frac{X_1 + \dots + X_n}{n} - E[X] \right| \geq \epsilon \right\} \leq \frac{\text{Var}[X]}{n \cdot \epsilon^2}.$$

Insbesondere strebt diese Wahrscheinlichkeit gegen 0 für $n \rightarrow \infty$.

Beweis. Sei $Y := \frac{1}{n} (X_1 + \dots + X_n)$. Dann gilt:

$$E[Y] = E \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \cdot \sum_{i=1}^n E[X_i] = \frac{1}{n} \cdot nE[X] = E[X]$$

und

$$\text{Var}[Y] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{n \cdot \text{Var}[X_n]}{n^2} = \frac{\text{Var}[X]}{n}. \quad (\text{Unabhängigkeit von } X_i \text{'s})$$

Aus der Tschebyschev-Ungleichung folgt für jedes $\epsilon > 0$

$$\Pr \{ |Y - E[X]| \geq \epsilon \} \leq \frac{\text{Var}[Y]}{\epsilon^2} = \frac{\text{Var}[X]}{n \cdot \epsilon^2}.$$

□

²⁰Denn dann $E[X] = 3n/4$ und $\text{Var}[X] = n(3/4)(1/4) = 3n/16$.

► **Beispiel 4.70: (Wahlvorhersage)** Gesucht ist der Umfang n von Stichproben für die Vorhersage des Stimmenanteil p einer Partei mit *Prämisse*

$$\Pr \{ \text{Fehler für Vorhersage von } p \text{ maximal } 0,01 \} \geq 0,95.$$

Seien X_1, \dots, X_n *unabhängige* Zufallsvariablen mit den Werten 0 (gegen Partei), 1 (für Partei) mit Wahrscheinlichkeiten $1 - p$ bzw. p . Dann ist $Y = \frac{1}{n} (X_1 + \dots + X_n)$ die Zufallsvariable, die den Stimmenanteil in n Stichproben ergibt. Es gilt $E[Y] = \frac{1}{n} \cdot np = p$ und

$$\text{Var}[Y] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{\text{Var}[X_1]}{n} = \frac{p(1-p)}{n} \leq \frac{1}{n}.$$

Nun wollen wir ein n bestimmen, so dass

$$\Pr \{ |Y - p| \geq \epsilon \} \leq \delta$$

mit $\epsilon = 0,01 = 10^{-2}$ und $\delta = 0,05$ gilt. Aus Tschebyschev-Ungleichung folgt

$$\Pr \{ |Y - p| \geq \epsilon \} \leq \frac{\text{Var}[Y]}{\epsilon^2} = \frac{1}{4n\epsilon^2}.$$

Da diese Wahrscheinlichkeit nicht größer als $\delta = 0,05$ sein darf, folgt die Bedingung

$$n \geq \frac{1}{4\epsilon^2\delta} = \frac{1}{4(10^{-2})^2(5/100)} = 10000 \cdot \frac{100}{4 \cdot 5} = 50000.$$

Es reicht also $n = 50.000$ zu nehmen, um die Prämisse zu erfüllen.

Das *starke* Gesetz der großen Zahlen besagt, dass für eine unendliche Folge von Zufallsvariablen X_1, X_2, X_3, \dots , die unabhängig und identisch verteilt sind sowie den selben Erwartungswert μ haben, gilt:

$$\Pr \left\{ \lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mu \right\} = 1$$

d.h. die repräsentative Stichprobe konvergiert fast sicher gegen μ .

4.11.3 Chernoff-Ungleichungen

In üblicher (kontinuierlicher) Stochastik ist das sogenannte “Central Limit Theorem” von großer Bedeutung. In der Informatik aber benutzt man stattdessen die Chernoff-Ungleichungen. Diese Ungleichungen sind Spezialfälle der Markoff Ungleichung, angewandt auf Summen von unabhängigen 0-1 Zufallsvariablen.

Die sogenannte “Murphy-Regel” (Murphy’s Rule) besagt: Wenn man erwartet, dass einige Sachen schief gehen könnten, dann wird mit Sicherheit irgendetwas schief gehen. Der folgende Satz formalisiert die Regel.

Satz 4.71. Seien A_1, A_2, \dots, A_n unabhängige Ereignisse, und X sei die Anzahl der Ereignisse die tatsächlich vorkommen. Die Wahrscheinlichkeit, dass keines der Ereignisse vorkommen wird, ist $\leq e^{-E[X]}$, d.h.

$$\Pr \{ X = 0 \} \leq e^{-E[X]}$$

Beweis. Sei X_i die Indikatorvariable für das i -te Ereignis, $i = 1, \dots, n$. Dann ist $X = X_1 + X_2 + \dots + X_n$. Es gilt:

$$\begin{aligned}
 \Pr\{X = 0\} &= \Pr\{\overline{A_1 \cup A_2 \cup \dots \cup A_n}\} && \text{(Definition von } X\text{)} \\
 &= \Pr\{\overline{A_1} \cap \overline{A_2} \cap \dots \cap \overline{A_n}\} && \text{(De Morgan-Regel)} \\
 &= \prod_{i=1}^n \Pr\{\overline{A_i}\} && \text{(Unabhängigkeit von } A_i\text{'s)} \\
 &= \prod_{i=1}^n (1 - \Pr\{A_i\}) \\
 &\leq \prod_{i=1}^n e^{-\Pr\{A_i\}} && \text{(da } 1 + x \leq e^x \text{ für alle } x \in \mathbb{R} \text{ gilt)} \\
 &= e^{-\sum_{i=1}^n \Pr\{A_i\}} && \text{(Algebra von Exponenten)} \\
 &= e^{-\sum_{i=1}^n E[X_i]} && \text{(Erwartungswert der Indikatorvariablen)} \\
 &= e^{-E[X]} && \text{(Linearität des Erwartungswertes)}
 \end{aligned}$$

□

► *Beispiel 4.72*: Wir konstruieren einen Mikroprozessor und wissen, dass jeder Transistor nur mit Wahrscheinlichkeit 10^{-5} beschädigt sein kann. Das klingt gut. Aber heutzutage enthält ein Mikroprozessor ca. 10^6 (und sogar mehr) Transistoren. Deshalb ist die erwartete Anzahl der beschädigten Transistoren in unserem Mikrochip gleich 10. Laut Satz 4.71 wird der Mikroprozessor nur mit Wahrscheinlichkeit e^{-10} (kleiner als 1 zu 3 Millionen!) defekt-frei sein!

Der Satz oben sagt Folgendes: Wenn $E[X]$ die erwartete Anzahl der tatsächlich vorkommenden Ereignisse aus A_1, A_2, \dots, A_n ist, dann wird mit Wahrscheinlichkeit $\Pr\{X \geq 1\} \geq 1 - e^{-E[X]}$ mindestens eines der Ereignisse vorkommen. Nun betrachten wir den allgemeinen Fall: Wie groß ist die Wahrscheinlichkeit $\Pr\{X \geq k\}$? Die Antwort ist mit folgendem Satz gegeben:

Satz 4.73. (Chernoff's Ungleichungen) Seien X_1, \dots, X_n unabhängige Bernoulli-Variablen mit $\Pr\{X_i = 1\} = p_i$ und sei $X = X_1 + \dots + X_n$. Sei $\mu = E[X] = p_1 + \dots + p_n$. Dann gilt:

1. für jedes $\delta > 0$

$$\Pr\{X \geq (1 + \delta)\mu\} \leq F(\mu, \delta) \quad \text{mit} \quad F(\mu, \delta) := \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu \quad (4.11)$$

2. für jedes $\delta > 2e - 1$

$$\Pr\{X \geq (1 + \delta)\mu\} \leq 2^{-(1+\delta)\mu} \quad (4.12)$$

3. für jedes $0 < \delta < 1$

$$\Pr\{X \geq (1 + \delta)\mu\} \leq e^{-\mu\delta^2/3} \quad (4.13)$$

4. für jedes $R > \mu$

$$\Pr\{X \geq R\} \leq e^{(1 - \ln \frac{R}{\mu}) \cdot R - \mu} \quad (4.14)$$

5. für jedes $0 < \delta < 1$

$$\Pr\{X < (1 - \delta)\mu\} \leq e^{-\mu\delta^2/2} \quad (4.15)$$

Beweis. Wir beweisen nur die ersten vier Ungleichungen (der Beweis der letzten ist analog). Markov's Ungleichung gibt uns für jedes $a > 0$

$$\Pr\{X \geq t\} = \Pr\{e^{a \cdot X} > e^{a \cdot t}\} \leq e^{-a \cdot t} \cdot E[e^{a \cdot X}],$$

wobei (wegen der (Un)abhängigkeit von X_i 's)

$$E[e^{a \cdot X}] = E[e^{a \cdot \sum_{i=1}^n X_i}] = E\left[\prod_{i=1}^n e^{a \cdot X_i}\right] = \prod_{i=1}^n E[e^{a \cdot X_i}]$$

gilt. Wenn wir die Parameter t und a auf

$$\begin{aligned} t &:= (1 + \delta)\mu \\ a &:= \ln(1 + \delta) \end{aligned}$$

setzen und die Ungleichung

$$E[(1 + \delta)^{X_i}] = p_i(1 + \delta) + (1 - p_i) = 1 + \delta \cdot p_i \leq e^{\delta \cdot p_i},$$

benutzen, dann bekommen wir

$$\Pr\{X \geq (1 + \delta) \cdot \mu\} \leq (1 + \delta)^{-(1+\delta) \cdot \mu} \cdot \prod_{i=1}^n E[(1 + \delta)^{X_i}],$$

wobei

$$\prod_{i=1}^n E[(1 + \delta)^{X_i}] \leq \prod_{i=1}^n e^{\delta \cdot p_i} = e^{\sum_{i=1}^n \delta \cdot p_i} = e^{\delta \mu}$$

ist, und die Ungleichung (4.11) folgt.

Die Ungleichung (4.12) folgt unmittelbar aus (4.11), da für $\delta > 2e - 1$ gilt

$$(1 + \delta)^{1+\delta} \geq (2e)^{1+\delta} \geq 2^{1+\delta} e^\delta.$$

Um (4.13) zu zeigen, reicht es zu zeigen, dass für alle $0 < x < 1$ die Ungleichung

$$F(\mu, x) = e^{\mu(x - (1+x) \ln(1+x))} \leq e^{-\mu x^2/3}$$

oder äquivalent, dass für alle $0 < x < 1$ die Ungleichung $f(x) \leq 0$ mit

$$f(x) := x - (1+x) \ln(1+x) + x^2/3$$

gilt. Dafür berechnen wir die Ableitungen von $f(x)$:

$$\begin{aligned} f'(x) &= 1 - \frac{1+x}{1+x} - \ln(1+x) + \frac{2}{3}x = -\ln(1+x) + \frac{2}{3}x \\ f''(x) &= -\frac{1}{1+x} + \frac{2}{3}. \end{aligned}$$

Wir sehen, dass $f''(x) < 0$ für $0 \leq x < 1/2$ und $f''(x) > 0$ für $x > 1/2$. Das bedeutet, dass $f'(x)$ im Intervall $[0, 1]$ zuerst fällt und dann wächst. Da $f'(0) = 0$ und $f'(1) < 0$, muss $f'(x) \leq 0$ in ganzem Intervall $[0, 1]$ gelten, woraus (wegen $f(0) = 0$) $f(x) \leq 0$ für alle x in diesem Intervall folgt.

Um (4.14) zu zeigen, reicht es $\delta := R/\mu - 1$ in (4.11) zu nehmen:²¹

$$\begin{aligned} \Pr\{X \geq R\} = \Pr\{X \geq (1+\delta)\mu\} &\leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu \\ &= \frac{e^{R-\mu}}{(R/\mu)^R} \\ &= e^{(1-\ln \frac{R}{\mu}) \cdot R - \mu}. \end{aligned}$$

□

Eine schöne Eigenschaft der Chernoff-Ungleichung ist die Tatsache, dass wir weder die Wahrscheinlichkeiten von einzelnen Ereignissen A_i noch ihre Anzahl n wissen brauchen: Es reicht nur zu wissen, dass die Ereignisse unabhängig sind und wie die erwartete Anzahl $\mu = E[X]$ der tatsächlich vorkommenden Ereignisse aussieht.



Der größte Nachteil dieser Ungleichungen ist aber die Tatsache, dass die Ereignisse *unabhängig* sein müssen!

► **Beispiel 4.74: (Experimentelle Bemessung des Erwartungswertes)** Wir nehmen an, dass ein Zufallsexperiment X vorliegt, dessen Erwartungswert $\mu = E[X]$ wir experimentell messen möchten. (Wir wissen die eigentliche Verteilung von X nicht.) Unser „Toleranzintervall“ ist $T = [\mu - \delta, \mu + \delta]$; je kleiner δ ist desto besser. Die Tschebyschev Ungleichung liefert das Ergebnis

$$\Pr\{X \notin T\} = \Pr\{|X - E[X]| > \delta\} \leq \epsilon := \frac{\text{Var}[X]}{\delta^2}.$$

²¹Beachte, dass $\delta > 0$ gilt, da $R > \mu$ gefördert wird.

Wenn aber δ klein ist oder wenn unser Experiment instabil ist (d.h. die Varianz von X groß ist), dann sind wir verloren. Wirklich? Nein, wir können „boosten“, d.h. das Experiment n mal wiederholen. Sei X_i das Ergebnis des i -ten Experiments. Wir betrachten den arithmetischen Mittel

$$Y = \frac{X_1 + X_2 + \dots + X_n}{n}$$

und beachten, dass $E[Y] = E[X]$ und $\text{Var}[Y] = \text{Var}[X]/n$ gilt (siehe Beweis von Satz 4.69). Die Tschebyschev Ungleichung liefert jetzt die Abschätzung

$$\Pr\{Y \notin T\} = \Pr\{|Y - E[Y]| > \delta\} \leq \frac{\text{Var}[Y]}{\delta^2} = \frac{\text{Var}[X]}{n \cdot \delta^2} = \frac{1}{n} \cdot \epsilon$$

und große Abweichungen vom Erwartungswert sind n mal unwahrscheinlicher geworden.

Die Fehlerwahrscheinlichkeit fällt also proportional zu n . Wir können aber diese Wahrscheinlichkeit noch scheller gegen 0 treiben, wenn wir anstatt des arithmetischen Mittels den *Median* nehmen.²² Sei also

$$M = \text{der Median von } X_1, \dots, X_n.$$

Wir wissen bereits, dass X mit Wahrscheinlichkeit mindestens $p := 1 - \epsilon = 1 - \text{Var}[X]/\delta^2$ in T liegt. Wenn aber der Median M außerhalb des Toleranzintervalls liegt, dann liegen mindestens $n/2$ Einzelschätzungen außerhalb. Wenn also Z_i die Indikatorvariable für das Ereignis „ $X_i \notin T$ “ ist, dann impliziert $M \notin T$, dass die Summe $Z = \sum_{i=1}^n Z_i$ mindestens $n/2$ sein muss. Aber nur $E[Z] \leq (1-p) \cdot n = \epsilon \cdot n$ außerhalb liegende Einzelschätzungen zu erwarten sind. Wenn wir also $\beta = (1-2 \cdot \epsilon)/(2\epsilon)$ wählen, dann gilt $(1+\beta) \cdot \epsilon \cdot n = n/2$. Wir wenden die Chernoff Ungleichung an und erhalten

$$\begin{aligned} \Pr\{M \notin T\} &\leq \Pr\{Z \geq n/2\} \leq \Pr\{Z \geq (1+\beta)E[Z]\} \\ &\leq e^{-\epsilon \cdot n \cdot \beta^2/3} \leq e^{-\Omega(\epsilon \cdot n)} \end{aligned}$$

und die Fehlerwahrscheinlichkeit in diesem Fall fällt sogar negativ exponentiell.

Im Satz 4.73 ist die Unabhängigkeit der Zufallsvariablen sehr wichtig. Trotzdem kann man eine ähnliche Schranke auf für Summen von *abhängigen* Zufallsvariablen zeigen.

Satz 4.75. Seien Y_1, \dots, Y_n (nicht unbedingt unabhängige!) binomiell verteilte Zufallsvariablen, und $Y = \sum_{i=1}^n Y_i$. Sei $\mu = E[Y]$. Dann gilt für jedes $B > \mu$

$$\Pr\{Y \geq B\} \leq n \cdot e^{(1 - \ln \frac{B}{n\mu}) \cdot \frac{B}{n}}$$

Beweis. Aus $Y \geq B$ folgt $Y_i \geq B/n$ für mindestens ein i . Wir können also (4.14) benutzen und erhalten

$$\Pr\{Y \geq B\} \leq n \cdot \max_i \Pr\{Y_i \geq B/n\} \leq n \cdot e^{(1 - \ln \frac{B}{n\mu}) \cdot \frac{B}{n}}.$$

□

²²Der Median einer Menge S von $|S| = n$ Zahlen (n ungerade) ist die „ $n/2$ -größte Zahl“ $x \in S$. D.h. es muss $|\{y \in S : y \leq x\}| = |\{z \in S : x \leq z\}|$ gelten. Ist z.B. $S = \{1, 3, 8, 10, 1000\}$, so ist $x = 8$ der Median.

Die oben erwähnten Chernoff-Ungleichungen betrachten nur Bernoulli-Variablen, die nur die Werte 0 oder 1 annehmen können. Zum Schluss geben wir (ohne Beweis) noch eine allgemeinere Version der Chernoff-Ungleichung an.

Satz 4.76. Seien X_1, \dots, X_n unabhängige Zufallsvariablen mit $0 \leq X_i \leq 1$. Sei $X = \sum_{i=1}^n X_i$. Dann gilt:

$$\Pr \{X - E[X] \geq c \sqrt{n}\} \leq e^{-c^2/2}.$$

▷ *Beispiel 4.77: (Job Scheduling)* Wir wollen n Jobs auf m gleich schnellen Prozessoren aufteilen. Die Länge (=Abfertigungszeit) des i -ten Jobs ist irgendeine Zahl L_i im Intervall $[0, 1]$. Wir wollen Jobs so verteilen, dass keiner der Prozessoren viel länger als die Durchschnittsbelastung

$$L = \frac{1}{m} \sum_{i=1}^n L_i$$

aller Prozessoren belastet wird. Eine optimale Verteilung zu finden ist sehr schwer. Das Problem ist noch schwieriger, wenn wir die Joblängen L_i im Voraus nicht kennen.

Stattdessen wenden wir die folgende einfachste “randomisierte” Strategie an: Für jeden Job i wählen wir rein zufällig einen Prozessor j und weisen i dem Prozessor j zu. Es wird sich herausstellen, dass diese “dumme Affenstrategie” eigentlich nicht so schlecht ist, auch wenn wir weder die Anzahl n der Jobs noch ihre Laufzeiten kennen!

Zuerst betrachten wir einen beliebigen (aber festen) Prozessor $j \in \{1, \dots, m\}$. Für diesen Prozessor sei X_i die Zeit, die der Prozessor braucht, um den i -ten Job abzufertigen, d.h.

$$X_i = \begin{cases} L_i & \text{falls der } i\text{-te Job dem Prozessor } j \text{ zugewiesen war} \\ 0 & \text{sonst.} \end{cases}$$

Die gesamte Laufzeit des Prozessors j ist also $X = \sum_{i=1}^n X_i$. Da jeder der n Jobs dem Prozessor j mit gleicher Wahrscheinlichkeit $1/m$ zugewiesen wird, ist die erwartete Laufzeit des Prozessors genau

$$E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \frac{1}{m} \cdot L_i = L.$$

Nach Satz 4.76 haben wir

$$\Pr \{X \geq L + c \sqrt{n}\} \leq e^{-c^2/2}$$

Damit wissen wir, dass *jeder einzelne* Prozessor nur mit Wahrscheinlichkeit $\leq e^{-c^2/2}$ länger als $L + c \sqrt{n}$ beschäftigt sein wird. Da wir insgesamt nur m Prozessoren haben, ist die Wahrscheinlichkeit, dass *mindestens* ein Prozessor länger als $L + c \sqrt{n}$ belästigt sein wird, nicht größer als $m \cdot e^{-c^2/2}$. Also gilt:

$$\Pr \{ \text{Kein Prozessor wird länger als } L + c \sqrt{n} \text{ laufen} \} \geq 1 - m \cdot e^{-c^2/2}$$

Für Summen von unabhängigen ± 1 -wertigen Zufallsvariablen hat Chernoff-Ungleichung die folgende Form:

Satz 4.78. Seien X_1, \dots, X_n unabhängige ± 1 -wertige Zufallsvariablen mit $\Pr\{X_i = -1\} = \Pr\{X_i = +1\} = 1/2$. Dann gilt für jedes $\lambda > 0$:

$$\Pr\{|X_1 + \dots + X_n| \geq \lambda\} \leq 2e^{-\lambda^2/2n}$$

Welche der drei Ungleichungen (Markov, Tschebyschev oder Chernoff) ist besser? Natürlich die Chernoff-Ungleichung, da sie eine *exponentiell* kleine obere Schranke gibt. Man muss aber beachten, dass diese Ungleichung nur für Zufallsvariablen X gilt, die Summen *unabhängiger Bernoulli-Variablen* sind, während es für Tschebyschev reicht, dass X^2 einen endlichen Erwartungswert hat. Schließlich, reicht es für Markov, dass X nicht negativ ist. Zusammengefasst:

$$\Pr\{X \geq (1 + \epsilon)E[X]\} \leq \begin{cases} \frac{1}{1 + \epsilon} & \text{Markov, wenn } X \geq 0 \\ \frac{\text{Var}[X]}{\epsilon^2 E[X]^2} & \text{Tschebyschev, wenn } E[X^2] < \infty \\ e^{-\epsilon^2 E[X]/3} & \text{Chernoff, wenn } X = X_1 + \dots + X_n, \\ & X_i \text{ Bernoulli und unabhängig} \end{cases}$$

4.12 Das Urnenmodell – Hashing*

Das sogenannte “Wortebuchsproblem” in der Informatik ist folgende. Wir haben eine sehr große Menge (das Universum) von möglichen Dateien (z.B. alle mögliche Namen). Wir wollen eine *beliebige* (aber relativ kleine) m -elementige Teilmengen $S \subseteq U$ so abspeichern zu können, dass die Frage “ist $x \in S$ ” für beliebiges $x \in U$ schnell beantwortet läßt. Das “Hashing”-Verfahren nimmt ein Array mit n ($n \geq m$) Zellen, wählt eine zufällige Hashfunktion $h : U \rightarrow \{1, \dots, n\}$ und speichert jedes $x \in S$ in der Zelle $h(x)$. Da $|U|$ viel größer als Anzahl n der Zellen ist, wird es bestimmt Zellen i geben, in deren sehr viele (mindestens $|U|/n$) Elementen des Universums abgebildet werden. Dann wäre eine *feste* Hashfunktion h für Teilmengen $S \subseteq h^{-1}(i)$ nutzlos. Um das zu vermeiden, wählt man deshalb die Hashfunktion h *zufällig* aus.

Dieses Verfahren—wie auch viele andere stochastische Prozesse—lassen sich gut in einem sogenannten “Urnenmodell” darstellen.

Wir haben m Kugeln und n Urnen, und werfen jede Kugel zufällig und unabhängig in diese Urnen. Jede Kugel kann mit gleicher Wahrscheinlichkeit $1/n$ in jede der n Urnen landen. Also

$$\begin{aligned} m &= \text{Anzahl der Kugeln} \\ n &= \text{Anzahl der Urnen} \\ \frac{1}{n} &= \Pr\{\text{eine Kugel flieg in eine bestimmte Urne}\} \end{aligned}$$

Man kann dann verschiedene Fragen stellen. Zum Beispiel, wie viele Kugeln werden (im Durchschnitt) in einer bestimmten Urne landen, wie viele Urnen (im Durchschnitt) werden leer bleiben, usw. Mit unseren jetztigen Kenntnissen können wir solche Fragen ziemlich leicht beantworten.

Erwartete Anzahl der Kugeln in einer Urne Sei

$$X = \text{Anzahl der Kugel in der } \textit{ersten} \text{ Urne.}$$

Dann ist $X = X_1 + \dots + X_m$ wobei X_i die Indikatorvariable für das Ereignis “ i -te Kugel fliegt in die erste Urne” ist. Die Linearität des Erwartungswertes ergibt:

$$E[X] = \sum_{i=1}^m E[X_i] = \sum_{i=1}^m \frac{1}{n} = \frac{m}{n}.$$

Erwartete Anzahl der Urnen mit genau einer Kugel Sei nun

$Y =$ Anzahl der Urnen mit genau einer Kugel

Dann ist $Y = Y_1 + \dots + Y_n$ wobei Y_j die Indikatorvariable für das Ereignis “ j -te Urne enthält genau eine Kugel”. Das Ereignis $Y_j = 1$ tritt genau dann ein, wenn eine der m Kugeln in die j -te Urne fliegt und die verbleibenden $m - 1$ Kugeln diese Urne vermeiden. Deshalb gilt

$$\begin{aligned} E[Y_j] &= \Pr\{Y_j = 1\} \\ &= \sum_{i=1}^m \Pr\{\text{nur Kugel } i \text{ fliegt in Urne } j\} \\ &= m \cdot \frac{1}{n} \left(1 - \frac{1}{n}\right)^{m-1} \end{aligned}$$

und somit auch

$$E[Y] = \sum_{j=1}^n E[Y_j] = m \left(1 - \frac{1}{n}\right)^{m-1} \sim m e^{-(m-1)/n}.$$

Erwartete Anzahl der Würfe bis eine leere Urne getroffen wird Angenommen, k Urnen sind bereits besetzt (enthalten mindestens eine Kugel). Wie lange müssen wir dann noch werfen bis die Kugel in eine leere Urne fliegt? Sei T_k die entsprechende Zufallsvariable,

$T_k =$ Anzahl der Versuche bis eine Kugel in eine leere Urne fliegt

Dann ist

$$\begin{aligned} E[T_k] &= \sum_{i=0}^{\infty} \Pr\{\text{Anzahl der Versuche} > i\} \quad (\text{Satz 4.56}) \\ &= \sum_{i=0}^{\infty} \Pr\{\text{alle ersten } i \text{ Kugeln fliegen in besetzten Urnen}\} \\ &= \sum_{i=0}^{\infty} \left(\frac{k}{n}\right)^i \\ &= \frac{1}{1 - k/n} = \frac{n}{n - k} \quad (\text{geometrische Reihe}) \end{aligned}$$

Man kann auch anders überlegen. Die Wahrscheinlichkeit eine leere Urne zu treffen ist $p = (n - k)/n$. Das ist also ein Bernoulli-Experiment mit der Erfolgswahrscheinlichkeit p , und wir wollen die erwartete Anzahl $E[T_k]$ der Versuche bis zum erstem Erfolg bestimmen. Wir werden bald sehen (im Abschnitt 4.10), dass T_k eine geometrisch verteilte Zufallsvariable ist und solche Zufallsvariablen den Erwartungswert $1/p = (n - k)/n$ haben.

Das Coupon Collector Problem Es gibt eine Serie von n Sammelbildern; in jede Runde kauft ein Sammler rein zufällig ein Bild. Was ist die erwartete Anzahl der Runden, bis der Sammler *alle* n Bilder hat? Dieses Problem ist als “Coupon Collector Problem” bekannt. Diese Frage kann man wiederum in einem Urnenmodell stellen. Kugeln sind nun die Runden und Urnen sind die Bilder. Die Frage ist, wieviel Kugeln müssen wir werfen, bis es *keine* Urne leer bleibt? Sei

$X =$ Anzahl der Versuche bis keine Urne leer wird.

Um $E[X]$ zu berechnen, summieren wir für alle $i = 1, \dots, n$ die erwartete Anzahl der Versuche, bis der Ball erstmals in i -te Urne fliegt:

$$\begin{aligned} E[X] &= \sum_{k=0}^{n-1} E[T_k] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \cdot \sum_{k=0}^{n-1} \frac{1}{n-k} \\ &= n \cdot \sum_{j=1}^n \frac{1}{j} = nH_n \quad (\text{harmonische Reihe}) \\ &\approx n \ln n. \end{aligned}$$

Man kann auch anders überlegen. Sei

$Z =$ die Anzahl der *leeren* Urnen

Dann ist $Z = Z_1 + \dots + Z_n$, wobei Z_j die Indikatorvariable für das Ereignis “ j -te Urne bleibt leer” ist. Nun haben wir

$$E[Z_j] = \Pr\{Z_j = 1\} = \left(1 - \frac{1}{n}\right)^m$$

da jede Kugel die j -te Urne mit Wahrscheinlichkeit $1 - 1/n$ vermeidet und $Z_j = 1$ genau dann, wenn *alle* m Kugeln dies tun. Damit ist

$$E[Z] = \sum_{j=1}^n E[Z_j] = n \left(1 - \frac{1}{n}\right)^m \sim n \cdot e^{-m/n}$$

Wenn wir also $m > n \ln n$ Kugeln in n Urnen werfen, dann kann man *keine* leere Urne mehr erwarten: In diesem Fall ist $E[Z] \sim n \cdot e^{-m/n} < n \cdot e^{-\ln n} = 1$.

Anzahl der “überfüllten” Urnen Wir haben bereits m Bälle in n Urnen geworfen. Wir sagen, dass eine Urne *überfüllt* ist, falls sie mehr als k Kugeln enthält (k ist ein Parameter). Für welche k wird es im Durchschnitt *keine* überfüllte Urnen geben? Einfachheit halber betrachten wir nur den Fall $m = n$ (genauso viele Kugeln wie Urnen).

Dazu betrachten wir die Zufallsvariable $X = X_1 + \dots + X_n$, wobei X_i die Indikatorvariable für das Ereignis “ i -te Urne hat mehr als k Bälle” ist. Dann ist X genau die Anzahl der überfüllten Urnen. Wir wollen ein k bestimmen, für das $E[X] < 1$ gilt. Dazu betrachten wir die Ereignisse

$A_{i,j} =$ “ i -te Urne enthält *genau* j Kugeln”

Dann ist $\Pr\{A_{i,j}\}$ gleich die Anzahl $\binom{n}{j}$ der Möglichkeiten, die j Bälle auszuwählen, mal die Wahrscheinlichkeit $\left(\frac{1}{n}\right)^j$, dass alle diese Bälle in Urne i fliegen. mal die Wahrscheinlichkeit $\left(1 - \frac{1}{n}\right)^{n-j}$,

dass keiner der verbleibenden $n - j$ Baller in diese Urne fliegt:

$$\begin{aligned} \Pr \{A_{i,j}\} &= \binom{n}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{n-j} \\ &\leq \binom{n}{j} \left(\frac{1}{n}\right)^j \\ &\leq \left(\frac{ne}{j}\right)^j \left(\frac{1}{n}\right)^j = \left(\frac{e}{j}\right)^j \end{aligned}$$

und damit auch

$$E[X_i] = \Pr \{X_i = 1\} = \sum_{j=k}^n \Pr \{A_{i,j}\} \leq \sum_{j=k}^n \left(\frac{e}{j}\right)^j \leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \dots\right).$$

Fur $k \rightarrow \infty$ konnen wir die Summe in Klammern ignorieren (sie strebt gegen 1) und wir haben eine (asymptotische) Ungleichung

$$E[X] = \sum_{i=1}^n E[X_i] \leq n \cdot \left(\frac{e}{k}\right)^k.$$

Es reicht also k so auszuwahlen, dass $\left(\frac{e}{k}\right)^k < 1/n$ gilt. Nach der Logarithmieren, muss die Ungleichung $k(1 - \ln k) < -\ln n$ gelten. Falls wir

$$k := \frac{\ln n}{\ln \ln n}$$

wahlen, dann gilt:

$$k(1 - \ln k) = \frac{\ln n}{\ln \ln n} (1 - \ln \ln n + \ln \ln \ln n) \sim -\ln n.$$

Also, wenn wir n Kugeln in n Urnen werfen, dann konnen wir erwarten, dass keine Urne mehr als $\ln n$ Kugeln enthalten wird.

Mehrfaches hashing – Bloom-Filter Um den Speicherplatz (= Anzahl n der Speicherzellen = Anzahl der benotigten Urnen) zu sparen, benutzt man oft Hashing mit mehreren Hashfunktionen. Das entsprechende Verfahren ist als ‘‘Bloom-Filter’’ bekannt und hat seit 1970 viele Anwendungen gefunden (das Program `ispel` ist nur ein Beispiel).

Sei U ein Universum und sei \mathcal{H} die Menge aller Funktionen, die U auf die Menge $\{1, \dots, n\}$ abbilden. Ein Bloom-Filter reprasentiert eine Menge $S \subseteq U$ durch ein Boole’sches Array B der Lange n und benutzt dazu k rein zufallig und unabhangig voneinander gewahlten Hashfunktionen $h_1, \dots, h_k \in \mathcal{H}$:

Anfanglich ist $S = \emptyset$ und alle Zellen von B sind auf Null gesetzt. Ein Element $x \in U$ wird in die Menge S eingefugt, indem das Array B nacheinander an den Stellen $h_1(x), \dots, h_k(x)$ auf Eins gesetzt wird.

Um nachzuprufen, ob x ein Element von S ist, wird B an den Stellen $h_1(x), \dots, h_k(x)$ uberpruft. Wenn B an allen Stellen den Wert 1 besitzt, dann wird die Vermutung ‘‘ $x \in S$ ’’ ausgegeben und ansonsten wird die definitive Ausgabe ‘‘ $x \notin S$ ’’ getroffen.

Offensichtlich erhalten wir konventionelles Hashing aus Bloom-Filtern fur $k = 1$. Wir beachten, dass eine negative Antwort auf eine ‘‘ist $x \in S$?’’ Anfrage stets richtig ist. Eine positive Antwort kann

allerdings falsch sein und Bloom-Filter produzieren damit “falsche Positive”. D.h. $x \in U$ ist ein falsches Positives, wenn $x \notin S$, aber der Filter mit “Ja” geantwortet hat. Wann passiert das? Wenn in alle k Zellen $h_1(x), \dots, h_k(x)$ irgendwelche Elemente $y \neq x$ aus S gehasht sind.

Wie groß ist die Wahrscheinlichkeit p_+ einer falschen positiven Antwort, wenn eine Menge S der Größe $|S| = s$ durch ein Boole’sches Array der Größe n mit Hilfe von k zufällig aus \mathcal{H} gewählten Hashfunktionen repräsentiert wird? Die Abschätzung $p_+ \leq P$ mit

$$P := \left(1 - \left(1 - \frac{1}{n} \right)^{ks} \right)^k$$

war von Bloom angegeben und ist seit 1970 mehrmals in der Literatur wiederholt. Praktiker haben aber bemerkt, dass mit dieser Abschätzung irgendwas nicht stimmt: die *tatsächliche* Wahrscheinlichkeit p_+ einer falschen positiven Antwort oft *größer* als Bloom’s Abschätzung P war. Und das ist wirklich der Fall, wie das folgende Beispiel zeigt.

▷ **Beispiel 4.79 : (Gegenbeispiel zur Bloom’s Abschätzung)** Wir betrachten den Fall wenn $n = k = 2$ und $s = 1$. Sei $U = \{x, y\}$, $S = \{x\}$ und $h_1, h_2 : U \rightarrow \{1, 2\}$. Element y ist ein falsches Positives genau dann, wenn beide $h_1(y)$ und $h_2(y)$ in der Menge $\{h_1(x), h_2(x)\}$ liegen.

Sei A_i das Ereignis “ $h_i(y) \in \{h_1(x), h_2(x)\}$ ” Dann gilt $p_+ = \Pr \{A_1 \cap A_2\}$. Der Wahrscheinlichkeitsraum besteht aus 16 Elementarereignissen:

$h_1(x)$	$h_2(x)$		$h_1(y)$	$h_2(y)$
1	1		1	1
1	2		1	2
2	1	×	2	1
2	2		2	2

⇒ Wahrscheinlichkeit eine Eins (eine bereits besetzte Zelle) zu erwischen ist

$$\Pr \{A_i\} = \frac{3 + 3 + 3 + 3}{16} = \frac{12}{16} = \frac{3}{4}$$

Bloom sagt: $\Pr \{A_1 \cap A_2\} = \Pr \{A_1\} \cdot \Pr \{A_2\} = \frac{9}{16}$.

Das ist aber falsch, da tatsächlich gilt

$$\Pr \{A_1 \cap A_2\} = \frac{3 + 2 + 2 + 3}{16} = \frac{10}{16}.$$

Warum passiert das? Da $\Pr \{A_1 \cap A_2\} \neq \Pr \{A_1\} \cdot \Pr \{A_2\}$, d.h. die Ereignisse *nicht* unabhängig sind!



Teufel steckt oft in Details! Stochastische Abhängigkeit von Ereignissen ist oft nicht offensichtlich und man soll damit vorsichtig umgehen.

Wie soll aber eine richtige Abschätzung für p_+ aussehen? Um diese Frage zu beantworten, können wir wiederum das Urnen-Model benutzen.

Wir werfen $m = ks$ blauer Baller²³ in n Urnen mit

$$\Pr \{ \text{Ball } b \text{ fliegt in } j\text{-te Urne} \} = 1/n$$

fur alle Baller b und alle Urnen j . Wir sagen, dass eine Urne ‘blau’ ist, falls sie mindestens einen blauen Ball enthalt.

Danach werfen wir k rote Baller.²⁴ Uns interessiert das Ereignis

$$A = \text{alle rote Baller landen in blauen Urnen.}$$

Fur jede Teilmenge der Urnen $I \subseteq \{1, \dots, n\}$ betrachten wir das Ereignis

$$B_I = I \text{ ist genau die Menge der blauen Urnen.}$$

Nach der Formel von der totalen Wahrscheinlichkeit gilt

$$\Pr \{A\} = \sum_I \Pr \{A|B_I\} \cdot \Pr \{B_I\} = \sum_{i=1}^m \binom{m}{i} p_i q_i,$$

wobei $p_i = \Pr \{A|E_I\}$ und $q_i = \Pr \{B_I\}$ fur eine *fixierte* Teilmenge I der Urnen mit $|I| = i$ ist. Die Wahrscheinlichkeiten p_i sind leicht zu bestimmen:

$$p_i = \Pr \{A|B_I\} = \left(\frac{i}{n}\right)^k.$$

Was aber mit der Wahrscheinlichkeiten $q_i = \Pr \{B_I\}$? Aus der Definition von B_I folgt

$$\Pr \{B_I\} = \frac{\text{Anzahl aller surjektiven Abbildungen } f : [m] \rightarrow [i]}{\text{Anzahl aller Abbildungen } f : [m] \rightarrow [n]}.$$

Sei $\Delta_m(i)$ die Anzahl aller surjektiven Abbildungen $f : [m] \rightarrow [i]$. Dann gilt also

$$q_i = \frac{\Delta_m(i)}{n^m}.$$

Damit erhalten wir

$$p_+ = \Pr \{A\} = \frac{1}{n^m} \sum_{i=1}^n \binom{n}{i} \left(\frac{i}{n}\right)^k \cdot \Delta_m(i).$$

Ein Ausdruck fur die Zahlen $\Delta_m(i)$ ist nicht allzu schwer zu bekommen (ubungsaufgabe!):

$$\Delta_m(i) = \sum_{j=1}^i (-1)^j \binom{i}{j} j^m.$$

In unserem speziellen (im Beispiel betrachteten) Fall haben wir $m = ks = 2$ und $n = 2$. Da $\Delta_2(1) = 1$ und $\Delta_2(2) = 2$, ergibt das

$$\begin{aligned} \Pr \{A\} &= \frac{1}{2^2} \sum_{i=1}^2 \binom{2}{i} \left(\frac{i}{2}\right)^2 \cdot \Delta_2(i) \\ &= \frac{1}{4} \cdot 2 \cdot \left(\frac{1}{2}\right)^2 \cdot 1 + \frac{1}{4} \cdot 1 \cdot (1)^2 \cdot 2 = \frac{1}{8} + \frac{1}{2} = \frac{5}{8} = \frac{10}{16}. \end{aligned}$$

²³Jeder Wurf entspricht einem der Werte $h_i(x)$ fur $x \in S$ und $i \in [k] = \{1, \dots, k\}$.

²⁴Jeder solcher Wurf entspricht einem der Werte $h_i(y)$ fur $i \in [k]$.

4.13 Bedingter Erwartungswert*

Definition: Sei X eine Zufallsvariable und A ein Ereignis. Der *bedingte Erwartungswert* $E[X|A]$ von X unter der Bedingung A ist definiert durch:

$$E[X|A] = \sum_x x \cdot \Pr\{X = x | A\}.$$

Wegen $\Pr\{X = x | A\} \leq \Pr\{X = x\} / \Pr\{A\}$ ist mit dem Erwartungswert $E[X]$ auch $E[X|A]$ wohldefiniert und endlich (natürlich nur wenn $\Pr\{A\} \neq 0$ gilt).

▷ *Beispiel 4.80:* Wir würfeln einmal einen Spielwürfel und X sei die gewürfelte Augenzahl. Was ist dann $E[X|X \geq 4]$?

$$E[X|X \geq 4] = \sum_{i=1}^6 i \cdot \Pr\{X = i | X \geq 4\} = \sum_{i=4}^6 i \cdot \frac{\Pr\{X = i\}}{\Pr\{X \geq 4\}} = \sum_{i=4}^6 i \cdot \frac{(1/6)}{(1/2)} = \sum_{i=4}^6 i \cdot \frac{1}{3} = 5.$$

Beachte, dass in diesem Fall der (unbedingte) Erwartungswert viel kleiner ist: $E[X] = \sum_{i=1}^6 i \cdot \frac{1}{6} = 3,5$.

Der bedingte Erwartungswert $E[X|A]$ ist einfach der Erwartungswert von X in einem anderen Wahrscheinlichkeitsraum, wo die Wahrscheinlichkeiten durch das Ereignis A bestimmt sind. Deshalb gelten für $E[X|A]$ dieselben Regeln wie für $E[X]$. Insbesondere gilt der Linearitätssatz (Satz 4.46) auch für $E[X|A]$.

Wo wir vom bedingten Erwartungswert profitieren können, ist die Tatsache, dass er oft ermöglicht, komplizierte Berechnungen von dem Erwartungswert $E[X]$ auf einfachere Fälle zu reduzieren.

Satz 4.81. (Regel des totalen Erwartungswertes) Sei $X : \Omega \rightarrow S$ eine Zufallsvariable mit $|S| < \infty$. Ist A_1, \dots, A_n eine disjunkte Zerlegung des Wahrscheinlichkeitsraums Ω , so gilt

$$E[X] = \sum_{i=1}^n \Pr\{A_i\} \cdot E[X|A_i]$$

Beweis.

$$\begin{aligned}
 E[X] &= \sum_{x \in S} x \cdot \Pr\{X = x\} && \text{(Definition von } E[X]) \\
 &= \sum_{x \in S} x \cdot \sum_{i=1}^n \Pr\{A_i\} \cdot \Pr\{X = x \mid A_i\} && \text{(totale Wahrscheinlichkeit)} \\
 &= \sum_{x \in S} \sum_{i=1}^n x \cdot \Pr\{A_i\} \cdot \Pr\{X = x \mid A_i\} \\
 &= \sum_{i=1}^n \sum_{x \in S} x \cdot \Pr\{A_i\} \cdot \Pr\{X = x \mid A_i\} && \text{(Umordnung der Summen)} \\
 &= \sum_{i=1}^n \Pr\{A_i\} \sum_x x \cdot \Pr\{X = x \mid A_i\} \\
 &= \sum_{i=1}^n \Pr\{A_i\} \cdot E[X \mid A_i] && \text{(Definition von } E[X \mid A_i])
 \end{aligned}$$

□

Sind nun $X, Y : \Omega \rightarrow \mathbb{R}$ zwei Zufallsvariablen, so kann man für jedes y in dem Wertebereich von Y den bedingten Erwartungswert $E[X \mid Y = y]$ von X unter der Bedingung, dass das Ereignis $Y = y$ eingetroffen ist, betrachten.

Nehmen wir nun an, dass Y nicht fixiert ist. Dann kann man für jedes Elementarereignis $\omega \in \Omega$ zuerst den Wert $y = Y(\omega)$ bestimmen und dann den Wert $E[X \mid Y = y]$ berechnen. So bekommt man eine neue Zufallsvariable, die man mit $E[X \mid Y]$ bezeichnet.



Beachte, dass (im Unterschied zu $E[X]$) $E[X \mid Y]$ keine Zahl sondern eine *Zufallsvariable* ist! D.h. $E[X \mid Y]$ ist eine Zufallsvariable $f(Y)$, die die Werte $E[X \mid Y = y]$ für $Y = y$ annimmt. Oder anders gesagt, $f(y) := E[X \mid Y = y]$ definiert eine Abbildung $f : \mathbb{R} \rightarrow \mathbb{R}$, und $E[X \mid Y]$ ist dann die Abbildung $f(Y)$ von Ω nach \mathbb{R} .

