Merkwürdige Mathematik

Jochen Werner
jochen.christa@t-online.de

Inhaltsverzeichnis

1 Einleitung **4**

2 Was ist eine „Milchmädchenrechnung“? **4**

3 Die Collatz-Vermutung bzw. das $(3n+1)$-Problem **6**

4 Die schönsten mathematischen Sätze **8**

5 Mathematische Fangfragen **9**

6 Der Satz von Morley **11**

7 Wie kann man sich einen Teil der Dezimaldarstellung der Zahl π merken? **16**

8 Die Quadratur des Kreises **16**

9 Goldener Schnitt, DIN-Format **20**

10 Die Fibonacci-Zahlen **22**

11 Fra Luca Pacioli: Divina Proportione **23**

12 Was haben Goethe, Lichtenberg, Schopenhauer und Poe über Mathematiker gesagt? **32**

13 Das Buffonsche Nadelproblem **34**

14 Der Freundschaftssatz **35**

15 Das Königsberger Brückenproblem **39**

16 Das Rösselsprungproblem und Hamilton-Kreise **42**

17 Witze über MathematikerInnen **44**

18 Wahre Geschichten, über die nur MathematikerInnen lachen können **48**

19 Die Eulersche Polyederformel **49**

20 Der Fünffarbensatz **53**

21 Irrationale Zahlen **60**

22 Das Fermat-Weber Problem **63**

23 In welchen Jahren fällt der Himmelfahrtstag auf den 1. Mai? **65**

24 Das allgemeine Dreieck **66**

25 Das Napoleon-Dreieck, Napoleonische Probleme **68**

26 Das arithmetisch-geometrische Mittel **72**

27 Die Ungleichung vom geometrisch-arithmetischen Mittel **77**

28 Die Keplerschen Gesetze und ihre Herleitung **79**

29 Der Heiratssatz **83**
29.1 Der Satz von Birkhoff-von Neumann 88

30 Das Zuordnungsproblem **90**

31 Die archimedische Methode zur Berechnung der Zahl π **96**

32 Der Brent-Salamin-Algorithmus **101**

33 Das Geburtstagsparadoxon **109**

34 Das Ziegenproblem **111**

35 Das Gefangenenproblem **113**

36 Der Satz von Stone-Weierstraß **117**

37 Der Brouwersche Fixpunktsatz **120**

38 Der Fundamentalsatz der Algebra **125**

39 Das Gefangenendilemma und andere Zwei-Personen-Spiele **126**

40 Das Public goods game **135**

41 Die Berechnung der Quadratwurzel **137**

42 Das Fagnano-Problem **142**

43 Das Apfelmännchen **146**

44 Die erste Optimierungsaufgabe in der Geschichte der Mathematik **147**

45 Das Farkas Lemma **149**

46 Der Dualitätssatz der linearen Optimierung 152

47 Das Minimalkosten-Fluss-Problem 157

48 Das Max Flow-Min Cut Theorem von Ford-Fulkerson 161

49 Trennungssätze für konvexe Mengen im $\mathbb{R}^n$ 168

50 Die Lagrangesche Multiplikatorenregel 171

51 Das Problem der Dido 179

52 Das diskrete Problem der Dido 184

53 Der Satz von Kuhn-Tucker 188

54 Die kreisrunde Wiese von Bauer Lindemann 195

55 Wie rechnete Adam Ries(e)? 198

56 Die Neunerprobe 201

57 Wofür steht ISBN? Was ist eine IBAN? 204

58 Der Satz von Perron 209

59 Der Satz von Perron-Frobenius 213

60 Anwendungen des Satzes von Perron-Frobenius 220
60.1 Konvergenz der Potenzmethode 220
60.2 Das Leontief-Modell in den Wirtschaftswissenschaften 223
60.3 Das Leslie-Modell in der Populationsdynamik 224
60.4 Google's PageRank 227
60.5 Weitere Ranking-Probleme 232

61 Konvexe, quadratisch restringierte quadratische Optimierungsaufgaben 239

62 Das Minimum Covering Sphere Problem 243

63 Die Ungleichungen von Steinhagen und Jung 248

64 Magische Quadrate 255

65 Lateinische Quadrate 265
65.1 Orthogonale lateinische Quadrate 265
65.2 Die Vervollständigung lateinischer Quadrate: Der Satz von Evans-Smetaniuk 271

66 Sudoku 285
66.1 Mathematische Aussagen zu Sudoku 285
66.1.1 Die Anzahl vollständiger Sudokus 286
66.1.2 Die Anzahl im wesentlichen verschiedener vollständiger Sudokus . . . 292
66.1.3 Minimale Anzahl von Ziffern in einem eindeutigen Sudoku-Rätsel . . . 299

66.2 Verfahren zur Lösung eines Sudoku-Rätsels 301
66.2.1 Sudoku und (binäre) lineare Optimierung 301
66.2.2 Lösung eines Sudoku-Rätsels durch rekursives Backtracking 306

67 Die schnelle Fourier-Transformation 309

68 Geraden in der Ebene: Problem von Sylvester 319

69 Sperners Lemma und Brouwers Fixpunktsatz 320

70 Die Gammafunktion und der Satz von Bohr-Mollerup 328

71 Das Problem der feindlichen Brüder bzw. dichteste Packungen von Kreisen in einem Quadrat 336

Literaturverzeichnis 351

1 Einleitung

Ich stelle in dieser Sammlung, fast völlig ungeordnet, einiges aus der Mathematik zusammen, was sich meiner Meinung nach zu merken lohnen könnte, was also (für mich) *merkwürdig* ist. Einiges kann auch von mathematischen Laien verstanden werden, für anderes werden wenigstens elementare Kenntnisse der Mathematik benötigt.

2 Was ist eine „Milchmädchenrechnung“?

Bei Wikipedia findet man: Ein Milchmädchen ist laut Grimmschem Wörterbuch ein Mädchen, das die Milch besorgt und sie auch feil hat. Milchmädchen spielten in der Geschichte im Bereich der Milcherzeugung und -verarbeitung eine große Rolle. Ihre Aufgabe umfasste u.a. Melktätigkeiten, die Butterherstellung oder auch das Verkaufen von Milch und Milchprodukten auf dem Milchmarkt.

Milchmädchenrechnung wird abfällig die finanzielle Planung eines Vorhabens bezeichnet, bei der abzusehen ist, dass diese das Vorhaben niemals tragen wird bzw. bei der unterstellt wird, dass sie das Vorhaben nicht tragen kann, weil sie auf Trugschlüssen beruht. Der Begriff geht vermutlich auf die Fabel “Die Milchfrau” von Johann Wilhelm Ludwig Gleim (1719-1803) zurück. Erzählt wird die Geschichte einer Bauersfrau, die sich auf dem Weg zum Markt bereits vorstellt, was mit dem Erlös für die Milch alles machbar wäre, dann aber die Milch verschüttet. Wir zitieren nach `http://gutenberg.spiegel.de/`, dort suche man weiter nach Gleim und Fabeln.