Um den Wert $Z(\omega)$ dieser neuen Zufallsvariable $Z = f(Y)$ auf $\omega \in \Omega$ zu bestimmen, bestimmt man zuerst den Wert $y = Y(\omega)$ und nimmt den Erwartungswert $E[X \mid Y = y]$ als den Wert von $Z(\omega)$.

► *Beispiel 4.82* : Wir würfeln einmal zwei Spielwürfel und betrachten die Zufallsvariable $X = X_1 + X_2$, wobei X_i die Augenzahl des i -ten Würfels ist. Um $E[X \mid X_1]$ zu bestimmen, berechnen wir zuerst die entsprechende Abbildung $f(y) = E[X \mid X_1 = y]$:

$$\begin{aligned}
 f(y) = E[X \mid X_1 = y] &= \sum_{x=2}^{12} x \cdot \Pr\{X = x \mid X_1 = y\} \\
 &= \sum_{x=y+1}^{y+6} x \cdot \Pr\{X_2 = x - y\} \\
 &= (6y + \sum_{i=1}^6 i) \frac{1}{6} \\
 &= y + \frac{7}{2}.
 \end{aligned}$$

Damit ist $E[X | X_1] = X_1 + \frac{7}{2}$.

- *Beispiel 4.83* : Wir würfeln n mal einen Würfel, und sei X_i ($1 \leq i \leq 6$) die Anzahl der Wurfes, die die Augenzahl i ergeben. Dann ist

$$E[X_1 | X_6] = \frac{n - X_6}{5}.$$

Warum? Wir müssen die Werte von $Z := E[X | Y]$ auf Elementarereignissen $\omega \in \Omega$ bestimmen, wobei die Elementarereignisse in diesem Fall alle Strings $\omega \in \{1, \dots, 6\}^n$ sind. Zuerst bestimmen wir den Wert $y = X_6(\omega)$. Das ist genau die Anzahl der 6'en in ω . Wissen wir nun, dass genau y Wurfes ein "6" ergeben, so ist die Wahrscheinlichkeit, in jeder von $n - y$ verbleibenden Wurfes ein "1" zu bekommen, gleich $1/5$. Damit ist $E[X_1 | X_6] = (n - X_6)/5$, wie behauptet.

Genauso bekommt man z.B.

$$E[X_1 | X_2, X_3] = \frac{n - X_2 - X_3}{4}.$$

Satz 4.84. Sind die Zufallsvariablen X und Y unabhängig, so gilt $E[X | Y] = E[X]$.

Beweis. Wegen der Unabhängigkeit von X und Y gilt für jedes y

$$\begin{aligned} E[X | Y = y] &= \sum_x x \cdot \Pr\{X = x | Y = y\} \\ &= \sum_x x \cdot \Pr\{X = x\} \quad (\text{Unabhängigkeit}) \\ &= E[X]. \end{aligned}$$

Da $E[X | Y = y] = E[X]$ für jeden möglichen Wert y von Y gilt, muss auch $E[X | Y] = E[X]$ gelten. \square

- *Beispiel 4.85* : Sei $X = X_1 + X_2$ die Zufallsvariable aus dem Beispiel 4.82. Da $E[X | X_1]$ eine Zufallsvariable ist, können wir den ihre Erwartungswert berechnen:

$$E[E[X | X_1]] = E\left[X_1 + \frac{7}{2}\right] = E[X_1] + \frac{7}{2} = \left(\frac{1}{6} \sum_{i=1}^6 i\right) + \frac{7}{2} = \frac{7}{2} + \frac{7}{2} = E[X_1] + E[X_2] = E[X].$$

Eine interessante Eigenschaft von Zufallsvariablen $Z := E[X | Y]$ ist, dass die im vorigen Beispiel entdeckte Eigenschaft auch im Allgemeinen gilt!

Satz 4.86. (Regel vom doppelten Erwartungswert)

$$E[E[X | Y]] = E[X]$$

Beweis. Sei $f(y) := E[X | Y = y]$. Dann gilt:

$$\begin{aligned}
 E[E[X | Y]] &= E[f(Y)] \\
 &= \sum_y f(y) \Pr\{Y = y\} && \text{(Lemma 4.52)} \\
 &= \sum_y \left(\sum_x x \cdot \Pr\{X = x | Y = y\} \right) \Pr\{Y = y\} && \text{(Def. von } f(y) = E[X | Y = y]\text{)} \\
 &= \sum_y \left(\sum_x x \cdot \frac{\Pr\{X = x, Y = y\}}{\Pr\{Y = y\}} \right) \Pr\{Y = y\} \\
 &= \sum_y \sum_x x \cdot \Pr\{X = x, Y = y\} \\
 &= \sum_x x \sum_y \Pr\{X = x, Y = y\} \\
 &= \sum_x x \Pr\{X = x\} && \text{(Satz 4.22)} \\
 &= E[X].
 \end{aligned}$$

□

Definition: Eine Folge von Zufallsvariablen X_0, X_1, \dots heißt ein *Martingal*, falls für alle $i \geq 0$ gilt:

$$E[X_{i+1} | X_0, \dots, X_i] = X_i.$$

► *Beispiel 4.87:* Wir haben eine Urne mit b blauen Kugeln und r roten Kugeln. In jedem Schritt wir ziehen eine Kugel rein zufällig und ersetzen diese mit zwei neuen Kugeln von derselben Farbe. Sei X_i die Anteil der roten Kugeln in der Urne nach i Schritten. So ist insbesondere

$$X_0 = \frac{r}{r+b}.$$

Behauptung: X_0, X_1, \dots ist ein Martingal.

Beweis. Sei $n_i = r + b + i$ die Gesamtzahl der Kugeln nach i Schritten. Dann ist $X_i n_i$ die Anzahl der roten Kugeln nach i Schritten. Es ist klar, dass diese Anzahl nur von der Anzahl der roten Kugeln nach $i - 1$ abhängen kann. Also gilt

$$\begin{aligned}
 E[X_{i+1} | X_0, X_1, \dots, X_i] &= E[X_{i+1} | X_i] \\
 &= X_i \cdot \frac{X_i n_i + 1}{n_{i+1}} + (1 - X_i) \cdot \frac{X_i n_i}{n_{i+1}} \\
 &= \frac{X_i}{n_{i+1}} + \frac{X_i n_i}{n_{i+1}} \\
 &= \frac{X_i (n_i + 1)}{n_i + 1} \\
 &= X_i.
 \end{aligned}$$

□

Lemma 4.88. Ist X_0, X_1, \dots ein Martingal, so gilt $E[X_i] = E[X_0]$ für alle $i \geq 0$.

Beweis. Induktion über i . Es gelte $E[X_i] = E[X_0]$. Da wir ein Martingal haben, gilt

$$X_i = E[X_{i+1} | X_0, \dots, X_i].$$

Wir nehmen den Erwartungswert von beiden Seiten und benutzen die Regel von doppelten Erwartungswert (Satz 4.86);

$$E[X_i] = E[E[X_{i+1} | X_0, \dots, X_i]] = E[X_{i+1}].$$

□

4.14 Summen von zufälliger Länge – Wald's Theorem

In diesem Abschnitt betrachten wir die folgende Fragestellung:

Wie kann man die erwarteten totalen Kosten eines Schritt-für-Schritt Zufallsexperiments berechnen, wenn sowohl die Kosten in jedem Schritt wie auch die Anzahl der Schritte von den vorherigen Ereignissen *abhängen* können?

▷ *Beispiel 4.89:* Wir würfeln einen Spielwürfel bis eine 6 rauskommt und summieren die bis dahin erschienene Augenzahlen. Welchen Erwartungswert hat diese Summe?

Wir denken uns jedes Würfeln als eine Zufallsvariable: für jedes $i = 1, 2, \dots$ sei X_i die im i -ten Würfeln erschienene Augenzahl. (Wir setzen $X_i = 0$, falls die Augenzahl 6 vor dem Schritt i ausgewürfelt wurde.) Also nimmt jedes X_i die Werte $0, 1, \dots, 6$ an. Setze $T := \min\{i : X_i = 6\}$. Das ist auch eine Zufallsvariable, die die Werte in \mathbb{N}_+ annimmt. Wir interessieren uns also für die Zufallsvariable $Y := X_1 + X_2 + \dots + X_T = \sum_{i=1}^T X_i$. Wir wissen, dass $E[X_i] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3,5$ für jedes i , und dass²⁵ $E[T] = \frac{1}{(1/6)} = 6$ ist.

Es gibt ein Resultat – Wald's Theorem – das uns sofort die Antwort liefert: $E[Y] = 6 \cdot 3,5 = 21$. D.h. wenn wir einen Spielwürfel bis eine 6 rauskommt würfeln, dann wird die erwartete Summe der insgesamt ausgewürfelten Augenzahlen gleich 21 sein.

Im Allgemeinen haben wir ein System, das Schritt-für-Schritt funktioniert und in i -tem Schritt die Kosten X_i verursacht. Damit haben wir eine (potentiell unendliche) Folge X_1, X_2, \dots der verursachten Kosten. Die Kosten sind zufällig und können voneinander abhängen. Die Lebensdauer T des Systems ist auch zufällig (und kann auch von bis dahin verursachten Kosten abhängen). Wie bestimmt man in einer solchen Situation die erwartete Gesamtkosten bis das System stirbt? Wegen so vielen möglichen Abhängigkeiten, sieht die Frage sehr schwer aus – wir wissen ja nicht, wann das System tatsächlich stirbt.

Zuerst betrachten wir den Fall, wenn (i) die Kosten X_i gleichverteilt sind und (ii) die Lebensdauer T des Systems von der verursachten Kosten unabhängig ist.

Bemerkung 4.90. Oft sagt man “seien X_1, X_2, \dots unabhängige Kopien einer Zufallsvariable X ”. Das bedeutet natürlich nicht, dass alle X_i 's eine und dieselbe Zufallsvariable X ist. Unter dessen versteht

²⁵ T ist geometrisch verteilt

man, dass X_i 's beliebige Zufallsvariablen sein können—die einzige Bedingung ist, dass jede von dieser Variablen X_i dieselbe Verteilung wie X haben muss. Zum Beispiel, wenn wir die Gleichverteilung auf Ω mit $\Pr\{\omega\} = 1/|\Omega|$ für alle $\omega \in \Omega$ betrachten, dann ist $Y : \Omega \rightarrow S$ eine Kopie von $X : \Omega \rightarrow S$ genau dann, wenn $|Y^{-1}(a)| = |X^{-1}(a)|$ für alle $a \in S$ gilt. Es ist klar, dass dann $Y^{-1}(a) = X^{-1}(a)$ nicht unbedingt gelten muß!

Satz 4.91. Seien X_1, X_2, \dots unabhängige Kopien einer reellwertigen Zufallsvariable X , und sei T eine davon unabhängige Zufallsvariable mit den Werten in \mathbb{N} und einem endlichen Erwartungswert. Dann gilt

$$E[X_1 + X_2 + \dots + X_T] = E[T] \cdot E[X]$$

Beweis. Sei $Y = X_1 + X_2 + \dots + X_T$. Wegen der Unabhängigkeit von T und X_i 's gilt ²⁶

$$\begin{aligned} \Pr\{Y = y \mid T = t\} &= \Pr\{X_1 + \dots + X_t = y \mid T = t\} \\ &= \Pr\{X_1 + \dots + X_t = y\}, \end{aligned}$$

also

$$E[Y \mid T = t] = E[X_1 + \dots + X_t] = t \cdot E[X].$$

Wir wenden die Regel des totalen Erwartungswertes an und erhalten

$$\begin{aligned} E[Y] &= \sum_t \Pr\{T = t\} \cdot E[Y \mid T = t] \\ &= \sum_t \Pr\{T = t\} \cdot t \cdot E[X] \\ &= E[X] \cdot \sum_t t \cdot \Pr\{T = t\} \\ &= E[X] \cdot E[T]. \end{aligned}$$

□

Im vorigen Satz spielt Unabhängigkeit der Zufallsvariablen eine große Rolle. In dem nächsten Satz spielt dagegen die Unabhängigkeit keine Rolle mehr. Es reicht, dass (i) die Kosten nicht-negativ sind und (ii) die erwarteten Kosten in jedem Schritt, unter der Bedingung, dass System bis dahin noch lebt, alle gleich sind.

Satz 4.92. (Wald's Theorem) Sei X_1, X_2, \dots eine Folge von nicht-negativen Zufallsvariablen und $T : \Omega \rightarrow \mathbb{N}_+$ eine Zufallsvariable, alle mit endlichen Erwartungswerten, so dass

$$E[X_i \mid T \geq i] = \mu$$

für ein festes $\mu \in \mathbb{R}$ und alle $i \geq 1$ gilt. Dann ist

$$E[X_1 + X_2 + \dots + X_T] = \mu \cdot E[T]$$

²⁶Sind die Zufallsvariablen X, Y, Z unabhängig, dann sind auch die Zufallsvariablen $X + Y$ und Z unabhängig. Übungsaufgabe!

Beweis. Sei I_k die Indikatorvariable für das Ereignis $T \geq k$, d.h. $I_k = 1$ falls der Prozess mindestens k Schritte läuft, und $I_k = 0$ falls der Prozess vor Zeitpunkt k endet. Für jedes $k = 1, 2, \dots$ gilt:

$$\begin{aligned} E[X_k I_k] &= E[X_k I_k | I_k = 1] \cdot \Pr\{I_k = 1\} + E[X_k I_k | I_k = 0] \cdot \Pr\{I_k = 0\} && \text{(Satz 4.81)} \\ &= E[X_k \cdot 1 | I_k = 1] \cdot \Pr\{I_k = 1\} + E[X_k \cdot 0 | I_k = 0] \cdot \Pr\{I_k = 0\} \\ &= E[X_k | I_k = 1] \cdot \Pr\{I_k = 1\} + 0 \\ &= E[X_k | T \geq k] \cdot \Pr\{T \geq k\} \\ &= \mu \cdot \Pr\{T \geq k\} \end{aligned}$$

Somit erhalten wir:

$$\begin{aligned} \sum_{k=1}^{\infty} E[X_k I_k] &= \mu \cdot \sum_{k=0}^{\infty} \Pr\{T > k\} \\ &= \mu \cdot E[T] && \text{(diskretwertige Zufallsvariablen, Satz 4.56)} \end{aligned}$$

Da $\mu \cdot E[T]$ endlich ist und alle $E[X_k I_k]$ nicht-negativ sind, ist die Reihe $\sum_{k=1}^{\infty} E[X_k I_k]$ absolut konvergent und wir können den Satz 4.58 (unendliche Linearität des Erwartungswertes) anwenden:

$$E[X_1 + X_2 + \dots + X_T] = E\left[\sum_{k=1}^{\infty} X_k I_k\right] = \sum_{k=1}^{\infty} E[X_k I_k] = \mu \cdot E[T]$$

□

Korollar 4.93. (Wald's Theorem - einfachste Form) Sei $T : \Omega \rightarrow \mathbb{N}_+$ und seien X_1, X_2, \dots unabhängige Kopien einer nicht-negativen Zufallsvariable X . Sind die Erwartungswerte von T und X endlich, so gilt

$$E[X_1 + X_2 + \dots + X_T] = E[T] \cdot E[X]$$

Beweis. In diesem Fall gilt $E[X_i | T \geq i] = E[X_1] = E[X]$, da der erste Versuch jedenfalls stattfinden muss und jeder Versuch X_i die gleiche Verteilung wie der erste Versuch X_1 hat (falls man zu diesem Versuch X_i überhaupt kommt!). □

▷ *Beispiel 4.94 : (Runs: Anzahl der Versuche)* Wir versuchen ein System (z.B. einen Computer) mit n Komponenten konstruieren. In jedem Schritt setzen wir eine neue Komponente ein. Aber die Komponenten sind unsicher ("no name" Produkte) und jedes Einsetzen kann das ganze System mit Wahrscheinlichkeit p kaputt machen. Falls dies geschieht, müssen wir die Konstruktion neu anfangen. Wir nehmen an, dass jede neue Komponente mit gleicher Wahrscheinlichkeit p das (bereits vorhandene) System zerstören kann. Was ist die erwartete Anzahl der Schritte bis das System fertig ist?

Wir können die Elementarereignisse als Strings von Einsen und Nullen kodieren, wobei eine 1 heißt "die Komponente war sicher gesetzt, dass System lebt weiter", und eine 0 heißt "die Komponente hat das System zerstört". D.h. wir haben eine Folge Y_1, Y_2, \dots von unabhängigen 0-1-Zufallsvariablen mit

$$\Pr\{Y_i = 0\} = p \quad \Pr\{Y_i = 1\} = q (= 1 - p)$$

und die Frage lautet: Wie lange müssen wir warten bis ein Block aus n Einsen kommt?

Wir können die Folge Y_1, Y_2, \dots in Blöcke (Versuche) aufteilen, wobei ein nicht erfolgreicher Versuch die Form²⁷ $1^k 0$ mit $0 \leq k < n$ hat; ein erfolgreicher Versuch hat die Form 1^n . Zum Beispiel für $n = 3$ hat die Folge 110100111 vier Versuche, wobei die ersten drei nicht erfolgreich sind:

$$\underbrace{110}_{X_1=3} \underbrace{10}_{X_2=2} \underbrace{0}_{X_3=1} \underbrace{111}_{X_4=3}$$

Wenn wir die Länge eines Versuchs mit X bezeichnen, so bekommen wir eine Folge X_1, X_2, \dots von unabhängigen Kopien dieser Zufallsvariable. Uns interessiert also die folgende Zufallsvariable

$$S := X_1 + X_2 + \dots + X_T$$

wobei

$$T := \min\{i : i\text{-ter Versuch war erfolgreich}\}.$$

Wald's Theorem (Korollar 4.93) sagt, dass dann

$$E[S] = E[T] \cdot E[X]$$

gilt. Die Zufallsvariable T beschreibt den ersten Versuch, der erfolgreich war. Da die Erfolgswahrscheinlichkeit für jede eingesetzte Komponente $1 - p$ ist und ein erfolgreicher Versuch aus n erfolgreich angesetzten Komponenten besteht, ist die Erfolgswahrscheinlichkeit $(1 - p)^n$. Damit haben wir²⁸

$$E[T] = \frac{1}{(1 - p)^n} = \frac{1}{q^n} \quad \text{mit } q := 1 - p$$

Andererseits gilt nach Satz 4.56:²⁹

$$E[X] = \sum_{i=0}^{\infty} \Pr\{X > i\} = \sum_{i=0}^{n-1} \Pr\{X > i\}.$$

Da³⁰ $\Pr\{X > i\} = (1 - p)^i = q^i$, folgt

$$E[X] = \sum_{i=0}^{n-1} q^i = \frac{1 - q^n}{1 - q} = \frac{1 - q^n}{p}.$$

Damit ist die erwartete Anzahl $E[S]$ der Schritte (bis das System fertig ist):

$$E[S] = E[X] \cdot E[T] = \frac{1 - q^n}{p} \cdot \frac{1}{q^n} = \frac{1 - q^n}{pq^n} = \frac{1}{p} \left(\frac{1}{q^n} - 1 \right).$$

Zum Beispiel, wenn nur eine 1% Chance besteht, dass eine eingefügte Komponente das System zerstört (also $p = 0,01$), dann ist die erwartete Anzahl der Schritte für ein System mit $n = 10$ Komponenten ungefähr 10. Für $n = 100$ sind das schon ungefähr 173 Schritte. Aber für $n = 1000$ sind das bereits satte 2.316.257 Schritte! Fazit: Man sollte das System modular ausbauen: 10 Module, je mit 10 Untermodulen, je aus 10 Komponenten.

²⁷ $1^k 0$ bezeichnet eine Folge aus k Einsen mit einer Null am Ende.

²⁸Geometrisch verteilte Zufallsvariable

²⁹ $\Pr\{X > n\} = 0$, da kein einzelner Versuch länger als n Schritte dauert.

³⁰Siehe Abschnitt 4.10.

4.15 Irrfahrten und Markov-Ketten

Ein stochastischer Prozess – als *Markov-Kette* (oder *Irrfahrt* oder *random walk*) bekannt – kommt in der Physik wie auch in der Informatik ziemlich häufig vor. Die Situation ist folgende.

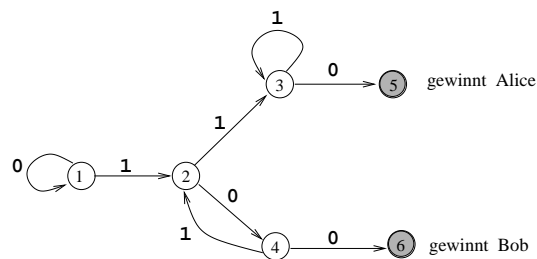
Wir haben ein Zufallsexperiment X_1, X_2, \dots , das in mehreren Schritten abläuft. In jedem Schritt liegt das Ergebnis des Experiments in einer (uns bekannten) endlichen Menge S (die “Zustandsmenge”). Die Bedingung ist die sogenannte “Gedächtnislosigkeit” des Experiments: In jedem Schritt i hängen die Ereignisse “ $X_i = s$ ” nur von der Ereignissen “ $X_{i-1} = r$ ” ab. D.h. die Wahrscheinlichkeit, in welchen Zustand das System (unser Experiment) im i -ten Schritt übergehen wird, hängt nur von dem Zustand ab, in dem das System sich gerade (im Schritt $i - 1$) befindet (und hängt nicht von der Vergangenheit ab).³¹

Die Theorie der Markov-Ketten ist umfangreich. Wir beschränken uns auf ein paar Ansätzen und Beispielen. Einige weitere Ansätze basieren sich auf dem Matrizenkalkül und wir werden sie im Abschnitt 5.13.6 behandeln.

► *Beispiel 4.95*: Zwei Spieler (Alice und Bob) spielen das folgende Spiel. Zunächst wählt Alice einen String $a \in \{0, 1\}^k$ und Bob einen String $b \in \{0, 1\}^k$. Danach werfen sie eine faire 0-1 Münze bis einer dieser zwei Strings erscheint. Derjenige gewinnt, dessen String als erstes erscheint.

Da in jedem Wurf 0 und 1 mit *gleicher* Wahrscheinlichkeit $\frac{1}{2}$ kommt, sagt uns die “Intuition”, dass die Gewinnchancen auch gleich sein sollten. Die genaue Analyse zeigt aber,³² dass das nicht unbedingt der Fall sein soll!

Dazu betrachten wir beispielsweise das Spiel mit $k = 3$, $a = 110$ und $b = 100$. Dieses Spiel kann man als eine Markov-Kette mit 6 Zuständen betrachten:



Wie groß sind die Chancen, dass Alice gewinnt? Dazu betrachten wir für alle Zustände i die Wahrscheinlichkeit p_i , ausgehend aus dem Zustand i in den Zustand 5 zu gelangen. Gesucht ist also p_1 . Durch Zerlegung der Wahrscheinlichkeiten nach dem ersten Sprung aus i heraus können

³¹Eigentlich haben wir Markov-Ketten – die Entscheidungsbäume – bereits (schweigend) eingeführt und benutzt. Das sind die einfachsten Markov-Ketten: Das System startet in dem Wurzel des Baums, wählt zufällig einen Nachbarn, läuft zu diesem Nachbarn, dann wieder wählt zufällig einen Nachbarn und läuft zu ihm, usw. bis ein Blatt erreicht ist. D.h. das System bewegt sich nur in einer Richtung, Zyklen sind nicht erlaubt. In allgemeinen Markov-Ketten, die wir jetzt betrachten, kann das System sich in beiden Richtungen bewegen.

³²Das Phänomen selbst ist noch überraschender: Falls Bob sein String b *zuerst* auswählen und Alice zeigen muss, dann kann Alice *immer* einen String a finden, so dass sie mit *größerer* Wahrscheinlichkeit gewinnen wird! Dieses Phänomen ist unter dem Namen “best bets for simpletons” bekannt. Mehr dazu kann man z.B. in meinem Buch (Abschnitt 24.4) finden.

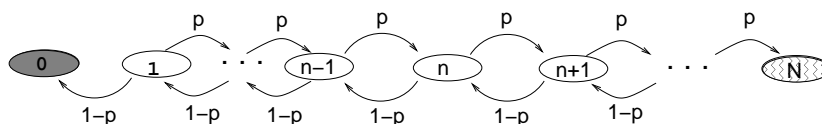
wir das folgende Gleichungssystem aufstellen:

$$\begin{aligned} p_1 &= p_2 \\ p_2 &= \frac{1}{2}p_3 + \frac{1}{2}p_4 \\ p_3 &= 1 \\ p_4 &= \frac{1}{4}. \end{aligned}$$

Durch Auflösen folgt $p_1 = 5/8$. Damit sind Alice's Gewinnchancen um $8/5 > 4/3$ größer als die von Bob.

Irrfahrten mit absorbierenden Zuständen

In einem Teich befinden sich $N + 1$ Steine $0, 1, \dots, N$. Ein Frosch sitzt anfänglich auf irgendeinem Stein $n \neq 0$. Der Stein 0 ist für den Frosch gefährlich – da steht ein Storch, der sofort den Frosch fängt, wenn er auf diesen Stein springt.



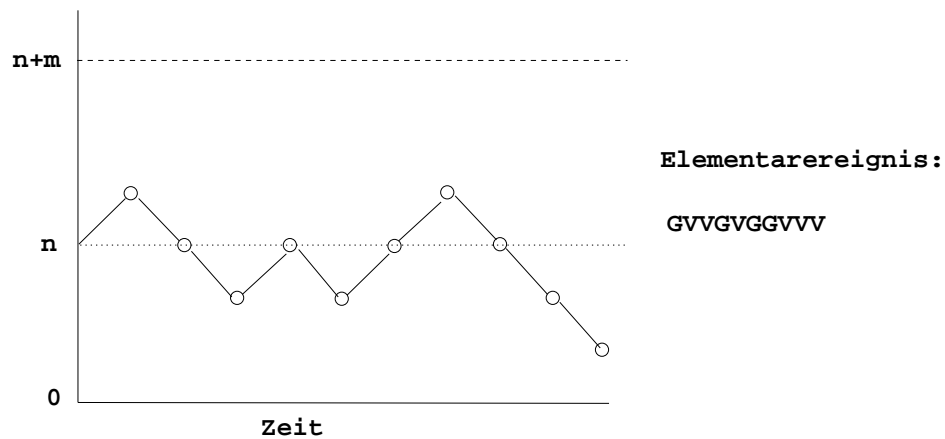
Der Frosch beginnt von Stein n mit Wahrscheinlichkeit p nach rechts und mit mit Wahrscheinlichkeit $q = 1 - p$ nach links springen. Wenn er den Stein N erreicht hat, bleibt er da für immer (das ist sein Zuhause, da ist er sicher). Uns interessiert die Wahrscheinlichkeit

$$w = \Pr \{ \text{Frosch überlebt} \},$$

dass der Frosch den (sicheren) Stein N erreicht.

Beachte, dass die Situation äquivalent zu dem Casino-Spiel ist, das wir im Abschnitt 3.10 betrachtet haben (Gambler's Ruin). Ein Spieler namens Theo Retiker nimmt in einem Casino an einem Spiel mit Gewinnwahrscheinlichkeit $0 < p \leq 1/2$ teil. Zum Beispiel wirft man eine (nicht unbedingt faire) Münze, dessen Seiten mit rot und blau gefärbt sind, und wir gewinnen, falls rot kommt.

Wir nehmen an, dass Theo in jeder Spielrunde nur 1 € einsetzen kann. Geht die Runde zu Theos Gunsten aus, erhält er den Einsatz zurück und zusätzlich denselben Betrag aus der Bank (Gewinn = 1 €). Endet die Runde ungünstig, verfällt der Einsatz (Gewinn = -1 €). Theo Retiker kommt ins Casino mit n Euro (Anfangskapital) und sein Ziel ist am Ende N Euro in der Tasche haben, d.h. er will $N - n$ Euro gewinnen (dann will er aufhören); in diesem Fall sagen wir, dass Theo gewinnt. Er spielt bis er $m = N - n$ Euro gewinnt oder bis er alle seine mitgenommenen n Euro verliert.



Beachte, dass Theo und Frosch denselben Prozess modellieren: Theo will N Euro am Ende haben und der Frosch will den Stein N (sein Haus) erreichen; Theo ist bankrot, wenn er nichts mehr in der Tasche hat, und der Frosch ist jedenfalls “bankrot”, wenn er den Stein 0 (mit dem Storch) erreicht. Deshalb gilt

$$w = \Pr \{ \text{Theo gewinnt} \} .$$

Wir haben bereits bewiesen (siehe Satz 3.87), dass für $p = 1/2$ die Gewinnchancen für Theo (oder die Überlebenschancen für den Frosch) immerhin

$$w = \frac{n}{N}$$

betragen. So sind z.B. fifty-fifty Chance, dass Theo sein Anfangskapital verdoppeln kann.

Wir haben aber auch gezeigt, dass sich die Situation dramatisch verändert, wenn $p < 1/2$ ist, d.h. wenn man in einem amerikanischen Casino spielt oder wenn der Frosch mit größere Wahrscheinlichkeit nach links (nahe zum Storch) als nach rechts springt. Dann gilt nämlich

$$w < e^{-(N-n)} .$$

D.h. sind dann z.B. Theos Chancen, 10€ zu gewinnen, kleiner als 2^{-10} auch wenn er eine Million Euro mit sich mitnimmt; dann wird er alles mit Wahrscheinlichkeit $1 - 2^{-10}$ verlieren!

Nun wollen wir wissen, wieviele Spielrunden Theo erwarten kann, bis er gewinnt oder alles verliert. Sei

- T = die Anzahl der Spielrunden, bis das Spiel zu Ende ist
- G = das Geld, das Bob am Ende gewinnt oder verliert.

Da Theo mit Wahrscheinlichkeit w gewinnt und mit Wahrscheinlichkeit $1 - w$ verliert, ist

$$E[G] = (N - n) \cdot w - n \cdot (1 - w) = Nw - n.$$

Wir wollen aber die erwartete Spieldauer $E[T]$ bestimmen. Dazu beachten wir, dass $G = X_1 + \dots + X_T$, wobei X_i das in der i -ten Spielrunde gewonnene (+1) oder verlorene (-1) Kapital ist. Die Zufallsvariablen X_1, \dots, X_T sind unabhängige Kopien einer Zufallsvariable X , die das gewonnene Kapital in einer Spielrunde angibt. Da $E[X] = (+1) \cdot p + (-1) \cdot (1 - p) = 2p - 1$, liegt es nahe, das Wald's Theorem anzuwenden, was uns die Gleichheit

$$E[G] = (2p - 1)E[T] \tag{4.16}$$

liefern würde; dann wäre die gesuchte erwartete Spieldauer $E[T] = E[G]/(2p - 1)$.

Die Sache hat aber einen Haken: Die Zufallsvariable X ist nicht positiv – ihr Wertebereich ist $\{-1, +1\}$. Deshalb können wir nicht das Wald’s Theorem direkt anwenden. In solchen Fällen wendet man einen einfachen Trick: Verschiebe die Zufallsvariablen nach rechts, um nicht-negative Zufallsvariablen zu bekommen. In unserem Fall können wir anstatt X die Zufallsvariable $Y := X + 1$ betrachten. Dann ist $E[Y] = E[X] + 1 = 2p$. Da Y nicht-negativ ist, können wir die einfachste Form des Wald’s Theorems anwenden und erhalten

$$E \left[\sum_{i=1}^T Y_i \right] = 2pE[T]$$

Andererseits, gilt

$$E \left[\sum_{i=1}^T Y_i \right] = E \left[\sum_{i=1}^T X_i + \sum_{i=1}^T 1 \right] = E \left[\sum_{i=1}^T X_i + T \right] = E \left[\sum_{i=1}^T X_i \right] + E[T] = E[G] + E[T]$$

Somit haben wir die Gleichung $E[G] + E[T] = 2pE[T]$ bewiesen, woraus (4.16) unmittelbar folgt. Damit haben wir bewiesen, dass für $p < 1/2$

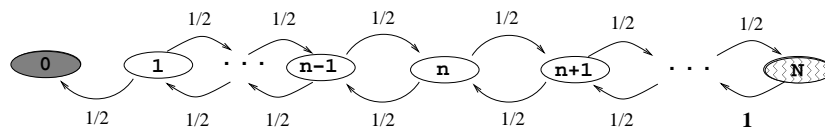
$$E[\text{Anzahl der Spielrunden}] = \frac{N \cdot \Pr \{ \text{Theo gewinnt} \}}{2p - 1}$$

gilt. Was passiert aber, wenn das Spiel fair ist, d.h. wenn $p = 1/2$ ist? Wie lange kann dann Theo spielen? Dieser Fall ist komplizierter und wir nur erwähnen (ohne Beweis³³) was dabei rauskommt:

$$E[\text{Anzahl der Spielrunden}] = n(N - n).$$

Irrfahrten mit reflektierenden Zuständen

Wir betrachten wiederum die Springerei vom Frosch in einem Teich mit N Steinen $0, 1, \dots, N$, wobei wiederum der Stein 0 gefährlich ist – da ist der Storch. Der Unterschied nun ist, dass neben dem letzten Stein N ein Baum steht, so dass der Frosch am diesem Stein “reflektiert” wird. D.h. erreicht er den Stein N , so springt er im nächsten Sprung mit Sicherheit zurück auf den einzigen Nachbarn $N - 1$. Einfachheitshalber, nehmen wir nun auch an, dass $p = 1/2$ gilt: Befindet er sich zu irgend einem Zeitpunkt auf einem Stein, der einen linken und einen rechten Nachbarstein hat, so springt er jeweils mit Wahrscheinlichkeit $1/2$ auf einen der beiden.



Der Frosch sitzt anfänglich auf irgendeinem Stein n .

Frage: Wie lange wird der Frosch springen, bis er gefressen wird? D.h. nach wievielen Sprüngen wird der Frosch erwartungsgemäß zum ersten mal Stein 0 erreichen, wenn er am Stein n startet? Uns interessiert also

$$x_n = E[\text{Anzahl der Sprünge, wenn am Stein } n \text{ gestartet}].$$

³³Der Beweis ist ähnlich wie der von der Gleichung (4.17): Es reicht in diesem Beweis die Randbedingung mit $E_n = N_{n-1} + 1$ und $E_n = 0$ ersetzen.

Wir wollen zeigen, dass

$$x_n = n \cdot (2N - n) \quad (4.17)$$

für alle $i = 1, \dots, n$ gilt. Damit ist³⁴ zum Beispiel $x_1 = 2N - 1$ bzw. $x_N = N^2$.

Beweis. Es gelten die folgende 3 Aussagen:

$$x_0 = 0, \quad x_N = x_{N-1} + 1$$

sowie

$$x_n = 1 + \frac{1}{2} \cdot x_{n-1} + \frac{1}{2} \cdot x_{n+1} \quad \text{für } n = 1, \dots, N-1.$$

Die ersten beiden Aussagen sind klar, die dritte gilt, weil der Frosch in Position $1 \leq n \leq N-1$ einen Sprung macht und dann mit Wahrscheinlichkeit $1/2$ in Position $n-1$ bzw. $n+1$ ist.³⁵ Wenn wir in der dritten Gleichung auf beiden Seiten $\frac{1}{2} \cdot x_n$ subtrahieren, dann durch 2 multiplizieren und umstellen, erhalten wir die Aussage

$$x_n - x_{n-1} = x_{n+1} - x_n + 2 \quad \text{für } n = 1, \dots, N-1.$$

Diese Gleichungen können wir anders hinschreiben. Wenn wir die Differenzen

$$D_n := x_n - x_{n-1}$$

für $n = 1, \dots, N$ betrachten, dann ist $D_N = x_N - x_{N-1} = 1$ und:

$$D_n = D_{n+1} + 2.$$

Damit ist $D_N = 1, D_{N-1} = 3, D_{N-2} = 5, D_{N-3} = 7, \dots, D_{N-n} = 2n + 1, \dots, D_1 = 2N - 1$ oder in einer "normalen" Reihenfolge

$$D_n = 1 + 2(N - n) \quad \text{für alle } n = 1, 2, \dots, N$$

Wie bekommen wir aus den D_n -Werten die uns eigentlich interessierenden x_n -Werte? Dazu beobachten wir:

$$D_1 + D_2 + \dots + D_n = (x_1 - x_0) + (x_2 - x_1) + \dots + (x_n - x_{n-1}) = x_n - x_0 = x_n.$$

Also ist

$$\begin{aligned} x_n &= \sum_{k=1}^n D_k = \sum_{k=1}^n (1 + 2(N - k)) = 2N \cdot n + n - 2 \cdot \sum_{k=1}^n k \\ &= 2N \cdot n + n - 2 \cdot \frac{n(n+1)}{2} = n \cdot (2N - n). \end{aligned}$$

□

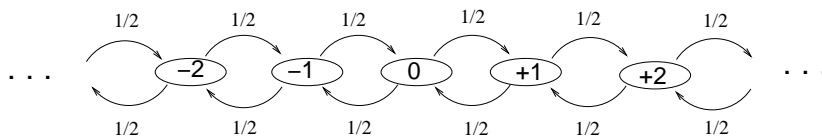
³⁴Mit diesem Modell kann man einen randomisierten Algorithmus entwerfen, der die sogenannte 2-SAT Problem in quadratischer Zeit löst.

³⁵Strenggenommen müssten wir zunächst zeigen, dass die Erwartungswerte endlich sind, damit wir solche Gleichungen schreiben können, aber wir wollen uns diese kleine mathematische Nachlässigkeit gönnen.

Irrfahrten in \mathbb{Z}^d

”A drunk man will find his way home, but a drunk bird may get lost forever”
- Shizuo Kakutani

Zuerst betrachten wir den Fall $d = 1$. In einem Teich befinden sich Steine $\dots, -2, -1, 0, 1, 2, \dots$ in einer Reihe. Ein Frosch sitzt anfänglich auf Stein 0. Dann beginnt er mit gleicher Wahrscheinlichkeit $1/2$ entweder nach rechts oder nach links springen.



Frage 1: Mit welcher Wahrscheinlichkeit wird der Frosch nach n Sprüngen vom Anfangsstein 0 um mindestens t Steine entfernt sein?

Wenn wir die Sprünge durch $+1$ (nach rechts) und -1 (nach links) bezeichnen, dann besteht unser Wahrscheinlichkeitsraum aus allen Worten der Länge n über dem Alphabet $\{-1, +1\}$, und jedes Wort dieselbe Wahrscheinlichkeit $1/2^n$ hat. Sei nun X der Stein, auf dem sich der Frosch nach n Sprüngen befindet. Dann ist

$$X = X_1 + X_2 + \dots + X_n$$

wobei

$$X_i = \begin{cases} +1 & \text{falls der Frosch im } i\text{-ten Schritt nach rechts springt} \\ -1 & \text{sonst} \end{cases}$$

Da $E[X_i] = (-1) \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$ für alle i , gilt $E[X] = \sum_{i=1}^n E[X_i] = 0$. Also ist die *erwartete* Entfernung nach beliebig viel Sprüngen gleich Null. Aber das sagt uns nicht die ganze Wahrheit: Es ist doch klar, dass zum Beispiel nach einem Sprung wird der Frosch um 1 von Null entfernt sein. Das ist noch ein Beispiel dafür, dass der Erwartungswert allein uns überhaupt nichts sagt. Wir müssen die Abweichungswahrscheinlichkeit von diesem Wert bestimmen!

Die Entfernung vom Stein 0 ist durch Zufallsvariable $|X|$ gegeben. Wir wollen also die Wahrscheinlichkeit dafür, dass $|X|$ größer als eine gegebene Zahl t ist, bestimmen. Tschebyschew's Ungleichung gibt uns

$$\Pr\{|X| \geq t\} = \Pr\{|X - E[X]| \geq t\} \leq \frac{\text{Var}[X]}{t^2}$$

wobei

$$\begin{aligned}
 \text{Var}[X] &= E[X^2] - E[X]^2 && \text{(Definition von Var}[X]) \\
 &= E[X^2] && \text{(da } E[X] = 0) \\
 &= E\left[\left(\sum_{i=1}^n X_i\right)^2\right] \\
 &= E\left[\sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j\right] \\
 &= \sum_{i=1}^n E[X_i^2] + \sum_{i \neq j} E[X_i X_j] && \text{(Linearität des Erwartungswertes)} \\
 &= \sum_{i=1}^n 1 + \sum_{i \neq j} 0 && (X_i^2 = 1 \text{ und } E[X_i X_j] = E[X_i] E[X_j] = 0) \\
 &= n.
 \end{aligned}$$

Also gilt für jedes $\alpha > 0$

$$\Pr\{|X| \geq \alpha \sqrt{n}\} \leq \frac{n}{(\alpha \sqrt{n})^2} = \frac{1}{\alpha^2}.$$

Zum Beispiel, wenn der Frosch $n = 10^6$ (eine Million!) Sprünge macht, dann wird er nur mit Wahrscheinlichkeit $\leq 0,01$ um mehr als 10.000 Steine von ursprünglichem Stein 0 entfernt sein.³⁶

Frage 2: Mit welcher Wahrscheinlichkeit wird der Frosch zurück zum ursprünglichen Stein 0 kommen?

Während seiner Sprungerei wird der Frosch gelegentlich den Ursprungsstein 0 besuchen; wir bezeichnen das als ‘Hausbesuch’.

Behauptung 4.96. Sei p die Wahrscheinlichkeit, dass der Frosch sein Haus *nicht* mehr wieder besucht. Ist $p > 0$, so ist die *erwartete* Anzahl der Hausbesuche gleich $1/p$.

Beweis. Jedes Mal wenn der Frosch sein Haus (Stein 0) verlässt, wird er mit Wahrscheinlichkeit p nie mehr zurück kommen, und mit Wahrscheinlichkeit $1 - p$ wird er doch zurück kommen. Falls er zurück kommt, dann ist er wieder in dieselben Situation. Wir können also den Prozess als eine geometrische Verteilung mit der Erfolgswahrscheinlichkeit p betrachten,³⁷ und wie wir bereits wissen, ist dann die erwartete Anzahl der Versuche bis zum ersten Erfolg gleich $1/p$. \square

Sei Y_n die Anzahl der Hausbesuche in der ersten $2n$ Schritten.

Behauptung 4.97.

$$E[Y_n] = \Theta(\sqrt{n})$$

³⁶Wir nehmen also $\alpha = 10$.

³⁷‘Erfolg’ hier ist ‘nie mehr zurück’. Misserfolg ist also ein Hausbesuch.

Beweis. Sei A_t das Ereignis, dass der Frosch zum Zeitpunkt $2t$ (d.h. nach *genau* $2t$ Sprünge) sein Haus besucht, und sei X_t die Indikatorvariable für A_t . Dann gilt: $Y_n = X_1 + X_2 + \dots + X_n$. Nach der Linearität des Erwartungswertes gilt:

$$E[Y_n] = E\left[\sum_{t=1}^n X_t\right] = \sum_{t=1}^n E[X_t] = \sum_{t=1}^n \Pr\{A_t\}.$$

Für jedes $t = 1, \dots, n$ besteht unser Wahrscheinlichkeitsraum aus allen endlichen Worten $a = (a_1, \dots, a_m)$ über dem Alphabet $\{-1, +1\}$; eine $+1$ bzw. -1 in i -ter Position bedeutet, dass der Frosch im i -ten Schritt nach rechts bzw. nach links springt. Das Ereignis A_t besteht aus allen Worten $a = (a_1, \dots, a_m)$ der Länge $m = 2t$ mit der Eigenschaft, dass die Summe $a_1 + a_2 + \dots + a_m$ gleich 0 ist. Oder anders gesagt, A_t enthält alle Worte der Länge $2t$ mit genau t Buchstaben $+1$. Deshalb ist die Wahrscheinlichkeit $\Pr\{A_t\}$ nach der Stirling-Formel³⁸ asymptotisch gleich:

$$\Pr\{A_t\} = \frac{\binom{2t}{t}}{2^{2t}} \sim \frac{1}{\sqrt{\pi t}}.$$

Somit haben wir

$$E[Y_n] \sim \sum_{t=1}^n \frac{1}{\sqrt{\pi t}} = \frac{1}{\sqrt{\pi}} \sum_{t=1}^n \frac{1}{\sqrt{t}} = \Theta(\sqrt{n}).$$

Die letzte Abschätzung kann man mit Integral-Kriterium (siehe Satz 3.11) zeigen. Sei $f(x) = 1/\sqrt{x}$ und $F(x) = 2\sqrt{x}$. Da

$$F'(x) = 2 \cdot \frac{1}{2} x^{\frac{1}{2}-1} = 1/\sqrt{x} = f(x)$$

gilt, ist $F(x)$ eine Stammfunktion für $f(x)$. Da $f(x)$ monoton fallend (mit wachsendem x) ist, liefert uns das Integral-Kriterium die Abschätzung:

$$2\sqrt{n+1} - 2 = F(n+1) - F(1) \leq \sum_{t=1}^n \frac{1}{\sqrt{t}} \leq F(n) - F(0) = 2\sqrt{n}$$

□

Da $E[Y_n] = \Theta(\sqrt{n})$, strebt mit $n \rightarrow \infty$ die erwartete Anzahl $\Theta(\sqrt{n})$ der Hausbesuche gegen Unendlich. Die Behauptung 4.96 sagt uns, dass in diesem Fall $p = 0$ gelten muss. D.h. in diesem Fall wird der Frosch mit Wahrscheinlichkeit 1 sein Haus irgenwanmal besuchen.