Die Milchfrau

Auf leichten Füßen lief ein artig Bauerweib,
Geliebt von ihrem Mann, gesund an Seel’ und Leib,
Früh Morgens nach der Stadt, und trug auf ihrem Kopfe
Vier Stübchen süße Milch, in einem großen Topfe;

Sie lief und wollte gern: «Kauft Milch!» am ersten schrei'n:
Denn, dachte sie bei sich, die erste Milch ist theuer;
Will's Gott, so nehm' ich heut' sechs baare Groschen ein!
Dafür kauf' ich mir dann ein halbes Hundert Eier;
Mein Hühnchen brütet sie mir all' auf einmal aus:
Gras eine Menge steht um unser kleines Haus;
Die kleinen Küchelchen, die meine Stimme hören,
Die werden herrlich da sich letzen, und sich nähren;
Und ganz gewiß! der Fuchs, der müßte listig seyn,
Ließ' er mir nicht so viel, daß ich ein kleines Schwein
Dafür ertauschen könnte! Seht nur an!
Wenn ich mich etwas schon darauf im Geiste freue,
So denk' ich nur dabei an meinen lieben Mann!
Zu mästen kostet's mir ja nur ein wenig Kleie!
Hab' ich das Schweinchen fett, dann kauf' ich eine Kuh
In meinen kleinen Stall, ein Kälbchen wohl dazu;
Das Kälbchen will ich dann auf meine Weide bringen,
Und munter hüpft's und springt's, wie da die Lämmer springen!

«Sei!» sagt sie und springt auf! und von dem Kopfe fällt
Der Topf; das baare Geld,
Und Kalb und Kuh und Reichthum und Vergnügen
Sieht nun das arme Weib vor sich in Scherben liegen!
Erschrocken bleibt sie stehn und sieht die Scherben an;
«Die schöne weiße Milch», sagt sie, «auf schwarzer Erde!»
Weint, geht nach Haus', erzählt's dem lieben Mann,
Der ihr entgegen kommt, mit ernstlicher Gebehrde;
«Kind», sagt der Mann, «schon gut! Bau' nur ein ander Mal
Nicht Schlösser in die Luft! Man bauet seine Qual!
Geschwinder drehet sich um sich kein Wagenrad,
Als sie verschwinden in den Wind!
Wir haben all' das Glück, das unser Junker hat,
Wenn wir zufrieden sind!»

Eine andere Herkunftserklärung ist die folgende: Zu der Zeit, als Milch noch in Kannen von Bauernhöfen geholt wurde, sagte man einigen Milchmädchen (Milchverkäuferinnen) nach, die Kannen mit Wasser aufzufüllen, wenn die Milch knapp wurde. Da sie natürlich dennoch die volle Summe als Geldbetrag veranschlagten, entwickelte sich der Begriff Milchmädchenrechnung. Die erste Herkunftserklärung zielt insbesondere auf den Aspekt des Selbstbetruges ab, während letztere einen Betrug gegenüber anderen ausdrückt. Die Fabel "Die Milchfrau und die Milchkanne" (La Laitière et le Pot au Lait) von Jean de La Fontaine (1621-1695) wird auch[1] als Ursprung genannt[2]. Das Original findet man unter `http://www.jdlf.com/lesfables/livrevii/`

[1]Für mich ist diese Fabel der eigentliche Ursprung des Wortes "Milchmädchenrechnung", denn die Fabel von Gleim ist lediglich eine Nachdichtung.

[2]Diese Fabel wurde übrigens zusammen mit fünf weiteren Fabeln von Jacques Offenbach (1819-

lalaitiereetlepotaulait, eine Übersetzung unter http://www.seniorentreff.de/diskussion/archiv4/a133.html.

3 Die Collatz-Vermutung bzw. das $(3n+1)$-Problem

Auf die folgende Weise bilde man eine Folge natürlicher Zahlen. Man starte mit einer natürlichen Zahl n. Ist diese gerade, so sei der Nachfolger $n/2$, andernfalls $3n+1$. Man beobachtet, dass diese Folge irgendwann mit 1 bzw. mit einem Zyklus $\{4,2,1\}$ endet. Dass dies für jeden Startwert n der Fall ist, das ist gerade die Collatz-Vermutung. Näheres zur Geschichte dieser Vermutung findet man bei http://de.wikipedia.org/wiki/Collatz-Problem. Formal definiere man die Funktion $f\colon \mathbb{N} \longrightarrow \mathbb{N}$ durch

$$f(n) := \begin{cases} 3n+1 & \text{falls } n \text{ ungerade,} \\ \dfrac{n}{2} & \text{falls } n \text{ gerade.} \end{cases}$$

Die (noch unbewiesene[3]) Collatz-Vermutung besteht darin, dass für jedes $n \in \mathbb{N}$ ein $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ existiert derart, dass $f^{(k)}(n) = 1$, wobei $f^{(k)}(n)$ die k-te Iterierte von f angewandt auf n (und $f^{(0)}(n) := n$) ist. Laut S. Wagon (1985) ist die Collatz[4]-Vermutung für alle $n \leq 3 \cdot 10^{12}$ richtig.

Ist $n = 23$, so erhält man z. B. die Folge der Iterierten

$$\{f^{(0)}(23), \ldots, f^{(15)}(23)\} = \{23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5, 16, 8, 4, 2, 1\}.$$

Für $n = 27$ benötigt man z. B. 111 Iterationen, d. h. es ist $f^{(111)}(27) = 1$. Hierbei werden erstaunlich große Zahlen erreicht, z. B. ist $f^{(78)}(27) = 9232$.

Das erste $k \in \mathbb{N}$ mit $f^{(k)}(n) = 1$ wird von S. Wagon *total stopping time* von n genannt, während das erste k mit $f^{(k)}(n) < n$ einfach nur *stopping time* genannt und mit $\sigma^*(n)$ bezeichnet wird. Die Collatz-Vermutung ist richtig, wenn $\sigma^*(n) < \infty$ für alle $n \in \mathbb{N}$ (vollständige Induktion!). Z. B. ist $\sigma^*(23) = 8$ und $\sigma^*(27) = 96$. Weiter ist einfach einzusehen, dass $\sigma^*(n) \leq 3$, falls $n \equiv 1 \pmod 4$. Denn in diesem Falle erhält man mit einem $m \in \mathbb{N}$ die Iterierten

$$n = 1+4m, \quad 3n+1 = 4+12m, \quad \frac{3n+1}{2} = 2+6m, \quad \frac{3n+1}{4} = 1+3m < 1+4m = n.$$

Bei S. Wagon (1985) wird ohne Beweis ausgesagt, dass $\sigma^*(n) \leq 13$, falls $n \equiv m \pmod{256}$ und

$$m \notin \{27, 31, 47, 63, 71, 91, 103, 111, 127, 155, 159, 167, 191, 207, 223, 231, 239, 251, 255\}.$$

1880) für eine Singstimme und Klavierbegleitung vertont.

[3] Auf SPIEGEL ONLINE vom 5.6.2011 kann man nachlesen, dass der Hamburger Mathematiker Gerhard Opfer glaubt, die Collatz-Vermutung mit funktionentheoretischen Methoden Anfang Mai 2011 bewiesen zu haben. Dies ist aber wohl nicht der Fall, siehe z. B. http://www.rzbt.haw-hamburg.de/dankert/spezmath/html/collatzproblem.html

[4] Nach Lothar Collatz (1910-1990) benannt, der in den 1930er Jahren wohl das Problem aufgestellt hat. Etwas zur Motivation kann man bei G. J. Wirsching (2001) nachlesen.

Dies ist nicht schwer zu beweisen. Z. B. erhält man für $n \equiv 23 \pmod{256}$ die folgenden Iterierten:

$$\begin{aligned} f^{0)}(n) &= 23 + 256m, \\ f^{(1)}(n) &= 70 + 768m, \\ f^{(2)}(n) &= 35 + 384m, \\ f^{(3)}(n) &= 106 + 1152m, \\ f^{(4)}(n) &= 53 + 576m, \\ f^{(5)}(n) &= 160 + 1728m, \\ f^{(6)}(n) &= 80 + 864m, \\ f^{(7)}(n) &= 40 + 432m, \\ f^{(8)}(n) &= 20 + 216m, \end{aligned}$$

für alle diese n ist also $\sigma^*(n) = 8$.

Es gibt viele weitere einfach formulierbare, (noch unbewiesene) zahlentheoretische Vermutungen. Z. B. besteht die von Christian Goldbach in einem Brief an Leonhard Euler geäußerte Vermutung in der Aussage:

- *Jede ungerade Zahl größer als 5 kann als Summe dreier Primzahlen geschrieben werden.*

Dies ist also die eigentliche Goldbachsche Vermutung. Als starke Goldbachsche Vermutung versteht man die Aussage:

- *Jede gerade Zahl größer als 2 kann als Summe zweier Primzahlen geschrieben werden.*

Aus der starken folgt die eigentliche Goldbachsche Vermutung. Denn ist $n > 5$ ungerade, so ist $n - 3$ gerade und größer als 2, so dass aus $n = (n - 3) + 3$ die Behauptung folgt.

Nicht bewiesen ist bisher auch die Vermutung, dass es unendlich viele Primzahlzwillinge (das sind Primzahlen mit dem Abstand 2) oder gar Primzahlvierlinge (bestehend aus zwei Primzahlzwillingspaaren im Abstand 4, also aus vier Primzahlen der Form p, $p + 2$, $p + 6$ und $p + 8$). Im Netz findet man Aussagen über die größten bekannten Primzahlzwillinge.

Der große Fermatsche Satz besagt, dass die Gleichung $a^n + b^n = c^n$ für ganzzahlige a, b, c ungleich 0 und natürliche Zahlen n größer als 2 keine Lösung besitzt. Diese Aussage wurde von Pierre de Fermat 1637 gemacht. Sie war bis 1994 eine Vermutung, erst dann wurde sie von Andrew Wiles bewiesen[5].

[5]Der erste Band des Buches *The Art of Computer Programming* von Donald E. Knuth nennt sich *Fundamental Algorithms*. Den Übungsaufgaben wird ein Schwierigkeitsgrad zwischen 0 und 50 zugeordnet. Zur Illustration erscheint der große Fermatsche Satz als Übungsaufgabe, der in der ersten Auflage von 1968 den höchsten Schwierigkeitsgrad 50, in der zweiten Auflage von 1997 aber nur noch den Schwierigkeitsgrad 45 erhält!

4 Die schönsten mathematischen Sätze

Im Mathematical Intelligencer Volume 10 Number 4 (1988) wurde dazu aufgerufen, den schönsten mathematischen Satz zu wählen, siehe auch P. Basieux (2007, S. 11). Hierbei wurden 24 Sätze vorgeschlagen. Im Mathematical Intelligencer Volume 12 Number 3 (1990), siehe auch `http://fma2.math.uni-magdeburg.de/~bessen/topten.html`, wird das Ergebnis veröffentlicht. Gewinner ist die Aussage bzw. Formel $e^{i\pi} = -1$ oder auch $e^{i\pi} + 1 = 0$, in der die wichtigsten mathematischen Konstanten 0, 1, die imaginäre Einheit i, die Euler-Zahl e und die Kreiszahl π vorkommen. Eine ungewöhnliche Darstellung dieser Formel findet man in Abbildung 1. Diese und ähnliche Bilder findet

Abbildung 1: Die schönste Formel

man, wenn man bei Google Bilder mit *Euler tattoo* sucht. Um diese Formel zu beweisen, muss man sich darüber verständigen, wie die (komplexe) Exponentialfunktion und die Zahl π definiert sind. In der Analysis wird i. Allg. die Exponentialfunktion über ihre Reihendarstellung, danach die reellen trigonometrischen Funktionen $\cos(\cdot)$ und $\sin(\cdot)$ als Real- bzw. Imaginärteil von $e^{i\cdot}$ definiert. Für $x \in \mathbb{R}$ ist also

$$\cos(x) := \Re(e^{ix}), \qquad \sin(x) := \Im(e^{ix}),$$

so dass die Eulersche Formel $e^{ix} = \cos(x) + i\sin(x)$ sozusagen per definitionem gilt. Bei O. Forster wird weiter $\pi/2$ als (eindeutig bestimmte) erste Nullstelle der Funktion cos im Intervall $[0, 2\pi]$ definiert (was natürlich nicht gerade eine intuitiv eingängige Definition ist). Hieraus kann dann leicht die genannte Formel bewiesen werden, von der (siehe Wikipedia) Spötter sagen, sie besage nichts anderes als: "Wenn man sich umdreht, schaut man in die andere Richtung". Die Silbermedaille geht an die Eulersche Polyederformel[6], die Bronzemedaille an die schon bei Euklid vorkommende Aussage, dass es unendlich viele Primzahlen gibt[7]. Einige weitere der schönsten mathematischen Sätze werden auch in unserer Sammlung vorkommen, wie z. B. der Brouwersche Fixpunktsatz auf Platz 6 (siehe Abschnitt 37) oder der Fundamentalsatz der Algebra (siehe

[6]Der Eulersche Polyedersatz, benannt nach Leonhard Euler, besagt: Seien E die Anzahl der Ecken, F die Anzahl der Flächen und K die Anzahl der Kanten eines beschränkten, konvexen Polyeders im $\mathbb{R}^3$, so ist $E - K + F = 2$. Hierauf gehen wir in Abschnitt 19 genauer ein.

[7]Sechs Beweise für die Unendlichkeit der Primzahlen findet man im ersten Kapitel des sehr schönen Buches von Martin Aigner und Günther M. Ziegler mit dem (deutschen) Titel "Das BUCH der Beweise", das in verschiedenen englischen und deutschen Auflagen erschienen ist.

Abschnitt 38). Der Satz von Stone-Weierstraß (siehe Abschnitt 36) kommt auch in der Sammlung Großer Sätze und Schöner Beweise bei J. Naas, W. Tutschke (2009) vor.

5 Mathematische Fangfragen

Die beiden folgenden Aufgaben habe ich dem Buch "Humor in der Mathematik" von F. Wille (1982) entnommen.

- Man strecke seinem Gegenüber beide Hände mit gespreizten Fingern entgegen und frage: »Wieviele Finger sind das?«

  Antwort: »10.«

  Frage: »Und wieviele Finger haben 10 Hände?«

Die Antwort hierauf wird sehr oft 100 sein, was bedeutet, dass $10 \cdot 5$ nicht richtig berechnet wurde.