Was passiert aber, wenn wir die Steine nicht in eine Linie (\mathbb{Z}) sondern als ein Gitter (\mathbb{Z}^2) anordnen? Einfachheit halber nehmen wir an, dass dann der Frosch sich in jedem Schritt entlang der beiden Achsen (x - und y -Achse) bewegen kann. In anderen Worten kann man die Situation mit zwei Fröschen vorstellen, die sich unabhängig von einander entlang der x -Achse und y -Achse springen. Ein Hausbesuch kommt nur dann zustande, wenn *beide* Frösche sich im Punkt $(0,0)$ treffen. In diesem Fall ist

$$E[X_t] = \Pr\{A_t\} = \Theta(1/\sqrt{t}) \cdot \Theta(1/\sqrt{t}) = \Theta(1/t)$$

und somit³⁹

$$E[Y_n] = \sum_{t=1}^n E[X_t] = \Theta(1/1 + 1/2 + \dots + 1/n) = \Theta(\log n).$$

³⁸ $\binom{n}{n/2} \sim 2^{n+1} / \sqrt{2\pi n}$.

³⁹Harmonische Reihe

Da der Erwartungswert $E[Y_n] = \Theta(\log n)$ gegen ∞ strebt, werden auch in diesem Fall die Frösche mit Wahrscheinlichkeit 1 sein Haus wieder besuchen.

Fazit: Ein betrunkenen Frosch wird auf jedem Fall ein Weg nach Hause finden! Wie ist aber mit einem betrunkenen Vogel?

Der Vogel bewegt sich in \mathbb{Z}^3 . In diesem Fall haben wir

$$E[X_t] = \Pr\{A_t\} = \Theta((1/\sqrt{t})^3) = \Theta(1/t^{3/2}).$$

Es ist aber bekannt (siehe Lemma ??), dass die Reihe $\sum_{t=1}^{\infty} t^{-3/2}$ konvergiert. Also gibt es eine Zahl L , so dass $\lim_{n \rightarrow \infty} E[Y_n] < L$. Laut der Behauptung 4.96 ist dann eine $p \geq 1/L > 0$ Chance, dass der Vogel *nie mehr* nach Hause kommt!

4.16 Statistisches Schätzen: Die Maximum-Likelihood-Methode *

Ein Fischteichbesitzer möchte seinen Fischbestand N schätzen. Er markiert dazu einige Fische. In einem späteren Fang findet er dann markierte wie unmarkierte Fische. Der Teichbesitzer überlegt: Der Anteil der markierten Fische im Fang wird vermutlich die Verhältnisse im Teich widerspiegeln. Ist also

$$\begin{aligned} r &= \text{Anzahl der markierten Fische} \\ n &= \text{Anzahl der Fische im Fang} \\ x &= \text{Anzahl der markierten Fische im Fang,} \end{aligned}$$

so ist zu erwarten, dass r/N und x/n einen ähnlichen Wert haben. Dies macht $(rn)/x$ zu einem plausiblen Schätzer für N .

Zu demselben Resultat führt ein allgemeines statistisches Prinzip – bekannt als *Maximum-Likelihood-Prinzip* – das besagt:

Wähle als Schätzer von N diejenige ganze Zahl \hat{N} , für die das beobachtete Ereignis maximale Wahrscheinlichkeit bekommt.

Um das Prinzip in unserem Fall anzuwenden, machen wir die Annahme, dass die Anzahl X der markierten Fische in dem Fang eine hypergeometrisch verteilte Zufallsvariable ist (siehe “Ergänzungen”). Gesucht ist dasjenige N , das

$$L_x(N) = \binom{r}{x} \binom{N-r}{n-x} / \binom{N}{n}$$

(die Statistiker sprechen von der *Likelihoodfunktion*) maximiert. Da

$$\binom{a}{b} / \binom{a-1}{b} = \frac{(a)_b}{b!} \cdot \frac{b!}{(a-1)_b} = \frac{a}{a-1} \cdot \frac{a-1}{b-2} \cdot \frac{a-2}{b-3} \cdots \frac{a-b+1}{a-b} = \frac{a}{a-b}$$

eine einfache Rechnung ergibt

$$\frac{L_x(N-1)}{L_x(N)} = \frac{N^2 - Nr - Nn + Nx}{N^2 - Nr - Nn + nr}.$$

Daher gilt $L_x(N-1) \leq L_x(N)$ genau dann, wenn $Nx \leq nr$. Die Likelihoodfunktion $L_x(N)$ wächst also für kleine Werte und fällt für große Werte von N . Der Wechsel findet bei nr/x statt. Als Maximum-Likelihood-Schätzer von N erhalten wir

$$\hat{N} = \left\lfloor \frac{nr}{x} \right\rfloor.$$

Die allgemeine Situation ist folgende. Sei $X : \Omega \rightarrow S$ eine Zufallsvariable mit einer endlichen Menge S (mit endlichem Wertebereich S), und die Verteilung von X beinhalte einen unbekanntem (möglicherweise mehrdimensionalen) Parameter a , deren Werte in einer uns bekannten Menge A liegen.⁴⁰ D.h. wir haben die Wahrscheinlichkeiten

$$\Pr_a \{X = x\} \quad \text{für alle } x \in S \text{ und alle } a \in A.$$

Lesen wir diese Wahrscheinlichkeiten als Funktionen von $a \in A$ für ein festes $x \in S$ (das als beobachteter Wert der Zufallsvariablen X interpretiert wird), dann haben wir die *Likelihood-Funktion* (zur Beobachtung x), d.h.,

$$L_x(a) = \Pr_a \{X = x\} \quad \text{für alle } a \in A$$

In der Likelihood-Methode will man einen Wert $\hat{a} \in A$ finden, der die Funktion $L_x(a)$ maximiert,

$$L_x(\hat{a}) = \max_{a \in A} L_x(a).$$

Intuitiv ist die Methode ganz vernünftig: Wir probieren einen solchen Wert \hat{a} des (unbekannten) Parameters a zu finden, der mit größter Wahrscheinlichkeit den beobachteten Wert x der Zufallsvariablen X erzeugt hat.

Man kann die allgemeine Situation anschaulich als eine Matrix M mit $m = |A|$ Zeilen und $n = |S|$ Spalten vorstellen: In Zeile zu $a \in A$ und Spalte zu $x \in S$ steht die Wahrscheinlichkeit $\Pr_a \{X = x\}$. Ist der beobachtete Wert x der Zufallsvariable X vorhanden, so sucht man den größten Wert in der Spalte zu x .

► *Beispiel 4.98*:⁴¹ Ein Spieler spielt mit einer fairen Münze oder mit einer präparierten Münze. Bei der präparierten Münze ist die Wahrscheinlichkeit 0.75 für den Ausgang „Wappen“. Schließlich sei bekannt, dass der Spieler zu Anfang die faire Münze mit Wahrscheinlichkeit 0.5 wählt und die benutzte Münze nicht mehr wechselt.

Wie kann man von dem Spielverlauf auf die benutzten Münzen zurückschließen? Wir nehmen an, dass der Spieler seine Münze *nicht wechselt*. Wir haben also einen (uns unbekanntem) Parameter a , der in der Menge $A = \{\text{fair, präpariert}\}$ liegt. Ein Spiel ist eine Reihenfolge (X_1, \dots, X_n) von Bernoulli-Experimenten, wobei für alle $i = 1, \dots, n$ entweder $\Pr \{X_i = 1\} = 1/2$ (falls der Spieler mit einer fairen Münze spielt) oder $\Pr \{X_i = 1\} = 3/4$ (falls der Spieler mit einer präparierten Münze spielt) gilt. Dementsprechend ist die Zufallsvariable $X = X_1 + X_2 + \dots + X_n$ entweder binomial $B(n, 1/2)$ -verteilt oder binomial $B(n, 3/4)$ -verteilt. Damit sind uns die Wahrscheinlichkeiten $\Pr_a \{X = x\}$ für alle $x \in S = \{0, 1, \dots, n\}$ und $a \in A = \{\text{fair, präpariert}\}$ bekannt. Wenn wir n Münzwürfe mit dem Resultat $x \in S$ (x -maligem Auftreten des Wappens) beobachten, dann erhalten wir

$$\Pr \{X = x \mid \text{faire Münze}\} = \binom{n}{x} \left(\frac{1}{2}\right)^n = \frac{1}{2^n} \binom{n}{x}$$

⁴⁰Zum Beispiel kann A eine Menge der Wahrscheinlichkeitsverteilungen sein.

⁴¹Siehe auch Beispiel 4.68.

und

$$\Pr \{X = x \mid \text{unfaire Münze}\} = \binom{n}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{n-x} = \frac{3^x}{4^n} \binom{n}{x}.$$

Damit ist

$$L_x(a) = \frac{1}{2^n} \binom{n}{x}, \quad \text{falls } a = \text{fair}$$

und

$$L_x(a) = \frac{3^x}{4^n} \binom{n}{x}, \quad \text{falls } a = \text{präpariert}$$

Da

$$\frac{1}{2^n} > \frac{3^x}{4^n} \iff 2^n > 3^x \iff x < \frac{n}{\log_2 3},$$

wählen wir “der Spieler hat wahrscheinlich die faire Münze benutzt” falls $x < n/(\log_2 3)$, da dann $L_x(\text{fair}) > L_x(\text{präpariert})$; und wählen “der Spieler hat wahrscheinlich die präpariert Münze benutzt”, falls $x > n/(\log_2 3)$.

- ▷ *Beispiel 4.99*: Ein Spieler spielt mit einer möglichst präparierten Münze. Sei $a > 0$ die (uns unbekannte) Wahrscheinlichkeit für den Ausgang ‘Wappen’. Sei außerdem X die Anzahl der Würfe, bis ein Wappen kommt. Wir bitten den Spieler die Münze mehrmals zu werfen, bis ein Wappen kommt. Er macht das, und sei x die Anzahl Würfe mit dem Ausgang ‘Kopf’. Die Likelihood-Funktion (zur Beobachtung x) ist

$$L_x(a) = \Pr_a \{X = x\} = a(1-a)^x \quad \text{für alle } a \in A = (0, 1]$$

Wir müssen also den Maximum von $L_x(a)$ über alle $a \in (0, 1]$ bestimmen. Dazu leiten wir diese Funktion nach a ab:

$$L_x(a)' = (1-a)^x - ax(1-a)^{x-1} = (1-a)^{x-1}[1 - a(x+1)]$$

Dann ist $L_x(a)' = 0$ genau dann, wenn $a = 1$ oder $a = \frac{1}{x+1}$ gilt. Wegen $L_x(1) = 0 < L_x(\frac{1}{x+1})$ (falls⁴² $x \geq 1$) kann $a = 1$ als globale Maximalstelle ausgeschlossen werden. Aufgrund von

$$a < \frac{1}{x+1} \iff 1 - a(x+1) > 0$$

hat $L_x(a)'$ im Punkt $\hat{a} := \frac{1}{x+1}$ einen Vorzeichenwechsel von Minus nach Plus, und \hat{a} ist globale Maximalstelle von L_x .

Wenn die Zufallsvariablen X_1, \dots, X_n als stochastisch unabhängig vorausgesetzt werden (wie im Fall eines Zufallsexperiments, das aus n unabhängigen Einzelexperimenten besteht), dann haben wir

$$\Pr_a \{X_1 = x_1, \dots, X_n = x_n\} = \prod_{i=1}^n \Pr_a \{X_i = x_i\}$$

und die Likelihood-Funktion (zu gegebenen $x = (x_1, \dots, x_n)$) ist gleich

$$L_x(a) = \prod_{i=1}^n \Pr_a \{X_i = x_i\}.$$

⁴²Wenn $x = 0$ (Wappen ist nach dem ersten Wurf erschienen), dann $\hat{a} = 1 = \frac{1}{x+1}$ auch die globale Maximalstelle von L_x ist.

Um das Maximierungsproblem zu vereinfachen, betrachtet man statt $L(a)$ die Funktion (die *Log-Likelihood-Funktion*)

$$\ell(a) = \ln L(a)$$

(wenn im Voraus bekannt ist, dass der Fall $L(a) = 0$ nicht auftritt). Die Log-Likelihood-Funktion hat oftmals eine einfachere Gestalt als die Likelihood-Funktion, z.B. werden Produkte zu Summen, allerdings um den Preis, dass jetzt Logarithmen auftreten.

4.17 Die probabilistische Methode*

Bis jetzt haben wir die Stochastik als eine Theorie betrachtet, die uns manche “reellen” Modelle mit Zufall analysieren lässt. Es gibt aber auch eine andere Seite der Stochastik: Man kann mit ihrer Hilfe Aussagen auch in Situationen, wo Zufall keine Rolle spielt, treffen!

Die Hauptidee der sogenannten *probabilistischen Methode* ist folgende: Will man die *Existenz* eines Objekts mit bestimmten Eigenschaften zeigen, so definiert man einen entsprechenden Wahrscheinlichkeitsraum und zeigt, dass ein zufällig gewähltes Element mit *positiver Wahrscheinlichkeit* die gewünschte Eigenschaft hat.

Ein Prototyp dieser (sehr mächtigen) Methode ist das folgende “Mittel-Argument”:

Sind $x_1, \dots, x_n \in \mathbb{R}$ und die Ungleichung

$$\frac{x_1 + \dots + x_n}{n} \geq a \quad (4.18)$$

gilt, so muss es mindestens ein j mit

$$x_j \geq a \quad (4.19)$$

geben.

Die Nützlichkeit dieses Argument liegt in der Tatsache, dass es oft viel leichter ist, die Ungleichung (4.18) zu beweisen, als ein x_j , für das die Ungleichung (4.19) gilt, zu finden.

Wir demonstrieren die probabilistische Methode auf ein paar Beispielen.⁴³

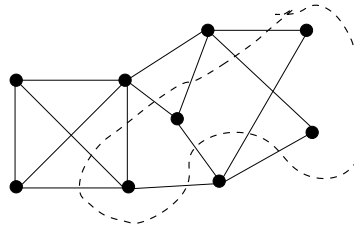
Sei $G = (V, E)$ ein ungerichteter Graph und $S \subseteq V$ eine Menge seiner Knoten. Dann ist S eine *Clique*, falls je zwei Knoten $u \neq v \in S$ mit einer Kante aus E verbunden sind. Sind *keine* zwei Knoten in S mit einer Kante aus E verbunden, so heißt S *unabhängige Menge*.

Gibt es Graphen mit nur sehr kleinen Cliques wie auch mit nur sehr kleinen unabhängigen Mengen? Solche Graphen nennt man *Ramsey-Graphen*.

Satz 4.100. (Erdős 1947) Es gibt Graphen mit n Knoten, die weder Cliques noch unabhängige Mengen der Größe $2 \log n$ besitzen.

Beweis. Um die Existenz solchen (sehr merkwürdigen!) Graphen zu beweisen, betrachten wir Zufallsgraphen über der Knotenmenge $V = \{1, \dots, n\}$: Wir werfen für jede potentielle Kante $\{u, v\}$ eine faire Münze und setzen die Kante ein, wenn das Ergebnis “Wappen” ist.

⁴³Mehr (zum Teil überraschenden) Anwendungen kann man in dem Buch *The Probabilistic Method* von Noga Alon und Joel Spencer finden.

Abbildung 4.6: Eine Clique S_1 der Größe 4 und eine unabhängige Menge S_2 der Größe 4

Wir fixieren eine Knotenmenge $S \subseteq V$ der Größe k und sei A_S das Ereignis “ S ist eine Clique oder eine unabhängige Menge”. Es ist

$$\Pr \{A_S\} = 2 \cdot 2^{-\binom{k}{2}},$$

denn entweder ist S eine Clique und alle $\binom{k}{2}$ Kanten sind vorhanden oder S ist eine unabhängige Menge und keine der $\binom{k}{2}$ Kanten ist vorhanden. Wir sind vor Allem an der Wahrscheinlichkeit p_k interessiert, dass ein Zufallsgraph G eine Clique der Größe k oder eine unabhängige Menge der Größe k besitzt. Da wir nur $\binom{n}{k}$ k -elementigen Mengen $S \subseteq V$ haben, gilt:

$$p_k \leq \binom{n}{k} \cdot 2 \cdot 2^{-\binom{k}{2}} < \frac{n^k}{k!} \cdot \frac{2 \cdot 2^{k/2}}{2^{k^2/2}}.$$

Wir setzen $k = 2 \cdot \log_2 n$ und erhalten $n^k = 2^{k^2/2}$. Da andererseits $2 \cdot 2^{k/2} < k!$ für $k \geq 4$, folgt somit $p_k < 1$ für $k \geq 4$: Es gibt somit Graphen, die nur Cliques oder unabhängige Mengen der Größe höchstens $2 \log_2 n - 1$ besitzen. Wir haben somit die Existenz eines Objekts durch Zählen nachgewiesen. \square

Bis heute sind keine *expliziten Graph-Konstruktionen* bekannt, die vergleichbare Ergebnisse liefern! Andererseits, liefert uns Satz 4.100 keinen *expliziten* Ramsey-Graphen – er zeigt lediglich nur die *Existenz* solchen Graphen. Bis heute ist es nur gelungen, explizite Graphen zu konstruieren, die weder Cliques noch unabhängige Mengen der Größe \sqrt{n} besitzen.

Als unser nächstes Beispiel betrachten wir den klassischen Satz von Turán. Ein Gremium V besteht aus n Personen. Das Problem ist, dass manche von ihnen einander kennen. Um die Entscheidung nicht durch Bekanntschaften beeinflussen, will man eine möglichst große Teilmenge $S \subseteq V$ der Personen zu finden, so dass keine zwei Personen aus S sich kennen.

Dieses Problem kann man mit Hilfe von Graphen darstellen. Wir haben einen ungerichteten Graph $G = (V, E)$, wobei Kanten den Bekanntschaften entsprechen. Man will eine möglichst große unabhängige Menge $S \subseteq V$ finden.

Das berühmte Satz⁴⁴ von Turán sagt, dass *jeder* Graph mit n Knoten und $|E| \leq nk/2$ Kanten eine unabhängige Menge S der Größe $|S| \geq n/(k+1)$ enthalten muss. Der Beweis ist nicht trivial. Andererseits, man kann “fast” dasselbe Resultat mit Hilfe der Zufall sehr leicht beweisen.

Satz 4.101. Jeder Graph $G = (V, E)$ mit $|V| = n$ Knoten und $|E| \leq nk/2$ Kanten muss eine unabhängige Menge S der Größe $|S| \geq n/2k$ enthalten

⁴⁴Dieser Satz hat die ganze *Extremale Graphentheorie* initiiert.

Beweis. Um die Menge S zu finden, konstruieren wir zuerst eine *zufällige* Menge $U \subseteq V$: Wir nehmen eine Münze, für die das Ergebnis Wappen mit Wahrscheinlichkeit p kommt (p ist ein Parameter, den wir später bestimmen wollen). Wir werfen dann für jeden potentiellen Knoten $i \in V$ eine faire Münze und nehmen diesen Knoten i in U , wenn das Ergebnis Wappen ist.

Für jeden Knoten $i \in V$ sei X_i die Indikatorvariable für das Ereignis " $i \in U$ " (i ist gewählt). Für jede Kante $e = \{i, j\} \in E$ sei Y_e die Indikatorvariable für das Ereignis " $i \in U$ und $j \in U$ " (beide Endknoten von e sind gewählt). Die Summe $X := \sum_{i \in V} X_i$ ist dann die Anzahl $|U|$ der gewählten Knoten und $Y := \sum_{e \in E} Y_e$ ist die Anzahl der Kanten, die in die Menge U gewählten Knoten verbinden. Da für jeden Knoten $i \in V$ gilt: $E[X_i] = \Pr\{i \in U\} = p$, und für jede Kante $e = \{i, j\}$ gilt: $E[Y_e] = \Pr\{i \in U \text{ und } j \in U\} = p^2$, haben wir (nach der Linearität des Erwartungswertes)

$$E[X] = \sum_{i \in V} E[X_i] = p \cdot |V| = pn$$

und

$$E[Y] = \sum_{e \in E} E[Y_e] = |E|p^2 \leq \frac{nk}{2}p^2.$$

Die Linearität des Erwartungswertes liefert uns wieder:

$$E[X - Y] \geq np - \frac{nk}{2}p^2.$$

Um diesen Ausdruck zu maximieren, wählen wir $p = 1/k$, und erhalten

$$E[X - Y] \geq \frac{n}{k} - \frac{n}{2k} = \frac{n}{2k}.$$

Also muss es mindestens eine Auswahl der Menge $U \subseteq V$ geben, für die $X - Y \geq n/2k$ gilt. Dies bedeutet, dass die Menge U um mindestens $n/2k$ *mehr* Knoten als Kanten enthält. Wir können also diese Kanten vernichten indem wir aus jeder solchen Kante einen Endknoten einfach aus der Menge U entfernen. Die verbleibende Menge $S \subseteq U$ wird immer noch $|S| \geq n/2k$ Knoten enthalten. Da wir alle Kanten aus U entfernt haben, ist die Menge S auch unabhängig. \square

Ein bipartiter Graph $G = (L \cup R, E)$ heißt ein (n, α, c) -*Expander*, wenn $|L| = |R| = n$ und für *alle* $S \subseteq L$ mit $|S| \leq \alpha \cdot |L|$ gilt $|\Gamma(S)| > c \cdot |S|$, wobei $\Gamma(S)$ die Menge der mit S benachbarten Knoten ist. Ein solcher Graph ist links *d-regulär*, falls jeder Knoten $u \in L$ den Grad d besitzt.

Es ist klar, dass es keinen *d-regulären* (n, α, c) -Expander mit $c > d$ geben kann, da dann $|\Gamma(S)| \leq d|S|$ für *jede* Teilmenge S gilt. Andererseits, braucht man in vielen Anwendungen (wie Tornado Codes) *d-reguläre* (n, α, c) -Expander mit einem Expansionsgrad $c \geq (1 - \epsilon)d$ für ein kleines $\epsilon > 0$. Mit der probabilistischen Methode kann man zeigen, dass solche Graphen auch tatsächlich existieren.

Satz 4.102. Für jedes $d \geq 3$ gibt es links *d-reguläre* $(n, \alpha, d - 2)$ -Expander mit $\alpha = \Omega(1/d^4)$.

Beweis. Sei $G = (L \cup R, E)$ ein durch folgenden Prozess zufällig erzeugter bipartiter Graph mit $|L| = |R| = n$: Zu jedem Knoten $u \in L$ wird zufällig gleichverteilt d Nachbarn von u je mit Wahrscheinlichkeit $1/n$ gewählt.

Für $k \leq \alpha n$ sei

$$p_k = \Pr\{\exists S \subseteq L \text{ mit } |S| = k \text{ und } |\Gamma(S)| < (d - 2)k\}.$$

Um den Satz zu beweisen, reicht es zu zeigen, dass $\sum_{k=1}^{\alpha n} p_k < 1$ gilt.

Fixiere eine beliebige Teilmenge $S \subseteq L$ mit $|S| = k$. Wir stellen uns vor, dass die Knoten in S seine Nachbarn einnacheinander wählen. Die Knoten in S können höchstens dk verschiedene Nachbarn wählen. Deshalb folgt aus $|\Gamma(S)| < (d-2)k$, dass mindestens $2k$ Knoten aus S einen bereits von anderen Knoten in S gewählten Nachbarn wählen müssen. Wir nennen solche Nachbarn *populär*. Die Wahrscheinlichkeit, dass ein Knoten $u \in S$ einen bereits von anderen Knoten aus S gewählten Nachbarn für sich aussucht, ist höchstens⁴⁵

$$\frac{\#\{\text{bereits gewählten Nachbarn}\}}{n} \leq \frac{dk}{n}.$$

Deshalb gilt

$$\Pr \{ \text{es gibt mindestens } 2k \text{ populäre Nachbarn} \} \leq \binom{dk}{2k} \left(\frac{dk}{n} \right)^{2k}.$$

Warum? Da es höchstens $\binom{dk}{2k}$ Möglichkeiten gibt, die $2k$ Knoten (aus höchstens dk von der Knoten aus S wählbaren Nachbarn) als "Kandidaten" für populäre Nachbarn zu markieren, und $(dk/n)^{2k}$ die obere Schranke für die Wahrscheinlichkeit, dass alle diese $2k$ Kandidaten auch tatsächlich populär sein werden, ist. Da wir $\binom{n}{k}$ Möglichkeiten für die Teilmenge S haben, folgt daraus:⁴⁶

$$\begin{aligned} p_k &\leq \binom{n}{k} \binom{dk}{2k} \left(\frac{dk}{n} \right)^{2k} \\ &\leq \left(\frac{en}{k} \right)^k \left(\frac{edk}{2k} \right)^{2k} \left(\frac{dk}{n} \right)^{2k} \\ &= \left(\frac{cd^4 k}{n} \right)^k \end{aligned}$$

mit $c = e^3/4$. Für $\alpha = 1/(cd^4)$ und $k \leq \alpha n$ haben wir also, dass stets $p_k \leq 4^{-k}$ gilt. Deshalb gilt

$$\begin{aligned} \Pr \{ G \text{ ist kein } (n, \alpha, d-2)\text{-Expander} \} &\leq \sum_{k=1}^{\alpha n} p_k \leq \sum_{k=1}^{\alpha n} 4^{-k} \leq \sum_{k=0}^n 4^{-k} - 1 \\ &= \frac{1 - (1/4)^{n+1}}{1 - (1/4)} - 1 \leq \frac{4}{3} - 1 \leq 1/3. \end{aligned}$$

□

⁴⁵Diese Abschätzung ist sehr grob, aber sie reicht uns vollkom aus. Um eine exaktere Abschätzung zu bekommen, konnte man eine ähnliche Analyse wie in Abschnitt 4.12 durchziehen.

⁴⁶Hier benutzen wir die Abschätzungen (siehe Lemma 1.43):

$$\left(\frac{n}{k} \right)^k \leq \binom{n}{k} < \left(\frac{en}{k} \right)^k.$$

4.18 Aufgaben

4.1. Gegeben sind zwei Ereignisse A und B mit $\Pr\{A\} = 0,7$, $\Pr\{B\} = 0,6$ und $\Pr\{A \cap B\} = 0,5$. Berechne:

- | | | |
|-----------------------------------|---|-----------------------------|
| (a) $\Pr\{A \cup B\}$ | (b) $\Pr\{\bar{A}\}$ | (c) $\Pr\{\bar{B}\}$ |
| (d) $\Pr\{\bar{A} \cup \bar{B}\}$ | (e) $\Pr\{\bar{A} \cap \bar{B}\}$ | (f) $\Pr\{A \cap \bar{B}\}$ |
| (g) $\Pr\{\bar{A} \cap B\}$ | (h) $\Pr\{(A \cap \bar{B}) \cup (\bar{A} \cap B)\}$ | |

4.2. Ein Prüfer hat 18 Standardfragen, von denen er in jeder Prüfung 6 zufällig auswählt, wobei jede Auswahl die gleiche Wahrscheinlichkeit besitzt. Ein Student kennt die Antwort von genau 10 Fragen. Wie groß ist die Wahrscheinlichkeit, dass er die Prüfung besteht, wenn er dazu mindestens drei Fragen richtig beantworten muss?

4.3. Von zehn Zahlen sind fünf positiv und fünf negativ. Zwei Zahlen werden zufällig ohne Zurücklegen gezogen und multipliziert. Ist es günstiger, auf ein positives oder ein negatives Produkt zu setzen?

4.4. Peter schlägt Paul ein Spielchen vor: "Du darfst 3 Würfel werfen. Tretten dabei Sechser auf, so hast du gewonnen. Wenn keine Sechser vorkommen, habe ich gewonnen." Paul überlegt rasch, dass für jeden Würfel die Wahrscheinlichkeit $1/6$ beträgt. Die Wahrscheinlichkeit, dass der erste oder der zweite oder der dritte eine Sechse aufweisen ist also $(1/6) + (1/6) + (1/6) = 1/2$. Das Spiel scheint ihm sehr fair zu sein. Würdest Du auch so überlegen?

4.5. (De Méré's Paradox)⁴⁷ Wir würfeln 3 mal einen Spielwürfel und betrachten zwei Ereignisse

$$A = \{\text{die Summe der Augenzahlen ist } 11\}$$

$$B = \{\text{die Summe der Augenzahlen ist } 12\}$$

Bestimme die Wahrscheinlichkeiten $\Pr\{A\}$ und $\Pr\{B\}$. Sind sie gleich?

4.6. Seien A und B zwei *unabhängige* Ereignisse mit $\Pr\{A\} = \Pr\{B\}$ und $\Pr\{A \cup B\} = 1/2$. Bestimme $\Pr\{A\}$.

4.7. Wir sagen, dass ein Ereignis A ist für einen anderen Ereignis B *attraktiv* (bzw. *abstoßend*), falls $\Pr\{B | A\} > \Pr\{B\}$ (bzw. $\Pr\{B | A\} < \Pr\{B\}$) gilt. Zeige folgendes:

- (a) A ist für B attraktiv $\iff B$ ist für A attraktiv.
- (b) A ist für B weder attraktiv noch abstoßend $\iff A$ und B sind unabhängig.
- (c) A ist für B attraktiv $\iff \Pr\{B | A\} > \Pr\{B | \bar{A}\}$.
- (d) A ist für B attraktiv $\implies A$ ist abstoßend für \bar{B} .

4.8. Es werden drei verschiedenfarbige Würfeln geworfen. Betrachte die folgenden Ereignisse:

A: Es fällt mindestens eine 1.

B: Jeder Würfel hat eine andere Augenzahl.

C: Die drei Augenzahlen stimmen überein.

Berechne $\Pr\{A\}$, $\Pr\{B\}$, $\Pr\{C\}$ und $\Pr\{A \cap B\}$. Sind A und B unabhängig?

4.9. Zeige: Wenn $\Pr\{A | B\} = \Pr\{A\}$ ist, dann ist auch $\Pr\{\bar{A} | B\} = \Pr\{\bar{A}\}$

4.10. Wir haben drei Münzen. Eine Münze (die WW-Münze) hat auf beiden Seiten das Wappen, die Zweite (die KK-Münze) hat auf beiden Seiten den Kopf, und die dritte (die WK-Münze) hat das Wappen auf einer und den Kopf auf der anderen Seite. Sie ziehen rein zufällig eine der drei Münzen, werfen diese Münze, und es

⁴⁷Diese Frage hat der französische Edelmann *De Méré* an seinem Freund *Pascal* in 17. Jahrhundert gestellt.

kommt Wappen. (Wir nehmen an, dass (ausser der Markierung) die Münzen fair sind, d.h. jede Seite kann mit gleicher Wahrscheinlichkeit $1/2$ kommen.) Was ist die Wahrscheinlichkeit dafür, dass die WK-Münze gezogen war? *Hinweis:* Die Antwort ist nicht $1/2$.

4.11. Ein Spieler wettet auf eine Zahl von 1 bis 6. Drei Würfel werden geworfen und der Spieler erhält 1 oder 2 oder 3 Euro, wenn 1 bzw. 2 bzw. 3 Würfel die gewettete Zahl zeigen. Wenn die gewettete Zahl überhaupt nicht erscheint, dann muss der Spieler ein Euro abgeben.

Wieviele Euro gewinnt (oder verliert) der Spieler im Mittel pro Spiel? Ist das Spiel fair?

4.12. Karl und Lina führen mit einem Glücksrad ein Spiel durch. Dieses enthält N gleich große Sektoren, die mit den Zahlen 0 bis $N - 1$ beschriftet sind. Man vereinbart, dass derjenige gewonnen hat, der zuerst die Zahl 0 dreht. Dann ist das Spiel beendet. Wir nehmen an, dass Karl beginnt.

Sei $p = 1/N$ die Wahrscheinlichkeit, dass nach einer Drehung das Rad auf 0 stehen bleibt, und sei $q = 1 - p$. Sei K das Ereignis, dass Karl in maximal k Runden gewinnt und sei L das Ereignis, dass Lina in maximal k Runden gewinnt. Um wieviel kleiner sind Linas Gewinnchancen?

4.13. (Russisches Roulette) Beim russischen Roulette steckt in der Trommel einer Pistole genau eine Kugel. Der "Spieler" versetzt die Trommel in Rotation und drückt dann ab. Zwei Gangster duellieren sich dergestalt, dass sie sich aus 2 m gegenseitig beschießen, wobei jeder genau eine Kugel im Magazin hat. Der beleidigte Gangster darf zuerst schießen, dann kommt der jeweils nächste an die Reihe. Vor jedem Schuss wird das Magazin in eine zufällige Umdrehung versetzt.

Wie groß ist die Überlebenschance für den 1. (beleidigten) Gangster bzw. für den 2. Gangster, wenn man 5 Runden vereinbart und das Magazin 8 Kugeln fassen kann, aber nur eine enthält?

4.14. Kurt gewinnt gegen Christian 60 % aller Tennis-Spiele; die Wahrscheinlichkeit, dass er einen einzelnen Satz gewinnt, beträgt somit 0,6.

Mit welcher Wahrscheinlichkeit gewinnt Kurt das Match aus der Serie Best-of-three? (Kurt muss zwei Sätze gewinnen, die Reihenfolge der Siege ist egal, aber wenn Kurt die ersten beiden Sätze gewinnt, dann ist das Spiel vorbei.)

4.15. Es gibt 4 Münzen 1, 2, 3, 4 in einer Kiste. Die Wahrscheinlichkeit, dass die i -te Münze den Kopf ergibt ist $1/i$. Wir wählen eine der vier Münzen rein zufällig und werfen sie solange, bis ein Kopf kommt.

- Gebe einen Wahrscheinlichkeitsraum für dieses Experiment. (Die Summe der Wahrscheinlichkeiten von Elementarereignissen muss 1 sein!)
- Was ist die Wahrscheinlichkeit dafür, dass der Kopf erstmals nach dem *zweiten* Wurf kommt?
- Angenommen, Kopf ist genau nach zwei Wurfen gekommen. Was ist dann die Wahrscheinlichkeit, dass die i -te Münze gewählt war?

4.16. 5 Urnen enthalten verschiedenfarbige Kugeln wie folgt:

Urne	1	2	3	4	5
Anzahl rote	4	3	1	2	3
Anzahl grüne	2	1	7	5	2

Es wird eine beliebige Urne ausgewählt und ihr eine beliebige Kugel entnommen. Mit welcher Wahrscheinlichkeit wurde die erste Urne gewählt unter der Voraussetzung, dass die gezogene Kugel rot war?

4.17. Wir haben zwei Urnen. Die erste Urne enthält 10 Kugeln: 4 rote und 6 blaue. Die zweite Urne enthält 16 rote Kugeln und eine unbekante Anzahl b von blauen Kugeln. Wir ziehen rein zufällig und unabhängig eine Kugel aus jeder der beiden Urnen. Die Wahrscheinlichkeit, dass beide Kugeln dieselbe Farbe tragen sei 0,44. Bestimme die Anzahl b der blauen Kugeln in der zweiten Urne.

4.18. Dreißig Studierende haben sich auf die Klausur in "Mathematische Grundlagen der Informatik" vorbereitet. Die Hälfte dieser 30 haben die Übungsaufgaben regelmäßig bearbeitet, die andere Hälfte aber nicht. Wir nehmen an, die Wahrscheinlichkeit eine Aufgabe in der Klausur zu lösen ist $p = 0,8$ für diejenigen, welche

die Übungsaufgaben bearbeitet haben. Für die anderen sei die Wahrscheinlichkeit immerhin noch $q = 0,6$. Wir nehmen auch an, dass es keine Korrelationen zwischen Fragen gibt: Beantwortet man eine Frage richtig bzw. falsch, so hat das keinen Einfluss auf andere Antworten.

In der Klausur werden 20 Fragen gestellt.⁴⁸ Es besteht, wer mindestens 13 davon richtig löst.

- (a) Wie groß ist die Wahrscheinlichkeit, dass ein zufällig herausgegriffener Studierender diese Klausur besteht?
- (b) Falls Sie einen zufällig gewählten Studierenden treffen und er oder sie gibt an, durchgefallen zu sein, wie groß ist dann die Wahrscheinlichkeit, dass er oder sie die Übungsaufgaben regelmäßig bearbeitet hat?

Hinweis: Seien K das Ereignis, die Klausur zu bestehen; A_i , die i -te Aufgabe richtig zu lösen und U , geübt zu haben. Gegeben sind also die beiden bedingten Wahrscheinlichkeiten $\Pr\{A_i | U\} = p = 0,8$ und $\Pr\{A_i | \bar{U}\} = q = 0,6$ für alle $i = 1, 2, \dots, 20$. Da es keine Korrelationen zwischen Fragen gibt, ist z.B. die Wahrscheinlichkeit, genau k Fragen (richtig) zu beantworten gleich $\binom{20}{k} r^k (1-r)^{20-k}$, wobei $r = p$, falls man geübt hat, und $r = q$ sonst.

4.19. Ein Studentenclub hat $n \geq 2$ Räume. Ein Student sucht dort seine Freundin. Er weiß, dass die Wahrscheinlichkeit, dass die Freundin den Klub besucht, gleich a ist, und die Wahrscheinlichkeit, dass sie in einem bestimmten Raum aufhält, für alle Räume gleich ist. Für $i = 1, \dots, n-1$ sei

- p_i = die Wahrscheinlichkeit, dass er seine Freundin im Raum $i+1$ antrifft, wenn er bereits i Räume erfolglos nach ihr abgesucht hat;
- q_i = die Wahrscheinlichkeit, dass er seine Freundin überhaupt im Klub antrifft, wenn er bereits i Räume erfolglos nach ihr abgesucht hat?

Bestimme diese Wahrscheinlichkeiten. Welche von diesen Wahrscheinlichkeiten sind für wachsendes i wachsend welche sind fallend?

4.20. Die folgende Daten waren einmal in *Wall Street Journal* veröffentlicht. Für eine i Jahre alte Frau in der U.S.A. mit $i \in \{20, 30, 40, 50, 60\}$ ist die Wahrscheinlichkeit, an einer bestimmten Krankheit (in der Zeitung war das Krebs) *in den nächsten 10 Jahren* zu erkranken, gleich 0,5%, 1,2%, 3,2%, 6,4%, 10,8%. Für dieselben Altersgruppen sind die Wahrscheinlichkeiten, *irgendwann* an dieser Krankheit zu erkranken, gleich 39,6%, 39,5%, 39,1%, 37,5%, 34,2%. Das sieht merkwürdig aus! Benutze die vorherige Aufgabe, um diese Daten zu erklären.

4.21. Ein Sortiment von 20 Teilen gilt als "gut", wenn es höchstens 2 defekte Teile enthält, als "schlecht", wenn es mindestens 4 defekte Teile enthält. Weder der Käufer noch der Verkäufer weiß, ob das gegebene Sortiment gut oder schlecht ist. Deshalb kommen sie überein, 4 zufällig herausgegriffene Teile zu testen. Nur wenn alle 4 in Ordnung sind, findet der Kauf (des ganzen Sortiments) statt. Der Verkäufer trägt bei diesem Verfahren das Risiko, ein gutes Sortiment nicht zu verkaufen, der Käufer das Risiko, ein schlechtes Sortiment zu kaufen.

Wer trägt das größere Risiko?

4.22. Von einem Spiel ist bekannt, dass man mit einer Wahrscheinlichkeit von $p = 0,1$ gewinnt. Man spielt so lange, bis man einen Gewinn erzielt. Dann beendet man seine Teilnahme am Spiel.

Wie lange muss man spielen (Anzahl der Spiele), wenn man mit einer Wahrscheinlichkeit von 0,75 einen Gewinn erzielen möchte? *Hinweis:* Geometrische Verteilung.

4.23. Sei $X : \Omega \rightarrow \mathbb{N}$ eine diskrete Zufallsvariable mit $E[X] > 0$. Zeige:

$$\frac{E[X]^2}{E[X^2]} \leq \Pr\{X \neq 0\} \leq E[X].$$

4.24. Zeige, dass die Markov-Ungleichung optimal ist. D.h. für eine natürliche Zahl k finde eine nicht-negative Zufallsvariable X , so dass $\Pr\{X \geq k \cdot E[X]\} = 1/k$ gilt.

⁴⁸Rein hypothetisch ... In Wirklichkeit werden wir viel weniger Fragen stellen.

4.25. Sei $X = \sum_{i=1}^n X_i$ eine Summe von Bernoulli-Variablen. Die Variablen X_i müssen *nicht* unbedingt unabhängig sein! Zeige:

$$E[X^2] = \sum_{i=1}^n \Pr\{X_i = 1\} E[X | X_i = 1].$$

Hinweis: Zuerst zeige, dass $E[X^2] = \sum_{i=1}^n E[X \cdot X_i]$ gilt.

4.26. Ein vereinfachtes Model der Börse geht davon aus, dass in einem Tag eine Aktie mit dem aktuellen Preis a mit Wahrscheinlichkeit p um Faktor $r > 1$ bis zum ar steigen wird und mit Wahrscheinlichkeit $1 - p$ bis zum a/r fallen wird.

Angenommen, wir starten mit dem Preis $a = 1$. Sei X der Preis der Aktie nach d Tagen. Bestimme $E[X]$ und $\text{Var}[X]$.

4.27. Seien X_1, \dots, X_n unabhängige Bernoulli-Variablen mit $\Pr\{X_i = 1\} = p_i$ and $\Pr\{X_i = 0\} = 1 - p_i$. Sei $X = \sum_{i=1}^n X_i \pmod{2}$. Zeige:

$$\Pr\{X = 1\} = \frac{1}{2} \left[1 - \prod_i (1 - 2p_i) \right].$$

Hinweis: Betrachte die Zufallsvariable $Y = Y_1 \cdots Y_n$ mit $Y_i = 1 - 2X_i$. Was ist $E[Y]$?

4.28. Seien $f, g : \mathbb{R} \rightarrow \mathbb{R}$ beliebige Funktionen. Zeige: Sind $X, Y : \Omega \rightarrow \mathbb{R}$ zwei unabhängige Zufallsvariablen, so sind die Zufallsvariablen $f(X)$ und $g(Y)$ unabhängig.

4.29. Seien A_1, \dots, A_n beliebige Ereignisse. Seien $a = \sum_{i=1}^n \Pr\{A_i\}$ und $b = \sum_{i < j} \Pr\{A_i \cap A_j\}$. Zeige

$$\Pr\{\bar{A}_1 \cdots \bar{A}_n\} \leq \frac{a + 2b}{a^2} - 1.$$

Hinweis: Sei X die Anzahl der tatsächlich vorkommenden Ereignisse. Benutze Tschebyschev's Ungleichung, um $\Pr\{X = 0\} \leq a^{-2} E[(X - a)^2]$ zu zeigen.

4.30. Sei $X : \Omega \rightarrow \{0, 1, \dots, M\}$ eine Zufallsvariable und $a = M - E[X]$. Zeige, dass

$$\Pr\{X \geq M - b\} \geq \frac{b - a}{b}$$

für jedes $1 \leq b \leq M$ gilt.

4.31. Für zwei Vektoren $u, v \in \{0, 1\}^n$ ist ihr Skalarprodukt als $\langle u, v \rangle = \sum_{i=1}^n u_i v_i \pmod{2}$ definiert. Sei nun \mathbf{x} ein rein Zufällig gewählter Vektor in $\{0, 1\}^n$.

Zeige: $\Pr\{\langle \mathbf{x}, v \rangle = 1\} = 1/2$ für jedes $v \neq (0, \dots, 0)$, und $\Pr\{\langle \mathbf{x}, v \rangle = \langle \mathbf{x}, w \rangle\} = 1/2$ für alle $v \neq w$.

4.32. Sei X nicht-negative ganzzahlige Zufallsvariable.

Zeige: $E[X^2] \geq E[X]$ und $\Pr\{X = 0\} \geq 1 - E[X]$.

4.33. Ist $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable und $f : \mathbb{R} \rightarrow \mathbb{R}$ eine Funktion, so ist auch $Y := f(X)$ eine Zufallsvariable. Wie sieht ihr Erwartungswert $E[Y]$ aus? Nach der Definition ist $E[Y] = \sum_y y \cdot \Pr\{Y = y\}$. Zeige, dass $E[f(X)] = \sum_x f(x) \cdot \Pr\{X = x\}$ gilt.