- »In einem dunklen Zimmer (die elektrische Birne ist gerade durchgebrannt) befindet sich eine Schublade mit 12 braunen Strümpfen und 12 schwarzen Strümpfen. Sie liegen wahllos durcheinander in der Schublade. Welches ist die geringste Zahl von Strümpfen, die du (im Dunkeln) herausgreifen musst, um mit Sicherheit ein Paar gleichfarbiger Strümpfe zu erhalten?«

Es ist erstaunlich, wie oft man eine falsche Antwort auf diese einfache Frage erhält. Eine ähnliche Aufgabe findet sich unter dem Stichwort *Sockensalat* bei S. Morris (2007).

Die folgende Frage lässt sich leicht im Kopf lösen. Auch hier erhält man erstaunlich oft falsche Antworten. Mein Vater stellte diese Frage gerne, daher die alte Währungsangabe.

- »Flasche und Korken kosten zusammen 1.10 DM. Die Flasche kostet 1 DM mehr als der Korken. Was kostet der Korken?«

Die nächste Aufgabe ist sehr wahrscheinlich einmal in einer Quizsendung des Fernsehens vorgekommen. Jedenfalls erhielt ich vor vielen Jahren eine entsprechende Anfrage von einem Schneidermeister aus Einbeck.

- Man denke sich um die Erde, die man sich als Kugel vorstelle, längs des Äquators ein Seil gespannt. Anschließend verlängere man das Seil um einen Meter und denke es sich wieder gleichmäßig um den Äquator gespannt. Was ist der Abstand des Seils zur Oberfläche der Kugel?

Intuitiv denkt man, dass dies ein winziger Abstand sein müsste. In Wahrheit spielt der Radius der Kugel überhaupt keine Rolle und der Abstand ist $1/(2\pi) \approx 0.1592$ Meter[8].

[8]Denn angenommen, der Radius der Erdkugel sei r Meter, die Länge eines längs des Äquators gespannten Seiles wäre dann $2\pi r$ Meter. Verlängert man das Seil um 1 Meter, so ist also $2\pi r + 1$ Meter seine Länge. Der Abstand x des so verlängerten Seiles zur Oberfläche der Erdkugel berechnet sich aus $2\pi(r + x) = 2\pi r + 1$, woraus man $x = 1/(2\pi)$ erhält.

Eine ganz ähnliche Aufgabe kommt in Jules Vernes *Fünf Wochen im Ballon* (siehe `http://gutenberg.spiegel.de/buch/4033/1`) vor:

- Dieser Gelehrte gab ihm eines Tages, in der Hoffnung, sich ihm angenehm zu machen, folgende Aufgabe zu lösen: Wenn die Zahl der von dem Doctor auf seinen Reisen um die Welt zurückgelegten Meilen gegeben ist, einen wie viel weiteren Weg hat – in Anbetracht der Verschiedenheit der Radien – sein Kopf gemacht, als seine Füße?

Die folgende Aufgabe entnehme ich dem gerade eben erwähnten Buch von S. Morris (2007).

- Hier ist ein kleiner Ausschnitt aus einem Gespräch, das ein Interviewer bei einer Umfrage mit Frau Fischer führte.

  Frage: Wieviele Kinder haben Sie?
  Frau F.: Zwei.
  Frage: Und wie alt, bitte?
  Frau F.: Eins ist fünf, das andere zwei.
  Frage: Ist eins davon ein Junge?
  Frau F.: Ja.

  Wie hoch ist die Wahrscheinlichkeit, dass Frau Fischers anderes Kind ebenfalls ein Junge ist?

Die Antwort $\frac{1}{2}$ ist falsch, richtig ist $\frac{1}{3}$. Denn bei zwei Kindern (wir gehen davon aus, dass Junge und Mädchen gleich wahrscheinlich sind) gibt es vier gleich wahrscheinliche Kombinationen, wobei das ältere Kind vorangestellt wird: (1) Junge-Junge, (2) Junge-Mädchen, (3) Mädchen-Junge, (4) Mädchen-Mädchen. Die letzte fällt weg, da wir wissen, dass ein Kind ein Junge ist. Die Wahrscheinlichkeit eines zweiten Jungen beträgt also $\frac{1}{3}$. Hätte der Interviewer aber auch noch wissen wollen, ob das fünfjährige Kind ein Junge ist und wäre diese Frage bejaht worden, so wäre mit Wahrscheinlichkeit $\frac{1}{2}$ das zweijährige Kind ebenfalls ein Junge. Denn jetzt sind nur noch die Kombinationen (1) und (2) möglich.

Die Beantwortung der folgenden Aufgabe ist bei Anwendung eines geeigneten Gedankenexperiments sehr einfach. Durch Einführung irrelevanter Informationen könnte man noch etwas mehr Verwirrung stiften.

- Ein Mönch bricht um 9.00 am Morgen auf, um einen Berg zu besteigen. Die Spitze des Berges erreicht er um 21.00 am Abend. Die Nacht verbringt er meditierend. Um 9.00 am nächsten Morgen bricht er wieder auf und geht genau denselben Weg wie beim Aufstieg wieder zurück. Er lässt sich Zeit und erreicht den Ausgangspunkt seiner Wanderung um 21.00. Weshalb gibt es eine Stelle auf dem Pfad, den der Mönch genau zur selben Uhrzeit beim Auf- und beim Abstieg erreicht?

Dass die Aussage richtig ist, ist völlig klar, wenn man sich vorstellt, dass am zweiten Tag um 9.00 eine weitere Person, meinetwegen ein Mönch, zum Berg aufsteigt. Diese Person und der absteigende Mönch müssen sich auf dem Pfad irgendwo treffen.

Eine letzte sehr einfache Frage:

- Gibt es einen Spielplan für die 18 Mannschaften der ersten Bundesliga, nach dem sich für *jede* Mannschaft Heim- und Auswärtsspiele abwechseln?

6 Der Satz von Morley

Der folgende Satz von Morley wird von H. S. M. Coxeter (1969, 23 ff.) einer der überraschendsten Sätze der elementaren Geometrie genannt, gelegentlich spricht man auch von "marvel of marvels" oder "Morley's miracle".

Die Innenwinkel eines beliebigen Dreiecks werden jeweils in drei gleich große Winkel unterteilt. Zu jeder Dreiecksseite betrachtet man den Schnittpunkt derjenigen zwei Teilungslinien, die von den Endpunkten dieser Seite ausgehen und zu dieser Seite benachbart sind. Das Morley-Dreieck entsteht durch Verbinden der drei erhaltenen Schnittpunkte. Der Satz von Morley besteht in der folgenden Aussage:

- Unabhängig von der Form des ursprünglichen Dreiecks ist das Morley-Dreieck gleichseitig.