4.34. Zeige, dass im Allgemeinen die Divisionsregel $E\left[\frac{X}{Y}\right] = \frac{E[X]}{E[Y]}$ auch für unabhängigen Zufallsvariablen X, Y *nicht* gilt!

4.35. Sei $A \subseteq \Omega$ ein Ereignis und X sei die Indikatorvariable für A , d.h. $X(\omega) = 1$ für $\omega \in A$ und $X(\omega) = 0$ für $\omega \notin A$. Zeige, dass $\text{Var}[X] = \Pr\{A\} \cdot \Pr\{\bar{A}\}$.

4.36. Ein Mann hat n Schlüssel aber nur eine davon passt zu seinem Tür. Der Mann probiert die Schlüssel zufällig. Sei X die Anzahl der Versuche bis der richtige Schlüssel gefunden ist. Bestimme den Erwartungswert $E[X]$, wenn der Mann den bereits ausprobierten Schlüssel

(a) am Bund lässt (also kann er ihn noch mal probieren);

(b) vom Bund nimmt

4.37. Seien \mathbf{A} und \mathbf{B} zwei zufällige Teilmengen der Menge $[n] = \{1, \dots, n\}$ mit $\Pr\{x \in \mathbf{A}\} = \Pr\{x \in \mathbf{B}\} = p$ für alle $x \in [n]$. Ab welchem p können wir nicht mehr erwarten, dass die Mengen \mathbf{A} und \mathbf{B} disjunkt sind?

4.38. Wir verteilen m Bonbons an n Kinder. Jedes der Kinder fängt mit gleicher Wahrscheinlichkeit ein Bonbon. Wie viele Bonbons müssen geworfen werden, bis jedes Kind ein Bonbon gefangen hat?

4.39. Es ist rutschig und wenn das Kind einen Schritt nach vorn versucht, dann kommt es tatsächlich mit einer Wahrscheinlichkeit $2/3$ um einen Schritt nach vorn. Allerdings rutscht es mit Wahrscheinlichkeit $1/3$ einen Schritt zurück. Alle Schritte seien hierbei voneinander unabhängig.

Das Kindergarten sei von dem Kind 100 Schritte weit entfernt. Zeige, dass das Kind nach 500 Schritten mit wenigstens 90% Wahrscheinlichkeit angekommen ist. Schätze hierzu entsprechende Wahrscheinlichkeit mit Hilfe des Satzes von Tschebyshev ab.

4.40. Seien A_1, \dots, A_n beliebige Ereignisse. Seien weiterhin $a = \sum_{i=1}^n \Pr\{A_i\}$ und $b = \sum_{i < j} \Pr\{A_i \cap A_j\}$. Zeige:

$$\Pr\{\bar{A}_1 \cdots \bar{A}_n\} \leq \frac{a + 2b}{a^2} - 1.$$

Hinweis: Sei X die Anzahl der Ereignisse, die tatsächlich vorkommen. Benutze die Tschebyshev-Ungleichung, um zu zeigen, dass $\Pr\{X = 0\} \leq a^{-2} \mathbb{E}[(X - a)^2]$ gilt. Benutze die Linearität des Erwartungswertes, um den rechten Term auszurechnen.

4.41. Wie oft muss eine faire Münze mindestens geworfen werden, damit mit einer Wahrscheinlichkeit von wenigstens $3/4$ die relative Häufigkeit von Kopf von erwarteten Wert $p = 1/2$ um weniger als $0,1$ abweicht?

Kapitel 5

Lineare Algebra

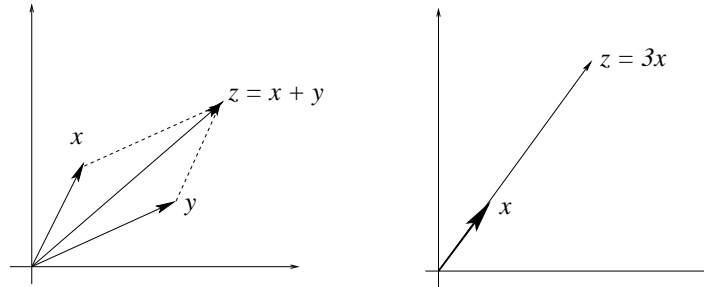
Contents

5.1	Lineare Vektorräume	254
5.2	Basis und Dimension	256
5.3	Skalarprodukt und Norm	258
5.4	Dimensionsschranke und ihre Anwendungen*	260
5.5	Matrizen	263
5.6	Rang einer Matrix	270
5.7	Lösbarkeit der linearen Gleichungssysteme	272
5.8	Gauß-Verfahren	275
5.9	Inversen von Matrizen	279
5.10	Orthogonalität	282
5.11	Determinanten	285
5.12	Eigenwerte und Eigenvektoren	289
5.13	Einige Anwendungen des Matrizenkalküls*	292
5.13.1	Matrizenkalkül und komplexe Zahlen	292
5.13.2	Diskrete Fourier-Transformation	293
5.13.3	Fehlerkorrigierende Codes	298
5.13.4	Expandergraphen	301
5.13.5	Expander-Codes	305
5.13.6	Markov-Ketten	307
5.14	Aufgaben	315

Falls nichts anderes gesagt ist, bezeichnet \mathbb{F} im Folgenden einen *beliebigen* Körper, deren Elemente wir “Zahlen” nennen werden.

5.1 Lineare Vektorräume

Ein Vektor $\mathbf{x} \in \mathbb{F}^n$ ist eine Folge $\mathbf{x} = (x_1, \dots, x_n)$ von Zahlen $x_i \in \mathbb{F}$. Vektoren kann man komponentenweise addieren und mit einer Zahl¹ (oder Skalar) $\lambda \in \mathbb{F}$ multiplizieren: $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \dots, x_n + y_n)$ und $\lambda \mathbf{x} = (\lambda x_1, \dots, \lambda x_n)$.



Nicht jede Teilmenge von Vektoren in \mathbb{F}^n ist unter diesen zwei Operationen abgeschlossen. Abgeschlossene Teilmengen heißen *Vektorräume*. D.h. eine Teilmenge $V \subseteq \mathbb{F}^n$ ist ein *Vektorraum* über den Körper \mathbb{F} , falls folgendes gilt:

- $\mathbf{u} \in V$ und $\lambda \in \mathbb{F} \Rightarrow \lambda \mathbf{u} \in V$;
- $\mathbf{u}, \mathbf{v} \in V \Rightarrow \mathbf{u} + \mathbf{v} \in V$.

Insbesondere ist \mathbb{F}^n ein Vektorraum, und der Nullvektor $\mathbf{0} = (0, \dots, 0)$ ist im *jeden* Vektorraum enthalten (da $0 \cdot \mathbf{u} = \mathbf{0}$ gilt). Deshalb kann man Vektorräume bekommen indem man eine beliebige (nicht leere) Teilmenge $A \subseteq \mathbb{F}^n$ der Vektoren nimmt und die Menge

$$\text{span}(A) = \left\{ \sum_{i=1}^k \lambda_i \mathbf{v}_i : k \in \mathbb{N}, \mathbf{v}_1, \dots, \mathbf{v}_k \in A, \lambda_1, \dots, \lambda_k \in \mathbb{F} \right\}$$

aller endlichen Linearkombinationen von Vektoren in A betrachtet. Man sagt auch, dass $\text{span}(A)$ von den Vektoren in A *erzeugt* (oder *aufgespannt*) Vektorraum ist.



Gilt $V = \text{span}(A)$ für eine *endliche* Menge A , so sagt man, dass V ein *endlich erzeugter* Vektorraum ist. In diesem Kapitel werden wir nur solche Vektorräume betrachten. D.h. unter “Vektorraum” werden wir immer einen endlich erzeugten Vektorraum verstehen!

Satz 5.1. Sei A eine endliche Menge der Vektoren in \mathbb{F}^n . Der von A erzeugte Vektorraum $\text{span}(A)$ bleibt unverändert, wenn man:

1. Einen Vektor \mathbf{v} von A mit einer Zahl $x \neq 0$ multipliziert.
2. Einen Vektor \mathbf{v} von A durch die Summe $\mathbf{v} + \mathbf{u}$ von \mathbf{v} mit einem anderen Vektor $\mathbf{u} \in A$ ersetzt.

Beweis. Zu 1: Den Koeffizient λ zu \mathbf{v} in jeder Linearkombination von A kann man durch λ/x ersetzen.

Zu 2: Die Koeffizienten λ und μ zu \mathbf{v} und \mathbf{u} in jeder Linearkombination von A kann man durch die Koeffizienten λ und $\mu - \lambda$ ersetzen. \square

¹Komponentenweise Multiplikation von \mathbf{x} mit einer Zahl λ entspricht dem “Skalierung” von \mathbf{x} – Verlängerung oder Verkürzung um Faktor λ .

Sei $A = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ ein Erzeugungssystem von V , d.h. $\text{span}(A) = V$. Welche Vektoren (wenn überhaupt) kann man aus A weglassen, ohne die Erzeugendeneigenschaft zu zerstören? Ein Vektor $\mathbf{v}_n \in A$ ist “überflüssig”, wenn $\text{span}(A \setminus \{\mathbf{v}_n\}) = \text{span}(A)$ gilt, also wenn \mathbf{v}_n in $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_{n-1})$ liegt. Dann gibt es also Zahlen $\lambda_1, \dots, \lambda_{n-1}$ mit

$$\mathbf{v}_n = \lambda_1 \mathbf{v}_1 + \dots + \lambda_{n-1} \mathbf{v}_{n-1}$$

bzw.

$$\mathbf{0} = \lambda_1 \mathbf{v}_1 + \dots + \lambda_n \mathbf{v}_n$$

mit $\lambda_n \neq 0$ (nämlich $\lambda_n = -1$). Anders ausgedrückt: Der Nullvektor lässt sich als Linearkombination der \mathbf{v}_i 's darstellen, wobei nicht alle Koeffizienten gleich Null sind. Diese Beobachtung führt uns zum folgenden wichtigen Konzept der linearen Algebra

Definition: Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m \in V$ heißen *linear unabhängig*, wenn aus

$$\lambda_1 \mathbf{v}_1 + \dots + \lambda_m \mathbf{v}_m = \mathbf{0} \quad \text{mit } \lambda_i \in \mathbb{F}$$

stets folgt: $\lambda_1 = \lambda_2 = \dots = \lambda_m = 0$.

Mit anderen Worten, $\mathbf{v}_1, \dots, \mathbf{v}_m \in V$ sind linear unabhängig, wenn der Nullvektor $\mathbf{0} = (0, \dots, 0)$ nur die triviale Darstellung $\mathbf{0} = 0\mathbf{v}_1 + \dots + 0\mathbf{v}_m$ zulässt.

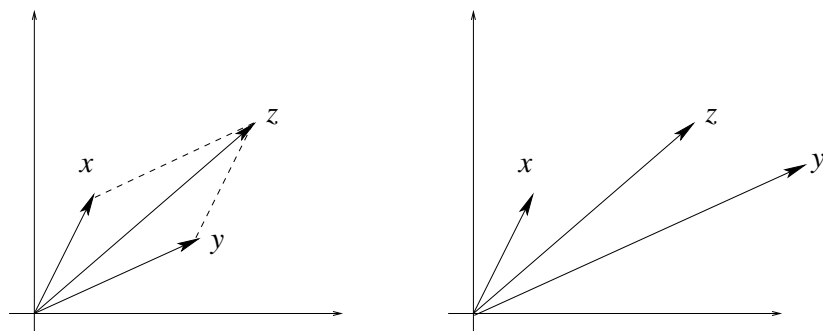


Abbildung 5.1: Vektoren $\mathbf{x}, \mathbf{y}, \mathbf{z}$ sind linear abhängig, da $\mathbf{z} = \mathbf{x} + \mathbf{y}$ gilt. Vektoren $\mathbf{x}, \mathbf{y}', \mathbf{z}$ sind auch linear abhängig, da $\mathbf{z} = \mathbf{x} + \frac{1}{2}\mathbf{y}'$ gilt.

▷ *Beispiel 5.2:* Seien²

$$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Dann sind die Vektoren $\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_4$ linear abhängig, da

$$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{2} \cdot \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Aber die ersten drei Vektoren $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ sind linear unabhängig. (Warum?)

²Vektoren schreibt man entweder waagrecht (als “Zeilen”) oder senkrecht (als “Spalten”).

5.2 Basis und Dimension

Ein Erzeugungssystem A für $V = \text{span}(A)$ ist *minimal*, falls $\text{span}(A \setminus \{\mathbf{v}\}) \neq V$ für alle $\mathbf{v} \in A$ gilt (d.h. wir können die Menge B nicht verkleinern, ohne die Erzeugungseigenschaft zu zerstören). Solche minimale Erzeugungssysteme nennt man *Basisen* von V .

Zum Beispiel die Menge $B = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ mit

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

ist eine Basis von $V = \mathbb{F}^3$; man nennt diese Basis *Standardbasis*.

Es ist klar, dass ein Vektorraum viele verschiedene Basisen besitzen kann. Es ist deshalb interessant, dass nichtdestotrotz *alle diese Basisen denselben Anzahl von Vektoren haben müssen!* Diese wichtige Eigenschaft der Vektorräume liefert uns der folgender Satz.

Satz 5.3. (Basisaustauschsatz von Steinitz) Sei B eine Basis von V und $\mathbf{x} \in V$, $\mathbf{x} \neq \mathbf{0}$. Dann gibt es ein $\mathbf{v} \in B$, so dass $(B \setminus \{\mathbf{v}\}) \cup \{\mathbf{x}\}$ auch eine Basis von V ist.

Beweis. Sei $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ und sei $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{v}_i$ die Darstellung von \mathbf{x} . Da $\mathbf{x} \neq \mathbf{0}$, muss es mindestens ein k mit $\lambda_k \neq 0$ geben. Nehmen wir o.B.d.A. an, dass $k = 1$ ist. Wir wollen zeigen, dass dann $B' = \{\mathbf{x}, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ auch eine Basis von V ist. D.h. wir müssen zeigen dass: (i) $\text{span}(B') = V$, und (ii) B' linear unabhängig ist.

(i) $\text{span}(B') = V$: Aus $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{v}_i$ folgt, dass

$$\mathbf{v}_1 = \frac{1}{\lambda_1} \mathbf{x} - \sum_{i=2}^n \frac{\lambda_i}{\lambda_1} \mathbf{v}_i$$

Da sich somit der entfernte Vektor \mathbf{v}_1 als Linearkombination aus $\text{span}(B')$ darstellen lässt, muss $\text{span}(B)$ in $\text{span}(B')$ enthalten sein. Da offensichtlich auch $\text{span}(B') \subseteq \text{span}(B)$ gilt (\mathbf{x} liegt ja in $\text{span}(B)$), muss die Gleichheit $\text{span}(B') = \text{span}(B)$ und damit auch $\text{span}(B') = V$ gelten.

(ii) $B' = \{\mathbf{x}, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ ist linear unabhängig: Sei

$$\begin{aligned} \mathbf{0} &= \mu_1 \mathbf{x} + \sum_{i=2}^n \mu_i \mathbf{v}_i & (5.1) \\ &= \mu_1 \sum_{i=1}^n \lambda_i \mathbf{v}_i + \sum_{i=2}^n \mu_i \mathbf{v}_i \\ &= \mu_1 \lambda_1 \mathbf{v}_1 + \sum_{i=2}^n (\lambda_1 + \mu_i) \mathbf{v}_i. \end{aligned}$$

Da B linear unabhängig ist, müssen die sämtliche Koeffizienten gleich 0. Insbesondere muss auch $\mu_1 \lambda_1 = 0$ gelten. Da aber $\lambda_1 \neq 0$, muss $\mu_1 = 0$ sein. Dann gilt jedoch (nach (5.1))

$$\mathbf{0} = \sum_{i=2}^n \mu_i \mathbf{v}_i.$$

Da B linear unabhängig ist,³ müssen alle $\mu_i = 0$ sein. Also ist auch B' linear unabhängig. \square

Korollar 5.4. (Dimension) Ist V ein endlich erzeugter Vektorraum, so besteht jede Basis von V aus der gleichen Anzahl von Vektoren.

Diese Zahl heißt *Dimension* von V , in Zeichen $\dim(V)$.

Beweis. Seien $A = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ und $B = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ zwei Basen von V . Nehmen wir an, dass $m < n$ gilt. Nach dem Basisaustauschsz können wir m (oder weniger) Vektoren aus A durch die Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m$ ersetzen, so dass die neue Menge A' immer noch eine Basis ist. Da m kleiner als n ist, muss mindestens ein Vektor $\mathbf{u} \in A \setminus B$ auch in der Menge A' bleiben. Da aber $A' \supseteq B \cup \{\mathbf{u}\}$ und B eine Basis war, gilt $\mathbf{u} \in \text{span}(B)$ und damit kann A' keine Basis sein. Widerspruch. \square

Der nächster Satz erklärt warum eigentlich eine Basis auch "Basis" genannt wird. Genau wie für anderen mathematischen Strukturen, heißen zwei Vektorräume U und W *isomorph*, falls es eine bijektive Abbildung $f : U \rightarrow W$ gibt, so dass $f(\lambda \mathbf{u}) = \lambda f(\mathbf{u})$ und $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$ für alle $\lambda \in \mathbb{F}$ und $\mathbf{u}, \mathbf{v} \in U$ gilt.

Satz 5.5. Sei V ein Vektorraum über einen Körper \mathbb{F} .

1. Ist $B \subseteq V$ eine Basis von V , so lässt sich jedes Vektor $\mathbf{u} \in V$, $\mathbf{u} \neq \mathbf{0}$ auf *genau eine* Weise als Linearkombination der Vektoren aus B darstellen.
2. Ist $d = \dim(V)$, so ist V isomorph zu dem Vektorraum \mathbb{F}^d .

Beweis. Zu 1: Sei $B = \{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ eine Basis und $\mathbf{0} \neq \mathbf{u} \in V$ beliebig. Wegen $V = \text{span}(B)$ gibt es einen Vektor $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{F}^d$ mit $\mathbf{u} = \sum_{i=1}^d x_i \mathbf{v}_i$. Es gelte außerdem $\mathbf{u} = \sum_{i=1}^d y_i \mathbf{v}_i$ mit $(y_1, \dots, y_d) \in \mathbb{F}^d$. Dann ist

$$\mathbf{0} = \mathbf{u} - \mathbf{u} = \sum_{i=1}^d x_i \mathbf{v}_i - \sum_{i=1}^d y_i \mathbf{v}_i = \sum_{i=1}^d (x_i - y_i) \mathbf{v}_i. \quad (5.2)$$

Nun ist B als Basis linear unabhängig, also muss $x_i = y_i$ für alle $i = 1, \dots, d$ gelten.

Zu 2: Sei $B = \{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ eine Basis von V . Wie wir bereits gezeigt haben, gibt es dann für jedes $\mathbf{u} \in V$ genau einen Vektor $f(\mathbf{u}) = (x_1, \dots, x_d) \in \mathbb{F}^d$ mit $\mathbf{u} = \sum_{i=1}^d x_i \mathbf{v}_i$. Damit ist die Abbildung $f : V \rightarrow \mathbb{F}^d$ bijektiv. Ausserdem, für alle $\mathbf{u}, \mathbf{w} \in V$ mit $f(\mathbf{u}) = (x_1, \dots, x_d)$ und $f(\mathbf{w}) = (y_1, \dots, y_d)$ gilt:

$$\mathbf{u} + \mathbf{w} = \sum_{i=1}^d x_i \mathbf{v}_i + \sum_{i=1}^d y_i \mathbf{v}_i = \sum_{i=1}^d (x_i + y_i) \mathbf{v}_i$$

und somit auch

$$f(\mathbf{u} + \mathbf{w}) = (x_1 + y_1, \dots, x_d + y_d) = (x_1, \dots, x_d) + (y_1, \dots, y_d) = f(\mathbf{u}) + f(\mathbf{w}).$$

Da für alle $\lambda \in \mathbb{F}$ die auch Gleichheit $f(\lambda \mathbf{u}) = \lambda f(\mathbf{u})$ (offensichtlich) gilt, ist $f : V \rightarrow \mathbb{F}^d$ ein Isomorphismus. \square

³Und jede (nicht leere) Teilmenge von linear unabhängigen Vektoren muss auch (trivialer Weise) linear unabhängig sein.

Die Zahlen x_i in $f(\mathbf{u}) = (x_1, \dots, x_n)$ heißen *Koordinaten* von Vektor $\mathbf{u} \in V$ bezüglich der Basis B . Beachte, dass die Koordinaten von der gewählten Basis abhängen: Verschiedene Basen erzeugen verschiedene Bijektionen $f : V \rightarrow \mathbb{F}^d$.

Im \mathbb{F}^d mit Standardbasis

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, \mathbf{e}_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

hat man für ein Vektor $\mathbf{u} \in \mathbb{F}^d$

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} = u_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \dots + u_d \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

d.h. in diesem Fall gilt: $f(\mathbf{u}) = \mathbf{u}$.

5.3 Skalarprodukt und Norm

Das *Skalarprodukt* (oder *inneres Produkt*) zweier Vektoren ist die Zahl

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

Zur Schreibweise: Anstatt von $\langle \mathbf{x}, \mathbf{y} \rangle$ schreibt man oft $\mathbf{x} \cdot \mathbf{y}$. Das Skalarprodukt hat folgende Eigenschaften (zeige das!):

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, und $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ genau dann, wenn $\mathbf{x} = \mathbf{0}$.
2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$.
3. $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$.

Die *Länge* (oder *euklidische Norm*) eines Vektors $\mathbf{x} = (x_1, \dots, x_n)$ ist definiert durch:

$$\|\mathbf{x}\| := \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}.$$

Die folgende Ungleichung hat sich als sehr nützlich erwiesen.

Satz 5.6. (Cauchy–Schwarz-Ungleichung) Für alle Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ gilt:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

D.h.

$$\left(\sum_{i=1}^n x_i y_i \right)^2 \leq \left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i^2 \right). \quad (5.3)$$

Beweis. Für jedes $\lambda \in \mathbb{R}$ gilt:

$$\begin{aligned} 0 \leq \langle \lambda \mathbf{x} - \mathbf{y}, \lambda \mathbf{x} - \mathbf{y} \rangle &= \langle \lambda \mathbf{x}, \lambda \mathbf{x} - \mathbf{y} \rangle - \langle \mathbf{y}, \lambda \mathbf{x} - \mathbf{y} \rangle \\ &= \lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle - 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle. \end{aligned}$$

Wir setzen $\lambda := \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$ und erhalten:

$$0 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\langle \mathbf{x}, \mathbf{x} \rangle^2} \langle \mathbf{x}, \mathbf{x} \rangle - 2 \frac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\langle \mathbf{x}, \mathbf{x} \rangle} + \langle \mathbf{y}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \frac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

woraus $\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2$ folgt. \square

Mit dem Skalarprodukt lässt sich auch der Begriff des *Winkels* zwischen zwei Vektoren erklären. Für $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$ zeigt die Cauchy-Schwarz Ungleichung, dass $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$, was wir in der Form

$$-1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \leq 1$$

schreiben können. Es gibt dann genau ein $\alpha \in [0, \pi]$ mit

$$\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \cos(\alpha). \quad (5.4)$$

Man nennt α den Winkel zwischen \mathbf{x} und \mathbf{y} .

Satz 5.7. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Dann gilt:

1. $\|\mathbf{x}\| \geq 0$, und $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$
2. $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (Dreiecksungleichung)
4. $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, falls \mathbf{x} und \mathbf{y} orthogonal sind, d.h. $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ gilt (Pythagoras-Theorem)

Beweis. (1) ist trivial. (2) gilt, denn

$$\|\lambda \mathbf{x}\| = \sqrt{\langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle} = \sqrt{\lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle} \stackrel{(*)}{=} \sqrt{\lambda^2} \cdot \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\lambda| \cdot \|\mathbf{x}\|$$

wobei (*) aus $\sqrt{a^2} = |a|$ für alle $a \in \mathbb{R}$ folgt. Zum (3) und (4):

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle + 2 \langle \mathbf{x}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 \langle \mathbf{x}, \mathbf{y} \rangle \quad (\text{damit ist (4) bewiesen}) \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| \quad (\text{Cauchy-Schwarz Ungleichung}) \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \end{aligned}$$

\square

Diese Eigenschaften charakterisieren unsere Anschauung von Länge oder Abstand, deshalb heißt auch

$$\text{dist}(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$$

der Abstand von \mathbf{x} und \mathbf{y} . Die Dreiecksungleichung für die euklidische Norm liefert und die Dreiecksungleichung für die Abstand:

$$\text{dist}(\mathbf{x}, \mathbf{z}) \leq \text{dist}(\mathbf{x}, \mathbf{y}) + \text{dist}(\mathbf{y}, \mathbf{z}).$$

5.4 Dimensionsschranke und ihre Anwendungen*

Ein wichtiges Korollar aus dem Basisaustauschsatz von Steinitz ist auch die Tatsache, dass keine linear unabhängige Menge in V mehr als $\dim(V)$ Vektoren haben kann.

Korollar 5.8. (Dimensionsschranke) Sind die Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m$ linear unabhängig in V , so gilt $m \leq \dim(V)$.

Auf dieser Eigenschaft der linear unabhängigen Vektoren basiert sich die sogenannte *Methode der linearen Algebra*, die viele Anwendungen in der Diskreten Mathematik gefunden hat. Wir demonstrieren die Methode mit zwei Beispielen.

▷ *Beispiel 5.9 : (“Oddtown”)* Eine kleine Stadt Namens “Eventown”⁴ hat n (erwachsene) Einwohner. Da in so kleinerer Stadt nicht viel los ist, haben die Einwohner eine Aktivität gefunden: sie probieren möglichst viele verschiedene⁵ Clubs (oder Vereine) zu bilden. Da zu viele Clubs schwer zu koordinieren sind, hat das Rathaus eine Regelung herausgegeben:

- (i) die Anzahl der Mitglieder in jedem Club muss *gerade* sein,
- (ii) je zwei Clubs müssen auch *gerade* Anzahl gemeinsamer Mitglieder haben.

Frage: Wieviele Clubs können die Einwohner mit dieser Regelung bilden? Die Antwort ist einfach: falls alle Einwohner verheiratet sind, können sie mindestens $2^{\lfloor n/2 \rfloor}$ Clubs bilden – es reicht, dass jeder Mann auch seine Frau in einer Club mitnimmt. Das ist viel zu viel für eine so kleine Stadt! Um die Ordnung im Stadt wieder herzustellen, ist das Rathaus gezwungen, irgendwie die Anzahl der Clubs drastisch zu reduzieren. Es ist aber nicht erlaubt, die Regelung komplett neu umzuschreiben – erlaubt ist nur *ein einziges Wort* zu ändern. Ist das überhaupt möglich?

Der Bürgermeister hat während seines Studiums die Vorlesung “Mathematische Grundlagen” besucht und hat das Korollar 5.8 gekannt. Deshalb hat er den folgenden Vorschlag gemacht: ersetze einfach das Wort “*gerade*” in (i) durch “*ungerade*”. Er behauptet, dass dann die Einwohner höchstens n verschiedene Clubs bilden können! Das Rathaus war vor diesem Vorschlag so begeistert, das es auch die Name der Stadt von “Eventown” auf “Oddtown” geändert hat. Die Frage ist nun, ob der Bürgermeister Recht hat? Und wenn ja, dann warum?

Das kann man mittels linearer Algebra beweisen. Seien $A_1, \dots, A_m \subseteq \{1, \dots, n\}$ alle mögliche Clubs, die in Oddtown gebildet werden können. Formell sieht die neue Regelung folgender Maßen aus:

- (i') für alle i muss $|A_i|$ *ungerade* sein,
- (ii) für alle $i \neq j$ muss $|A_i \cap A_j|$ *gerade* sein.

Behauptung: $m \leq n$.

Beweis. Sei $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{in}) \in \text{GF}(2)^n$ der Inzidenzvektor für den Club $A_i \subseteq \{1, \dots, n\}$, d.h. $v_{ij} = 1$ genau dann, wenn der j -te Einwohner zum Club A_i gehört. Wenn wir modulo 2

⁴engl. “gerade” = “even”

⁵“Verschiedene” bedeutet hier, dass keine zwei Clubs *dieselben* Mitglieder hat.

rechnen, dann kann man die Regeln (i') und (ii) so umschreiben:

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

Wir wollen zeigen, dass dann die Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m$ linear unabhängig sein müssen: aus der Dimensionsschranke (Korollar 5.8) folgt dann unmittelbar die Ungleichung $m \leq \dim(\text{GF}(2)^n) = n$.

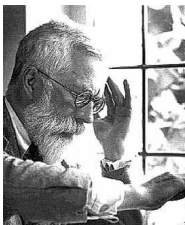
Um die Unabhängigkeit der Vektoren \mathbf{v}_i zu zeigen, sei $\sum_{i=1}^m \lambda_i \mathbf{v}_i = \mathbf{0}$. Dann gilt für jedes $j = 1, \dots, n$:

$$0 = \langle \mathbf{0}, \mathbf{v}_j \rangle = \sum_{i=1}^m \lambda_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \lambda_j \underbrace{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}_{=1} = \lambda_j$$

und damit $\lambda_j = 0$ für alle j . Deshalb sind die Vektoren linear unabhängig und die Ungleichung $m \leq n$ ist bewiesen. \square

Wir geben auch eine mehr respektable Anwendung der Dimensionsschranke an. Der folgender Satz, bekannt als *Fisher-Ungleichung* ist ein der Kernsätze in der sogenannten *Design Theory*. Statistiker R. A. Fisher hat diesen Satz in 1940 für den speziellen Fall, wenn $k = 1$ und alle Mengen gleich groß sind. Später, de Bruijn and Erdős (1948), R. C. Bose (1949) und Majumdar (1953) haben den Satz bis zu folgender allgemeiner Form gebracht. Alle ursprüngliche Beweise dieses Satzes waren echt kompliziert.

Es ist deshalb überraschend, wie einfach kann man diesen Satz mit Hilfe der linearen Algebra beweisen.



Fisher-Ungleichung

Seien A_1, \dots, A_m verschiedene Teilmengen von $\{1, \dots, n\}$ mit der Eigenschaft, dass

$$|A_i \cap A_j| = k$$

für ein festes k und alle $i \neq j$ gilt. Dann gilt: $m \leq n$.

Beweis. Seien $\mathbf{v}_1, \dots, \mathbf{v}_m \in \{0, 1\}^n$ Inzidenzvektoren von A_1, \dots, A_m , d.h. Vektor \mathbf{v}_i hat Einsen in Positionen j mit $j \in A_i$ und sonst Nullen. Beachte, dass die Skalarprodukt $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ nichts anderes als die Kardinalität $|A_i \cap A_j|$ des Durchschnitts von A_i und A_j ist.

Unser Ziel ist zu zeigen, dass die Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m$ linear unabhängig in \mathbb{R}^n sind, dann folgt die Behauptung $m \leq \dim(\mathbb{R}^n) = n$ aus der Dimensionsschranke (Korollar 5.8).

Wir führen einen Widerspruchsbeweis. Nehmen wir an, dass die Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_m$ linear abhängig sind. Dann gibt es reelle Zahlen $\lambda_1, \dots, \lambda_m$ mit $\sum_{i=1}^m \lambda_i \mathbf{v}_i = \mathbf{0}$ und $\lambda_i \neq 0$ für mindestens ein i . Weiterhin gilt

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \begin{cases} |A_i| & \text{falls } i = j \\ k & \text{falls } i \neq j \end{cases}$$

Es folgt:

$$\begin{aligned}
 0 = \langle \mathbf{0}, \mathbf{0} \rangle &= \left(\sum_{i=1}^m \lambda_i \mathbf{v}_i \right) \left(\sum_{j=1}^m \lambda_j \mathbf{v}_j \right) = \sum_{i=1}^m \lambda_i^2 \langle \mathbf{v}_i, \mathbf{v}_i \rangle + \sum_{1 \leq i \neq j \leq m} \lambda_i \lambda_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle \\
 &= \sum_{i=1}^m \lambda_i^2 |A_i| + k \cdot \sum_{1 \leq i \neq j \leq m} \lambda_i \lambda_j \\
 &= \sum_{i=1}^m \lambda_i^2 |A_i| + k \cdot \left(\sum_{i=1}^m \lambda_i \right)^2 - k \cdot \sum_{i=1}^m \lambda_i^2 \\
 &= \sum_{i=1}^m \lambda_i^2 (|A_i| - k) + k \cdot \left(\sum_{i=1}^m \lambda_i \right)^2.
 \end{aligned}$$

Es ist klar, dass $|A_i| \geq k$ für alle i gilt und $|A_i| = k$ für *höchstens ein* i gelten kann (da sonst würde die Eigenschaft $|A_i \cap A_j| = k$ verletzt). Wir wissen, dass nicht alle Zahlen $\lambda_1, \dots, \lambda_m$ gleich Null sind. Sind es mindestens zwei davon ungleich Null, so ist bereits die erste Summe ungleich Null. Ist nur eine von dieser Zahlen ungleich Null, so muss die zweite Summe ungleich Null sein. In beiden Fällen bekommen wir ein Widerspruch. \square

In vielen Anwendungen in der Informatik braucht man einige Objekte mit spezifischen Eigenschaften *explizit* zu konstruieren. Und die lineare Algebra kann hier oft hilfreich sein. Wir demonstrieren dies auf einem Beispiel.

Für einen Graphen $G = (V, E)$ sei $\rho(G)$ die größte Zahl r , so dass der Graph G eine Clique oder eine unabhängige Menge der Grösse r besitzt. Existenz von Graphen auf n Knoten mit $\rho(G) \leq 2 \log n$ haben wir bereits in Abschnitt 4.17 mit Hilfe von sogenannten “Probabilistischen Methode” bewiesen. Aber die *Konstruktion* von solchen (sehr merkwürdigen!) Graphen ist sehr schwierig. Es ist auch schwierig, Graphen mit $\rho(G) \leq n^\epsilon$ mit $\epsilon < 1$, zu konstruieren: Bislang sind nur explizite Graphen G mit $\rho(G) = n^{1/2}$ bekannt. Solche Graphen sind mit Hilfe der sogenannten “Lineare-Algebra-Methode” konstruiert. Um diese Methode zu demonstrieren, werden wir nun einen Graphen mit $\rho(G) = n^{1/3}$ konstruieren.

Sei $n = \binom{t}{3}$ und betrachte den Graphen $G = (V, E)$, deren Knoten alle 3-elementige Teilmengen von $\{1, \dots, t\}$ sind. Zwei Mengen (Knoten) A und B sind adjazent genau dann, wenn $|A \cap B| = 1$ gilt.

Satz 5.10. (Nagy 1972) Der Graph G hat $n = \Theta(t^3)$ Knoten und enthält weder eine Clique noch eine unabhängige Menge mit $t + 1$ Knoten.

Beweis. Sei A_1, \dots, A_m eine Clique in G . Dann gilt $|A_i \cap A_j| = 1$ für alle $1 \leq i \neq j \leq m$. Nach Fisher’s Ungleichung muss dann $m \leq t$ gelten.

Sei nun A_1, \dots, A_m eine unabhängige Menge in G . Dann gilt $|A_i \cap A_j| \in \{0, 2\}$ für alle $1 \leq i \neq j \leq m$. D.h. alle $|A_i| = 3$ sind ungerade und alle $|A_i \cap A_j|$ sind gerade Zahlen. Der Beispiel mit dem Odd-Town sagt uns, dass auch in diesem Fall $m \leq t$ gelten muss. \square

5.5 Matrizen

Endliche Mengen von Vektoren betrachtet man üblicherweise als “Matrizen.” Eine Matrix ist ein rechteckiges Zahlenschema. Eine $m \times n$ Matrix über einem Körper besteht aus m waagrecht verlaufenden Zeilen und n senkrechten verlaufenden Spalten:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

mit $a_{ij} \in \mathbb{F}$. Wenn die Zahlen m und n bekannt sind, schreibt man oft kurz $A = (a_{ij})$ oder⁶ $A = (a_{i,j})$.

Die *transponierte* Matrix von einer $m \times n$ -Matrix $A = (a_{ij})$ ist die $n \times m$ -Matrix $A^T = (c_{ij})$ mit $c_{ij} = a_{ji}$. D.h. man bekommt eine transponierte Matrix A^T indem man die Matrix über die Hauptdiagonale “umkippt”: Zeilen von A sind dann die Spalten von A^T und Spalten von A sind dann die Zeilen von A^T .

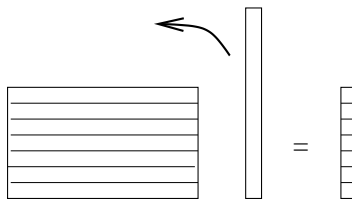
Vektoren $\mathbf{x} \in \mathbb{F}^n$ kann man auch als Matrizen betrachten: entweder als einen Zeilenvektor ($1 \times n$ Matrix) oder als einen Spaltenvektor ($n \times 1$ Matrix).

Eine $n \times n$ *Einheitsmatrix* hat die Form (wenn die Dimension n aus dem Kontext klar ist, werden wir einfach E anstatt E_n schreiben):

$$E_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

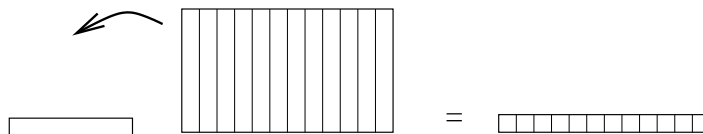
Eine $m \times n$ Matrix $A = (a_{ij})$ kann man mit einem Spaltenvektor \mathbf{x} der Länge n *von rechts* wie auch mit einem Zeilenvektor \mathbf{y} der Länge m *von links* multiplizieren:

- $A\mathbf{x}$ ist ein Spaltenvektor, deren Einträge die Skalarprodukte von \mathbf{x} mit der *Zeilen* von A sind. Anschaulich:



- $\mathbf{y}A$ ist ein Zeilenvektor, deren Einträge die Skalarprodukte von \mathbf{y} mit der *Spalten* von A sind. Anschaulich:

⁶Man trennt die Indizes i und j durch die Komma nur dann, wenn man mögliche Verwechslungen vermeiden will. Will man zum Beispiel das Element in i -ter Zeile und $(n - j)$ -ter Spalte angeben, so schreibt man “ $a_{i,n-j}$ ”, nicht “ a_{in-j} ”.



- Multipliziert man A von beiden Seiten, so bekommt man eine Zahl \mathbf{yAx} , d.h. das Skalarprodukt von \mathbf{yA} und \mathbf{x} :

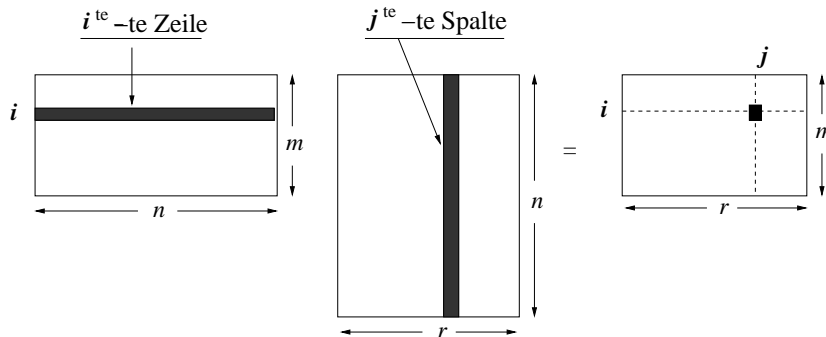
$$\mathbf{yAx} = \langle \mathbf{yA}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{Ax} \rangle = \sum_{i=1}^m \sum_{j=1}^n a_{ij} y_i x_j.$$

Nun schauen wir, wie man zwei Matrizen addieren und multiplizieren kann.

- Die Summe zweier $m \times n$ -Matrizen $A = (a_{ij})$ und $B = (b_{ij})$ ist die Matrix $A + B = (c_{i,j})$ mit $c_{ij} = a_{ij} + b_{ij}$.
- Die λ -Fache ($\lambda \in \mathbb{F}$) von $A = (a_{ij})$ ist die Matrix $\lambda \cdot A = (c_{i,j})$ mit $c_{ij} = \lambda a_{ij}$.
- Wie sollte man zwei Matrizen *multiplizieren*? Hier könnte man verschiedene Definitionen wählen. Wir werden bald sehen, dass in Anwendungen die folgende (auf erstem Blick sehr unnatürliche) Definition von großer Bedeutung ist. Ist $A = (a_{ij})$ eine $m \times n$ -Matrix und $B = (b_{ij})$ eine $n \times r$ -Matrix, so ist ihr Matrixprodukt die $m \times r$ -Matrix $A \cdot B = (c_{i,j})$ mit⁷

$$c_{ij} = \sum_{k=1}^n a_{ik} \cdot b_{kj}.$$

Trotz aller Indices gibt es eine einfache Merkregel zur Berechnung von c_{ij} :



$$\begin{aligned} c_{ij} &= \text{Skalarprodukt der } i\text{-ten Zeile der ersten Matrix} \\ &\quad \text{mit der } j\text{-ten Spalte der zweiten Matrix} \\ &= \sum_{k=1}^n a_{ik} \cdot b_{kj} \end{aligned}$$

⁷Beachte, dass wir *nicht* komponentenweise multiplizieren. D.h. c_{ij} ist nicht als $c_{ij} = a_{ij} \cdot b_{ij}$ definiert!! In diesem Fall hätten wir ein "Traumprodukt" – viele Studenten können nur darüber träumen. Man muss aber sagen, dass (im Gegenteil zur allgemeinen Auffassung, dass das Traumprodukt nutzlos ist) auch diese Definition des Matrixprodukts (auch bekannt als Hadamard-Produkt) interessante Anwendungen (insbesondere in Kombinatorik) haben *kann*.



Daran erkennt man im übrigen sehr deutlich, dass man nicht irgendwelche Matrizen multiplizieren kann: Die Spaltenzahl der ersten Matrix muss mit der Zeilenzahl der zweiten Matrix gleich sein!

Wenn man die Spalten von B als Vektoren $\mathbf{b}_1, \dots, \mathbf{b}_r$ betrachtet, so ist AB die Matrix, deren Spalten die Matrix-Vektor Produkte $A\mathbf{b}_1, \dots, A\mathbf{b}_r$ sind.

Bemerkung 5.11. Der (einziger!) Grund, warum die Multiplikation von Matrizen so “unnatürlich” definiert ist, ist folgender. Eine Abbildung $L : \mathbb{F}^n \rightarrow \mathbb{F}^m$ heißt *linear*, falls für alle $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ und $\lambda, \mu \in \mathbb{F}$ gilt

$$L(\lambda\mathbf{x} + \mu\mathbf{y}) = \lambda L(\mathbf{x}) + \mu L(\mathbf{y}).$$

Man kann zeigen (wir werden das nicht tun), dass es für jede lineare Abbildung $L : \mathbb{F}^n \rightarrow \mathbb{F}^m$ eine $m \times n$ Matrix A mit $L(\mathbf{x}) = A \cdot \mathbf{x}$ gibt. Also sind lineare Abbildungen nichts anderes als Matrix-Vektor Produkte.

Seien nun $h : \mathbb{F}^m \rightarrow \mathbb{F}^k$ und $g : \mathbb{F}^n \rightarrow \mathbb{F}^m$ zwei lineare Abbildungen, und A und B die zur diesen Abbildungen gehörenden $m \times k$ und $n \times m$ Matrizen, d.h. $h(\mathbf{y}) = A\mathbf{y}$ und $g(\mathbf{x}) = B\mathbf{x}$. Dann ist die Komposition $f(\mathbf{x}) = h(g(\mathbf{x}))$ durch das Produkt AB diesen beiden Matrizen definiert: $f(\mathbf{x}) = A(B\mathbf{x}) = (AB)\mathbf{x}$. Das ist eine der wichtigsten Eigenschaften des Matrizenprodukts überhaupt!

Satz 5.12. (Rechenregeln)

Distributivgesetze	$(A + B) \cdot C = A \cdot C + B \cdot C$ $A \cdot (B + C) = A \cdot B + A \cdot C$
Homogenität	$\lambda \cdot (A \cdot B) = (\lambda \cdot A) \cdot B = A \cdot (\lambda \cdot B)$
Assoziativgesetz	$A \cdot (B \cdot C) = (A \cdot B) \cdot C$
Transponieren	$(A + B)^\top = A^\top + B^\top$ $(A \cdot B)^\top = B^\top \cdot A^\top$ (beachte die Reihenfolge!)
Einheitsmatrix	$E_m \cdot A = A$ und $A \cdot E_n = A$

Beweis. Übungsaufgabe. □



Matrizenmultiplikation ist i.A. nicht kommutativ, d.h. $A \cdot B = B \cdot C$ gilt i.A. nicht! Gegenbeispiel:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

► **Beispiel 5.13 : (Unsterbliche Hasen)** Wir betrachten folgendes Modell der Entwicklung einer Hasenpopulation: Im ersten Lebensjahr sind Hasen noch nicht erwachsen. Wenn sie ab dem zweiten Lebensjahr erwachsen sind, dann wirft jedes Hasenpaar jedes Jahr ein Paar Junge. Annahme: Hasen sterben nie!

Frage: Wie entwickelt sich die Hasenpopulation im Lauf der Jahre?

Wir bezeichnen mit a_n die Anzahl der Hasenpaare im ersten Lebensjahr (junge Hasen), und mit b_n die Anzahl der erwachsenen Hasenpaare, jeweils im Jahr Nummer n . Es gelten die Rekursionsformeln:

1. $a_{n+1} = b_n$: es gibt für jedes im Jahr n erwachsene Hasenpaar ein Paar Junge im nächsten Jahr.
2. $b_{n+1} = b_n + a_n$: zu denn schon im Jahr n erwachsenen Hasenpaaren kommen im folgenden Jahr noch die erwachsen gewordenen Junghasen hinzu.

In Matrixform ist also

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = A \cdot \begin{bmatrix} a_n \\ b_n \end{bmatrix} \quad \text{mit} \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Die Hasenpopulation im Jahr n wird damit durch

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} = A \cdot \begin{bmatrix} a_{n-1} \\ b_{n-1} \end{bmatrix} = A \cdot A \cdot \begin{bmatrix} a_{n-2} \\ b_{n-2} \end{bmatrix} = \dots = A^{n-1} \cdot \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$$

beschrieben.

Weil die Hasen in unserem Model nicht sterben, ist es einfach, die Entwicklung auch auf andere Art zu beschreiben: Wir führen hier nur Buch über die *Gesamtzahl* c_n der Hasenpaare im Jahr n . Die Zahl b_n der im Jahr n erwachsenen Hasen ist einfach die Gesamtzahl c_{n-1} der Hasen, die schon im Jahr zuvor da waren. Damit ergibt sich die Formel

$$c_{n+2} = c_{n+1} + c_n,$$

den im Jahr $n+1$ kommen zu den a_{n+1} Hasenpaaren des Vorjahres noch die von den erwachsenen, also schon im Jahr n vorhandenen Hasenpaaren geborenen Junghasen zu. Die Folge $c_1, c_2, c_3, c_4, \dots$ ist also vollständig durch die Rekursionsformel und die Angabe von c_1, c_2 festgelegt. Die Rekursionsformel lässt sich auch als

$$\begin{bmatrix} c_{n+1} \\ c_{n+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_n \\ c_{n+1} \end{bmatrix}$$

schreiben. Der Spezialfall $c_1 = c_2 = 1$ heißt auch *Fibonacci-Folge*. Jedes c_n lässt sich auch explizit durch die beeindruckende Formel

$$c_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right)$$

definieren (siehe Beispiel 3.84). Das kann man auch mit Induktion über n beweisen (Übungsaufgabe).

- ▷ **Beispiel 5.14: (Anwendung in Ökonomie)** Wir betrachten eine Anzahl von Gütern (Waren, Dienstleistungen), durchnummeriert von 1 bis n . Einen Vektor $x = (x_1, \dots, x_n)$ deuten wir als Mengenangaben für diese Güter.

Durch eine $n \times n$ -Matrix $A = (a_{ij})$ beschreiben wir, wieviel von den jeweiligen Gütern benötigt wird, um eines dieser Güter herzustellen. Der Koeffizient a_{ij} gibt an, wieviel von dem Gut mit der Nummer i bei der Produktion einer Einheit von Gut j verbraucht wird.

Beispielsweise betrachten wir die Produktion von Stahl. Wir nummerieren in der Reihenfolge: 1 = Steinkohle, 2 = Stahl. Dann besagt die Matrix⁸

$$A = \begin{bmatrix} 0 & 3 \\ 0,1 & 0 \end{bmatrix}$$

folgendes:

- Zur Produktion einer Tonne Stahl werden drei Tonnen Steinkohle benötigt.
- Zur Förderung von einer Tonne Steinkohle werden 100 kg Stahl (in Form der Abnutzung von Geräten) benötigt.

In der Ökonomie sind Modelle mit weitaus mehr Gütern gebräuchlich, verwandte Matrizen werden für ganze Volkswirtschaften erstellt (mit einigen hundert Zeilen und Spalten).

Eine Fragestellung für das obige Model: Wieviel muss produziert werden, wenn 1000 t Stahl verkauft werden sollen? Im einfachsten Beispiel oben müssen dazu 3000 t Steinkohle bereitgestellt werden, für deren Förderung wiederum 300 t Stahl verbraucht werden, zu dessen Produktion 900 t Steinkohle nötig waren, zu deren Förderung 90 t Stahl gebraucht werden, zu dessen Produktion 270 t Steinkohle ... Insgesamt sind wir soweit schon bei 1390 t Stahl und 4170 t Steinkohle, und es ist klar, dass sich die Rechnung im Prinzip beliebig fortsetzt, allerdings absehbar mit immer kleineren, am Ende bedeutungslosen Mengen.

Etwas formaler: Um $b \in \mathbb{R}^n$ zu verkaufen muss natürlich b produziert werden, aber zusätzlich wird Ab benötigt, zur Produktion von Ab wiederum A^2b , zur Produktion von A^2b zusätzlich A^3b , und so weiter, insgesamt

$$x := b + Ab + A^2b + \dots = \sum_{k=0}^{\infty} A^k b.$$

Die exakte Bestimmung von x (ohne den Grenzwert zu berechnen) läuft so:

$$\begin{aligned} x &= b + Ab + A^2b + A^3b + \dots \\ Ax &= Ab + A^2b + A^3b + \dots \\ x - Ax &= b \end{aligned}$$

Also ergibt sich die gesuchte nötige Gesamtproduktion x als Lösung des linearen Gleichungssystems

$$(E_n - A)x = b.$$

Nun geben wir ein Beispiel, wie die Matrixmultiplikation in Graphentheorie angewandt sein kann. In vielen Anwendungen tauchen die folgenden Probleme auf. Gegeben sei ein (gerichteter oder ungerichteter) Graph $G = (V, E)$ und eine natürliche Zahl $n \geq 1$. Für zwei Knoten i und j interessiert man sich, ob es einen Weg von i nach j der Länge n gibt.⁹ Wenn ja, wie viele solcher Wege gibt es

⁸Matrizen wie

$$A = \begin{bmatrix} 0,2 & 3 \\ 0,1 & 0,1 \end{bmatrix}$$

sind qualitativ plausibler (schließlich wird auch bei der Stahlproduktion wieder Stahl verbraucht ...)

⁹Zur Erinnerung: Ein Weg von u nach v in einem ungerichteten Graphen ist eine Folge u_0, u_1, \dots, u_l jeweils benachbarter Knoten. Beachte, dass der Weg einen Knoten wie auch als auch eine Kante mehrmals durchlaufen kann!

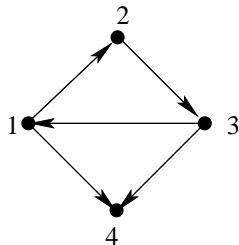
insgesamt? Dazu kann man gut die Matrizenalgebra verwenden: man nimmt die sogenannte ‘‘Adjazenzmatrix’’ $A = (a_{ij})$ von G und berechnet ihre n -te Potenz A^n . Wir mssen aber zuerst definieren, was die Adjazenzmatrix berhaupt ist.

Sei G ein Graph mit der Knotenmenge $V = \{1, \dots, n\}$. Die $n \times n$ Matrix $A = (a_{ij})$ mit

$$a_{ij} = \begin{cases} 1 & \text{falls } i \text{ und } j \text{ benachbart in } G \text{ sind} \\ 0 & \text{sonst} \end{cases}$$

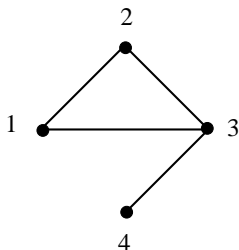
het Adjazenzmatrix von G . Ist der Graph ungerichtet, so ist seine Adjazenzmatrix symmetrisch.

▸ *Beispiel 5.15*: Ein gerichteter Graph mit seiner Adjazenzmatrix:



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

▸ *Beispiel 5.16*: Ein ungerichteter Graph G mit seiner Adjazenzmatrix A und die Produktmatrix $A^2 = A \cdot A$:



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

$$A^2 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

Mit A^n_{ij} werden wir den Eintrag in der i -ten Zeile und j -ten Spalte von A^n bezeichnen:

$$A^n_{ij} := \text{Eintrag in der } i\text{-ten Zeile und } j\text{-ten Spalte von } A^n$$

Satz 5.17. Sei G ein (gerichteter oder ungerichteter) Graph mit den Knoten $1, \dots, m$ und $A = (a_{i,j})$ seine Adjazenzmatrix. Sei A^n die n -te Potenz von A . Dann ist fr jede $1 \leq i, j \leq m$

$$A^n_{ij} = \text{Anzahl der Wege der Lnge } n \text{ in } G \text{ von } i \text{ nach } j$$

Beweis. Wir beweisen den Satz mittels Induktion ber n .

Fr $n = 1$ ist die Behauptung trivialerweise erfllt, denn es gilt $A^1 = A$ und die Adjazenzmatrix zeigt die Zahl aller Kanten, also aller Wege der Lnge 1 zwischen zwei Knoten an.

Sei die Behauptung bereits für alle Potenzen A^r , $r = 1, \dots, n-1$ bewiesen. Nach der Definition der Matrixmultiplikation ist der Eintrag A_{ij}^n in der i -ten Zeile und j -ten Spalte von $A^n = A \cdot A^{n-1}$ genau das Skalarprodukt der i -ten Zeile von A mit der j -ten Spalte von A^{n-1} , d.h.

$$A_{ij}^n = \sum_{k=1}^n a_{ik} \cdot A_{kj}^{n-1},$$

wobei die A_{kj}^{n-1} Koeffizienten der Matrix A^{n-1} bezeichnen.

Für jedes k , $1 \leq k \leq m$, ist nun der Summand $a_{ik} \cdot A_{kj}^{n-1}$ genau dann von Null verschieden, wenn $a_{ik} = 1$ gilt und demzufolge $a_{ik} \cdot A_{kj}^{n-1} = A_{kj}^{n-1}$ ist. Da nach Induktionsvoraussetzung A_{kj}^{n-1} die Zahl der Wege der Länge $n-1$ angibt, die von k nach j führen, und sich aufgrund der Existenz der Kante (i, k) ($a_{ik} = 1$) jeder dieser Wege zu einem Weg der Länge n von i nach j fortsetzen lässt, trägt der Summand $a_{ik} \cdot A_{kj}^{n-1}$ genau die Anzahl der Wege der Länge $1 + (n-1) = n$ von i über k nach j zur Summe A_{ij}^n bei. Da über alle Zwischenknoten k ($1 \leq k \leq m$), summiert wird, gibt A_{ij}^n wie behauptet die Zahl sämtlicher Wege der Länge n an, die in G von i nach j führen. \square

Eine der Grundaufgaben in der linearen Algebra betrifft das Lösen linearer Gleichungssysteme. Sind

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

dann ist

$$A\mathbf{x} = \mathbf{b}$$

eine prägnante Abkürzung für das *Gleichungssystem* mit n Unbekannten und m Gleichungen

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

Gesucht ist dann die Lösungsmenge aller $x \in \mathbb{R}^n$, für die die m Gleichungen *gleichzeitig* erfüllt sind.

Die wichtigsten Fragen über lineare (wie auch über allgemeine) Gleichungssysteme sind:

- Ist $A\mathbf{x} = \mathbf{b}$ überhaupt lösbar?
- Falls lösbar, wieviele Lösungen \mathbf{x} dann $A\mathbf{x} = \mathbf{b}$ hat?
- Falls lösbar, wie kann man die Lösungen \mathbf{x} von $A\mathbf{x} = \mathbf{b}$ finden?

Um diese (wie auch viele andere) Fragen über Matrizen zu beantworten, erwies sich das Konzept des "Rangs" von Matrizen als sehr hilfreich.

5.6 Rang einer Matrix

Sei A eine $m \times n$ Matrix über einen Körper \mathbb{F} :

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Der *Spaltenraum* von A ist der von den Spalten von A erzeugter Vektorraum

$$V = \{A\mathbf{x} : \mathbf{x} \in \mathbb{F}^n\} \subseteq \mathbb{F}^m.$$

Der *Zeilenraum* von A ist der von den Zeilen von A erzeugter Vektorraum

$$W = \{\mathbf{y}A : \mathbf{y} \in \mathbb{F}^m\} \subseteq \mathbb{F}^n.$$

Dann heißt $\dim(V)$ der *Zeilenrang* und $\dim(W)$ der *Spaltenrang* von A .

Eine wichtige Beobachtung ist, dass (nach Satz 5.1) der Zeilenraum (bzw. Spaltenraum) unter folgenden *Elementartransformationen* unverändert bleiben:

- (i) Permutation von Zeilen (bzw. Spalten);
- (ii) Addition eines skalaren Vielfachen einer Zeile (bzw. Spalte) zu einer anderen Zeile (bzw. Spalte).



Aber vorsichtig: Permutation von Spalten (bzw. Zeilen) *kann* den Zeilenraum (bzw. Spaltenraum) verändern! Um das zu sehen, nehmen wir zum Beispiel die Matrix

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

und vertauschen die erste mit der zweiten Spalte:

$$A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Dann liegt der Vektor $(0, 1, 1)$ in den Zeilenraum von A aber nicht in den Zeilenraum von A' .

Trotzdem, man kann leicht zeigen (Übungsaufgabe!), dass die *Dimensionen* $\dim(V)$ und $\dim(W)$ der Vektorräume auch bei der Permutationen der Spalten wie auch Zeilen unverändert bleiben. Diese Eigenschaft erlaubt uns den folgenden Satz zu zeigen.

Satz 5.18. (Rang) Sei A eine $m \times n$ Matrix über \mathbb{R} . Dann gilt:

$$\text{Spaltenrang}(A) = \text{Zeilenrang}(A).$$

Diese Zahl heißt *Rang* von A und ist mit $\text{rk}(A)$ bezeichnet.

Beweis. Induktion nach n . Basis $n = 1$ (nur eine Spalte) ist trivial: Der Zeilenrang wie auch der Spaltenrang sind in diesem Fall beide gleich 1.

Induktionsschritt: Wir nehmen an, dass die Behauptung für alle Matrizen mit $n - 1$ Spalten gilt. Ist die gegebene Matrix $A \neq \mathbf{0}$, so kann man sie durch Anwendung der Elementartransformationen auf die Form

$$B = \begin{bmatrix} b_{11} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{bmatrix}$$

mit $b_{11} \neq 0$ bringen: Durch Permutationen erreicht man zunächst $b_{11} \neq 0$ und annulliert dann die restlichen Elemente der ersten Zeile und Spalte durch geeignete Additionen. Genauer addiert man zunächst das $(-b_{1j}/b_{11})$ -fache der ersten Spalte zu der j -ten Spalte ($j = 2, \dots, n$) und verfährt dann analog mit den Zeilen.

Bezeichnet man mit V' und W' die entsprechenden Vektorräume für die $(m - 1) \times (n - 1)$ Matrix A' , so folgt

$$\dim(V) = \dim(V') + 1, \quad \dim(W) = \dim(W') + 1$$

und nach der Induktionsannahme $\dim(V') = \dim(W')$ folgt die Behauptung. \square

Matrixmultiplikation erlaubt uns den Rang einer Matrix anders charakterisieren:

Satz 5.19. Sei A eine $n \times m$ Matrix über einen Körper \mathbb{F} . Dann ist $\text{rk}(A)$ die kleinste Zahl r für die es eine $n \times r$ Matrix B und eine $r \times m$ Matrix C mit $A = B \cdot C$ gibt.

Beweis. Ist der Spaltenrang von $A = (a_{ij})$ gleich r , so gibt es r Spalten, die alle verbleibenden Spalten erzeugen. O.B.d.A. können wir annehmen, dass dies die ersten r Spalten sind (sonst permutiere sie entsprechend). Sei B die $n \times r$ Matrix die aus diesen ersten r Spalten von A besteht. Seien $\mathbf{b}_1, \dots, \mathbf{b}_n \in \mathbb{F}^r$ die Zeilen von B . Da für $r + 1 \leq j \leq m$ die j -te Spalte von A eine Linearkombination der ersten r Spalten sein soll, muss es einen Vektor $\mathbf{c}_j \in \mathbb{F}^r$ mit $a_{ij} = \langle \mathbf{b}_i, \mathbf{c}_j \rangle$ geben. Ist $j \leq r$, so gilt dasselbe mit $\mathbf{c}_j = (0, \dots, 0, 1, 0, \dots, 0)$, wobei die Eins in j -ter Position steht. Damit ist jeder Eintrag a_{ij} von A als Skalarprodukt zweier Vektoren aus \mathbb{F}^r dargestellt:

$$a_{ij} = \langle \mathbf{b}_i, \mathbf{c}_j \rangle \tag{*}$$

Betrachten man nun die Vektoren $\mathbf{c}_1^\top, \dots, \mathbf{c}_m^\top$ als Spalten einer $r \times m$ Matrix C , so kann man die Bedingung (*) als $A = B \cdot C$ schreiben. \square

Bemerkung 5.20. Satz 5.19 erlaubt uns den Rang mit Hilfe von Skalarprodukt zu beschreiben. Für eine Matrix $A = (a_{ij})$ über \mathbb{F} ist $\text{rk}(A)$ die kleinste Zahl r , so dass es Vektoren $\mathbf{v}_i \in \mathbb{F}^r$ mit der Eigenschaft $a_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ zu den Zeilen und Spalten zugewiesen werden können.

Die folgenden Eigenschaften des Rangs bezüglich der Matrixaddition und Matrixmultiplikation sind zwar nicht scharf, sind aber oft sehr nützlich.

Satz 5.21. 1. Sind A und B zwei $m \times n$ -Matrizen, so gilt:

$$\operatorname{rk}(A) - \operatorname{rk}(B) \leq \operatorname{rk}(A + B) \leq \operatorname{rk}(A) + \operatorname{rk}(B). \quad (5.5)$$

2. Ist A eine $m \times n$ -Matrix und B eine $n \times p$ -Matrix, so gilt

$$\operatorname{rk}(A) + \operatorname{rk}(B) - n \leq \operatorname{rk}(A \cdot B) \leq \min \{ \operatorname{rk}(A), \operatorname{rk}(B) \}. \quad (5.6)$$

Beweis. (5.5) folgt aus der Charakterisierung von Rang durch Skalarprodukt.

Zu (5.6): Der Spaltenraum¹⁰ von $A \cdot B$ ist eine Teilmenge des Spaltenraums von A , und der Zeilenraum von $A \cdot B$ ist eine Teilmenge des Zeilenraums von B . \square

Eine (quadratische!) $n \times n$ Matrix heißt *singulär*, falls $\operatorname{rk}(A) < n$ gilt. Gilt $\operatorname{rk}(A) = n$, so sagt man, dass A einen *vollen Rang* hat; solche Matrizen nennt man auch *regulär*. Eine nützliche Merkregel ist:

$$A \text{ ist regulär (hat vollen Rang)} \iff \text{aus } \mathbf{Ax} = \mathbf{0} \text{ folgt } \mathbf{x} = \mathbf{0}.$$

▷ *Beispiel 5.22* : Jede Dreiecksmatrix

$$A = \begin{bmatrix} a_{11} & * & * & \dots & * \\ 0 & a_{22} & * & \dots & * \\ 0 & 0 & a_{33} & \dots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

wobei $a_{ii} \neq 0$ für alle $i = 1, \dots, n$ und mit * gefüllten Einträge aus beliebigen Zahlen bestehen, hat vollen Rang.

5.7 Lösbarkeit der linearen Gleichungssysteme

Für eine $m \times n$ Matrix A über einem Körper \mathbb{F} , sei

$$\operatorname{Im} A = \{ \mathbf{Ax} : \mathbf{x} \in \mathbb{F}^n \} \subseteq \mathbb{F}^m$$

der Spaltenraum von A . Man kann sich leicht überzeugen (Übungsaufgabe!), dass $\operatorname{Im} A$ ein Vektorraum in \mathbb{F}^m ist. Also ist das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ lösbar genau dann, wenn der Koeffizientenvektor \mathbf{b} in diesem Vektorraum liegt. Das system ist *universell* lösbar, falls $\mathbb{F}^m \subseteq \operatorname{Im} A = \mathbb{F}^m$. Das Gleichungssystem ist *eindeutig* lösbar, falls $\mathbf{Ax} = \mathbf{b}$ für genau ein $\mathbf{x} \in \mathbb{F}^n$ gilt.

Satz 5.23. Für eine $m \times n$ Matrix A ist die Gleichungssystem $\mathbf{Ax} = \mathbf{b}$

1. lösbar $\iff \operatorname{rk}(A) = \operatorname{rk}(A|\mathbf{b})$
2. universell lösbar $\iff \operatorname{rk}(A) = m$
2. eindeutig lösbar $\iff \operatorname{rk}(A) = n = m$

¹⁰D.h. der von Spalten aufgespannte Vektorraum.

Beweis. Seien $\mathbf{v}_1, \dots, \mathbf{v}_n$ die Spaltenvektoren von A .

$$\begin{aligned} \mathbf{Ax} = \mathbf{b} \text{ ist lösbar} &\iff \exists \mathbf{x} \in \mathbb{F}^n \text{ mit } \mathbf{Ax} = \mathbf{b} \\ &\iff \mathbf{b} \in \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_n) \\ &\iff \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_n) = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_n, \mathbf{b}) \\ &\iff \text{rk}(A) = \text{rk}(A|\mathbf{b}). \end{aligned}$$

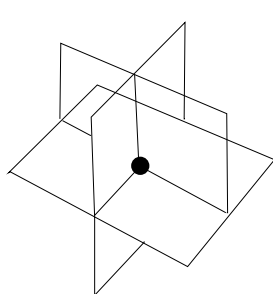
$$\begin{aligned} \text{rk}(A) = m &\iff \text{jeder Vektor } \mathbf{b} \in \mathbb{R}^m \text{ ist eine Linearkombination von } \mathbf{v}_1, \dots, \mathbf{v}_n \\ &\iff \mathbf{Ax} = \mathbf{b} \text{ ist universell lösbar.} \end{aligned}$$

$$\begin{aligned} \text{rk}(A) = n \text{ und } m = n &\iff \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \text{ ist eine Basis von } \mathbb{F}^n = \mathbb{F}^m \\ &\iff \mathbf{b} \text{ lässt sich auf genau eine Weise als Linearkombination} \\ &\quad \text{der Vektoren } \mathbf{v}_1, \dots, \mathbf{v}_n \text{ darstellen (Satz 5.5)} \\ &\iff \mathbf{Ax} = \mathbf{b} \text{ ist eindeutig lösbar.} \end{aligned}$$

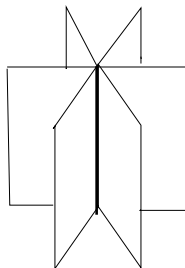
□

Korollar 5.24. Die Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ über \mathbb{R} hat entweder (i) genau eine Lösung, oder (ii) unendlich viele Lösungen, oder (iii) keine Lösung.

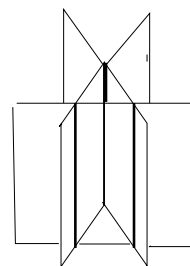
In dreidimensionalem Raum \mathbb{R}^3 definiert jede Gleichung $ax + by + cz = d$ eine Ebene E . Die Lösung für ein System aus 3 solchen Gleichungen ist genau der Durchschnitt $E_1 \cap E_2 \cap E_3$ der entsprechenden Ebenen. Diese Ebenen können sich schneiden in entweder (i) einem Punkt (= genau eine Lösung), oder (ii) in einer Geraden (= unendlich viele Lösungen). Wenn die Geraden $E_1 \cap E_2$, $E_1 \cap E_3$ und $E_2 \cap E_3$ parallel sind, dann gibt es keine Lösung für das System. Das kann man wie folgt veranschaulichen:



genau eine Loesung



unendlich viele Loesungen



keine Loesung

Lemma 5.25. (Farkas Lemma) Sei A eine $m \times n$ Matrix über einen Körper und $\mathbf{b} \in \mathbb{F}^m$. Dann ist die Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ lösbar \iff das Gleichungssystem $\mathbf{y}^T A = 0$ keine Lösung $\mathbf{y} \in \mathbb{F}^m$ mit $\langle \mathbf{y}, \mathbf{b} \rangle = 0$ hat.

Beweis.

$$\begin{aligned}
 \mathbf{Ax} = \mathbf{b} &\iff \text{span}(\mathbf{A}, \mathbf{b}) = \text{span}(\mathbf{A}) \\
 &\iff \text{span}(\mathbf{A}, \mathbf{b})^\top = \text{span}(\mathbf{A})^\top \\
 &\iff \text{span}(\mathbf{A})^\top \setminus \text{span}(\mathbf{A}, \mathbf{b})^\top = \emptyset \\
 &\iff \text{es gibt kein } \mathbf{y} \text{ mit } \mathbf{y}^\top \mathbf{A} = 0 \text{ und } \langle \mathbf{y}, \mathbf{b} \rangle = 0.
 \end{aligned}$$

□



Dieses Lemma hat eine interessante logische Struktur: $\exists x P(x) \iff \forall y \neg Q(y)$. D.h. irgenwas gibt es *genau dann, wenn* irgenwas anderes nicht gibt! Solche Aussagen sind in der Mathematik selten.

Is $\mathbf{b} = \mathbf{0}$ der Nullvektor, so heißt das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ *homogen*. Die Menge der Lösungen des homogenen Gleichungssystems $\mathbf{Ax} = \mathbf{0}$ bezeichnet man mit $\text{Ker } A$:

$$\text{Ker } A = \{\mathbf{x} \in \mathbb{F}^n : \mathbf{Ax} = \mathbf{0}\}.$$

Beachte, dass $\text{Ker } A$ ein linearer Unterraum von \mathbb{F}^n ist – ein solcher Vektorraum heißt ein *affiner* Raum.



Ist $\mathbf{b} \neq \mathbf{0}$, so muss $U = \{\mathbf{x} \in \mathbb{F}^n : \mathbf{Ax} = \mathbf{b}\}$ nicht unbedingt ein Vektorraum sein! Ist zum Beispiel $\mathbf{b} + \mathbf{b} \neq \mathbf{b}$, so gehört $\mathbf{x} + \mathbf{x}$ nicht zu U :

$$A(\mathbf{x} + \mathbf{x}) = \mathbf{Ax} + \mathbf{Ay} = \mathbf{b} + \mathbf{b} \neq \mathbf{b}.$$

Wie gross ist $|\text{Ker } A|$? Diese Frage kann mit dem folgenden Satz beantworten.

Satz 5.26. (Dimensionsformel für lineare Abbildungen) Für jede $m \times n$ Matrix A über einem Körper \mathbb{F} gilt:

$$\dim(\text{Ker } A) + \dim(\text{Im } A) = n. \quad (5.7)$$

Beweis. Betrachte die durch $L(x) = \mathbf{Ax}$ definierte Abbildung $L : \mathbb{F}^n \rightarrow \mathbb{F}^m$. Diese Abbildung ist *linear*, da für alle $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ und $\lambda, \mu \in \mathbb{F}$ die Gleichung $L(\lambda \mathbf{x} + \mu \mathbf{y}) = \lambda L(\mathbf{x}) + \mu L(\mathbf{y})$ gilt.

Ist $\text{Im } L = \{\mathbf{0}\}$, so ist $\mathbb{F}^n = \text{Ker } L$ und wir sind fertig. Nehmen wir also an, dass $\text{Im } L \neq \{\mathbf{0}\}$. Sei $\mathbf{w}_1, \dots, \mathbf{w}_s \in \mathbb{F}^m$ mit $s = \dim(\text{Im } L)$ eine Basis von $\text{Im } L$, und nehme s Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_s \in \mathbb{F}^n$ für die gilt $L(\mathbf{v}_1) = \mathbf{w}_1, \dots, L(\mathbf{v}_s) = \mathbf{w}_s$. Sei auch $\mathbf{u}_1, \dots, \mathbf{u}_r \in \mathbb{F}^n$ eine Basis von $\text{Ker } L$; es gilt also $L(\mathbf{u}_1) = \dots = L(\mathbf{u}_r) = \mathbf{0}$. Es reicht zu zeigen, dass $B = \{\mathbf{v}_1, \dots, \mathbf{v}_s, \mathbf{u}_1, \dots, \mathbf{u}_r\}$ eine Basis von \mathbb{F}^n bildet. Wir müssen zeigen: B ist linear unabhängig, und $\text{span}(B) = \mathbb{F}^n$.

Behauptung: B ist linear unabhängig. Betrachte eine beliebige Linearkombination

$$\sum_{i=1}^s \lambda_i \mathbf{v}_i + \sum_{j=1}^r \mu_j \mathbf{u}_j = \mathbf{0}$$

Da L linear ist, gilt

$$\mathbf{0} = \sum_{i=1}^s \lambda_i L(\mathbf{v}_i) + \sum_{j=1}^r \mu_j \overbrace{L(\mathbf{u}_j)}{=0} = \sum_{i=1}^s \lambda_i \mathbf{w}_i$$

Da $\mathbf{w}_1, \dots, \mathbf{w}_s$ linear unabhängig sind, es folgt $\lambda_1 = \dots = \lambda_s = 0$. Also ist $\sum_{j=1}^r \mu_j \mathbf{u}_j = \mathbf{0}$. Aber $\mathbf{v}_1, \dots, \mathbf{v}_s$ bilden eine Basis von $\text{Ker } L$ und sind deshalb linear unabhängig. D.h. es muss auch $\mu_1 = \dots = \mu_s = 0$ gelten.

Behauptung: $\text{span}(B) = \mathbb{F}^n$. Nehmen wir einen beliebigen Vektor $\mathbf{v} \in \mathbb{F}^n$. Da $L(\mathbf{v}) \in \text{Im } L$ und die Vektoren $\mathbf{w}_1, \dots, \mathbf{w}_s$ eine Basis von $\text{Im } L$ bilden, kann man $L(\mathbf{v})$ als ihre Linearkombination $L(\mathbf{v}) = \sum_{i=1}^s \lambda_i \mathbf{w}_i$ darstellen. Da L linear ist, gilt

$$L(\mathbf{v}) = \sum_{i=1}^s \lambda_i \mathbf{w}_i = \sum_{i=1}^s \lambda_i L(\mathbf{v}_i) = L(\mathbf{x}) \quad \text{wobei } \mathbf{x} := \sum_{i=1}^s \lambda_i \mathbf{v}_i.$$

Da $\mathbf{0} = L(\mathbf{v}) - L(\mathbf{x}) = L(\mathbf{v} - \mathbf{x})$, liegt Vektor $\mathbf{v} - \mathbf{x}$ in $\text{Ker } L$. Da die Vektoren $\mathbf{u}_1, \dots, \mathbf{u}_r$ eine Basis von $\text{Ker } L$ bilden, kann man $\mathbf{v} - \mathbf{x}$ als ihre Linearkombination $\mathbf{v} - \mathbf{x} = \sum_{i=1}^r \mu_i \mathbf{u}_i$ darstellen. Damit haben wir den Vektor \mathbf{v} als Linearkombination

$$\mathbf{v} = \mathbf{x} - \sum_{i=1}^r \mu_i \mathbf{u}_i = \sum_{i=1}^s \lambda_i \mathbf{v}_i + \sum_{i=1}^r \mu_i \mathbf{u}_i$$

der Vektoren aus B dargestellt und $\mathbf{v} \in \text{span}(B)$ gezeigt. □

Korollar 5.27. Sei A eine $m \times n$ -Matrix über einem Körper \mathbb{F} . Dann ist die Menge L aller Lösungen des homogenen Gleichungssystems $A\mathbf{x} = \mathbf{0}$ ein Unterraum von \mathbb{R}^n mit $\dim(L) = n - \text{rk}(A)$.

Homogene Gleichungssysteme sind gut, um die lineare Unabhängigkeit bzw. lineare Abhängigkeit von Vektoren $\mathbf{v}_1, \dots, \mathbf{v}_n$ zu zeigen: Fasse die Vektoren als Spalten einer Matrix A und schaue, welche Lösungen \mathbf{x} das Gleichungssystem $A\mathbf{x} = \mathbf{0}$ hat. Jede Lösung \mathbf{x} gibt uns eine Linearkombination der Spalten, die gleich $\mathbf{0}$ ein muss. Also sind $\mathbf{v}_1, \dots, \mathbf{v}_n$ linear unabhängig genau dann, wenn es keine weitere Lösungen von $A\mathbf{x} = \mathbf{0}$ ausser $\mathbf{x} = \mathbf{0}$ gibt.

5.8 Gauß-Verfahren

Um die Lösbarkeit von $A\mathbf{x} = \mathbf{b}$ zu bestimmen reicht es also die Rangs von A und $(A|\mathbf{b})$ zu vergleichen. Wie bestimmt man aber den Rang einer Matrix? Diese Frage kann man leicht lösen, indem man die Matrix in eine sogenannte "Zeilenstufenform" überführt.

Eine Matrix A ist in einer *Zeilenstufenform*, falls sie die folgende Form hat:

$$A = \begin{bmatrix} 0 & \dots & 0 & \bullet & * & * & * & * & * & * & \dots & * & \dots & * \\ 0 & \dots & 0 & 0 & 0 & \bullet & * & * & * & * & \dots & * & \dots & * \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \bullet & * & \dots & * & \dots & * \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bullet & * & \dots & * \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & \dots & 0 \end{bmatrix}$$

Die \bullet 's bezeichnen von Null verschiedene Matrixeinträge (*Pivotelemente*) und die mit *'s gefüllte Zone besteht aus beliebigen Zahlen.

Zum Beispiel

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix} \text{ und } \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{sind in Zeilenstufenform,}$$

$$\begin{bmatrix} 0 & 0 & 3 \\ 1 & 2 & 1 \end{bmatrix} \text{ und } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & 7 & 8 \end{bmatrix} \quad \text{aber nicht.}$$

Behauptung 5.28. Ist eine Matrix A in Zeilenstufenform, so gilt

$$\text{rk}(A) = \text{Anzahl der Pivotelemente } \bullet \text{ in } A$$

Beweis. Seien $\mathbf{v}_1, \dots, \mathbf{v}_r$ die Zeilen von A mit Pivotelementen und sei $\lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \dots + \lambda_r \mathbf{v}_r = \mathbf{0}$ eine Linearkombination. Da die *Spalte* zu dem Pivotelement in der ersten Zeile \mathbf{v}_1 sonst nur Nullen hat, muss $\lambda_1 = 0$ gelten. Also gilt $\lambda_2 \mathbf{v}_2 + \dots + \lambda_r \mathbf{v}_r = \mathbf{0}$. Nach demselben Argument muss $\lambda_2 = 0$ gelten, usw. Somit haben wir, dass die Zeilen $\mathbf{v}_1, \dots, \mathbf{v}_r$ linear unabhängig sind und damit muss der Zeilenrang (und damit auch $\text{rk}(A)$) mindestens r sein. Andererseits, kann der Zeilenrang (und damit auch $\text{rk}(A)$) nicht größer als r sein, da alle andere Zeilen nur aus Nullen bestehen. \square

Die Überführung einer beliebigen Matrix $A = (a_{ij})$ in Zeilenstufenform geht folgendermaßen vor:

1. Ist in der ersten Spalte ein Eintrag $\neq 0$, so kann man die entsprechende Zeile durch Vertauschung mit der ersten Zeile an die oberste Position bringen.
2. Danach addiert man Vielfache der ersten Zeile zu den folgenden, so dass überall sonst in der ersten Spalte nur noch Nullen stehen.
3. Man wendet darauf das Verfahren auf die Matrix an, die entsteht, wenn man die erste Zeile und die erste Spalte streicht.

Während der Überführung einer Matrix in einer Zeilenstufenform erhalten wir (im Allgemeinen) eine *andere* Matrix A' . Satz 5.1 sagt uns aber, dass der Rang von A bleibt dabei *unverändert*.

▷ *Beispiel 5.29* :

$$A = \begin{bmatrix} 1 & 3 & -4 \\ 3 & 9 & -2 \\ 4 & 12 & -6 \\ 2 & 6 & 2 \end{bmatrix}$$

$$\begin{array}{ccc|l}
 1 & 3 & -4 & \\
 3 & 9 & -2 & z_2 - 3z_1 \\
 4 & 12 & -6 & z_3 - 4z_1 \\
 2 & 6 & 2 & z_4 - 3z_1 \\
 \hline
 1 & 3 & -4 & \\
 0 & 0 & 10 & \\
 0 & 0 & 10 & z_3 - z_2 \\
 0 & 0 & 10 & z_4 - z_2 \\
 \hline
 \mathbf{1} & 3 & -4 & \\
 0 & 0 & \mathbf{10} & \\
 0 & 0 & 0 & \\
 0 & 0 & 0 &
 \end{array}$$

Deshalb ist der Zeilenrang, und deshalb auch der Rank, von A gleich 2.

Um eine Lösung \mathbf{x} von $A\mathbf{x} = \mathbf{b}$ zu bestimmen, kann man nun folgendermaßen angehen – dieses einfache Verfahren ist als *Gauß-Verfahren* bekannt. Dieses Verfahren besteht aus folgenden drei Schritten:

Schritt 1: *Forwärtselimination*: Bringe die erweiterte Matrix $[A|\mathbf{b}]$ zu einer Zeilenstufenform $[A'|\mathbf{b}']$.

Schritt 2: *Lösbarkeitsentscheidung*: Hat mindestens eine Zeile von $[A'|\mathbf{b}']$ ohne Pivotelement einen Eintrag $\neq 0$ in der letzten Spalte (zu \mathbf{b}), so ist $\text{rk}(A|\mathbf{b}) > \text{rk}(A) = r$, und das Gleichungssystem ist nicht lösbar (siehe Satz 5.23). Ist das nicht der Fall, so ist das Gleichungssystem lösbar und wir können den nächsten Schritt machen.

Schritt 3: *Rückwärtssubstitution*: Die zu Spalten ohne Pivotelemente gehörenden Variablen sind die *freien Variablen*. Man löst dann das Gleichungssystem nach den zu den Pivotelementen gehörenden *abhängigen Variablen* aus und bestimmt diese nacheinander in Abhängigkeit von den freien Variablen.

► *Beispiel 5.30* :

$$[A|\mathbf{b}] = \left[\begin{array}{ccc|c}
 1 & 3 & -4 & 3 \\
 3 & 9 & -2 & -11 \\
 4 & 12 & -6 & -6 \\
 2 & 6 & 2 & -10
 \end{array} \right]$$

$$\begin{array}{cccc|c}
 1 & 3 & -4 & & 3 \\
 3 & 9 & -2 & -11 & z_2 - 3z_1 \\
 4 & 12 & -6 & -6 & z_3 - 4z_1 \\
 2 & 6 & 2 & -10 & z_4 - 3z_1 \\
 \hline
 1 & 3 & -4 & & 3 \\
 0 & 0 & 10 & -20 & \\
 0 & 0 & 10 & -18 & z_3 - z_2 \\
 0 & 0 & 10 & -16 & z_4 - z_2 \\
 \hline
 1 & 3 & -4 & & 3 \\
 0 & 0 & 10 & -20 & \\
 0 & 0 & 0 & 2 & \\
 0 & 0 & 0 & 4 & z_4 - 2z_3 \\
 \hline
 1 & 3 & -4 & & 3 \\
 0 & 0 & 10 & -20 & \\
 0 & 0 & 0 & 2 & \\
 0 & 0 & 0 & 0 & 0
 \end{array}$$

Lösbarkeitsentscheidung: Das Gleichungssystem $A\mathbf{x} = \mathbf{b}$ ist *nicht* lösbar, da $b'_3 = 2 \neq 0$ ist.

► *Beispiel 5.31* :

$$[A|\mathbf{b}] = \left[\begin{array}{cccc|c} 1 & -4 & 2 & 0 & 2 \\ 2 & -3 & -1 & -5 & 14 \\ 3 & -7 & 1 & -5 & 16 \\ 0 & 1 & -1 & -1 & 2 \end{array} \right]$$

$$\begin{array}{cccc|c}
 1 & -4 & 2 & 0 & 2 \\
 2 & -3 & -1 & -5 & 14 & z_2 - 2z_1 \\
 3 & -7 & 1 & -5 & 16 & z_3 - 3z_1 \\
 0 & 1 & -1 & -1 & 2 \\
 \hline
 1 & -4 & 2 & 0 & 2 \\
 0 & 5 & -5 & -5 & 10 & z_2 \leftrightarrow z_4 \\
 0 & 5 & -5 & -5 & 10 & z_2 - z_3 \\
 0 & 1 & -1 & -1 & 2 & z_3 - 5z_4 \\
 \hline
 1 & -4 & 2 & 0 & 2 \\
 0 & 1 & -1 & -1 & 2 \\
 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0
 \end{array}$$

Lösbarkeitsentscheidung: Hier ist $\text{rk}(A) = 2 = \text{rk}(A|\mathbf{b})$, also ist das System lösbar.

Rückwärtssubstitution: Die freien Variablen sind x_3 und x_4 . Wir setzen $x_3 = \lambda$, $x_4 = \mu$ und erhalten

$$\begin{aligned}
 x_2 &= 2 + x_3 + x_4 = 2 + \lambda + \mu \\
 x_1 &= 2 + 4x_2 - 2x_3 = 2 + 4(2 + \lambda + \mu) - 2\lambda = 10 + 2\lambda + 4\mu
 \end{aligned}$$

Alle Lösungen ergeben sich daher als

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 10 + 2\lambda + 4\mu \\ 2 + \lambda + \mu \\ \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} 10 \\ 2 \\ 0 \\ 0 \end{bmatrix} + \lambda \cdot \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \mu \cdot \begin{bmatrix} 4 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

5.9 Inversen von Matrizen

Eine $n \times n$ -Matrix A heißt *invertierbar*, wenn es eine $n \times n$ -Matrix A^{-1} gibt, so dass:

$$A^{-1} \cdot A = A \cdot A^{-1} = E_n.$$

A^{-1} heißt dann die *Inverse* von A .

Lemma 5.32. (Eigenschaften von Inversen)

$$\begin{aligned} (A^{-1})^{-1} &= A && \text{Inversion ist involutorisch} \\ (A \cdot B)^{-1} &= B^{-1} \cdot A^{-1} && \text{Inverse eines Produkts (beachte die Reihenfolge!)} \\ (A^T)^{-1} &= (A^{-1})^T && \begin{aligned} &\text{Inverse der Transponierten} \\ &= \text{Transponierte der Inversen} \end{aligned} \end{aligned}$$

$$A \text{ invertierbar} \iff \text{rk}(A) = n \iff \text{aus } Ax = \mathbf{0} \text{ folgt } x = \mathbf{0}$$

Beweis. Die erste Eigenschaft ist trivial: multipliziere beide Seiten mit A^{-1} .

Um die zweite Eigenschaft $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$ zu zeigen, benutzen wir einfach die Definition von $(A \cdot B)^{-1}$: das muss eine Matrix C (falls eine solche existiert) mit der Eigenschaft $(AB)C = C(AB) = E$ sein. Falls A^{-1} und B^{-1} existieren, so kann man $C := B^{-1} \cdot A^{-1}$ nehmen.

Die dritte Eigenschaft $(A^T)^{-1} = (A^{-1})^T$ ist wieder trivial.

Ist $\text{rk}(A) = n$, so bilden die Spalten von A eine Basis von \mathbb{F}^n . Deshalb hat das Gleichungssystem $A \cdot X = B$ eine Lösung X für beliebige $n \times n$ Matrix B , und damit auch für $B = E$. Ist nun A invertierbar, so folgt aus Satz 5.21(2): $\text{rk}(A) \geq \text{rk}(A \cdot A^{-1}) = \text{rk}(E) = n$. \square

Die inverse A^{-1} einer $n \times n$ Matrix A (falls sie überhaupt existiert) kann man mit dem sogenannten *Gauß-Jordan Verfahren* bestimmen.

Schritt 0: A mit Einheitsmatrix E nach rechts erweitern: $A \mapsto [A|E]$.