In Abbildung 2 veranschaulichen wir uns den Satz von Morley. Elementare Beweise

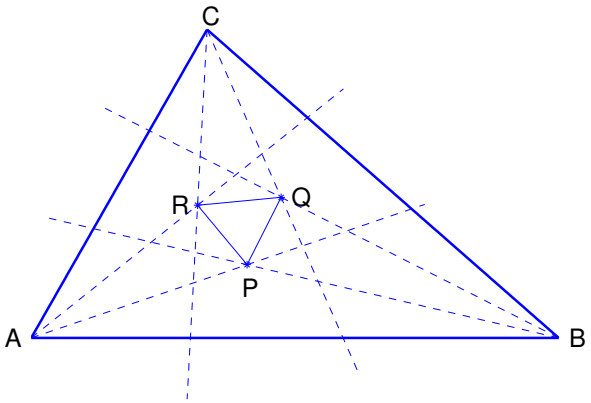


Abbildung 2: Der Satz von Morley

dieses Satzes, der aus dem Jahre 1899 stammt, findet man u. a. in dem genannten Buch von Coxeter, ferner in Aufsätzen von D. J. Newman (1996), H. Geiges (2001) und G. Wanner (2004). Durch den Aufsatz von H. Geiges (2001) wird man auf die Arbeit von A. Connes (1998) (1982 erhielt er die Fields Medaille) aufmerksam gemacht. Unter Zugrundelegung der Aufsätze von Connes und Geiges wollen wir einen elementaren analytischen Beweis des Satzes von Morley angeben.

Die Aussage des folgenden Lemmas wird von Connes eine klassische Charakterisierung gleichseitiger Dreiecke genannt, es ist Lemma 3 bei H. Geiges (2001).

Lemma *Seien P, Q, R drei Punkte in der komplexen Zahlenebene $\mathbb{C}$, ferner sei*

$$\eta := e^{i2\pi/3}.$$

Dann ist das Dreieck $\triangle PQR$ genau dann ein positiv (d. h. im Gegenuhrzeigersinn) orientiertes gleichseitiges Dreieck, wenn

$$P + \eta Q + \eta^2 R = 0.$$

Beweis: Das Dreieck $\triangle PQR$ ist genau dann gleichseitig und positiv orientiert, wenn das Dreieck $\triangle 0(Q-P)(R-P)$ dieselbe Eigenschaft hat. Dies wiederum ist gleichbedeutend damit, dass man $R-P$ aus $Q-P$ durch eine Drehung (bezüglich des Nullpunktes) um $60°$ im mathematisch positiven Sinne erhält, dass also $R - P = e^{i\pi/3}(Q - P)$ bzw.

$$Q - P = e^{-i\pi/3}(R - P) = -\eta(R - P).$$

Unter Benutzung von $1 + \eta + \eta^2 = 0$ ist dies wiederum äquivalent zu

$$P + \eta Q + \eta^2 R = (-\eta - \eta^2)P + \eta Q + \eta^2 R = \eta((Q - P) + \eta(R - P)) = 0,$$

was zu beweisen war. □

Es folgt ein einfach zu formulierender Hilfssatz über Dreiecke in der komplexen Zahlenebene.

Hilfssatz *Seien $A, B, C \in \mathbb{C}$ paarweise verschieden und nicht kollinear, also nicht auf einer Geraden. Sei $\alpha := \sphericalangle CAB$, $\beta := \sphericalangle ABC$ und $\gamma := \sphericalangle BCA$. Dann ist*

$$(1 - e^{i2\alpha})A + e^{i2\alpha}(1 - e^{i2\beta})B + e^{i2(\alpha+\beta)}(1 - e^{i2\gamma})C = 0.$$

Beweis: Sei $a := |B - C|$, $b := |C - A|$ und $c := |A - B|$. Offenbar ist

$$C = A + \frac{b}{c}e^{i\alpha}(B - A), \quad A = B + \frac{c}{a}e^{i\beta}(C - B), \quad B = C + \frac{a}{b}e^{i\gamma}(A - C).$$

Hiermit erhalten wir

$$\begin{aligned}
& (1 - e^{i2\alpha})A + e^{i2\alpha}(1 - e^{i2\beta})B + e^{i2(\alpha+\beta)}(1 - e^{i2\gamma})C \\
= \; & (1 - e^{i2\alpha})A + (e^{i2\alpha} - e^{i2(\alpha+\beta)})B + (e^{i2(\alpha+\beta)} - 1)C \\
= \; & (A - C) + e^{i2\alpha}(B - A) + e^{i2(\alpha+\beta)}(C - B) \\
= \; & -\frac{b}{c}e^{i\alpha}(B - A) + {}^{i2\alpha}\,(B - A) + e^{i2(\alpha+\beta)}\Big(-\frac{a}{c}e^{-i\beta}\Big)(B - A) \\
= \; & e^{i\alpha}\Big(-\frac{b}{c} + e^{i\alpha} - \frac{a}{c}e^{i(\alpha+\beta)}\Big)(B - A) \\
= \; & \frac{e^{i\alpha}}{c}\Big(-b + ce^{i\alpha} + ae^{-i\gamma}\Big)(B - A).
\end{aligned}$$

Damit ist die Behauptung äquivalent zu

$$-b + ce^{i\alpha} + ae^{-i\gamma} = 0$$

bzw.

$$-b + c\cos\alpha + a\cos\gamma = 0, \qquad c\sin\alpha - a\sin\gamma = 0.$$

Wegen des Kosinussatzes ist

$$\cos\alpha = \frac{b^2+c^2-a^2}{2bc}, \qquad \cos\gamma = a^2+b^2-c^2 2ab.$$

Daher ist

$$-b + c\cos\alpha + a\cos\gamma = -b + \frac{b^2+c^2-a^2}{2b} + \frac{a^2+b^2-c^2}{2b} = 0.$$

Ferner ist wegen des Sinussatzes

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}.$$

Daher ist auch

$$c\sin\alpha - a\sin\gamma = 0.$$

Die Behauptung ist damit bewiesen. □

Wir betrachten nichtkonstante, affine Abbildungen $f\colon \mathbb{C} \longrightarrow \mathbb{C}$ von $\mathbb{C}$ in sich, also Abbildungen der Form $f(z) := az + b$ mit $a, b \in \mathbb{C}$ und $a \neq 0$. Wir schreiben $\rho(f) := a$ für den Rotationsanteil und $\tau(f) := b$ für den Translationsanteil von f. Für $a \neq 0, 1$ hat f den eindeutigen Fixpunkt

$$F(f) := \frac{b}{1-a} = \frac{\tau(f)}{1-\rho(f)}.$$

Der folgende Satz ist das entscheidende Theorem in der Arbeit von A. Connes (1998), siehe auch Satz 4 bei H. Geiges (2001), aus welchem durch eine geeignete Spezialisierung die Aussage des Satzes von Morley folgt.

Satz *Seien $f_i(z) := a_i z + b_i$, $i = 1, 2, 3$, drei nichtkonstante (also $a_1a_2a_3 \neq 0$) affine Abbildungen, von denen f_1f_2, f_2f_3, f_3f_1 und $f_1f_2f_3$ keine Translationen sind. Sei $\omega := \rho(f_1f_2f_3) = a_1a_2a_3$. Dann sind die folgenden beiden Aussagen gleichwertig:*

(a) Es ist $f_1^3 f_2^3 f_3^3$ die identische Abbildung.

(b) Es ist $\omega^3 = 1$ und $P + \omega Q + \omega^2 R = 0$, wobei $P := F(f_1f_2)$, $Q := F(f_2f_3)$ und $R := F(f_3f_1)$.