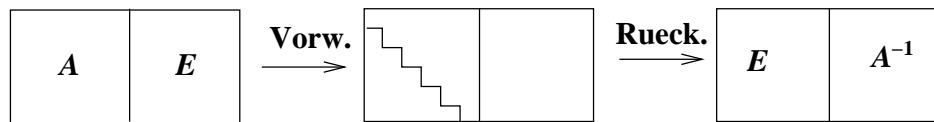
Schritt 1: Vorwärtselimination mit elementaren Zeilenumformungen bis zur Zeilenstufenform links von $\left| \right.$

Schritt 2: Invertierbarkeitstest

Fall 1. Links von $\left| \right.$ steht eine 0 in der Hauptdiagonale
 $\implies A$ ist nicht invertierbar, fertig.

Fall 2. Links von $|$ steht keine 0 in der Hauptdiagonale, dann

Schritt 3: Rückwärtselimination mit elementaren Zeilenumformungen bis links die Einheitsmatrix E steht. Rechts von $|$ steht dann A^{-1} .



Das Verfahren sieht irgendwie “magisch” aus – warum soll am Ende die Inverse rechts von $|$ stehen?

Für gegebene $n \times n$ Matrix A , gesucht ist eine $n \times n$ Matrix X mit $A \cdot X = E$. Sind $\mathbf{x}_1, \dots, \mathbf{x}_n$ die Spalten von X , so müssen wir die n Gleichungssysteme

$$A\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, A\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, A\mathbf{x}_x = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

lösen. Und das Gauß–Jordan Verfahren löst einfach diese Gleichungssysteme *simultan!* Da am Ende links von $|$ die Einheitsmatrix steht, müssen die Spalten der rechts stehenden Matrix die gesuchten Lösungen, d.h. Spalten von $X = A^{-1}$ sein.

► *Beispiel 5.33* : Gesucht ist die Inverse zu

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

A mit Einheitsmatrix E nach rechts erweitern:

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array} \right]$$

Ziehe 3-faches der ersten Zeile von 2. Zeile ab:

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array} \right]$$

Teile 2. Zeile durch -2 :

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & \frac{3}{2} & -\frac{1}{2} \end{array} \right]$$

Ziehe das 2-fache der 2. Zeile von der ersten ab:

$$\left[\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & \frac{3}{2} & -\frac{1}{2} \end{array} \right]$$

Also ist

$$A^{-1} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \quad \text{die Inverse zu} \quad A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

Probe:

$$A \cdot A^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

► *Beispiel 5.34* : Gegeben sei die Matrix

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 3 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$

Gesucht ist eine Matrix X mit

$$A \cdot X = E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Schritt 0: A mit Matrix B nach rechts erweitern: $A \mapsto [A|E]$.

Schritt 1: Vorwärtselimination mit elementaren Zeilenumformungen bis zur Zeilenstufenform links von |

$$\begin{array}{ccc|ccc} 2 & 0 & 1 & 1 & 0 & 0 \\ 3 & 1 & 2 & 0 & 1 & 0 & -3/2z_1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ \hline 2 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1/2 & -3/2 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & -z_2 \\ \hline 2 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1/2 & -3/2 & 1 & 0 \\ 0 & 0 & 1/2 & 3/2 & -1 & 1 \end{array}$$

Schritt 2: Invertierbarkeitstest. Links von | steht keine 0 in der Hauptdiagonale $\implies A$ ist invertierbar.

Schritt 3: Rückwärtselimination mit elementaren Zeilenumformungen bis links die Einheitsmatrix E steht. Rechts von | steht dann A^{-1} .

$$\begin{array}{ccc|ccc} 2 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1/2 & -3/2 & 1 & 0 & -z_3 \\ 0 & 0 & 1/2 & 3/2 & -1 & 1 \\ \hline 2 & 0 & 1 & 1 & 0 & 0 & -2z_3 \\ 0 & 1 & 0 & -3 & 2 & -1 \\ 0 & 0 & 1/2 & 3/2 & -1 & 1 \\ \hline 2 & 0 & 0 & -1 & 2 & -2 & \cdot 1/2 \\ 0 & 1 & 0 & -3 & 2 & -1 \\ 0 & 0 & 1/2 & 3/2 & -1 & 1 & \cdot 2 \\ \hline 1 & 0 & 0 & -1 & 1 & -1 \\ 0 & 1 & 0 & -3 & 2 & -1 \\ 0 & 0 & 1 & 3 & -2 & 2 \end{array}$$

Jetzt steht auf der linken Seite die Einheitsmatrix, also rechts die Lösung

$$A^{-1} = \begin{bmatrix} -1 & 1 & -1 \\ -3 & 2 & -1 \\ 3 & -2 & 2 \end{bmatrix}$$

5.10 Orthogonalität

Die Vektoren \mathbf{x} und \mathbf{y} heißen *orthogonal*, falls $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ gilt. Wenn $A \subseteq V$ eine Teilmenge eines Vektorraums V ist, dann ist

$$A^\perp = \{\mathbf{x} \in V : \langle \mathbf{x}, \mathbf{y} \rangle = 0 \text{ für alle } \mathbf{y} \in A\}$$

das *orthogonale Komplement* von A .

Satz 5.35. Für alle $A \subseteq V$ ist A^\perp ein Vektorraum.

Beweis. Übungsaufgabe. □

Sei V ein endlichdimensionaler Vektorraum über \mathbb{R} und sei $\mathbf{v}_1, \dots, \mathbf{v}_n$ seine Basis. Die Basis heißt *orthonormal*, wenn

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j. \end{cases}$$

In anderen Worten eine Basis ist orthonormal, wenn seine Vektoren die Länge 1 haben und senkrecht zu einanderem liegen. Die Frage, ob jeder endlicherzeugter Vektorraum V eine Orthonormalbasis besitzt lässt sich *positiv* mit folgendem Satz beantworten.

Satz 5.36. Jeder endlicherzeugter Vektorraum besitzt eine Orthonormalbasis.

Dies ist eine direkte Folgerung aus dem folgenden Satz.

Satz 5.37. (Gram–Schmidt-Orthogonalisierungsverfahren) Sei $W \subseteq V$ ein Unterraum von V und $\mathbf{w}_1, \dots, \mathbf{w}_m$ Orthonormalbasis von W . Ist $W \neq V$, dann gibt es ein $\mathbf{v} \in W^\perp$, so dass $\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{v}$ eine Orthonormalbasis von $\text{span}(\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{v})$ ist.

Beweis. Wähle zuert einen beliebigen Vektor $\mathbf{a} \in V \setminus W$, setze $\lambda_i := \langle \mathbf{a}, \mathbf{w}_i \rangle$ und betrachte den Vektor

$$\mathbf{b} := \mathbf{a} - \sum_{i=1}^m \lambda_i \mathbf{w}_i.$$

Da für jedes $i_0 \in \{1, \dots, m\}$ gilt:

$$\begin{aligned} \langle \mathbf{b}, \mathbf{w}_{i_0} \rangle &= \langle \mathbf{a}, \mathbf{w}_{i_0} \rangle - \sum_{i=1}^m \lambda_i \langle \mathbf{w}_i, \mathbf{w}_{i_0} \rangle = \langle \mathbf{a}, \mathbf{w}_{i_0} \rangle - \lambda_{i_0} \underbrace{\langle \mathbf{w}_{i_0}, \mathbf{w}_{i_0} \rangle}_{=1} \\ &= \langle \mathbf{a}, \mathbf{w}_{i_0} \rangle - \langle \mathbf{a}, \mathbf{w}_{i_0} \rangle = 0, \end{aligned}$$

der Vektor \mathbf{b} auf allen $\mathbf{w}_1, \dots, \mathbf{w}_m$ senkrecht steht (und damit sind die Vektoren $\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{b}$ linear unabhängig (siehe Aufgabe 36), und Vektor $\mathbf{v} = \frac{\mathbf{b}}{\|\mathbf{b}\|}$ leistet das Gewünschte. □

Satz 5.38. Sei $\mathbf{v}_1, \dots, \mathbf{v}_n$ eine Orthonormalbasis von V . Dann gilt für jedes $\mathbf{x} \in V$:

$$\mathbf{x} = \sum_{i=1}^n \langle \mathbf{x}, \mathbf{v}_i \rangle \mathbf{v}_i.$$



Dieser Satz erklärt, warum ist es gut, eine Orthonormalbasis $\mathbf{v}_1, \dots, \mathbf{v}_n$ zu haben: Dann kann man nämlich sehr einfach die Koeffizienten der Koordinatendarstellung von Vektoren $\mathbf{x} \in V$ bezüglich dieser Basis bestimmen!

Beweis. Sei $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{v}_i$ die (eindeutige, siehe Satz 5.5) Darstellung von \mathbf{x} . Die Koeffizienten λ_i sind dann leicht zu bestimmen:

$$\langle \mathbf{x}, \mathbf{v}_i \rangle = \lambda_i \underbrace{\langle \mathbf{v}_i, \mathbf{v}_i \rangle}_{=1} + \sum_{j \neq i} \lambda_j \underbrace{\langle \mathbf{v}_j, \mathbf{v}_i \rangle}_{=0} = \lambda_i$$

□

Sei V ein Vektorraum. Eine Teilmenge $U \subseteq V$ heißt *Unterraum* von V , falls U selbst ein Vektorraum ist.

Satz 5.39. Ist $U \subseteq V$ ein Unterraum, so gilt

$$\dim(U) + \dim(U^\perp) = \dim(V)$$

und jeder Vektor $\mathbf{v} \in V$ lässt sich *eindeutig* als die Summe

$$\mathbf{v} = \mathbf{u} + \mathbf{w} \quad \text{mit } \mathbf{u} \in U \text{ und } \mathbf{w} \in U^\perp$$

darstellen. Vector $\mathbf{u} = \mathbf{v} - \mathbf{w}$ heisst dann die *orthogonale Projektion* von \mathbf{v} auf U und ist mit $\mathbf{u} = \text{proj}_U(\mathbf{v})$ bezeichnet.

Beweis. Ist $U = \{\mathbf{0}\}$ oder $U = V$, so ist die Aussage trivial. Nehmen wir deshalb an, dass $\{\mathbf{0}\} \subset U \subseteq V$ ein echter nicht trivialer Unterraum von V ist. Sei $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ eine orthonormale Basis von U . Wende das Gram–Schmidt-Orthogonalisierungsverfahren an und erweitere sie bis zu einer orthonormalen Basis $\{\mathbf{u}_1, \dots, \mathbf{u}_r, \mathbf{v}_1, \dots, \mathbf{v}_s\}$ von V . Wir behaupten, dass dann $B = \{\mathbf{v}_1, \dots, \mathbf{v}_s\}$ eine (auch orthonormale) Basis von U^\perp ist.

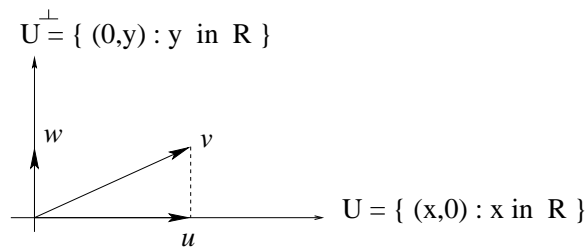
Zu zeigen: $U^\perp \subseteq \text{span}(B)$. Ist $\mathbf{x} \in U^\perp$, so gilt

$$\mathbf{x} = \sum_{i=1}^r \underbrace{\langle \mathbf{x}, \mathbf{u}_i \rangle}_{=0} \mathbf{u}_i + \sum_{j=1}^s \langle \mathbf{x}, \mathbf{v}_j \rangle \mathbf{v}_j = \sum_{j=1}^s \langle \mathbf{x}, \mathbf{v}_j \rangle \mathbf{v}_j.$$

Zu zeigen: B ist unabhängig. Ist $\sum_{j=1}^s \lambda_j \mathbf{v}_j = \mathbf{0}$, so gilt für jedes $i = 1, \dots, s$

$$0 = \langle \mathbf{0}, \mathbf{v}_i \rangle = \sum_{j=1}^s \lambda_j \langle \mathbf{v}_j, \mathbf{v}_i \rangle = \lambda_i \langle \mathbf{v}_i, \mathbf{v}_i \rangle = \lambda_i.$$

□

Abbildung 5.2: $\mathbf{u} = \text{proj}_U(\mathbf{v})$

Projektionen haben die Eigenschaft, dass $\mathbf{x} = \text{proj}_U(\mathbf{v})$ derjeniger Vektor in U ist, der die *kleinste* Abstand zu Vektor \mathbf{v} hat.

Satz 5.40. Sei $U \subseteq V$ ein Unterraum, $\mathbf{v} \in V$ und $\mathbf{u} = \text{proj}_U(\mathbf{v})$. Dann gilt: $\|\mathbf{v} - \mathbf{x}\| > \|\mathbf{v} - \mathbf{u}\|$ für alle $\mathbf{x} \in U$, $\mathbf{x} \neq \mathbf{u}$.

Beweis. Da $\mathbf{u} = \text{proj}_U(\mathbf{v})$, muss es ein $\mathbf{w} \in U^\perp$ mit $\mathbf{v} = \mathbf{u} + \mathbf{w}$ geben. Dies bedeutet insbesondere, dass der Vektor $\mathbf{v} - \mathbf{u} = \mathbf{w}$ in U^\perp liegen muss. Andererseits, liegt der Vektor $\mathbf{u} - \mathbf{x}$ in U , da beide Vektoren \mathbf{u} und \mathbf{x} in U liegen. Damit wissen wir, dass die Vektoren $\mathbf{v} - \mathbf{u}$ und $\mathbf{u} - \mathbf{x}$ orthogonal sind, d.h. $\langle \mathbf{v} - \mathbf{u}, \mathbf{u} - \mathbf{x} \rangle = 0$ gilt. Mit der Anwendung des Pythagoras-Theorems erhalten wir:

$$\begin{aligned} \|\mathbf{v} - \mathbf{x}\|^2 &= \|(\mathbf{v} - \mathbf{u}) + (\mathbf{u} - \mathbf{x})\|^2 \\ &= \|\mathbf{v} - \mathbf{u}\|^2 + \|\mathbf{u} - \mathbf{x}\|^2 && \text{(Pythagoras)} \\ &> \|\mathbf{v} - \mathbf{u}\|^2 && \text{(da } \mathbf{x} \neq \mathbf{u}) \end{aligned}$$

□

Ist $U \subseteq \mathbb{F}^n$ ein Unterraum, so wissen wir bereits (siehe Satz 5.39), dass jeder Vektor $\mathbf{v} \in V$ sich eindeutig als die Summe

$$\mathbf{v} = \mathbf{u} + \mathbf{w} \quad \text{mit } \mathbf{u} \in U \text{ und } \mathbf{w} \in U^\perp$$

darstellen lässt. Der Vektor $\mathbf{u} = \mathbf{v} - \mathbf{w}$ heisst dann die *orthogonale Projektion* von \mathbf{v} auf U und ist mit $\mathbf{u} = \text{proj}_U(\mathbf{v})$ bezeichnet. Nun wollen wir die Projektionen auch berechnen können. D.h. wir wollen eine Matrix A finden, so dass $\text{proj}_U(\mathbf{v}) = A\mathbf{v}$ gilt.

Satz 5.41. Für jeden Unterraum $U \subseteq \mathbb{F}^n$ gibt es eine $n \times n$ Matrix A , so dass $A = A^\top$ (A ist symmetrisch), $AA^\top = A$ (A ist idempotent) und für $\mathbf{v} \in \mathbb{F}^n$

$$\text{proj}_U(\mathbf{v}) = A\mathbf{v}$$

gilt.

Beweis. Sei $B = [\mathbf{b}_1, \dots, \mathbf{b}_k]$ mit $k = \dim(U)$ die $n \times k$ Matrix, deren Spalten $\mathbf{b}_1, \dots, \mathbf{b}_k$ eine Basis von U bilden. Betrachte die folgende $n \times n$ Matrix

$$A = B(B^\top B)^{-1}B^\top.$$

Aus $(X \cdot Y)^T = Y^T \cdot X^T$ und $(X^T)^T = X$ folgt, dass A eine symmetrische und idempotente Matrix ist.

Sei nun $\mathbf{u} = \text{proj}_U(\mathbf{v})$. Dann gibt es ein (eindeutiger) Vektor $\mathbf{w} \in U^\perp$ mit $\mathbf{v} = \mathbf{u} + \mathbf{w}$. Wir können den Vektor \mathbf{u} als eine Linearkombination $\mathbf{u} = B\mathbf{x} = \sum_{i=1}^k x_i \mathbf{b}_i$ der Basisvektoren darstellen. Wir wollen den Vektor $\mathbf{x} = (x_1, \dots, x_k)$ der Koeffizienten bestimmen. Da $\mathbf{v} = \mathbf{u} + \mathbf{w}$, wissen wir, dass

$$\mathbf{v} = \left(\sum_{i=1}^k x_i \mathbf{b}_i \right) + \mathbf{w}$$

gilt. Wir betrachten nun die Skalarprodukte von \mathbf{v} mit Basisvektoren:

$$\langle \mathbf{b}_j, \mathbf{v} \rangle = \sum_{i=1}^k x_i \langle \mathbf{b}_j, \mathbf{b}_i \rangle + \langle \mathbf{b}_j, \mathbf{w} \rangle = \sum_{i=1}^k x_i \langle \mathbf{b}_j, \mathbf{b}_i \rangle$$

In der Form von Matrizen können wir dies als

$$\begin{bmatrix} \langle \mathbf{b}_1, \mathbf{b}_1 \rangle & \langle \mathbf{b}_1, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_1, \mathbf{b}_k \rangle \\ \langle \mathbf{b}_2, \mathbf{b}_1 \rangle & \langle \mathbf{b}_2, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_2, \mathbf{b}_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}_k, \mathbf{b}_1 \rangle & \langle \mathbf{b}_k, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_k, \mathbf{b}_k \rangle \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} \langle \mathbf{b}_1, \mathbf{v} \rangle \\ \langle \mathbf{b}_2, \mathbf{v} \rangle \\ \vdots \\ \langle \mathbf{b}_k, \mathbf{v} \rangle \end{bmatrix}$$

umschreiben. Dies ist in der Form

$$(B^T B)\mathbf{x} = B^T \mathbf{v}.$$

Da B und \mathbf{v} gegeben sind, reicht es diesen Gleichungssystem auf \mathbf{x} auflösen. Da die Spalten von B linear unabhängig sind, hat die Matrix $C = B^T B$ den vollen Rank. (Warum? Übungsaufgabe!) Damit ist diese Matrix invertierbar und wir können das Gleichungssystem oben nach \mathbf{x} auflösen, indem wir beide Seiten mit $C^{-1} = (B^T B)^{-1}$ multiplizieren

$$\mathbf{x} = B(B^T B)^{-1} B^T \cdot \mathbf{v} = A \cdot \mathbf{v}.$$

Zusammen mit $\mathbf{u} = B\mathbf{x}$ folgt daraus die Behauptung:

$$\text{proj}_U(\mathbf{v}) = \mathbf{u} = B\mathbf{x} = B(B^T B)^{-1} B^T \mathbf{v} = A\mathbf{v}.$$

□

5.11 Determinanten

In diesem Abschnitt betrachten wir nur quadratische Matrizen, d.h. $n \times n$ -Matrizen. Mit jeder solchen Matrix A kann man eine Zahl – sogenannte “Determinante” $\det A$ in Verbindung bringen. Sei S_n die Menge aller Permutationen $\pi : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$. Ist

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

eine $n \times n$ -Matrix über \mathbb{F} , und in \mathbb{F} gilt $1 + 1 \neq 0$, so ist ihre *Determinante* $\det A$ definiert durch:

$$\det A := \sum_{\pi \in S_n} \sigma(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)} \quad (5.8)$$

wobei $\sigma(\pi)$ das Signum (+1 oder -1) von $\prod_{1 \leq i < j \leq n} (\pi(j) - \pi(i))$ ist. Die Determinante kann man auch rekursiv wie folgt ausrechnen:

$$n = 1: \quad \det A := a_{11}$$

$$n > 1: \quad \det A := \sum_{k=1}^n (-1)^{k+1} a_{k1} \cdot \det A_{k1} \quad (\text{Entwicklung nach der 1. Spalte})$$

Hier ist A_{k1} eine $(n-1) \times (n-1)$ -Untermatrix von A ohne ersten Spalte und k -ten Zeile.

Die Determinante von A bezeichnet man mit

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

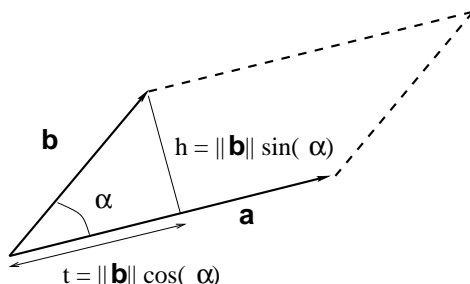
Zum Beispiel, ist $A = [\mathbf{a}, \mathbf{b}]$ mit Spalten $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ und $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, so ist

$$\det \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1$$

die Determinante von \mathbf{a} und \mathbf{b} . Die Vektoren \mathbf{a} und \mathbf{b} sind genau dann linear abhängig, wenn sie auf einer Geraden liegen (*kolinear* sind). Anderenfalls spannen \mathbf{a}, \mathbf{b} ein Parallelogramm auf; dieses hat die Fläche

$$F = \|\mathbf{a}\| \cdot h = \|\mathbf{a}\| \cdot \|\mathbf{b}\| \sin \alpha$$

wobei α der Winkel zwischen \mathbf{a} und \mathbf{b} ist.



Wenn wir nun das Quadrat von F betrachten, erhalten wir:

$$\begin{aligned} F^2 &= \|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2 \cdot \sin^2 \alpha = \|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2 \cdot (1 - \cos^2 \alpha) \quad (\text{Pythagoras}) \\ &= \|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2 \cdot \left(1 - \frac{\langle \mathbf{a}, \mathbf{b} \rangle^2}{\|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2}\right) \quad (5.4) \\ &= \|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2 - \langle \mathbf{a}, \mathbf{b} \rangle^2 \\ &= (a_1^2 + a_2^2)(b_1^2 + b_2^2) - (a_1 b_1 + a_2 b_2)^2 \\ &= (a_1 b_2 - a_2 b_1)^2. \end{aligned}$$

Also ist $F = |\det(\mathbf{a}, \mathbf{b})|$. Fazit: Determinante von A gibt uns die Fläche des von den Spalten von A ausgespannten Parallelogramms. Daraus folgt auch:

$$\det(\mathbf{a}, \mathbf{b}) = 0 \iff \mathbf{a} \text{ und } \mathbf{b} \text{ linear abhängig sind.}$$

Behauptung 5.42. Seien A, B beliebige $n \times n$ -Matrizen und E eine Einheitsmatrix. Dann gilt:

1. $\det E = 1$.
2. $\det A = 0$, wenn A eine Nullzeile enthält.
3. $\det \lambda A = \lambda^n \det A$.
4. $\det A^\top = \det A$.
5. Entsteht B aus A durch Vertauschung zweier Zeilen oder zweier Spalten, so gilt

$$\det B = -\det A. \quad (5.9)$$

6. $\det A \cdot B = \det A \cdot \det B$.

^aDie "Regel" $\det A + B = \det A + \det B$ ist für $n \geq 2$ falsch!

Die ersten vier Eigenschaften folgen unmittelbar aus der Definition 5.8. Die letzten zwei sind weniger trivial.

Eine der wichtigsten Eigenschaften der Determinante ist ihre *linearität*.

Satz 5.43. (Linearität von Determinanten) Die Determinante ist linear in den Zeilen, d.h. wenn wir die r -te Zeile als Vektor $\mathbf{a}_i = (a_{i1}, \dots, a_{in})$ abkürzen und wenn $\mathbf{a}_i = \lambda \mathbf{b}_i + \mu \mathbf{c}_i$ ist, so gilt

$$\det \begin{bmatrix} \vdots \\ \lambda \mathbf{b}_i + \mu \mathbf{c}_i \\ \vdots \end{bmatrix} = \lambda \det \begin{bmatrix} \vdots \\ \mathbf{b}_i \\ \vdots \end{bmatrix} + \mu \det \begin{bmatrix} \vdots \\ \mathbf{c}_i \\ \vdots \end{bmatrix}$$

wobei hier die Punkte \vdots andeuten sollen, dass dieser Teil des Vektors bei der Rechnung unverändert bleibt. Dasselbe gilt auch für Spalten.

Ist die Matrix A bereits in einer Zeilenstufenform

$$A = \begin{bmatrix} a_{11} & * & * & \dots & * \\ 0 & a_{22} & * & \dots & * \\ 0 & 0 & a_{33} & \dots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn}, \end{bmatrix}$$

dann gilt offensichtlich:

$$\det A = a_{11} \cdot a_{22} \cdot \dots \cdot a_{nn}.$$

Eine natürliche Frage deshalb ist, ob (und wenn ja, dann wie) sich die Determinante verändert, wenn man die Matrix mit Hilfe von Elementartransformationen zu einer Zeilenstufenform überführt?

Elementartransformationen sind zwei: (i) Permutation von Zeilen und (ii) Addition eines skalaren Vielfachen einer Zeile zu einer anderen Zeile. Wir wissen bereits, dass die erste Transformation nur die Vorzeichen der Determinante verändern kann: Entsteht B aus A durch Vertauschung zweier Zeilen oder zweier Spalten, so gilt $\det B = -\det A$. Was aber mit der zweiten Transformation? Überraschenderweise ist diese (zweite) Transformation noch harmloser: Sie verändert nicht mal die Vorzeichen!

Satz 5.44. Entsteht B aus A durch Addition einer mit λ multiplizierten Zeile zu einer anderen, so gilt $\det B = \det A$.

Beweis. Sei A eine $n \times n$ Matrix. Da durch das Transponieren einer Matrix Spalten in Zeilen umgeformt werden und $\det A^T = \det A$ gilt, genügt es den Spaltenfall zu betrachten.¹¹

Wird ein λ -faches der j -ten Spalte \mathbf{a}_j zu der i -ten Spalte \mathbf{a}_i , $i \neq j$ addiert (also wird \mathbf{a}_i durch $\mathbf{a}_i + \lambda\mathbf{a}_j$ ersetzt), so erhält man die Matrix

$$A' = [\mathbf{a}_1, \dots, \mathbf{a}_i + \lambda\mathbf{a}_j, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n]$$

Aufgrund der Linearität (Satz 5.43) kann die Determinante aufgeteilt werden in

$$\begin{aligned} \det A' &= \det [\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n] + \det [\mathbf{a}_1, \dots, \lambda\mathbf{a}_j, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n] \\ &= \det A + \lambda \det [\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n] \end{aligned}$$

Die zweite Matrix enthält zwei identische Spalten und hat somit Determinante Null. Warum? Da nach (5.9) muss ja

$$\det [\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n] = -\det [\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n]$$

gelten. □

Als Korollar bekommen wir einen einfachen Algorithmus zur Berechnung der Determinante.

Korollar 5.45. Wird eine $n \times n$ Matrix $A = (a_{ij})$ durch Elementartransformationen mit insgesamt k Zeilenvertauschungen zu einer Dreiecksmatrix $B = (b_{ij})$ gebracht, so gilt

$$\det A = (-1)^k b_{11} \cdot b_{22} \cdots b_{nn}.$$

Insbesondere gilt für den Betrag:

$$|\det A| = |b_{11} \cdot b_{22} \cdots b_{nn}|.$$

Im Falle einer singulären Matrix A ($\text{rk}(A) < n$) enthält die Zeilenstufenform $B = (b_{ij})$ wenigstens eine Null an der Diagonale ($b_{ii} = 0$ für mindestens ein i). Damit erhalten wir den folgenden Singularkriterium:

Korollar 5.46.

$$\det A = 0 \iff \text{rk}(A) < n$$

¹¹Nur um die Schreibweise zu vereinfachen.

Ausser das Determinanten die Singularität bzw. Regularität der Matrizen widerspiegeln, kann man sie auch für die Lösung linearer Gleichungssystemen $\mathbf{Ax} = \mathbf{b}$ mit regulären Koeffizientenmatrizen A benutzen. In diesem Fall existiert ja A^{-1} und $\mathbf{Ax} = \mathbf{b}$ ist dann universell und eindeutig lösbar und die Lösung ist $\mathbf{x} = A^{-1}\mathbf{b}$.

Satz 5.47. (Cramer'sche Regel) Ist A eine reguläre $n \times n$ -Matrix, so werden die Komponenten x_i der eindeutig bestimmten Lösung von $\mathbf{Ax} = \mathbf{b}$ gegeben durch

$$x_i = \frac{\det A_i(\mathbf{b})}{\det A} \quad \text{für} \quad i = 1, \dots, n,$$

wobei $A_i(\mathbf{b})$ die durch Ersetzen der i -ten Spalte von A durch \mathbf{b} entstandene Matrix ist.

5.12 Eigenwerte und Eigenvektoren

Definition: Sei A eine $n \times n$ -Matrix über \mathbb{R} . Ein Skalar $\lambda \in \mathbb{R}$ ist ein *Eigenwert* von A , falls es ein Vektor $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{x} \neq \mathbf{0}$ und $\mathbf{Ax} = \lambda\mathbf{x}$ gibt; der Vektor \mathbf{x} selbst heißt *Eigenvektor* zum Eigenwert λ .

Wegen $\lambda\mathbf{x} = \lambda E\mathbf{x}$ kann man das System $\mathbf{Ax} = \lambda\mathbf{x}$ auch in der Form $(A - \lambda E)\mathbf{x} = \mathbf{0}$ mit

$$A - \lambda E = \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix}$$

schreiben.

Man kann leicht zeigen (siehe Aufgabe 41), dass Eigenvektoren zu verschiedenen Eigenwerten linear unabhängig sein müssen. Daraus unmittelbar folgt, dass es zu jeder $n \times n$ Matrix höchstens n verschiedene Eigenwerte gibt. (Warum? Siehe Korollar 5.8)

Um Eigenwerte einer $n \times n$ Matrix A zu bestimmen, kann man Determinanten benutzen. Nämlich, ist λ Eigenwert von A genau dann, wenn die Gleichungssystem $(A - \lambda E)\mathbf{x} = \mathbf{0}$ eine nichttriviale Lösung $\mathbf{x} \neq \mathbf{0}$ hat, und das ist genau dann der Fall, wenn

$$\det(A - \lambda E) = 0 \quad (5.10)$$

ist (siehe Korollar 5.46). Das Polynom

$$p(\lambda) = \det(A - \lambda E) \quad (5.11)$$

heißt das *charakteristische Polynom* von A . Sind $\lambda_1, \dots, \lambda_n$ die Eigenwerte von A , so sind sie die Nullstellen dieses Polynoms und es gilt

$$\det(A - \lambda E) = (-1)^n (\lambda - \lambda_1) \cdots (\lambda - \lambda_n). \quad (5.12)$$

Daraus ergeben sich einige Verbindungen zwischen Eigenwerten und der Determinante wie auch der "Spur" einer Matrix. Die *Spur* (engl. "trace") einer $n \times n$ Matrix $A = (a_{ij})$ ist die Summe

$$\text{Tr}(A) = \sum_{i=1}^n a_{ii}$$

der Diagonalelemente.

Satz 5.48. Ist A eine $n \times n$ Matrix mit Eigenwerten $\lambda_1, \dots, \lambda_n$, so gilt:

$$\det A = \prod_{i=1}^n \lambda_i,$$

$$\operatorname{Tr}(A) = \sum_{i=1}^n \lambda_i.$$

Beweis. Um die erste Gleichung zu erhalten, setze einfach $\lambda = 0$ in (5.12).

Um die zweite Gleichung zu beweisen, schreibe die rechte Seite von (5.12) als ein Polynom (in λ) und beachte, dass der Koeffizient zu λ^{n-1} ist gleich

$$(-1)^n \sum_i i = 1^n - \lambda_i = (-1)^{n+1}(\lambda_1 + \lambda_2 + \dots + \lambda_n),$$

wobei das Koeffizient zu λ^{n-1} in $\det(A - \lambda E)$ ist gleich $(-1)^{n-1}(a_{11} + a_{22} + \dots + a_{nn})$. \square

Da wir uns nur über Nullstellen des charakteristischen Polynoms $\det(A - \lambda E)$ kümmern, reicht es die Matrix $A - \lambda E$ zu einer Zeilenstufenform $B = (b_{ij})$ zu reduzieren (Anzahl der Zeilenvertauschungen ist dabei uns absolut unwichtig!) und die Gleichung

$$b_{11} \cdot b_{22} \cdots b_{nn} = 0$$

nach λ auflösen. Da $\det(A - \lambda E)$ ein Polynom (mit λ als seine Variable) des Grades n ist, so kann A höchstens n verschiedene Eigenwerte haben. (Wir haben diesen Fakt bereits oben erwähnt.)

► *Beispiel 5.49 :*

$$A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

$$\begin{aligned} \det(A - \lambda E) &= \begin{vmatrix} 3 - \lambda & -1 \\ -1 & 3 - \lambda \end{vmatrix} = (3 - \lambda)(3 - \lambda) - (-1)(-1) = (\lambda^2 - 6\lambda + 9) - 1 \\ &= (\lambda - 2)(\lambda - 4) \end{aligned}$$

Somit sind $\lambda = 2$ und $\lambda = 4$ die Lösungen von $\det(A - \lambda E)$, also - die Eigenwerte von A .

► *Beispiel 5.50 :*

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad A - \lambda E = \begin{bmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix}$$

Um $|\det(A - \lambda E)|$ auszurechnen, transformieren wir die Matrix $A - \lambda E$ auf einer Zeilenstufenform. Zuerst vertauschen wir die 1. und 3. Spalten :

$$\begin{bmatrix} 1 & 1 & -\lambda \\ 1 & -\lambda & 1 \\ -\lambda & 1 & 1 \end{bmatrix}$$

Danach ziehen wir die 1. Zeile aus der 2. Zeile und addieren das λ -fache von der 1. Zeile zu der 3. Zeile:

$$\begin{bmatrix} 1 & 1 & -\lambda \\ 0 & -\lambda - 1 & \lambda + 1 \\ 0 & \lambda + 1 & -\lambda^2 + 1 \end{bmatrix}$$

Anschließend addieren wir die 2. Zeile zu der 3. Zeile und erhalten die Matrix, die bereits in einer Zeilenstufenform ist:

$$\begin{bmatrix} \mathbf{1} & 1 & -\lambda \\ 0 & -\lambda - \mathbf{1} & \lambda + 1 \\ 0 & & -\lambda^2 + \lambda + \mathbf{1} \end{bmatrix}$$

Daraus ergibt sich die Gleichung:

$$|\det(A - \lambda E)| = |1 \cdot (\lambda + 1) \cdot (-\lambda^2 + \lambda + 2)| = 0.$$

Aus dieser Gleichung erhalten wir, dass die Matrix A die beiden Eigenwerte -1 und 2 besitzt, wobei -1 zweifacher Eigenwert ist.

Zwei $n \times n$ Matrizen R und S heißen *ähnlich*, falls es eine invertierbare Matrix P mit

$$R = P^{-1}SP$$

gibt.

Satz 5.51. Ähnliche Matrizen besitzen denselben Spektrum, d.h. die Mengen ihrer Eigenwerte sind gleich.

Beweis.

$$\begin{aligned} \det(R - \lambda E) &= \det(P^{-1}SP - \lambda P^{-1}P) \\ &= \det(P^{-1}(S - \lambda E)P) \\ &= \det P^{-1} \cdot \det(S - \lambda E) \cdot \det P \\ &= \det(S - \lambda E) \cdot \det(P^{-1}P) \\ &= \det(S - \lambda E). \end{aligned}$$

□

Eine *Diagonalmatrix* ist eine $n \times n$ Matrix $D = (d_{ij})$ mit $d_{ij} = 0$ für alle $i \neq j$, d.h. $D = E \cdot \mathbf{d}$ mit $\mathbf{d} = (d_{11}, d_{22}, \dots, d_{nn})$.

Eine $n \times n$ Matrix ist *diagonalisierbar*, falls sie zu einer Diagonalmatrix ähnlich ist. Solche Matrizen sind gut, da sie handbare Darstellungen haben. Leider ist nicht jede Matrix diagonalisierbar. Dafür braucht man, dass Eigenwerte linear unabhängig sind. Da Eigenvektoren zu verschiedenen Eigenwerten linear unabhängig sein müssen (siehe Aufgabe 41), reicht es (für die Diagonalisierbarkeit), dass alle Eigenwerte verschieden sind.

Satz 5.52. Sei A eine $n \times n$ Matrix. Sind die Eigenvektoren von A linear unabhängig, so ist A diagonalisierbar.

Beweis. Sei Λ eine $n \times n$ Diagonalmatrix, deren Diagonaleinträge die Eigenwerte $\lambda_1, \dots, \lambda_n$ von A sind, und sei $V = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ die $n \times n$ Matrix, deren Spalten die Eigenvektoren sind. Aus $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$ für alle i folgt

$$A \cdot V = A \cdot [\mathbf{v}_1, \dots, \mathbf{v}_n] = [\lambda_1\mathbf{v}_1, \dots, \lambda_n\mathbf{v}_n] = [\mathbf{v}_1, \dots, \mathbf{v}_n] \cdot \Lambda = V \cdot \Lambda.$$

Die inverse V^{-1} existiert genau dann, wenn $\text{rk}(V) = n$ ist, d.h. wenn $\mathbf{v}_1, \dots, \mathbf{v}_n$ linear unabhängig sind. Und in diesem Fall haben wir $A = V^{-1}\Lambda V$. \square

Mit der Induktion kann man leicht zeigen, dass

$$A^k = V\Lambda^k V^{-1}$$

für alle $k = 0, 1, \dots$ gilt. Außerdem ist Λ^k eine Diagonalmatrix mit $\lambda_1^k, \dots, \lambda_n^k$ auf der Diagonale. Diese Beobachtungen geben uns eine geschlossene Form für die Potenzen A_k .

5.13 Einige Anwendungen des Matrizenkalküls*

Im diesen Abschnitt werden wir auf ein paar Beispiele demonstrieren, wie Matrizenkalkül in manchen Situationen helfen kann.

5.13.1 Matrizenkalkül und komplexe Zahlen

Komplexe Zahlen bilden einen Ring, indem die Gleichung $x^2 = -1$ eine Lösung $i = \sqrt{-1}$ hat, d.h. $i^2 = -1$ gilt.

Tatsächlich kennen wir schon einen anderen Ring R , in dem die Gleichung $x^2 = -1$ lösbar ist! Das ist der Ring von 2×2 -Matrizen mit üblichen Matrizenaddition und Multiplikation. Setzen wir nun

$$I := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$E := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

so erhalten wir

$$I^2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = E.$$

Der Ring R enthält zwar \mathbb{R} nicht direkt, aber die Matrizen $aE = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ mit $a \in \mathbb{R}$ verhalten sich beim Rechnen wie die reellen Zahlen selbst, d.h. $\{aE : a \in \mathbb{R}\}$ ist ein zu \mathbb{R} isomorpher Unterkörper in \mathbb{R} . Für diesen konkreten Ring R besteht also die vorhin ausgewählte Teilmenge \mathbb{C} genau aus allen Matrizen

$$aE + bI = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

mit $a, b \in \mathbb{R}$. Mit der Identifikation $E \mapsto 1$ und $I \mapsto i$ sind wir wieder bei der üblichen Schreibweise

$$a + ib \mapsto \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Wozu ist diese Matrixdarstellung der komplexen Zahlen gut? Sie erleichtert die Additions- und Multiplikationsregeln solcher Zahlen zu erinnern:

$$\begin{aligned}(a + ib) + (c + id) &= \begin{bmatrix} a & -b \\ b & a \end{bmatrix} + \begin{bmatrix} c & -d \\ d & c \end{bmatrix} \\ &= \begin{bmatrix} a + c & -(b + d) \\ b + d & a + c \end{bmatrix} \\ &= (a + c) + i(b + d)\end{aligned}$$

und

$$\begin{aligned}(a + ib) \cdot (c + id) &= \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \cdot \begin{bmatrix} c & -d \\ d & c \end{bmatrix} \\ &= \begin{bmatrix} ac - bd & -(ad + bc) \\ ad + bc & ac - bd \end{bmatrix} \\ &= (ac - bd) + i(ad + bc)\end{aligned}$$

Ist die komplexe Zahl $z = a + ib$ als $z = aE + bI = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ dargestellt, so stellt die Matrix

$$\bar{z} := aE - bI = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$$

die *konjugierte* Zahl dar. Ist $z = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, so ist

$$z^{-1} = \begin{bmatrix} \frac{a}{|z|^2} & \frac{b}{|z|^2} \\ -\frac{b}{|z|^2} & \frac{a}{|z|^2} \end{bmatrix} = \frac{\bar{z}}{|z|^2}$$

mit $|z| = \sqrt{a^2 + b^2}$ die multiplikative Inverse von z , da

$$z \cdot z^{-1} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \cdot \begin{bmatrix} \frac{a}{|z|^2} & \frac{b}{|z|^2} \\ -\frac{b}{|z|^2} & \frac{a}{|z|^2} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = I$$

gilt.

5.13.2 Diskrete Fourier-Transformation

Wie kann man zwei Polynome vom Grad n über einem Körper \mathbb{F} möglichst schnell multiplizieren? Es ist klar, dass $O(n^2)$ arithmetischen Operationen reichen aus. Geht es aber schneller? Auf dem ersten Blick sollte es nicht gehen, da jedes Paar von Koeffizienten multipliziert sein muss. Wie wir bald sehen werden, trügt der Schein! Wenn man diese Frage genauer anschaut, reichen sogar $O(n \log n)$ Operationen vollkom aus!

Zunächst machen wir ein paar Bemerkungen über Polynome.

1. Jedes Polynom $a(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1}$ kann man als Koeffizientenvektor $\mathbf{a} = (a_0, a_1, \dots, a_{n-1})$ betrachten. Damit ist für jedes $x \in \mathbb{F}$ der Wert $a(x)$ nichts anderes als das Skalarprodukt des Koeffizientenvektors \mathbf{a} mit dem Potenzenvektor $\mathbf{x} = (1, x, x^2, \dots, x^{n-1})$.

2. Jedes Polynom $a(x)$ vom Grad $n - 1$ ist durch seine Werte $y_1 = a(x_1), \dots, y_n = a(x_n)$ auf beliebigen n verschiedenen Punkten x_1, \dots, x_n eindeutig bestimmt.

Das Ziel der sogenannten “diskreten Fourier Transformation” ist das Rechnen mit Polynomen durch effiziente Matrix-Operationen zu ersetzen. Man will nämlich eine $n \times n$ Matrix M_n bestimmen, so dass

1. Die Matrix-Vektor Produkte $M_n \cdot \mathbf{x}$ und $M_n^{-1} \cdot \mathbf{x}$ viel weniger als n^2 arithmetische Operationen brauchen.
2. Für ein Polynom $a(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1}$ ist $\mathbf{y} = M_n \cdot \mathbf{a}$ die Auswertung

$$y_1 = a(x_1), y_2 = a(x_2), \dots, y_n = a(x_n)$$

von $a(x)$ auf n verschiedenen Punkten x_1, \dots, x_n .

In der diskreten Fourier-Transformation wählt man diese Evaluierungspunkte x_1, \dots, x_n sehr spezifisch aus. Man nimmt nämlich die sogenannte n -te Einheitswurzeln, d.h. die Lösungen der Gleichung $x^n = 1$ über \mathbb{F} .

Eine primitive n -te Einheitswurzel ist eine Zahl $\omega_n \in \mathbb{F} \setminus \{0\}$, so dass $\omega_n^n = 1$ und $\omega_n^k \neq 1$ für alle $k = 1, 2, \dots, n - 1$ gilt.

▷ *Beispiel 5.53* : Der Körper $\mathbb{F} = \mathbb{R}$ der reellen Zahlen hat nur ± 1 als Einheitswurzeln, weil aus $x^n = 1$ auch $|x| = 1$ folgt.

▷ *Beispiel 5.54* : In dem Körper \mathbb{C} der komplexen Zahlen sind die primitive n -te Einheitswurzeln genau die Zahlen $\omega_n^0, \omega_n^1, \dots, \omega_n^{n-1}$ mit

$$\omega_n = e^{\frac{2i\pi}{n}} = \cos \frac{2\pi k}{n} + i \sin \frac{2\pi}{n}.$$

Die Werte $\omega_n^0, \omega_n^1, \dots, \omega_n^{n-1}$ liegen auf dem Einheitskreis und bilden bezüglich der Multiplikation eine Gruppe mit n Elementen. Die Abbildung 5.3 zeigt die primitive 8-te Einheitswurzel $\omega_8 := e^{\frac{2i\pi}{8}}$ und ihre Potenzen. Man rechnet leicht nach, dass $\omega_8^2 = i$ eine primitive 4-te Einheitswurzel ist.

▷ *Beispiel 5.55* : Ist $\mathbb{F} = \mathbb{F}_q$ ein endlicher Körper mit q Elementen, so gibt es (nach dem sogenannten Satz von Artin) eine primitive n -te Einheitswurzel genau dann, wenn n ein Teiler von $q - 1$ ist.

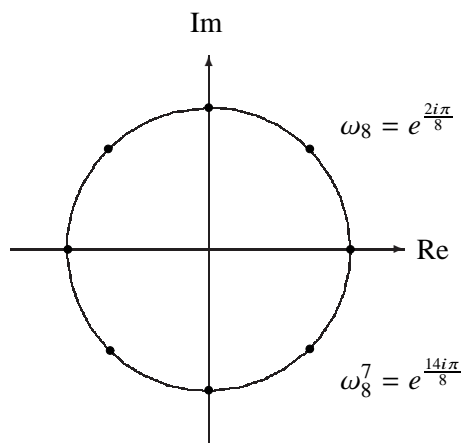


Abbildung 5.3: Die primitive 8-te Einheitswurzel

Lemma 5.56. Sei ω_n eine primitive n -te Einheitswurzel in einem Körper \mathbb{F} .

1. Die Zahlen $1, \omega_n, \omega_n^2, \dots, \omega_n^{n-1}$ sind verschieden und es gilt:

$$1 + \omega_n + \omega_n^2 + \dots + \omega_n^{n-1} = 0.$$

2. Ist $d = \text{ggT}(k, n)$ der größter gemeinsamer Teiler von k und n , so ist

$$\omega_n^k = \omega_{n/d}$$

eine primitive (n/d) -te Einheitswurzel. Insbesondere, ist ω_n^{-1} eine primitive n -te Einheitswurzel.

3. Is $n = 2m$, so gilt

$$\omega_n^k = -\omega_n^{m+k}$$

für alle $k = 0, 1, \dots, m$.

Beweis. Zu (1): Angenommen gebe es $0 \leq s < r \leq n-1$ mit $\omega_n^s = \omega_n^r$. Dann sollte auch $\omega_n^{r-s} = 1$ gelten, was aber unmöglich ist, da $\omega_n^k \neq 1$ für alle $k = 1, 2, \dots, n-1$ gelten muss.

Da $\sum_{k=0}^{n-1} \omega_n^k$ eine geometrische Reihe ist, erhalten wir

$$\sum_{k=0}^{n-1} \omega_n^k = \frac{\omega_n^n - 1}{\omega_n - 1} = 0.$$

Zu (2): Sei $m = n/d$. Wegen $n \mid km$ haben wir $(\omega_n^k)^m = 1$. Nehmen wir an, dass $(\omega_n^k)^j = 1$ für ein $j \in \{1, \dots, m-1\}$ gilt. Dann muss kj durch n teilbar sein und damit muss auch $(k/d)j$ durch $m = n/d$ teilbar sein. Da aber $m = n/d$ und k/d teilerfremd sind, muss (nach Lemma 2.8) auch j durch $m = n/d$ teilbar sein, ein Widerspruch.

Zu (3): Es gilt $(\omega_n^{k+m})^2 = (\omega_n^k)^2 \cdot (\omega_n^n)^2 = (\omega_n^k)^2$, da $\omega_n^n = 1$ gilt. Weiter ist $\omega_n^{k+m} \neq \omega_n^k$, denn sonst wäre $\omega_n^m = 1$ was der Tatsache widerspricht, dass ω_n eine primitive n -te Einheitswurzel ist. Da die Quadrate von ω_n^{k+m} und ω_n^k übereinstimmen, aber $\omega_n^{k+m} \neq \omega_n^k$ gilt, muss $\omega_n^{k+m} = -\omega_n^k$ gelten.

□

Man will nun ein gegebenes Polynom

$$a(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$$

auf n verschiedenen Punkten

$$1, \omega_n, \omega_n^2, \dots, \omega_n^{n-1}$$

auswerten, d.h. alle n Werte

$$y_i = a(\omega_n^i) \quad i = 0, 1, \dots, n-1$$

bestimmen. Die Hauptidee der sogenannten *diskreten Fourier-Transformation* beruht sich auf der Gleichung

$$a(x) = a_{\text{even}}(x^2) + x \cdot a_{\text{odd}}(x^2) \quad (5.13)$$

mit

$$\begin{aligned} a_{\text{even}}(x) &= a_0 + a_2x + a_4x^2 + a_6x^3 + \cdots + a_{2k}x^k + \cdots + a_{n-2}x^{n/2-1} \\ a_{\text{odd}}(x) &= a_1 + a_3x + a_5x^2 + a_7x^3 + \cdots + a_{2k+1}x^k + \cdots + a_{n-1}x^{n/2-1}. \end{aligned}$$

Will man nun ein Polynom $a(x)$ auf der i -ten Potenz ω_n^i der n -ten primitiven Einheitswurzel ω_n auswerten, so reicht es¹² die Polynome $a_{\text{even}}(x)$ und $a_{\text{odd}}(x)$ auf der i -ten Potenz $\omega_{n/2}^i$ der $(n/2)$ -ten primitiven Einheitswurzel $\omega_{n/2}$ auswerten:

$$a(\omega_n^i) = a_{\text{even}}(\omega_{n/2}^i) + \omega_n^i \cdot a_{\text{odd}}(\omega_{n/2}^i) \quad (5.14)$$

D.h. anstatt ein Polynom des Grades $n-1$ an n Punkten $1, \omega_n, \omega_n^2, \dots, \omega_n^{n-1}$ auszuwerten, reicht es zwei Polynome des Grades $n/2-1$ an $n/2$ Punkten $1, \omega_{n/2}, \omega_{n/2}^2, \dots, \omega_{n/2}^{n/2-1}$ auszuwerten.

Diese Beobachtung liefert uns einen rekursiven Auswertungsverfahren, den kann man auch als ein Matrix-Vektor Produkt kurz beschreiben. Dazu betrachten wir die Matrix (die sogenannte Fourier-Matrix)

$$\text{DFT}_n = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_n & \omega_n^2 & \dots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \dots & \omega_n^{2(n-1)} \\ 1 & \omega_n^3 & \omega_n^6 & \dots & \omega_n^{3(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \dots & \omega_n^{(n-1)^2} \end{pmatrix}$$

Die *diskrete Fourier-Transformation* eines Vektors $\mathbf{x} \in \mathbb{F}^n$ ist durch $\mathbf{y} = \text{DFT}_n \cdot \mathbf{x}$ definiert. Oft bezeichnet man diese Operation als

$$\mathbf{y} = \text{DFT}_n(\mathbf{x}).$$

Beachte, dass dann $\mathbf{y} = \text{DFT}_n(\mathbf{a})$ genau die Auswertung des Polynoms $a(x)$ auf n verschiedenen Punkten

$$1, \omega_n, \omega_n^2, \dots, \omega_n^{n-1}$$

ergibt.

¹²Da $\omega_n^2 = \omega_{n/2}$ gilt.

Für einen Vektor $\mathbf{x} = (x_0, x_1, \dots, x_{n-1})$ sei $\mathbf{x}_{\text{even}} = (x_0, x_2, \dots, x_{n-2})$ und $\mathbf{x}_{\text{odd}} = (x_1, x_3, \dots, x_{n-1})$. Aus der Beobachtung (5.14) – zusammen mit der (aus Lemma 5.56(3) folgender) Beobachtung, dass für $\frac{n}{2} \leq i < n$ die Gleichheit $\omega_n^i = \omega_n^{(i-n/2)+n/2} = -\omega_n^{i-n/2}$ gilt – bekommt man unmittelbar die folgende rekursive Gleichung für die Berechnung der diskreten Fourier-Transformation $\text{DFT}_n(\mathbf{x})$:

Satz 5.57. Sei n gerade. Dann gilt

$$\text{DFT}_n(\mathbf{x}) = \begin{pmatrix} \text{DFT}_{n/2} & \Delta_{n/2} \cdot \text{DFT}_{n/2} \\ \text{DFT}_{n/2} & -\Delta_{n/2} \cdot \text{DFT}_{n/2} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_{\text{even}} \\ \mathbf{x}_{\text{odd}} \end{pmatrix}$$

wobei $\Delta_{n/2}$ die $(n/2) \times (n/2)$ Diagonalmatrix mit den Einträgen $1, \omega_n, \omega_n^2, \dots, \omega_n^{n/2-1}$ auf der Diagonale ist.

Sei nun $T(n)$ die Anzahl der arithmetischen Operationen, die man für die Berechnung der diskreten Fourier-Transformation $\text{DFT}_n(\mathbf{x})$ braucht. Da sich die Vektorlänge in jedem Schritt halbiert, gilt

$$T(n) \leq 2 \cdot T\left(\frac{n}{2}\right) + O(n)$$

woraus nach Masters-Theorem (Satz 3.94) $T(n) = O(n \log n)$ folgt. Also ist dieses Verfahren um $\Omega(n/\log n)$ schneller als die triviale Berechnung von $\text{DFT}_n(\mathbf{x})$ mit $O(n^2)$ Operationen.

Die Matrix DFT_n hat einige sehr interessante Eigenschaften, die die Berechnung von $\text{DFT}_n(\mathbf{a})$ wie auch die Berechnung von der Inverse DFT_n^{-1} sehr erleichtern.

Lemma 5.58. Die Inverse DFT_n^{-1} von DFT_n ist gegeben durch

$$(\text{DFT}_n^{-1})_{i,j} = \frac{1}{n} \cdot \omega_n^{-i \cdot j}$$

Beweis. Wir multiplizieren die i -te Zeile von DFT_n mit der j -ter Spalte von DFT_n^{-1} und erhalten den Skalarprodukt

$$\sum_{k=0}^{n-1} \omega_n^{i \cdot k} \cdot \frac{\omega_n^{-k \cdot j}}{n} = \frac{1}{n} \cdot \sum_{k=0}^{n-1} \omega_n^{(i-j) \cdot k} = \frac{\omega_n^{(i-j)}}{n} \cdot \sum_{k=0}^{n-1} \omega_n^k.$$

Wenn $i = j$, dann ist dieser Skalarprodukt gleich $\frac{1}{n}(1 + 1 + \dots + 1) = 1$ und wenn $i \neq j$, dann ist er gleich 0 nach Lemma 5.56(1). \square

Wir wissen bereits, dass die diskrete Fourier-Transformation $\mathbf{y} = \text{DFT}_n(\mathbf{a})$ schnell berechenbar ist. Diese Eigenschaft ist in vielen Anwendungen benutzt. Wir zeigen nur, wie man damit Polynome schnell multiplizieren kann. Sei

$$c(x) = c_0 + c_1x + c_2x^2 + \dots + c_{2n-2}x^{2n-2} = a(x) \cdot b(x)$$

Produkt zweier Polynome

$$\begin{aligned} a(x) &= a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1}, \\ b(x) &= b_0 + b_1x + b_2x^2 + \dots + b_{n-1}x^{n-1}. \end{aligned}$$

Um das Produktpolynom $c(x) = a(x) \cdot b(x)$ zu bestimmen, reicht es den Koeffizienten-Vektor \mathbf{c} zu bestimmen. Wir können o.B.d.A annehmen, dass alle Koeffizienten-Vektoren \mathbf{a} , \mathbf{b} und \mathbf{c} die Länge $2n$ haben – dazu reicht es die Vektoren nach rechts mit n Nullen zu erweitern.

Satz 5.59.

$$\mathbf{c} = \text{DFT}_{2n}^{-1} (\text{DFT}_{2n}(\mathbf{a}) \cdot \text{DFT}_{2n}(\mathbf{b}))$$

wobei \cdot die komponentenweise Multiplikation der Vektoren bezeichnet.

Beweis. Vektoren $\mathbf{y} = \text{DFT}_{2n}(\mathbf{a})$ und $\mathbf{z} = \text{DFT}_{2n}(\mathbf{b})$ geben uns die Auswertungen der Polynome $a(x)$ und $b(x)$ auf $2n$ verschiedenen Punkten – den Potenzen $1, \omega_{2n}, \omega_{2n}^2, \dots, \omega_{2n}^{2n-1}$ eines primitiven $(2n)$ -ten Einheitswurzeln ω_{2n} . Damit ergibt das Vektor $\mathbf{w} = \mathbf{y} \cdot \mathbf{z}$ die Auswertung des Produktpolynoms $c(x) = a(x) \cdot b(x)$ auf diesen $2n$ verschiedenen Punkten. Da der Grad von $c(x)$ kleiner als die Anzahl dieser Punkte is, muss der Koeffizienten-Vektor \mathbf{c} eindeutig durch \mathbf{w} bestimmt sein, und kann als $\mathbf{c} = \text{DFT}_{2n}^{-1} \cdot \mathbf{w}$ berechnet sein. \square

5.13.3 Fehlerkorrigierende Codes

Für die Frage, wozu es denn gut ist, die Vektorräume auch über *endlichen* Körpern zu haben, ist die Codierungstheorie ein wichtiges Beispiel. Da dies Gegenstand einer eigener Vorlesung ist, besprechen wir hier nur Grundfragen zur Anknüpfung an lineare Algebra.

Nun werden an jede Art der Nachrichtübertragung zwei gegensätzliche Forderungen gestellt: Sie soll einerseits so wirtschaftlich (d.h. schnell) wie möglich sein und andererseits so sicher wie möglich sein. Mit Sicherheit ist dabei die Vermeidung von Übertragungsfehlern gemeint, nicht die Abhörsicherung, letztere ist das Feld der Kryptographie, die spezielle Arten der Codierung verwendet (wir haben bereits ein solches Verfahren – die RSA-Codes – im Abschnitt 2.4.1 gelernt).

Die allgemeine Situation besteht aus zwei Spielern: Alice (die Senderin) und Bob (der Empfänger). Alice will Nachrichten an Bob verschicken. Dazu wählen Alice und Bob eine geeignete Teilmenge der binären Strings $C \subseteq \{0, 1\}^n$ aus; C nennt man das *Code*. Dann kodiert Alice ihre Nachricht als ein String $\mathbf{y} = (y_1, \dots, y_n)$ aus C . Während der Übertragung können einige Symbolen y_i geändert sein. Würde man davon ausgehen, dass *alle* n Symbolen geändert sein könnten, wäre dann nichts mehr zu machen – keine Kodierung könnte dann helfen. Also geht man davon aus, dass höchstens bis zu irgendwelcher Anzahl t von Symbolen im String \mathbf{y} geändert sein könnten.

Bob bekommt nun einen String $\mathbf{x} = (x_1, \dots, x_n)$, der sich möglicherweise in bis zu t Positionen von der Originalnachricht \mathbf{y} unterscheidet. Es soll Bob möglich sein, die Originalnachricht \mathbf{y} aus \mathbf{x} wieder (eindeutig!) zu rekonstruieren.

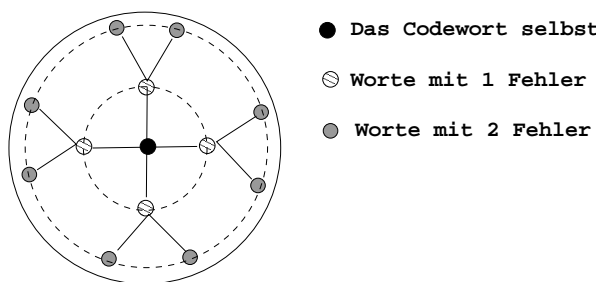
Um die Situation genauer zu beschreiben, brauchen wir den Begriff der “Hammingdistanz” zwischen Strings. Die *Hammingdistanz* $d(\mathbf{x}, \mathbf{y})$ zwischen Strings \mathbf{x}, \mathbf{y} ist die Anzahl der Stellen, in der diese Strings sich unterscheiden: $d(\mathbf{x}, \mathbf{y}) = |\{i : x_i \neq y_i\}|$. Das Code C heißt *t-fehlerkorrigierend*, falls die *minimale Distanz*

$$d(C) := \min\{d(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in C \text{ und } \mathbf{x} \neq \mathbf{y}\}$$

mindestens $2t + 1$ beträgt. Die Bedeutung hier ist klar: Ist $d(C) \geq 2t + 1$, dann kann die *Hammingkugel*

$$B(\mathbf{x}, t) = \{\mathbf{z} : \mathbf{z} \in \mathbb{F}_2^n \text{ mit } d(\mathbf{x}, \mathbf{z}) \leq t\}$$

nur einen Codewort enthalten.



Da während der Übertragung höchstens t Bits von \mathbf{y} (Alices Nachricht) geändert wurde, muss \mathbf{y} das *einzige* Codewort in $B(\mathbf{x}, t)$ sein. Das Problem dabei ist, dieses einzige Codewort zu finden.

Eine Möglichkeit wäre für *alle* Vektoren $\mathbf{z} \in B(\mathbf{x}, t)$ zu testen, ob \mathbf{z} in C liegt. Es sind aber $|B(\mathbf{x}, t)| = \sum_{i=0}^t \binom{n}{i} \approx n^t$ solchen Vektoren, und jeder von ihnen muss mit $|C|$ Vektoren aus C verglichen werden. Das braucht insgesamt $\approx n^t |C|$ Vergleiche. Wenn Alice viele verschiedene Nachrichten verschicken will, dann muss $|C|$ groß sein (da jede Nachricht ihr eigenes Codewort braucht), und die Dekodierung wird dann sehr zeitaufwändig!

Man muss also einen Code C finden, so dass:

1. C groß genug ist
2. $d(C) \geq 2t + 1$ gilt
3. Dekodierung (wie auch Kodierung) relativ leicht sind.

Um diese Ziele zu erreichen, betrachten wir die Strings als Vektoren in dem Vektorraum \mathbb{F}_2^n , wobei $\mathbb{F}_2 = \text{GF}(2)$ der Körper von Charakteristik 2 ist (Addition und Multiplikation modulo 2). Als Code wählen wir einen Vektorraum $C \subseteq \mathbb{F}_2^n$, so dass:

- (i) die Dimension $\dim(C) = k$ groß genug ist (und damit $|C| = 2^k$ auch groß ist), und
- (ii) jeder Vektor $\mathbf{x} \in C$, $\mathbf{x} \neq \mathbf{0}$ mindestens $2t + 1$ von Null verschiedene Bits hat.

Das (ii) äquivalent zu $d(C) \geq 2t + 1$ ist, folgt aus dem folgenden Behauptung. Für ein Vektor $\mathbf{x} \in C$ sei

$$|\mathbf{x}| = x_1 + x_2 + \dots + x_n$$

die Anzahl der Einsen in \mathbf{x} .

Behauptung 5.60. Sei $C \subseteq \mathbb{F}_2^n$ ein linearer Code, $|C| \geq 2$ und sei

$$w(C) = \min\{|\mathbf{x}| : \mathbf{x} \in C \text{ und } \mathbf{x} \neq \mathbf{0}\}.$$

Dann gilt: $d(C) = w(C)$.

Beweis. $d(C) \leq w(C)$: Sei $\mathbf{x} \in C$, $\mathbf{x} \neq \mathbf{0}$ mit $|\mathbf{x}| = w(C)$. Da $\mathbf{0} \in C$ und $\mathbf{x} \neq \mathbf{0}$, haben wir: $d(C) \leq d(\mathbf{x}, \mathbf{0}) = |\mathbf{x}| = w(C)$.

$d(C) \geq w(C)$: Seien nun \mathbf{x} und \mathbf{y} zwei verschiedene Vektoren aus C mit $d(\mathbf{x}, \mathbf{y}) = d(C)$. Da C ein Vektorraum ist, gehört auch der Vektor ¹³ $\mathbf{x} \oplus \mathbf{y}$ zu C . Da $\mathbf{x} \oplus \mathbf{y} \neq \mathbf{0}$ (die Vektoren \mathbf{x}, \mathbf{y} sind ja verschieden), haben wir $d(C) = d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} \oplus \mathbf{y}| \geq w(C)$. \square

¹³Wie es üblich ist, bezeichnen wir die Addition in $\text{GF}(2)$ durch \oplus , d.h. $x \oplus y = x + y \text{ mod } 2$.

Es bleibt also nur zu zeigen, wie man mit Hilfe eines linearen Codes C effizient kodieren und dekodieren kann. Dazu benutzt man sogenannte Generator- und Kontrollmatrizen.

1. Eine *Generatormatrix* (oder *Erzeugermatrix*) G des Codes ist eine $k \times n$ -Matrix, deren Zeilen eine Basis von C bilden; also gilt $C = \{\mathbf{y}^\top G : \mathbf{y} \in \mathbb{F}^k\}$.
2. Eine *Kontrollmatrix* (oder *Prüfmatrix*) H des Codes ist eine $(n - k) \times n$ -Matrix, deren Zeilen eine Basis von C^\perp bilden; also gilt $C = \{\mathbf{x} \in \mathbb{F}^n : H\mathbf{x} = \mathbf{0}\}$.

Die Zeilen von G und die Zeilen von H sind also paarweise orthogonal. Der Name¹⁴ “Kontrollmatrix” erklärt sich von selbst: Da $(C^\perp)^\perp = C$ ist, gehört ein $\mathbf{x} \in \mathbb{F}^n$ genau dann zu C , wenn $H \cdot \mathbf{x} = \mathbf{0}$ gilt. Es gilt also: Ist $C \subseteq \mathbb{F}^n$ ein linearer Code mit der Kontrollmatrix H , so gilt für alle $\mathbf{x} \in \mathbb{F}^n$

$$\mathbf{x} \in C \iff H \cdot \mathbf{x} = \mathbf{0}$$

D.h. Codeworte sind genau die Vektoren, die senkrecht zu *allen* Zeilen von H liegen.

Sind die Matrizen G und H gegeben, so ist die Kodierung (für Alice) sehr einfach: Um ihre Nachricht $\mathbf{v} \in \mathbb{F}_2^k$ zu kodieren, multipliziert sie die Generatormatrix G von links mit \mathbf{v}^\top und verschickt das Codewort $\mathbf{y} = \mathbf{v}^\top \cdot G$.

Bob bekommt einen Vektor \mathbf{x} , der sich möglicherweise in bis zu t Bits von der Originalnachricht \mathbf{y} unterscheidet. Um ihm zu dekodieren, berechnet Bob den Vektor $s(\mathbf{x}) := H \cdot \mathbf{x}$ (das sogenannte *Syndrom* von \mathbf{x}) und benutzt den folgenden Fakt. Sei $d(\mathbf{x}, C) = \min\{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in C\}$.

Lemma 5.61. Für jedes $\mathbf{x} \in \mathbb{F}_2^n$ mit $d(\mathbf{x}, C) \leq t$ gibt es *genau einen* Vektor $\mathbf{a} \in B(\mathbf{0}, t)$ mit $s(\mathbf{x}) = s(\mathbf{a})$ und $\mathbf{x} \oplus \mathbf{a} \in C$.

Beweis. Da wir $\leq t$ Fehlern haben können, wissen wir, dass es *mindestens ein* $\mathbf{a} \in B(\mathbf{0}, t)$ mit $\mathbf{x} \oplus \mathbf{a} \in C$ gibt. Dann ist $\mathbf{0} = H(\mathbf{x} \oplus \mathbf{a}) = H\mathbf{x} \oplus H\mathbf{a}$ und, da wir in dem Körper $GF(2)$ arbeiten, die Gleichheit $H\mathbf{x} = H\mathbf{a}$ folgt.

Um die Eindeutigkeit von \mathbf{a} zu beweisen, nehmen wir an, dass es auch ein anderes $\mathbf{b} \in B(\mathbf{0}, t)$ mit $\mathbf{b} \neq \mathbf{a}$ und $\mathbf{x} \oplus \mathbf{b} \in C$ gibt. Seien $\mathbf{u} = \mathbf{x} \oplus \mathbf{a}$ und $\mathbf{v} = \mathbf{x} \oplus \mathbf{b}$. Da C linear ist, muss auch der Vektor $\mathbf{u} \oplus \mathbf{v} = \mathbf{b} \oplus \mathbf{a}$ in C liegen. Da aber $\mathbf{b} \oplus \mathbf{a} \neq \mathbf{0}$, $\mathbf{0} \in C$ und C t -fehlerkorrigierend ist, muss dann $d(\mathbf{b} \oplus \mathbf{a}, \mathbf{0}) \geq 2t + 1$ gelten. Andererseits gilt:

$$d(\mathbf{b} \oplus \mathbf{a}, \mathbf{0}) = d(\mathbf{b}, \mathbf{a}) \leq d(\mathbf{b}, \mathbf{0}) + d(\mathbf{0}, \mathbf{a}) \leq 2t,$$

ein Widerspruch mit $d(C) \geq 2t + 1$. □

Also reicht es Bob in einer (im voraus vorbereiteten) Liste L den einzigen String $\mathbf{a} \in B(\mathbf{0}, t)$ mit $s(\mathbf{a}) = s(\mathbf{x})$ zu finden: Dann ist $\mathbf{x} \oplus \mathbf{a}$ der \mathbf{x} am nächsten liegendes Codewort und deshalb muss $\mathbf{y} = \mathbf{x} \oplus \mathbf{a}$ genau die von Alice geschickte Nachricht sein.

Zusammengefasst läuft die Kommunikation folgendermaßen ab:

1. Alice kodiert ihre Nachricht $\mathbf{v} \in \{0, 1\}^k$ als $\mathbf{y} = \mathbf{v}^\top G$ und verschickt sie.
2. Bob bekommt einen Vektor \mathbf{x} , der kann von \mathbf{y} in bis zu t Bits unterscheiden. Er sucht dann den einzigen Vektor $\mathbf{a} \in \{0, 1\}^n$ mit $|\mathbf{a}| \leq t$ Einsen, für den $H \cdot \mathbf{a} = H \cdot \mathbf{x}$ gilt. Dann ist $\mathbf{y} = \mathbf{x} \oplus \mathbf{a}$ die von Alice geschickte (kodierte) Nachricht.

¹⁴Auf englisch “parity-check matrix”

3. Bob weiss, dass die von Alice geschickte Originalnachricht \mathbf{v} eine Lösung \mathbf{z} des Gleichungssystems $\mathbf{z}^T G = \mathbf{y}$ ist. Er löst¹⁵ also dieses Gleichungssystem, um eine Lösung \mathbf{z} zu bekommen. Da die Zeilen von G eine *Basis* von C bilden, weiss Bob (nach Satz 5.5), dass dann $\mathbf{z} = \mathbf{v}$ gelten muss. Er hat also die von Alice geschickte Nachricht rekonstruiert.

▷ *Beispiel 5.62* : $\mathbb{F} = GF(2)$

$$C = \text{span}(\{(00000), (11010), (01101), (10111)\})$$

$d(C) = 3 \implies C$ ist t -fehlerkorrigierend mit $t = 1$. Kontrollmatrix:

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

Die folgende Tabelle gibt alle Vektoren aus $B(\mathbf{0}, 1)$ und die zugehörigen Syndrome wieder:

$$L = \begin{array}{|c|c|} \hline \mathbf{a} & s(\mathbf{a}) \\ \hline 00000 & 000 \\ 10000 & 100 \\ 01000 & 010 \\ 00100 & 001 \\ 00010 & 110 \\ 00001 & 011 \\ \hline \end{array}$$

Ist etwa $\mathbf{x} = (11110)$ die eintreffende Botschaft, so gehört hierzu das Syndrom $s(\mathbf{x}) = H \cdot \mathbf{x} = (001)$ mit dem eindeutigen $\mathbf{a} = (00100)$ und wir erhalten $\mathbf{y} = \mathbf{x} \oplus \mathbf{a} = (11010)$ als Nachricht.

▷ *Beispiel 5.63* : Das sogenannte *Hammingcode* $C \subseteq GF(2)^n$ ist ein lineares Code mit $n = 2^r - 1$ und Dimension $k = n - r$. Seine Kontrollmatrix H ist eine $r \times (2^r - 1)$ Matrix, deren Spalten alle binäre Codes der Zahlen $1, 2, \dots, n$ sind. Dieses Code ist 1-fehlerkorrigierend. (Warum? Übungsaufgabe.) In diesem Fall ist $|C| = 2^{n-r} = 2^{2^r-1} - r$ aber die Syndrom-Liste L ist sehr kurz: $|L| \leq n + 1 = 2^r$. Für $r = 3$ ist $|C| \geq 2^{2^3-1} - 3 = 2^7 - 3 = 124 - 3 = 121$ und $|L| \leq 2^3 = 8$. Die entsprechenden Generator- und Kontrollmatrizen sind (verifiziere, dass die Zeilen von G und die Zeilen von H paarweise orthogonal sind!)

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \quad H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

5.13.4 Expandergraphen

Eigenwerte haben viele Anwendungen in der Diskreten Mathematik und in der Informatik. Und der Grund dafür ist, dass die Eigenwerte der Adjazenzmatrizen ungerichteter Graphen $G = (V, E)$ einige

¹⁵Zum Beispiel mit dem Gauß-Verfahren (siehe Abschnitt 5.8).

wichtige Eigenschaften (wie der “Expansionsgrad”) widerspiegeln. Graphen mit großem Expansionsgrad sind in vielen Gebieten anwendbar: Kodierungstheorie (z.B. Tornado Codes), Komplexitätstheorie (z.B. gute Zufallsgeneratoren), Schaltkreiskomplexität (um gute untere Schranken zu beweisen), Konstruktion von WWW-Suchmaschinen, usw.

Sei $G = (V, E)$ ein ungerichteter Graph und $S \subseteq V$. Mit

$$\Gamma(S) = \{v \in V : uv \in E \text{ für ein } u \in S\}$$

bezeichnen wir die Menge aller Nachbarn von S , d.h. die Menge aller Knoten, die mit mindestens einer Knoten aus S benachbart sind. Ein Graph ist d -regulär, falls jeder Knoten genau d Nachbarn hat.

Definition: Sei G ein d -regulärer Graph $G = (V, E)$ mit $|V| = n$ Knoten. Dann ist G ein (n, d, c) -Expander, falls

$$|\Gamma(S) \setminus S| \geq c|S|$$

für jede Teilmenge $S \subseteq V$ mit $|S| \leq \frac{1}{2}|V|$ gilt.

Natürlich ist jeder zusammenhängender d -regulärer Graph ein (n, d, c) -Expander für irgendein $c > 0$. Andererseits, der vollständiger Graph K_n ist n -regulär und sehr stark expandiert, da $|\Gamma(S) \setminus S| \geq n - |S|$ gilt. In Anwendungen aber braucht man, dass der Graph so “dünn” wie möglich ist, d.h. die Anzahl $|E|$ der Kanten muss so klein wie möglich sein. Insbesondere, braucht man *explizite* Familien von (n, d, c) -Expandern mit $n \rightarrow \infty$, wobei beide Parameter d und $c > 0$ *konstant(!)* sind. Diese “Expansionseigenschaft” (jede Menge hat viele echten Nachbarn, obwohl der Graph sehr dünn ist) spielt eine wichtige Rolle in verschiedenen Feldern der diskreten Mathematik und Informatik.

Die Existenz von Expandergraphen kann man relativ leicht mittels sogenannter “probabilistischen Methode” nachweisen (siehe Abschnitt 4.17). Es gibt auch einige *explizite Konstruktionen* von Expandergraphen. Hier wir erwähnen nur zwei davon.

Konstruktion 1: Als Knotenmenge nehmen wir die Menge $V = \mathbb{Z}_m \times \mathbb{Z}_m$. Jeder der Knoten (x, y) ist verbunden mit 4 Knoten $(x + y, y), (x - y, y), (x, y + y), (x, x - y)$ (alle Operationen sind modulo m).

Konstruktion 2: Als Knotenmenge nehmen wir $V = \mathbb{Z}_p$, wobei p eine Primzahl ist. Jeder der Knoten x ist verbunden mit 3 Knoten $x + 1, x - 1, x^{-1}$ (hier wiederum ist $x \pm 1$ modulo p und x^{-1} ist die multiplikative Inverse von x in \mathbb{Z}_p); der Knoten $x = 0$ ist mit den Knoten $p - 1, 0$ und 1 verbunden.

Es ist aber oft sehr schwierig *nachzuweisen*, dass ein *bestimmter* Graph auch ein Expander ist. In einem solchen Nachweis spielen die Eigenwerte der Adjazenzmatrix eine große Rolle.

Sei $G = (V, E)$ ein ungerichteter Graph mit der Knotenmenge V , $|V| = n$, und sei $A = (a_{uv})$ seine Adjazenzmatrix, d.h.

$$a_{uv} = \begin{cases} 1 & \text{falls } uv \in E \\ 0 & \text{falls } uv \notin E. \end{cases}$$

Da der Graph ungerichtet ist, gilt: $uv \in E \iff vu \in E$. Also ist die Adjazenzmatrix symmetrisch ($A^T = A$ gilt). Für solchen Graphen gilt folgendes (ohne Beweis):

Satz 5.64. Sei A eine symmetrische $n \times n$ -Matrix über \mathbb{R} . Dann gilt:

1. A hat n reellen Eigenwerte $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.
2. Es gibt Eigenvektoren, die eine Orthonormalbasis von \mathbb{R}^n bilden, wobei der Eigenvektor zu λ_1 ist

$$\mathbf{v}_1 = \frac{1}{\sqrt{n}} \mathbf{1} = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right).$$

3. Die ersten zwei Eigenwerte sind gegeben durch:

$$\begin{aligned} \lambda_1 &= \max \left\{ \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2} : \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \right\}, \\ \lambda_2 &= \max \left\{ \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2} : \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \text{ und } \langle \mathbf{x}, \mathbf{1} \rangle = 0 \right\}. \end{aligned}$$

Für $S, T \subseteq V$ sei

$$e(S, T) = |\{uv : u \in S, v \in T \text{ und } uv \in E\}|$$

die Anzahl aller Kanten, die Knoten in S mit Knoten in T verbinden. Beachte, dass die Mengen S und T nicht unbedingt disjunkt sein müssen.

Lemma 5.65. Sei $G = (V, E)$ ein ungerichteter d -regulärer Graph auf $n = |V|$ Knoten. Sei $\lambda = \lambda_2$ der zweitgrößte Eigenwert von G . Dann gilt für jede Zerlegung $V = S \cup T$:

$$e(S, T) \geq \frac{(d - \lambda)|S||T|}{n} \quad (5.15)$$

Beweis. Sei $s = |S|$ und $t = |T| = n - s$. Betrachte den Vektor $\mathbf{x} : V \rightarrow \mathbb{R}$ mit

$$x_u = \begin{cases} -t & \text{falls } u \in S, \\ s & \text{falls } u \in T. \end{cases}$$

Der Vektor \mathbf{x} hat also die Form:

$$\mathbf{x} = (\overbrace{-t, -t, \dots, -t}^{|S| \text{ mal}}, \overbrace{s, s, \dots, s}^{|T| \text{ mal}})$$

Daraus folgt, dass Vektor \mathbf{x} senkrecht zum Vektor $\mathbf{1}$ steht: Dann ist

$$\langle \mathbf{x}, \mathbf{1} \rangle = \sum_{u \in S} (-t) + \sum_{u \in T} s = s(-t) + ts = 0.$$

Ausserdem gilt

$$\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \sum_{u \in S} (-t)^2 + \sum_{u \in T} s^2 = st^2 + ts^2 = st(s + t) = stn,$$

woraus nach Satz 5.64 die Ungleichung

$$\langle \mathbf{x}, A\mathbf{x} \rangle \leq \lambda \|\mathbf{x}\|^2 = \lambda stn \quad (5.16)$$

folgt. Andererseits, gilt

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \sum_{u \in V} x_u \left(\sum_{v: uv \in E} x_v \right) = 2 \sum_{uv \in E} x_u x_v.$$

Nach der Definition von \mathbf{x} gilt für jede Kante $uv \in E$:

$$x_u x_v = \begin{cases} t^2 & \text{falls } u, v \in S, \\ s^2 & \text{falls } u, v \in T, \\ -st & \text{falls } u \in S \text{ und } v \in T. \end{cases}$$

Daraus folgt:

$$\begin{aligned} 2 \sum_{uv \in E} x_u x_v &= 2 [t^2 \cdot e_S + s^2 \cdot e_T - st \cdot e(S, T)] \\ &= t^2 \cdot (2e_S) + s^2 \cdot (2e_T) - 2st \cdot e(S, T), \end{aligned}$$

wobei e_S bzw. e_T die Anzahl der Kanten bezeichnet, deren beide Endknoten in S bzw. in T liegen.

Wir wollen nun die Terme e_S und e_T durch $e(S, T)$ ersetzen, und dazu benutzen wir die Regularität. Wenn wir die Grade aller Knoten in S aufsummieren, dann kommt $2e_S + e(S, T)$ raus (wir müssen $2e_S$ anstatt e_S nehmen, da jede Kante mit *beiden* Endknoten in S nicht 1 sondern 2 zu dieser Summe beiträgt). Andererseits, ist diese Summe gleich $d|S| = ds$, da der Graph d -regulär ist. Deshalb gilt $2e_S + e(S, T) = ds$ oder äquivalent: $2e_S = ds - e(S, T)$. Genauso gilt $2e_T = dt - e(S, T)$. Wir setzen diese Werte in die obige Gleichung und erhalten

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x} \rangle &= t^2 \cdot (2e_S) + s^2 \cdot (2e_T) - 2st \cdot e(S, T) \\ &= t^2 \cdot (ds - e(S, T)) + s^2 \cdot (dt - e(S, T)) - 2st \cdot e(S, T) \\ &= (t^2 ds + s^2 dt) - e(S, T) \cdot (t^2 + 2st + s^2) \\ &= std(t + s) - e(S, T) \cdot (t + s)^2 \\ &= stdn - e(S, T) \cdot n^2, \quad (\text{da } s + t = n \text{ gilt}). \end{aligned}$$

Zusammen mit (5.16) ergibt dies

$$e(S, T) = \frac{dstdn - \langle \mathbf{x}, A\mathbf{x} \rangle}{n^2} \geq \frac{dstdn - \lambda stn}{n^2} = \frac{(d - \lambda)st}{n},$$

wie erwünscht. □

Eine wichtige Konsequenz aus Lemma 5.65 ist, dass d -reguläre Graphen mit $\lambda < d$ auch gute Expander sind.

Korollar 5.66. Sei $G = (V, E)$ ein ungerichteter d -regulärer Graph mit n Knoten, und sei $\lambda = \lambda_2$ der zweiter Eigenwert seiner Adjazenzmatrix. Dann ist G ein (n, d, c) -Expander mit

$$c \geq \frac{d - \lambda}{2d}.$$

Beweis. Sei $S \subseteq V$ mit $|S| \leq n/2$ und sei $\bar{S} = V \setminus S$. Eine Kante $uv \in E$ mit $u \in S$ kann zwischen S und \bar{S} liegen, nur wenn der andere Endpunkt v in $\Gamma(S) \setminus S$ liegt. Deshalb gilt $e(S, \bar{S}) \leq d \cdot |\Gamma(S) \setminus S|$. Zusammen mit (5.15) ergibt dies

$$|\Gamma(S) \setminus S| \geq \frac{(d-\lambda)|S|(n-|S|)}{dn} \geq \frac{d-\lambda}{d} \cdot |S| \left(1 - \frac{|S|}{n}\right) \geq \frac{d-\lambda}{2d} \cdot |S|.$$

□

5.13.5 Expander-Codes

Expandergraphen erlauben uns sehr effiziente fehlerkorrigierende Codes zu entwerfen (sogenannte “Tornado Codes”). Hier beschreiben wir nur die Hauptidee dieser Codes.

Sei $G = (L \cup R, E)$ ein bipartiter Graph mit $|L| = n$, $|R| = m$ und $E \subseteq L \times R$. Der Code $C \subseteq \{0, 1\}^n$ zum Graphen G ist wie folgt definiert. Jedem Knoten $u \in L$ ist eine Variable x_u zugeordnet. Ist $\mathbf{x} \in \{0, 1\}^n$ eine Belegung dieser Variablen, so heißt der Knoten $v \in R$ *zufrieden mit \mathbf{x}* , falls

$$\sum_{u \in \Gamma(v)} x_u \bmod 2 = 0$$

gilt (siehe Abbildung 5.4). Dann ist der Code zum Graphen G definiert durch:

$$C = \{\mathbf{x} \in \{0, 1\}^n : \text{alle Knoten in } R \text{ sind mit } \mathbf{x} \text{ zufrieden}\}.$$

Es ist klar, dass das Code C linear ist, d.h. der Vektor $\mathbf{x} \oplus \mathbf{y}$ für alle $\mathbf{x}, \mathbf{y} \in C$ in Code liegt. Außerdem ist das Code durch eine lineare Gleichungssystem mit m Gleichungen und n Variablen definiert, woraus $|C| \geq 2^{n-m}$ folgt.

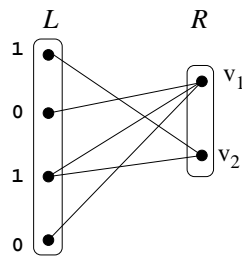


Abbildung 5.4: Vektor $\mathbf{x} = (1010)$. Knoten v_2 ist mit \mathbf{x} zufrieden, Knoten v_1 aber nicht.

Sei $\text{dist}(C)$ der minimale Hamming-Abstand zwischen zwei verschiedenen Vektoren in C . Ein bipartiter Graph $G = (L \cup R, E)$ heißt *links d -regulär*, falls jeder Knoten $u \in L$ denselben Grad d hat.

Lemma 5.67. Sei G ein links d -regulärer (α, c) -Expander mit $c \geq \frac{1}{2}d$ und sei $C \subseteq \{0, 1\}^n$ der Code zu G . Dann gilt:

$$\text{dist}(C) > \alpha n.$$

Beweis. Wir nehmen an, dass $\text{dist}(C) \leq \alpha n$ gilt. Dann muss es einen Vektor $\mathbf{x} \in C$ mit $|\mathbf{x}| \leq \alpha n$ Einsen geben (siehe Behauptung 5.60). Sei nun $S = \{u \in L : x_u = 1\}$. Dann gilt $|S| \leq \text{dist}(C) \leq \alpha n$.

Da G ein $(\alpha, d/2)$ -Expander ist, gilt damit auch $|\Gamma(S)| > \frac{1}{2}d|S|$. Wir behaupten, dass es mindestens einen Knoten $v_0 \in \Gamma(S)$ mit *genau einen* Nachbar in S geben muss. Wäre das nämlich nicht der Fall, so hätten wir

$$2 \cdot |\Gamma(S)| > 2 \cdot \frac{d}{2}|S| = d|S|$$

Kanten zwischen S und $\Gamma(S)$, was unmöglich ist, da jeder Knoten in S Grad d hat.

Da aber $x_u = 0$ für alle $u \notin S$, ist *genau ein* von der Bits x_u mit $u \in \Gamma(v_0)$ gleich 1. Deshalb ist $\sum_{u \in \Gamma(v_0)} x_u = 1$ und der Knoten v_0 kann nicht zufrieden mit dem Vektor \mathbf{x} sein, ein Widerspruch mit $\mathbf{x} \in C$. \square

Nach Lemma 5.67 können Expander Codes ziemlich viel (bis zu $\alpha n/2$) Übertragungsfehler korrigieren. Viel wichtiger aber ist, dass der Dekodierungs-Algorithmus für solche Codes verblüffend einfach ist.

Wir sagen, dass ein Knoten $u \in L$ von \mathbf{x} *kritisch* ist, falls mehr als die Hälfte der Nachbarn von u unzufrieden mit dem aktuellen Vektor \mathbf{x} sind.

WHILE $\exists u \in L$ und u kritisch DO ersetze das Bit x_u durch $x_u \oplus 1$.

Im Verlauf des Algorithmus werden verschiedene Bits geflippt. Es ist klar, dass in jedem Schritt (falls er unternommen wird!) wird sich die Anzahl der unzufriedenen Knoten in R verkleinern. D.h. der Algorithmus wird bestimmt terminieren. Das Problem aber ist, dass der resultierende Vektor \mathbf{x} muss nicht unbedingt in C liegen! Zum Beispiel, wenn G der im Abbildung 5.4 dargestellte Graph ist und $\mathbf{x} = (1010)$ ist, so wird der Algorithmus keinen Bit von \mathbf{x} flippen, da keiner der Knoten in L kritisch ist. Der Vektor \mathbf{x} aber gehört nicht zum C (da v_1 unzufrieden mit \mathbf{x} ist), aber der Algorithmus gibt \mathbf{x} aus.

Es ist deshalb interessant, dass der Algorithmus bereits korrekt funktionieren wird, wenn der Graph G ein Expander ist!

Lemma 5.68. Sei G ein links d -regulärer (α, c) -Expander mit $c > \frac{3}{4}d$ und sei $C \subseteq \{0, 1\}^n$ der Code zu G . Seien $\mathbf{y} \in C$ und $\mathbf{y}' \in \{0, 1\}^n$. Ist $\text{dist}(\mathbf{y}, \mathbf{y}') \leq \alpha n/2$, so wird der Algorithmus die Originalnachricht \mathbf{y} aus dem (beschädigten) Vektor \mathbf{y}' in höchstens $|R| = m$ Schritten rekonstruieren.