Beweis: Wegen

$$\begin{aligned}(f_1^3f_2^3f_3^3)(z) &= (f_1^3f_2^3)(a_3^3z + (a_3^2+a_3+1)b_3)\\ &= f_1^3(a_2^3a_3^3z + a_2^3(a_3^2+a_3+1)b_3 + (a_2^2+a_2+1)b_2)\\ &= a_1^3a_2^3a_3^3z + a_1^3a_2^3(a_3^2+a_3+1)b_3 + a_1^3(a_2^2+a_2+1)b_2 + (a_1^2+a_1+1)b_1\end{aligned}$$

ist

$$\rho(f_1^3f_2^3f_3^3) = (a_1a_2a_3)^3 = \omega^3$$

und

$$\tau(f_1^3f_2^3f_3^3) = a_1^3a_2^3(a_3^2+a_3+1)b_3 + a_1^3(a_2^2+a_2+1)b_2 + (a_1^2+a_1+1)b_1.$$

Da $f_1f_2f_3$ nach Voraussetzung keine Translation ist, ist $\omega = \rho(f_1f_2f_3) \neq 1$. Daher ist (a) äquivalent zu $\omega^3 = 1$ und

$$a_1^3a_2^3(a_3^2 + a_3 + 1)b_3 + a_1^3(a_2^2 + a_2 + 1)b_2 + (a_1^2 + a_1 + 1)b_1 = 0.$$

Jetzt nehmen wir an, es gelte (a) und zeigen, dass (b) richtig ist. Da $\omega^3 = 1$ und $\omega \neq 1$ ist $\omega = \eta$ oder $\omega = \eta^2$, wobei $\eta := e^{i2\pi/3}$. Wegen $1 + \eta + \eta^2 = 0$ und $1 + \eta^2 + (\eta^2)^2 = 0$ ist $1 + \omega + \omega^2 = 0$. Berücksichtigt man

$$P = \frac{a_1b_2 + b_1}{1 - a_1a_2}, \qquad Q = \frac{a_2b_3 + b_2}{1 - a_2a_3}, \qquad R = \frac{a_3b_1 + b_3}{1 - a_3a_1}$$

und setzt man

$$c := (1 - a_1a_2)(1 - a_2a_3)(1 - a_3a_1),$$

so erhält man (wir folgen der Darstellung von H. Geiges (2001)) unter Benutzung von $\omega^3 = 1$ und $1 + \omega + \omega^2 = 0$, dass

$$\begin{aligned}
-\omega^2a_1^2a_2c(P + \omega Q + \omega^2R) &= -\omega^2a_1^2a_2\Big[(a_1b_2 + b_1)(1 - a_2a_3)(1 - a_3a_1) \\
&\quad + (a_2b_3 + b_2)(1 - a_1a_2)(1 - a_3a_1)a_1a_2a_3 \\
&\quad + (a_3b_1 + b_3)(1 - a_1a_2)(1 - a_2a_3)a_1^2a_2^2a_3^2\Big] \\
&= -\omega^2a_1^2a_2\Big[b_1(1 - a_2a_3)[(1 - a_3a_1) \\
&\quad + a_3(1 - a_1a_2)a_1^2a_2^2a_3^2] \\
&\quad + b_2(1 - a_3a_1)a_1[(1 - a_2a_3) + (1 - a_1a_2)a_2a_3] \\
&\quad + b_3(1 - a_1a_2)a_1a_2^2a_3[(1 - a_3a_1) \\
&\quad + (1 - a_2a_3)a_1a_3]\Big] \\
&= -\omega^2a_1^2a_2\Big[b_1(1 - a_2a_3)(-a_1a_3 + a_1^2a_2^2a_3^3) \\
&\quad + b_2(1 - a_3a_1)a_1(1 - a_1a_2^2a_3) \\
&\quad + b_3(1 - a_1a_2)a_1a_2^2a_3(1 - a_1a_2a_3^2)\Big] \\
&= -\omega^2a_1^2a_2\Big[b_1(-a_1a_3 + a_1^2a_2^2a_3^3 + a_1a_2a_3^2 - a_1^2a_2^3a_3^4) \\
&\quad + b_2(a_1 - a_1^2a_3 - a_1^2a_2^2a_3 + a_1^3a_2^2a_3^2) \\
&\quad + b_3(a_1a_2^2a_3 - a_1^2a_2^3a_3 - a_1^2a_2^3a_3^3 + a_1^3a_2^4a_3^3)\Big] \\
&= b_1(a_1^2\omega^3 - a_1\omega^5 - a_1\omega^4 + \omega^6) \\
&\quad + b_2a_1^3(-a_2\omega^2 + \omega^3 + a_2^2\omega^3 - a_2\omega^4) \\
&\quad + b_3a_1^3a_2^3(-a_3\omega^2 + \omega^3 + a_3^2\omega^3 - a_3\omega^4) \\
&= b_1(a_1^2 - a_1\omega^2 - a_1\omega + 1) \\
&\quad + b_2a_1^3(-a_2\omega^2 + 1 + a_2^2 - a_2\omega) \\
&\quad + b_3a_1^3a_2^3(-a_3\omega^2 + 1 + a_3^2 - a_3\omega) \\
&= b_1(a_1^2 + a_1 + 1) + b_2a_1^3(a_2^2 + a_2 + 1)
\end{aligned}$$

$$
\begin{aligned}
&\quad + b_3 a_1^3 a_2^3 (a_3^2 + a_3 + 1) \\
&= \tau(f_1^3 f_2^3 f_3^3).
\end{aligned}
$$

Ist (a) richtig, so ist $\tau(f_1^3 f_2^3 f_3^3) = 0$ und folglich $P + \omega Q + \omega^2 R = 0$, also (b) richtig. Gilt umgekehrt (b), so folgt aus $\omega^3 = 1$, dass $\rho(f_1^3 f_2^3 f_3^3) = 1$.Da wegen der obigen Gleichungskette aus $P + \omega Q + \omega^2 R = 0$ sofort $\tau(f_1^3 f_2^3 f_3^3) = 0$ folgt, ist $f_1^3 f_2^3 f_3^3$ die Identität. Insgesamt ist die Gleichwertigkeit von (a) und (b) bewiesen. □

Beweis des Satzes von Morley Wir geben spezielle affin lineare Abbildungen von $\mathbb{C}$ in sich an, die die Voraussetzungen des vorigen Satzes erfüllen. Gegeben sei ein Dreieck $\triangle ABC$ wie in Abbildung 2. Man setze

$$
\alpha := \sphericalangle CAB, \qquad \beta := \sphericalangle ABC, \qquad \gamma := \sphericalangle BCA.
$$

Hiermit sei f_1 (bzw. f_2, f_3) eine Drehung im Gegenuhrzeigersinn um den Punkt A (bzw. B, C) um den Winkel $2\alpha/3$ (bzw. $2\beta/3$, $2\gamma/3$). Also ist

$$
f_1(z) = e^{i2\alpha/3}(z - A) + A, \qquad f_2(z) = e^{i2\beta/3}(z - B) + B, \qquad f_3(z) = e^{i2\gamma/3}(z - C) + C.
$$

Dann ist

$$
\omega := \rho(f_1 f_2 f_3) = e^{i2(\alpha+\beta+\gamma)/3} = e^{i2\pi/3} = \eta.
$$

Offensichtlich erfüllen die nichtkonstanten affinen Abbildungen f_1, f_2, f_3 die Voraussetzungen des letzten Satzes. Mit P, Q und R wie in Abbildung 2 und der Bezeichnung $F(f)$ für den Fixpunkt der nichtkonstanten Abbildung f, die keine Translation ist, ist

$$
P = F(f_1 f_2), \qquad Q = F(f_2 f_3), \qquad R = F(f_3 f_1).
$$

Denn: $P = F(f_1 f_2)$ bzw. $(f_1 f_2)(P) = P$ besagt, dass eine Drehung von P um B um den Winkel $2\beta/3$ und eine anschließende Drehung dieses so erhaltenen Punktes um A um den Winkel $2\alpha/3$, jeweils im Gegenuhrzeigersinn, wieder zum Punkt P zurückführt. Dies ist aber offensichtlich richtig. Wir zeigen jetzt noch, dass $f_1^3 f_2^3 f_3^3$ die identische Abbildung ist, also Bedingung (a) des Satzes erfüllt ist. Der Rotationsanteil ist $\rho(f_1^3 f_2^3 f_3^3) = \eta^3 = 1$ während der Translationsanteil gegeben ist durch (siehe den Beginn des Beweises obigen Satzes)

$$
\begin{aligned}
\tau(f_1^3 f_2^3 f_3^3) &= e^{i2(\alpha+\beta)}(e^{i4\gamma/3} + e^{i2\gamma/3} + 1)(1 - e^{i2\gamma/3})C \\
&\quad + e^{i2\alpha}(e^{i4\beta/3} + e^{i2\beta/3} + 1)(1 - e^{i2\beta/3})B \\
&\quad + (e^{i4\alpha/3} + e^{i2\alpha/3} + 1)(1 - e^{i2\alpha/3})A \\
&= e^{i2(\alpha+\beta)}(1 - e^{i2\gamma})C + e^{i2\alpha}(1 - e^{i2\beta})B + (1 - e^{i2\alpha})A
\end{aligned}
$$

und dieser verschwindet wegen des obigen Hilfssatzes. Also ist $f_1^3 f_2^3 f_3^3(z) = z$ bzw. $f_1^3 f_2^3 f_3^3$ die identische Abbildung. Eine Anwendung des obigen Satzes liefert $P + \eta Q + \eta^2 R = 0$, und dies bedeutet wegen des obigen Lemmas, dass das Dreieck $\triangle PQR$ ein positiv (d. h. im Gegenuhrzeigersinn) orientiertes gleichseitiges Dreieck ist. Damit ist der Satz von Morley bewiesen. □

7 Wie kann man sich einen Teil der Dezimaldarstellung der Zahl π merken?

Auf 15 Stellen kann man sich π durch den folgenden Spruch merken, wobei die Anzahl der Buchstaben in einem Wort die entsprechende Ziffer ergibt:

- How I want a drink, alcoholic of course, after the heavy lectures involving quantum mechanics.

Ein deutsches "Gedicht" für die ersten 23 Stellen ist

> Wie, o dies π
> macht ernstlich so viele Müh'.
> Lernt immerhin, Jünglinge, leichte Verselein,
> Wie so zum Beispiel dies dürfte zu merken sein!

Ein weiteres englisches Gedicht ist bei L. Berggren et al. (1997, S. 302) zu finden (übrigens in einem Aufsatz mit dem Titel "The Chronology of Pi"), es liefert 30 Stellen von π:

> Now I, even I, would celebrate
> In rhymes inapt, the great
> Immortal Syracusan, rivaled nevermore,
> Who in his wondrous lore,
> Passed on before
> Left men his guidance
> How to circles mensurate.

Wir zitieren J. Arndt, C. Haenel (1998, S. 30): "Das wohl längste Merkgedicht hat Michael Keith gebraut, indem er die im Englischen sehr bekannte Ballade von Edgar Allen Poe *The Raven* (Der Rabe) so modifiziert hat, dass sie nicht weniger als 740 Stellen von π liefert. ..." Das ganze Gedicht ist übrigens bei L. Berggren et al. (1997, S. 659) abgedruckt. Dort wird auch erklärt, dass die Ziffer 0 (sie kommt z. B. als 32. Nachkommastelle vor) in diesem Gedicht durch ein Wort mit 10 Buchstaben repräsentiert wird.

8 Die Quadratur des Kreises

Die Quadratur des Kreises ist ein klassisches Problem der Geometrie. Die Aufgabe besteht darin, nur mit Lineal und Zirkel aus einem gegebenen Kreis ein Quadrat mit demselben Flächeninhalt zu konstruieren. Dieses Problem wurde etwa 500 v. Chr. von einigen "alten Griechen" wie Anaxagoras, Antiphon, Hippokrates von Chios und Hippias gestellt und besteht in folgendem:

- Konstruiere allein mit Zirkel und Lineal in endlich vielen Schritten aus einem beliebigen Kreis ein flächengleiches Quadrat.

Unter "Konstruierbarkeit mit Zirkel und Lineal" versteht man hierbei, endlich oft eine der folgenden Operationen durchzuführen:

1. Verbinde zwei Punkte durch eine Gerade.
2. Bestimme den Schnittpunkt zweier Geraden.
3. Schlage einen Kreis mit einem gegebenen Radius um einen Punkt.
4. Bestimme den Schnittpunkt eines Kreises mit einem anderen Kreis oder einer Geraden.

Beispiel: Wir führen die Quadratur eines Rechtecks vor. Gegeben sei also das Rechteck $ABCD$, siehe Abbildung 3. Die folgende Konstruktion wird bei Euklid (Buch II,

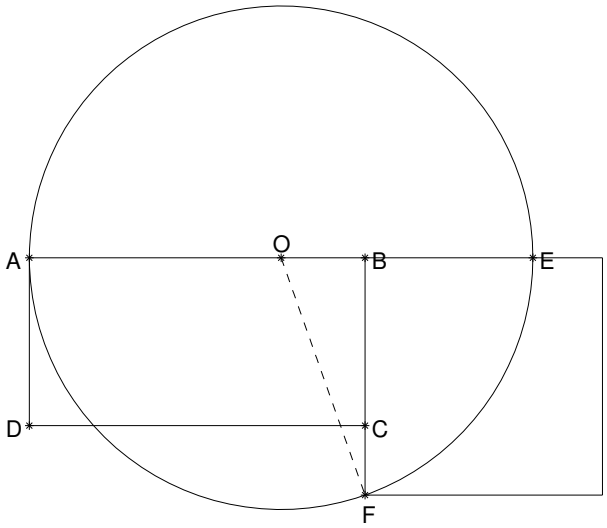


Abbildung 3: Quadratur eines Rechtecks

Proposition 14) angegeben:

1. Verlängere AB über B hinaus und bestimme E mit $|BE| = |BC|$.
2. Halbiere AE und erhalte O.
3. Mit O als Mittelpunkt beschreibe man einen Kreis mit dem Radius $|OA|$.
4. Sei F Schnittpunkt dieses Kreises mit der über C hinaus verlängerten Strecke BC.

Dann ist $|BF|$ die Seite des gesuchten Quadrats, also $|AB| \cdot |BC| = |BF|^2$. Dies wollen wir wie Euklid beweisen. Der Beweis erfolgt in vier Schritten.