Beweis. Wir betrachten den allgemeinen Schritt. Sei \mathbf{x} der bis zum diesen Schritt vom Algorithmus erzeugte Vektor, und sei $S = \{u \in L : x_u \neq y_u\}$ die Menge der aktuellen "Fehlerbits". Ausserdem, sei Z bzw. U die Menge aller Nachbarn von Knoten in S , die zufrieden bzw. unzufrieden mit dem (aktuellen) Vektor \mathbf{x} sind.¹⁶

Zuerst zeigen wir, dass jeder Nachbarn von S , der mit dem Vektor \mathbf{x} zufrieden ist, muss mindestens zwei Nachbarn in S haben, d.h.

$$|S \cap \Gamma(v)| \geq 2 \tag{5.17}$$

für alle $v \in \Gamma(S)$ gilt. Angenommen, es gibt einen Knoten $v \in Z$, der nur einen Nachbarn u in S hat. Da $\mathbf{y} \in C$, muss der Knoten v zufrieden mit \mathbf{y} sein, d.h. die Anzahl der Einsen in $\{y_u : u \in \Gamma(v)\}$ muss gerade sein. Aber auf diesen Bits unterscheidet \mathbf{x} vom \mathbf{y} an *genau einer* Stelle x_u . Deshalb kann v nicht zufrieden mit \mathbf{x} sein, ein Widerspruch mit $v \in Z$.

¹⁶Beachte, dass der Algorithmus nur die Mengen U und Z in jedem Schritt kennt, die Menge S ist ihm unbekannt.

Wir behaupten nun folgendes:

Solange $0 < |S| \leq \alpha n$ gilt, wird der Algorithmus mindestens ein Bit flippen.

Da $|S| \leq \alpha n$, muss S um Faktor $> \frac{3}{4}d$ expandieren, woraus die Ungleichung

$$|Z| + |U| = |\Gamma(S)| > \frac{3}{4}d|S| \quad (5.18)$$

folgt. Betrachte nun die Menge aller $d|S|$ Kanten zwischen S und $\Gamma(S) = Z \cup U$. Mindestens $|U|$ von diesen Kanten sind inzident mit den Knoten in U , und nach (5.17) mindestens $2|Z|$ diesen Kanten müssen inzident mit den Knoten in Z sein. Deshalb gilt

$$2|Z| + |U| \leq d|S|. \quad (5.19)$$

Aus (5.18) und (5.19) folgt

$$d|S| - |U| \geq 2|Z| > 2 \left(\frac{3}{4}d|S| - |U| \right)$$

oder äquivalent

$$|U| > \frac{1}{2}d|S|. \quad (5.20)$$

Also, mehr als $\frac{1}{2}d|S|$ der Nachbarn der $|S|$ Knoten sind unzufrieden. Deshalb muss es einen Knoten $u \in S$ geben, so dass u mehr als $d/2$ unzufriedenen Nachbarn hat, und der Algorithmus flippt den Bit x_u .

Es bleibt also zu zeigen, dass die Invariante $|S| \leq \alpha n$ stets gilt (bis S leer ist). Am Anfang haben wir $|S| = \text{dist}(\mathbf{y}, \mathbf{y}') \leq \alpha n/2$, und deshalb auch

$$|U| \leq |\Gamma(S)| \leq \frac{1}{2}\alpha dn. \quad (5.21)$$

Ausserdem, muss diese Ungleichung in *jedem* Schritt gelten, da sich in jedem Schritt (falls er geschieht) die Menge U verkleinert. Warum? Durch Flippen eines Bits x_u sind nur die Nachbarn von u betroffen und nach dem Flippen haben wir *mehr* zufriedene davon als unzufriedene. Also muss auch $|S| \leq \alpha n$ in jedem Schritt gelten, denn sonst (wenn $|S| > \alpha n$) hätten wir nach (5.20) $|U| > \frac{1}{2}d|S| > \frac{1}{2}\alpha dn$, ein Widerspruch mit (5.21).

Damit ist Lemma 5.68 bewiesen. □

5.13.6 Markov-Ketten

In diesem Abschnitt betrachten wir eine Situation, wo drei uns bereits bekannte Gebiete – Stochastik, Matrizenkalkül und Analysis – sich zusammen treffen.

Man hat ein System, das in jedem Zeitpunkt in einem der m Zustände $V = \{1, \dots, m\}$ sein kann. Sei p_{ij} die Wahrscheinlichkeit, dass das System im nächsten Schritt zum Zustand j übergeht, falls es jetzt in Zustand i ist. Da die p_{ij} 's Wahrscheinlichkeiten sein sollen, muss $0 \leq p_{ij} \leq 1$ für alle i, j und

$$\sum_{j=1}^m p_{ij} = 1$$

für alle i gelten. Wir haben also eine unendliche Folge der Zufallsvariablen $\langle X_n \rangle$, die Werte in $\{1, 2, \dots, m\}$ annehmen. Der Index n entspricht den Zeiteinheiten (=Schritten) und X_n stellt den Zustand des Systems nach n Schritten dar. Der wichtigste Unterschied der Markov-Ketten¹⁷ von allgemeinen stochastischen Prozessen ist ihre "Gedächtnislosigkeit": In jedem Schritt i hängen die Ereignisse " $X_{n+1} = j$ " nur von der Ereignissen " $X_n = i$ " ab. D.h. die Wahrscheinlichkeit, in welchen Zustand das System (unser Experiment) im $(n + 1)$ -ten Schritt übergehen wird, hängt nur von dem Zustand ab, in dem das System sich gerade (im Schritt n) befindet (und hängt nicht von der Vergangenheit ab). Mathematisch lässt sich diese Eigenschaft so auszudrücken:

$$\Pr \{X_{n+1} = j \mid X_n = i, X_{n-1} = k, \dots, X_0 = l\} = \Pr \{X_{n+1} = j \mid X_n = i\} = p_{ij}.$$

Das System kann man auch als ein *gerichteten* Graph $G = (V, E)$ darstellen, wobei die Kante von i nach j (falls sie vorhanden ist¹⁸) ist mit der Wahrscheinlichkeit p_{ij} markiert; die Zahl p_{ij} ist also die Wahrscheinlichkeit, dass beginnend in Knoten (= Zustand) i das System in einem Schritt zum Knoten (= Zustand) j übergehen wird. Die Matrix

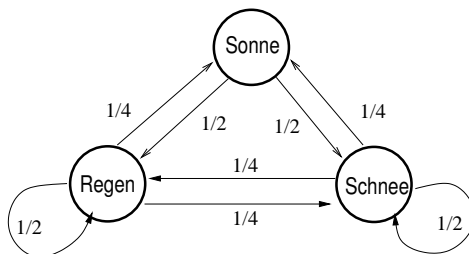
$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mm} \end{bmatrix}$$

heißt die *Übergangsmatrix*. Ein solches System heißt *Markov-Kette*.

Satz 5.69. Ist P die Übergangsmatrix einer Markov-Kette, so ist P_{ij}^n die Wahrscheinlichkeit, in n Schritten den Knoten j aus Knoten i zu erreichen.

Beweis. Völlig analog mit dem Beweis vom Satz 5.17. □

- *Beispiel 5.70* : Ein Land irgendwo am Rande der Welt ist durchaus schön, nur mit dem Wetter haben die Einwohner Pech gehabt: sie haben nie zwei Tage nacheinander die Sonne. Falls sie die Sonne haben, dann wird bestimmt am nächsten Tag keine Sonne sein, sondern es wird entweder regnen oder schneien, und zwar mit gleicher Wahrscheinlichkeit $1/2$. Wenn es gerade regnet oder schneit, dann gibt es eine $1/2$ Chance, dass genau so auch am nächsten Tag sein wird. Wenn am den Tag schneit oder regnet, dann ist die Chance, am nächsten Tag Sonne zu haben, nur $1/4$:



¹⁷Engl. *Markov chains*.

¹⁸Die nicht vorhandene Kanten (i, j) tragen also die Wahrscheinlichkeit $p_{ij} = 0$.

Damit haben wir drei mögliche Zustände 1 = "regnet", 2 = "sonnig", 3 = "schneit" und die Übergangsmatrix sieht folgendermaßen aus:

$$P = \begin{array}{l} \text{regnet} \\ \text{sonnig} \\ \text{schneit} \end{array} \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

Wenn man die Potenzen P^n für $n = 2, 3, \dots$ betrachtet, bekommt man sehr bald (bereits bei $n = 7$) die Matrix

$$P^7 = \begin{bmatrix} 0,4 & 0,2 & 0,4 \\ 0,4 & 0,2 & 0,4 \\ 0,4 & 0,2 & 0,4 \end{bmatrix}$$

D.h. langfristig gesehen (mit Abstand von mindestens einer Woche) wird sich das Wetterverhalten stabilisieren: egal welches Wetter heute ist, wird in einer Woche nur mit Wahrscheinlichkeit 0,2 sonnig sein, und mit gleicher Wahrscheinlichkeit 0,4 wird entweder regnen oder schneien.

Eine interessante Klasse sind Markov-Ketten mit sogenannten "absorbierenden" Zuständen. Ein Zustand i ist *absorbierend*, falls $p_{ii} = 1$ gilt. D.h. kommt einmal das System in Zustand i , so kann es den Zustand nicht mehr verlassen. Wenn zum Beispiel man mit einer Markov-Kette ein System von Lebewesen modelliert, dann hat dieses System einen absorbierenden Zustand "Tod" (wie in dem Beispiel mit dem Frosch und dem Storch im Abschnitt 4.15). Die nicht absorbierenden Zustände heißen auch *vorübergehende* Zustände.

Eine Markov-Kette heißt *absorbierend*, wenn sie folgenden zwei Bedingungen erfüllt:

1. Die Kette hat mindestens einen absorbierenden Zustand.
2. Es ist möglich, aus jedem nicht absorbierenden Zustand einen absorbierenden Zustand zu erreichen (vielleicht in mehr als einem Schritt).

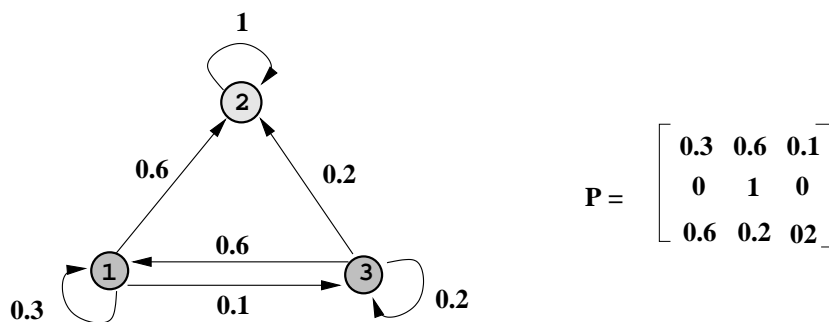
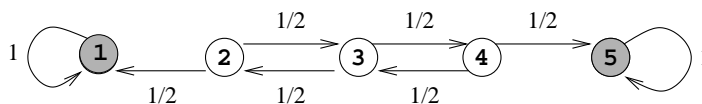


Abbildung 5.5: Eine absorbierende Markov-Kette mit 3 Zuständen und seine Übergangsmatrix P . Der Zustand 2 ist absorbierend.

► **Beispiel 5.71 : (Spaziergang eines Betrunkenen)** Ein stark betrunkenener Man spaziert entlang des aus vier Strecken bestehenden Park Avenue:



Wenn er an einer der Ecken 2, 3 oder 4 ist, so geht er nach links oder nach rechts mit gleicher Wahrscheinlichkeit $1/2$. Er bewegt sich so bis er entweder die Ecke 5 (das ist eine Bar) oder die Ecke 1 (da ist seine Wohnung) erreicht; falls er eine dieser zwei Ecken erreicht, bleibt er da. Wir können dieses System durch folgende Übergangsmatrix darstellen:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Die Kette ist absorbierend, da von nicht absorbierenden Zuständen 2, 3 und 4 kann man die absorbierende Zustände 1 und 5 erreichen.

Typische Fragen für gegebene absorbierende Markov-Kette sind:

1. Ist das System im nicht absorbierenden Zustand i , wie lange wird es durchschnittlich dauern bis das System in einen der absorbierenden Zustände übergeht?
2. Ist das System im nicht absorbierenden Zustand i , mit welcher Wahrscheinlichkeit wird dann das System in einen absorbierenden Zustand j übergehen?

Diese Frage kann man mit Hilfe der Matrizenalgebra relativ leicht beantworten.

Sei P die Übergangsmatrix einer absorbierenden Markov-Kette. Nummeriere die Zustände so, dass nicht-absorbierende Zustände zuerst kommen. Falls die Kette r absorbierenden und s nicht-absorbierenden Zustände hat, dann sieht die *kanonische Form* der Übergangsmatrix wie folgt aus:

$$P = \left[\begin{array}{c|c} Q & R \\ \hline \mathbf{0} & E \end{array} \right] \quad (5.22)$$

Hier ist E die $r \times r$ Einheitsmatrix und $\mathbf{0}$ ist die $r \times s$ Nullmatrix. Beachte, dass die n -te Potenz P^n von P die folgende Form hat (nachrechnen!):

$$P^n = \left[\begin{array}{c|c} Q^n & * \\ \hline \mathbf{0} & E \end{array} \right]$$

Damit ist Q^n_{ij} die Wahrscheinlichkeit, in n Schritten den vorübergehenden Zustand j aus vorübergehendem Zustand i zu erreichen. Zunächst zeigen wir, dass diese Wahrscheinlichkeiten mit wachsendem n gegen Null streben.¹⁹

Lemma 5.72. In einer absorbierenden Markov-Kette gilt $\lim_{n \rightarrow \infty} Q^n = \mathbf{0}$. D.h. mit Wahrscheinlichkeit 1 wird die Kette aus jedem vorübergehendem Zustand zu einem absorbierendem Zustand übergehen.

¹⁹Den Begriff der Konvergenz kann man leicht auf die Folgen von Matrizen übertragen. Dazu reicht es die ϵ -Umgebung $U_\epsilon(A)$ einer Matrix $A = (a_{ij})$ als die Menge aller Matrizen $B = (b_{ij})$ mit $|b_{ij} - a_{ij}| < \epsilon$ für alle i, j definieren.

Beweis. Aus jedem vorübergehenden Zustand i ist es möglich einen absorbierenden Zustand zu erreichen. Sei t_i die minimale Anzahl der Schritte aus Zustand i einen absorbierenden Zustand zu erreichen. Sei p_i die Wahrscheinlichkeit, dass das System aus Zustand i in t_i Schritten einen absorbierenden Zustand *nicht* erreichen wird. Dann gilt $p_i < 1$. Sei t die größte Zahl aus t_1, \dots, t_s und p die größte Zahl aus p_1, \dots, p_s . Die Wahrscheinlichkeit, in t Schritte nicht absorbiert zu sein, ist $\leq p$; für $2t$ Schritten ist diese Wahrscheinlichkeit $\leq p^2$, usw. Da $p < 1$, strebt die Folge $(Q^{it} : i = 1, 2, \dots)$ gegen der Nullmatrix $\mathbf{0}$ und damit muss ²⁰ auch die Folge $(Q^n : n = 1, 2, \dots)$ gegen der Nullmatrix $\mathbf{0}$ streben. \square

Satz 5.73. Sei P die Übergangsmatrix einer absorbierenden Markov-Kette mit n Zuständen und sei Q die nicht absorbierende Teilmatrix von P . Dann hat die Matrix $E - Q$ die Inverse $N = (E - Q)^{-1}$ und es gilt: $N = E + Q + Q^2 + \dots$ Sei weiterhin (v_1, \dots, v_s) die i -te Zeile von N . Dann gilt:

$$\sum_{j=1}^s v_j = \text{die erwartete Anzahl der Schritte bis die Kette in einen absorbierenden Zustand landet, wenn sie in Zustand } i \text{ startet}$$

Die Matrix $N = (E - Q)^{-1}$ nennt man die *fundamentale Matrix* der Markov-Kette.

Beweis. Sei $(E - Q)\mathbf{x} = \mathbf{0}$, d.h. $\mathbf{x} = Q \cdot \mathbf{x}$. Daraus folgt: $\mathbf{x} = Q^n \mathbf{x}$ für alle $n = 1, 2, \dots$. Da $\lim_{n \rightarrow \infty} Q^n = \mathbf{0}$ ist, muss $\lim_{n \rightarrow \infty} Q^n \mathbf{x} = \mathbf{0}$ und somit auch $\mathbf{x} = \mathbf{0}$ gelten. D.h. das Gleichungssystem $(E - Q)\mathbf{x} = \mathbf{0}$ kann nur triviale Lösung $\mathbf{x} = \mathbf{0}$ haben woraus folgt, dass die Matrix $E - Q$ invertierbar ist. Sei $N = (E - Q)^{-1}$ die inverse Matrix von $E - Q$. Beachte dass

$$(E - Q)(E + Q + Q^2 + \dots + Q^n) = E - Q^{n+1}$$

gilt. Multiplizieren wir beide Seiten mit $N = (E - Q)^{-1}$, so erhalten wir

$$E + Q + Q^2 + \dots + Q^n = N(E - Q^{n+1}).$$

Strebt nun n gegen ∞ , so strebt nach Lemma 5.72 die Folge der Matrizen Q^{n+1} gegen der Nullmatrix $\mathbf{0}$. Somit gilt:

$$N = E + Q + Q^2 + \dots$$

d.h. die Reihe $Q^0 + Q^1 + Q^2 + \dots$ mit $Q^0 = E$ konvergiert (und sogar absolut, da Einträge nicht-negativ sind) gegen den Grenzwert N :

$$N = \lim_{n \rightarrow \infty} \sum_{k=0}^n Q^k$$

Es bleibt zu zeigen, dass N_{ij} genau die erwartete Anzahl der Besuche des Zustands j ist, wenn das System in Zustand i startet. Somit ist die Summe $\sum_{j=1}^s v_j = \sum_{j=1}^s N_{ij}$ die erwartete Anzahl der Schritte bis die Kette in einen absorbierenden Zustand landet, wenn sie in Zustand i startet.

Seien nun i, j beliebig (aber fest) und sei X_k die Indikatorvariable für das Ereignis, dass das System den Zustand j in genau k Schritten erreicht, wenn es im Zustand i startet. Die Zufallsvariable $Y_n :=$

²⁰Die Folge $(Q^n : n = 1, 2, \dots)$ ist monoton fallende Folge von Matrizen.

$X_1 + X_2 + \dots + X_n$ beschreibt also die Anzahl der Besuche des Zustands j in ersten n Schritten, wenn das System in Zustand i startet. Dann gilt $\Pr\{X_k = 1\} = Q_{ij}^k$, $\Pr\{X_k = 0\} = 1 - Q_{ij}^k$ und

$$E[Y_n] = E\left[\sum_{k=0}^n X_k\right] = \sum_{k=0}^n E[X_k] = \sum_{k=0}^n \Pr\{X_k = 1\} = Q_{ij}^0 + Q_{ij}^1 + Q_{ij}^2 + \dots + Q_{ij}^n.$$

Andererseits ist der Grenzwert $\lim_{n \rightarrow \infty} E[Y_n]$ genau die erwartete Anzahl der Besuche des Zustands j , wenn das System in Zustand i startet. Nach der unendlichen Linearität des Erwartungswertes gilt:

$$\lim_{n \rightarrow \infty} E[Y_n] = E\left[\sum_{n=0}^{\infty} Y_n\right] = \sum_{n=0}^{\infty} E[Y_n] = Q_{ij}^0 + Q_{ij}^1 + Q_{ij}^2 + \dots = N_{ij}$$

□

Die nächste Frage ist, mit welcher *Wahrscheinlichkeit* wird das System aus einem vorübergehenden Zustand i in einen absorbierenden Zustand j übergehen? Die Antwort ist im folgenden Satz gegeben.
21

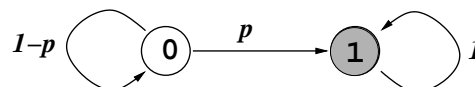
Satz 5.74. Sei $B = N \cdot R$. Dann ist B_{ij} die Wahrscheinlichkeit, dass das System im Zustand j absorbiert wird, wenn es in vorübergehenden Zustand i startet.

Beweis. Da $N = E + Q + Q^2 + \dots$, haben wir $B = NR = ER + QR + Q^2R + \dots$. Es gilt also:²²

$$\begin{aligned} B_{ij} &= \sum_{n=0}^{\infty} \sum_{k=1}^m Q_{ik}^n \cdot R_{kj} \\ &= \sum_{k=1}^m \sum_{n=0}^{\infty} Q_{ik}^n \cdot R_{kj} \\ &= \sum_{k=1}^m N_{ik} \cdot R_{kj} \\ &= NR_{ij}. \end{aligned}$$

□

► **Beispiel 5.75 : (Geometrische Verteilung)** Wir wiederholen das Bernoulli-Experiment X_1, X_2, \dots mit Erfolgswahrscheinlichkeit $p \neq 0$ oftmals und wollen die Erwartete Anzahl der Mißerfolge (= Anzahl der Nullen) vor dem ersten Erfolg (vor der ersten Eins) bestimmen. Die entsprechende Zufallsvariable ist $Y = \min\{T : X_T = 1\}$. Das Experiment kann man als die folgende absorbierende Markov-Kette darstellen:



²¹Hier ist R die $s \times r$ Teilmatrix aus der kanonischen Form (5.22) der Übergangsmatrix P .

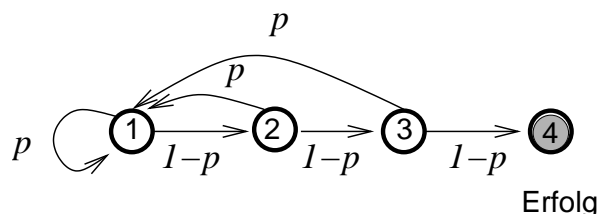
²²Da $Q^0 = E$

Die entsprechende Übergangsmatrix (bereits in einer kanonischer Form) ist

$$P = \left[\begin{array}{c|c} 1-p & p \\ \hline 0 & 1 \end{array} \right]$$

Somit ist $Q = 1 - p$, $E - Q = 1 - (1 - p) = p$ und $N = (E - Q)^{-1} = \frac{1}{p}$. Ausserdem gilt: $N \cdot R = \frac{1}{p} \cdot p = 1$, d.h. dass System wird mit Wahrscheinlichkeit 1 den Zustand 1 (Erfolg) erreichen.

- *Beispiel 5.76*: **(Runs)** Wir wiederholen wiederum das Bernoulli-Experiment X_1, X_2, \dots mit *Mißerfolgswahrscheinlichkeit* $p \neq 0$ vielmals und wollen die Erwartete Anzahl der Versuche vor dem ersten Erfolg bestimmen. Versuche sind Zahlen 0 und 1, wobei eine Null mit Wahrscheinlichkeit p kommt. Als Erfolg nehmen wir ein Run 111 aus drei nacheinander folgenden Einsen. Das Experiment kann man als die folgende absorbierende Marko-Kette darstellen:



Die entsprechende Übergangsmatrix (bereits in einer kanonischer Form) ist

$$P = \left[\begin{array}{ccc|c} p & 1-p & 0 & 0 \\ p & 0 & 1-p & 0 \\ p & 0 & 0 & 1-p \\ \hline 0 & 0 & 0 & 1 \end{array} \right]$$

Die Matrix Q ist also

$$Q = \left[\begin{array}{ccc} p & 1-p & 0 \\ p & 0 & 1-p \\ p & 0 & 0 \end{array} \right]$$

Nun benutzen wir das Program maple um die Inverse $N = (E - Q)^{-1}$ auszurechnen:²³

$$N = \left[\begin{array}{ccc} -\frac{1}{-1+3p-3p^2+p^3} + \frac{1}{1-2p+p^2} - \frac{1}{-1+p} & & \\ \frac{p(p-2)}{-1+3p-3p^2+p^3} + \frac{1}{1-2p+p^2} - \frac{1}{-1+p} & & \\ -\frac{p}{-1+3p-3p^2+p^3} + \frac{p}{1-2p+p^2} - \frac{1}{-1+p} & & \end{array} \right]$$

²³Wir können das tun, da wir bereits wissen, wie man die Inversen berechnen kann, und wir können diese Routinearbeit dem Computer überlassen.

und

$$N \cdot \mathbf{1} = \begin{bmatrix} -\frac{p^2 - 3p + 3}{(-1 + p)(1 - 2p + p^2)} \\ \frac{p - 2}{(-1 + p)(1 - 2p + p^2)} \\ 1 \\ -\frac{1}{(-1 + p)(1 - 2p + p^2)} \end{bmatrix}$$

Somit ist die erwartete Anzahl der Versuche (beginnend im Zustand 0), bis ein Run 111 kommt, gleich (die erste Zeile von $N \cdot \mathbf{1}$):

$$-\frac{p^2 - 3p + 3}{(-1 + p)(1 - 2p + p^2)} = \frac{1 - (1 - p)^3}{p(1 - p)^3},$$

eine Formel, die wir bereits mit Hilfe von Wald's Theorem (für beliebig langen Runs) bewiesen haben (siehe Beispiel 4.94).

► **Beispiel 5.77: (Spaziergang eines Betrunkenen – Fortsetzung)** Die kanonische Form für die Matrix dieser Markov-Kette ist:

$$P = \left[\begin{array}{ccc|cc} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right]$$

Also ist

$$Q = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}$$

und

$$E - Q = \begin{bmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{bmatrix}$$

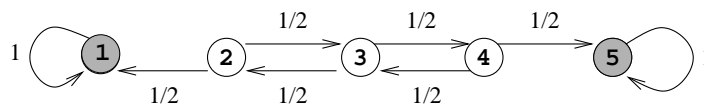
Berechnet man nun $(E - Q)^{-1}$ (z.B. mit Gauß-Jordan Verfahren, siehe Abschnitt 5.9), so bekommt man

$$N = (E - Q)^{-1} = \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix}$$

Damit ist

$$N \cdot \mathbf{1} = \begin{matrix} 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix}$$

Wenn der Betrunkene in einem der Zustände 2, 3 oder 4 startet, dann wird er im Durchschnitt 3, 4 oder 3 Schritte machen, bis er absorbiert (in der Kneipe oder zu Hause) wird.



Und mit welchen Wahrscheinlichkeiten wird der Betrunkene absorbiert? Aus der kanonischen Form der Übergangsmatrix für diese Markov-Kette haben wir:

$$R = \begin{matrix} 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Dann ist das Produkt $B = NR$ gleich

$$\begin{aligned} B = NR &= \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix} \cdot \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{bmatrix} \\ &= \begin{matrix} 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} \end{aligned}$$

Diese Matrix gibt uns die Wahrscheinlichkeiten, mit denen der Betrunkene im Zustand 1 (zu Hause) oder im Zustand 5 (die Kneipe) absorbiert wird, wenn er in einen der Zustände 2, 3 oder 4 startet. Startet er z.B. im Zustand 2, so wird er nur mit Wahrscheinlichkeit $3/4$ (und nicht mit Wahrscheinlichkeit $1/2$, wie man vermuten könnte!) nach Hause kommen und immerhin mit Wahrscheinlichkeit $1/4$ die Kneipe erreichen.

5.14 Aufgaben

5.1. Sei V ein Vektorraum über einem Körper \mathbb{F} mit $\text{char}(\mathbb{F}) \neq 2$. Seien \mathbf{u} und \mathbf{v} zwei linear unabhängige Vektoren in V . Zeige, dass dann auch die Vektoren $\mathbf{x} = \mathbf{u} - \mathbf{v}$ und $\mathbf{y} = \mathbf{u} + \mathbf{v}$ linear unabhängig sind.

5.2. Es sei \mathbb{F} ein Körper und A eine (beliebige) Menge. Die Menge \mathbb{F}^A aller Funktionen $f : X \rightarrow \mathbb{F}$ mit der Addition

$$(f + g)(x) := f(x) + g(x)$$

und der Skalarmultiplikation mit einer reellen Zahl λ gemäß

$$\lambda f(x) := \lambda \cdot f(x) \quad \forall x \in X$$

bildet einen Vektorraum über dem Körper \mathbb{F} .

(a) (**Unabhängigkeits-Kriterium**) Seien nun $f_1, \dots, f_m : A \rightarrow \mathbb{F}$ Funktionen, für die es Elemente a_1, \dots, a_m mit den folgenden Eigenschaften gibt:

- (i) $f_i(a_i) \neq 0$ für alle $1 \leq i \leq m$,
- (ii) $f_i(a_j) = 0$ für alle $1 \leq j < i \leq m$.

Zeige, dass dann f_1, \dots, f_m als Vektoren in \mathbb{F}^A linear unabhängig sind.

(b) Zeige, dass die Funktionen f, g, h linear unabhängig (in Vektorraum $\mathbb{R}^{\mathbb{R}}$) sind, wobei

$$f(x) = e^x, \quad g(x) = x^4, \quad h(x) = 4x$$

5.3. Beweise oder widerlege: Die drei Vektoren $1, \sqrt{2}, \sqrt{3}$ im Vektorraum \mathbb{R} über dem Körper \mathbb{Q} sind linear unabhängig. Sind die drei Vektoren linear unabhängig über dem Körper \mathbb{R} ?

5.4. Sei A eine $m \times n$ Matrix, und $\mathbf{b} \in \mathbb{R}^m$ ein Vektor, $\mathbf{b} \neq \mathbf{0}$. Ist dann die Menge $U = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n, A\mathbf{x} = \mathbf{0}\}$ ein Vektorraum? Ist dann die Menge $V = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n, A\mathbf{x} = \mathbf{b}\}$ ein Vektorraum?

5.5. Sei $L : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ die Abbildung definiert durch die Matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 1 & 1 & -2 \end{bmatrix}$$

Bestimme Basen und Dimensionen von $\text{Im } L$ und $\text{Ker } L$.

Hinweis: $\text{Im } L$ ist der Spaltenraum von $A \Rightarrow \dim(\text{Im } L)$ ist der Zeilenrang von der transponierten Matrix A^T ; diese Matrix bekommt man, wenn man die Spalten von A als Zeilen schreibt. Um den Zeilenrang (und damit auch den Rang) einer Matrix zu bestimmen, transformiert man die Matrix zu einer Zeilenstufenform.

5.6. Die Abbildung $L : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ sei definiert durch

$$L \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y - z \\ y + z \\ x + y - 2z \end{pmatrix}$$

Bestimme Basen und Dimensionen der Untervektorräume $\text{Im } L$ und $\text{Ker } L$.

5.7. Sei $L : V \rightarrow W$ eine lineare Abbildung. Dann gilt offensichtlich: L ist surjektiv $\iff \text{Im } L = W$. Zeige, dass

$$L \text{ ist injektiv} \iff \text{Ker } L = \{\mathbf{0}\}$$

5.8. Ein Vater und seine beide Söhne sind zusammen hundert Jahre alt, der Vater ist doppelt so alt wie sein ältester Sohn und dreissig Jahre älter als sein jüngster. Wie alt ist der Vater?

5.9. Wir wollen einen Obstsalat aus Äpfeln, Bananen und Orangen. Die Nährstoffanteile seien in folgender Tabelle angegeben:

	Eiweiß	Fett	Kohlenhydrate
Apfel	0,3	0,6	15
Bananen	1,1	0,2	22
Orangen	1,0	0,2	12

Stelle einen Obstsalat zusammen, der insgesamt 9 g Eiweiß, 5 g Fett und 194 g Kohlenhydrate enthält.

5.10. Löse die folgenden Gleichungssysteme:

$$\begin{bmatrix} 1 & 2 & 0 \\ 3 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 3 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 0 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 2 & 2 \\ 2 & 1 & 0 \\ 4 & 2 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

5.11. Löse die Gleichung:

$$\begin{bmatrix} -4 & x \\ -x & 4 \end{bmatrix}^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

5.12. Für welche Werte von a hat das folgende Gleichungssystem (i) *keine* bzw. (ii) *genau eine* bzw. (iii) *mehrere Lösungen*?

$$\begin{aligned}x + y + az &= 1 \\x + ay + z &= 1 \\ax + y + z &= 1\end{aligned}$$

5.13. Berechne $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^n$, wobei n eine natürliche Zahl sein soll.

5.14. Bestimme die inverse Matrizen von:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

5.15. Gegeben ist die Matrix $A = \begin{bmatrix} 1 & 3 \\ 4 & -3 \end{bmatrix}$. Gesucht ist ein Vektor $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, so dass gilt: $A\mathbf{x} = \mathbf{x}$.

5.16. Ermittle den Rang folgender Matrizen:

$$A = \begin{bmatrix} 1 & -1 & 3 & 2 \\ -2 & -1 & 4 & 3 \\ 0 & -3 & 10 & 7 \\ -7 & 1 & -1 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 5 & 3 & 4 & 2 \\ 2 & 1 & 0 & 3 & 1 \\ 1 & 14 & 9 & 9 & 5 \\ -1 & 4 & 3 & 1 & 1 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 3 & 0 & 1 & 4 \\ 5 & 1 & 2 & 3 & 1 \\ 2 & 2 & 1 & 0 & 3 \\ 4 & 2 & 1 & 4 & 2 \\ 2 & -4 & 1 & 2 & 6 \end{bmatrix}$$

5.17. Es sei $A = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$. Berechne

(a) $A^2 + 3A - 10E$

(b) $2A^2 - 3A + 5E$

5.18. Es sei $A = \begin{bmatrix} 1 & 2 \\ 4 & -3 \end{bmatrix}$. Berechne

(a) $A^2 + 2A - 11E$

(b) $2A^3 - 4A + 5E$

5.19. Bestimme die Inverse A^{-1} von

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 2 \end{bmatrix}$$

falls sie existiert.

5.20. Sei $B = PAP^{-1}$, wobei A und P invertierbare $n \times n$ Matrizen sind. Finde B^{-1} .

5.21. Seien A und B invertierbare $n \times n$ Matrizen über \mathbb{R} . Zeige, dass dann die Spalten von $A^{-1}B$ den ganzen Vektorraum \mathbb{R}^n erzeugen.

5.22. Nehmen wir an, dass die letzte Spalte von AB eine Nullspalte $\mathbf{0}$ ist, aber die Matrix B selbst keine Nullspalte enthält. Was kann man dann über die Spalten von A sagen?

5.23. Zeige: Wenn die Spalten von B linear abhängig sind, dann sind auch die Spalten von AB linear abhängig.

5.24. Sei $CA = E_n$ (die $n \times n$ Einheitsmatrix). Zeige, dass dann die Gleichungssystem $A\mathbf{x} = \mathbf{0}$ nur die triviale Lösung $\mathbf{x} = \mathbf{0}$ haben kann.

5.25. Sei $AD = E_m$ (die $m \times m$ Einheitsmatrix). Zeige, dass dann die Gleichungssystem $A\mathbf{x} = \mathbf{b}$ für alle $\mathbf{b} \in \mathbb{R}^m$ lösbar ist.

5.26. Sei $(B - C)D = \mathbf{0}$, wobei B und C $m \times n$ Matrizen sind, und die Matrix D invertierbar ist. Zeige, dass dann $B = C$ gelten muss.

5.27. Berechne die Inverse A^{-1} von

$$A = \begin{bmatrix} 0 & 1 & -1 \\ 4 & -3 & 4 \\ 3 & -3 & 4 \end{bmatrix}$$

falls sie existiert.

5.28. Berechne die Eigenwerte von

$$A = \begin{bmatrix} 1 & -3 & 5 \\ 0 & 1 & 0 \\ 0 & -2 & 2 \end{bmatrix}$$

5.29. Sei $G = (V, E)$ ein ungerichteter Graph mit $V = \{1, \dots, n\}$. Ein *Dreieck* in G ist eine Menge $\{i, j, k\}$ aus drei Knoten die paarweise adjazent sind, d.h. $\{i, j\} \in E$, $\{i, k\} \in E$ und $\{j, k\} \in E$. Sei $A = (a_{ij})$ die Adjazenzmatrix von G und $A^2 = (b_{i,j})$ Zeige, dass G ein Dreieck genau dann besitzt, wenn es ein Paar $1 \leq i < j \leq n$ mit $a_{ij} \neq 0$ und $b_{ij} \neq 0$ gibt.

Fazit: Da die Multiplikation zweier Matrizen relativ schnell ausgerechnet sein kann (in Zeit ungefähr n^2), haben wir damit einen Algorithmus zur bestimmung der Dreiecksfreiheit von Graphen, der wesentlich schneller als der trivialer Algorithmus (der alle $\binom{n}{3} \approx n^3$ Möglichkeiten ausprobiert) läuft.

5.30. Definiere die Matrizen H_{2m} , $m = 2, 4, 8, \dots$ rekursiv wie folgt:

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad H_{2m} = \begin{bmatrix} H_m & H_m \\ H_m & -H_m \end{bmatrix}.$$

Zeige, dass die Matrizen H_{2m} regulär sind, d.h. vollen Rang haben.

5.31. Sei \mathbb{F} ein Körper und \mathcal{M}_n die Menge aller $n \times n$ -Matrizen über \mathbb{F} . Zeige, dass dann $(\mathcal{M}_n, +, \cdot)$ ein Ring ist.

5.32. Die *Disjunktheitsmatrix* (engl. *disjointness matrix*) ist eine $2^n \times 2^n$ 0-1 Matrix D_n , deren Zeilen und Spalten mit Teilmengen einer n -elementigen Menge markiert sind, und

$$D_n(A, B) = 1 \iff A \cap B = \emptyset.$$

Zeige, dass D_n über jeden Körper \mathbb{F} den vollen Rang 2^n hat. *Hinweis:* Benutze die Induktion über n zusammen mit der folgender rekursiver Konstruktion von D_n :

$$D_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad D_n = \begin{bmatrix} D_{n-1} & D_{n-1} \\ D_{n-1} & 0 \end{bmatrix}$$

5.33. Die “Schneidungsmatrix” (engl. *intersection matrix*) ist eine $(2^n - 1) \times (2^n - 1)$ 0-1 Matrix Q_n deren Zeilen und Spalten mit *nicht leeren* Teilmengen einer n -elementigen Menge markiert sind, und $Q_n(A, B) = 1 \iff A \cap B \neq \emptyset$. Zeige, dass diese Matrix auch den vollen Rang über jeden Körper \mathbb{F} hat. *Hinweis:* Sei I_n die $2^n \times 2^n$ Matrix, die nur aus Einsen besteht. Dann ist $I_n - D_n$ genau die Matrix Q_n mit einer zusätzlicher Nullzeile und einer zusätzlicher Nullspalte.

5.34. Sei A eine $m \times n$ Matrix über dem Körper $\mathbb{F} = GF(2)$ (nur zwei Elemente 0 und 1 mit der Addition und Multiplikation modulo 2) und $\text{span}(A)$ sei ihr Zeilenraum. Ein Vektor $\mathbf{x} \in \mathbb{F}^n$ heißt *gerade* bzw. *ungerade*, falls die Anzahl $|\mathbf{x}| = x_1 + \dots + x_n$ der Einsen in \mathbf{x} gerade bzw. ungerade ist. Sei $\mathbf{1} = (1, \dots, 1)$. Zeige folgendes: $\mathbf{1} \notin \text{span}(A)$ genau dann, wenn es einen ungeraden Vektor $\mathbf{x} \in \mathbb{F}^n$ mit $A \cdot \mathbf{x} = \mathbf{0}$ gibt.

5.35. (Putnam Exam, 1991) Seien A und B zwei *verschiedene* $n \times n$ -Matrizen über \mathbb{R} mit $A^3 = B^3$ und $A^2B = B^2A$. Zeige, dass dann die Matrix $A^2 + B^2$ keine Inverse haben kann. *Hinweis:* Betrachte $(A^2 + B^2)(A - B)$.

5.36. Seien $\mathbf{v}_1, \dots, \mathbf{v}_k \in V \setminus \{\mathbf{0}\}$ mit $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ für alle $i \neq j$. Zeige, dass dann $\mathbf{v}_1, \dots, \mathbf{v}_k$ linear unabhängig sind.

5.37. Zeige, dass für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ gilt:

(a) $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$.

(b) Ist $\|\mathbf{x}\| = \|\mathbf{y}\|$, so sind die Vektoren $\mathbf{x} + \mathbf{y}$ und $\mathbf{x} - \mathbf{y}$ orthogonal.

5.38. Benutze die Cauchy–Schwarz Ungleichung um folgendes zu Zeigen: Ist $\mathbf{u} = (u_1, \dots, u_n)$ ein Vektor in \mathbb{R}^n , dann gilt: $|\mathbf{u}| \leq \sqrt{n} \cdot \|\mathbf{u}\|$, wobei $|\mathbf{u}| := |u_1| + \dots + |u_n|$ und $|u_i|$ der Betrag von u_i ist. *Hinweis:* Betrachte den Vektor $\mathbf{v} = (v_1, \dots, v_n)$ mit $v_i = 1$ falls $u_i > 0$, und $v_i = -1$ sonst. Beachte, dass $|\mathbf{u}| = \langle \mathbf{u}, \mathbf{v} \rangle$ und $\|\mathbf{v}\| = \sqrt{n}$.

5.39. Seien x_1, \dots, x_n reelle Zahlen und $\sigma : [n] \rightarrow [n]$ sei eine Permutation von $\{1, \dots, n\}$. Zeige, dass dann $\sum_{i=1}^n x_i \cdot x_{\sigma(i)} \leq \sum_{i=1}^n x_i^2$. *Hinweis:* Cauchy–Schwarz Ungleichung.

5.40. Sei V von den Vektoren $\mathbf{v}_1 = (1, 1, -1)$, $\mathbf{v}_2 = (-1, 2, 2)$ und $\mathbf{v}_3 = (1, 4, 0)$ erzeugter Unterraum von \mathbb{R}^3 . Wende den Gauß–Schmidt-Orthogonalisierungsverfahren, um die Orthonormalbasis von V zu bestimmen.

5.41. Zeige, dass Eigenvektoren zu verschiedenen Eigenwerten linear unabhängig sein müssen. *Hinweis:* Als erstes zeige folgendes: Ist $\mathbf{x} = \alpha \mathbf{y}$, so können \mathbf{x}, \mathbf{y} nicht Eigenvektoren zu verschiedenen Eigenwerten sein.

5.42. Beweise oder widerlege: Wenn alle Einträge einer quadratischen Matrix A nicht negativ sind, dann sind auch alle Eigenwerte von A nicht negativ.

5.43. Sei A eine reelwertige $n \times n$ Matrix und sei $B = A \cdot A^T$. Zeige, dass alle Eigenwerte von B nicht negativ sind.

5.44. Eine $n \times n$ Matrix heißt *symmetrisch*, falls $A^T = A$ gilt. Zeige, dass Eigenvektoren zu verschiedenen Eigenwerten einer symmetrischen Matrix orthogonal sind.

5.45. Sei $C \subseteq \mathbb{F}^n$ ein lineares Code der Dimension $k = \dim(C)$. Sei $G = [E_k | A]$ seine Generatormatrix. Zeige, dass dann die Kontrollmatrix die Form $H = [-A^T | E_{n-k}]$ hat.

5.46. Die *Vandermonde-Matrix* (engl. *Vandermonde matrix*) ist eine $n \times n$ Matrix X_n , deren i -te Zeile die Form $(1, x_i, x_i^2, \dots, x_i^{n-1})$ mit $x_i \in \mathbb{F}$ hat:

$$X_n = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix}$$

Zeige, dass

$$\det X_n = \prod_{1 \leq i < j \leq n} (x_j - x_i).$$

Hinweis: Induktion über n . Multipliziere jede Spalte mit x_1 und subtrahiere sie von der (nächsten) rechten Spalte, beginnend mit der rechten Spalte. Das sollte $\det(X_n) = (x_n - x_1) \cdots (x_2 - x_1) \det(X_{n-1})$ liefern.

5.47. Sei V ein Vektorraum der Dimension n über einem Körper \mathbb{F} , und sei $W \subseteq V$. Man sagt, dass die Vektoren aus W in *allgemeiner Lage* (in *general position*) sind, falls *jede*(!) $n = \dim(V)$ Vektoren aus W linear unabhängig sind. Sei $V = \mathbb{F}^n$ mit $|\mathbb{F}| \geq n$ und sei W die Menge aller Vektoren von der Form

$$m(a) := (1, a, a^2, \dots, a^{n-1}) \in \mathbb{F}^n \quad \text{mit} \quad a \in \mathbb{F}.$$

Zeige, dass die Vektoren aus W in allgemeiner Lage sind.

5.48. Eine $n \times n$ Matrix $A = (a_{i,j})$ heißt *streng diagonal dominant*, wenn

$$|a_{i,i}| > |a_{i,1}| + \dots + |a_{i,i-1}| + |a_{i,i+1}| + \dots + |a_{i,n}|$$

für alle Zeilen $i = 1, \dots, n$ gilt. Zeige, dass solche Matrizen regulär sind, d.h. den vollen Rang haben.