- Es ist $|AB| \cdot |BC| + |OB|^2 = |OE|^2$.

Denn: Wir berücksichtigen $|BE| = |BC|$ und erhalten

$$\begin{aligned} |OE|^2 - |OB|^2 &= \Big(\frac{|AB| + |BE|}{2}\Big)^2 - \Big(|AB| - \frac{|AB| + |BE|}{2}\Big)^2 \\ &= \Big(\frac{|AB| + |BC|}{2}\Big)^2 - \Big(\frac{|AB| - |BC|}{2}\Big)^2 \\ &= |AB| \cdot |BC|, \end{aligned}$$

womit die erste Behauptung bewiesen ist.

- Es ist $|AB| \cdot |BC| + |OB|^2 = |OF|^2$.

Denn: Dies folgt einfach aus der ersten Beziehung und aus $|OE| = |OF|$, da E und F auf einem Kreis um O mit dem Radius $|OA| = |OE|$ liegen.

- Es ist $|AB| \cdot |BC| + |OB|^2 = |BF|^2 + |OB|^2$.

Denn: Wegen des Satzes von Pythagoras ist $|OF|^2 = |BF|^2 + |OB|^2$, so dass die Aussage aus der zweiten Beziehung folgt.

- Es ist $|AB| \cdot |BC| = |BF|^2$.

Denn: Subtrahiere $|OB|^2$ von beiden Seiten der dritten Beziehung.

Ein alternativer Beweis könnte mit Hilfe des Sehnensatzes (zu finden übrigens bei Euklid Buch III, Proposition 35) erfolgen. Dieser sagt aus:

- Hat man zwei Sehnen durch einen Punkt im Innern eines Kreises, so ist das Produkt der beiden Segmente der einen Sehne gleich dem Produkt der beiden Segmente der anderen Sehne.

In Abbildung 4 geben wir an, wie wir den Sehnensatz anwenden. Also ist $|AB| \cdot |BE| = |BF| \cdot |BG|$. Wegen $|BE| = |BC|$ und $|BG| = |BF|$ (dies erkennt man durch eine Anwendung des Satzes von Pythagoras auf $\triangle OGB$ und $\triangle OFB$) folgt auch jetzt wieder $|AB| \cdot |BC| = |BF|^2$. □

Im Jahr 1882 bewies Ferdinand von[9] Lindemann, dass die Quadratur des Kreises nicht möglich ist. Wäre die Konstruktion möglich, so müsste man, ausgehend von einem Kreis mit dem Radius 1, in endlich vielen Schritten mit Zirkel und Lineal eine Strecke der Länge $\sqrt{\pi}$ konstruieren können. Diese Strecke ist genau dann konstruierbar, wenn die Zahl π bzw. eine Strecke der Länge π konstruierbar ist. Es sind jedoch nur algebraische Zahlen konstruierbar, also jene Zahlen, die eine Lösung (Nullstelle) eines Polynoms beliebigen Grades mit rationalen Koeffizienten sind. Zahlen, die nicht algebraisch sind, heißen transzendent und sind nicht konstruierbar. Ferdinand von Lindemann gelang der Beweis, dass π nicht algebraisch, sondern transzendent ist. Deshalb ist π nicht konstruierbar und die Quadratur des Kreises unmöglich.

Die Nutzlosigkeit der Suche nach Lösungen hat die Quadratur des Kreises als Metapher bekannt gemacht. Der Ausdruck wird einerseits als Synonym für ein Unterfangen

[9] Den Adeltitel erhielt er erst später.

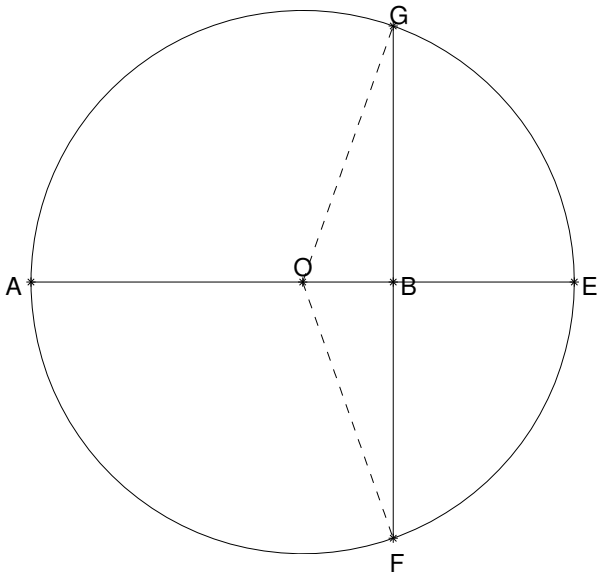


Abbildung 4: Der Sehnensatz: $|AB| \cdot |BE| = |BF| \cdot |BG|$

benutzt, das von vornherein zum Scheitern verurteilt ist[10]. Andererseits bezeichnet man das Ergebnis großer Anstrengungen scherzhaft als Quadratur des Kreises, wenn es einem unglaublichen Wunder gleichkommt.

Im *Zauberberg* von Thomas Mann kann man nachlesen, dass Hofrat Behrens die folgende Meinung vertritt[11]: ≪Wir haben die Analyse, wir haben die Aussprache,– ja Mahlzeit! Je mehr die Rasselbande sich ausspricht, desto lüsterner wird sie. Ich predige die Mathematik. [...] Die Beschäftigung mit der Mathematik, sage ich, ist das beste Mittel gegen die Kupidität. Staatsanwalt Paravant, der stark angefochten war, hat sich drauf geworfen, er hat es jetzt mit der Quadratur des Kreises und spürt große Erleichterung. Aber die meisten sind ja zu dumm und zu faul dazu, daß Gott erbarm.≫ Weiter wird von dem Staatsanwalt Paravant und seinen Bemühungen zur Quadratur des Kreises gesagt[12]: ≪Mit verdoppelter Inbrunst hatte er sich seitdem der klaräugigen Göttin in die Arme geworfen, von deren kalmierender Macht der Hofrat so Sittliches zu sagen wußte, und das Problem, dem bei Tag und Nacht all sein Sinnen gehörte, an das er all jene Persistenz, die ganze sportliche Zähigkeit wandte, mit der er ehemals, vor seiner oft verlängerten Beurlaubung, welche in völlige Quieszierung überzugehen drohte, die Überführung armer Sünder betrieben hatte - war kein anderes

[10]Oft wird aber auch, wenn der Ursprung der Redensart dem betreffenden nicht klar ist, ein sehr schwieriges Unterfangen gemeint. Hierfür gibt es unzählige Beispiele in den Zeitungen oder Reden von Politikern. Sehr oft ist es einer Person, die die Redensart "Quadratur des Kreises" benutzt, überhaupt nicht klar, was das eigentlich bedeutet. Vor vielen Jahren war ich als Assistent Beisitzer einer Prüfung, die mein Lehrer Lothar Collatz abnahm. Er fragte den Prüfling: „Lässt sich ein Quadrat konform auf einen Kreis abbilden?" Antwort: „Nein! Sonst wäre die Quadratur des Kreises möglich!" Unter `http://www.youtube.com/watch?v=UDDqjPeHj6s` zeigt der Autor eines Videos, dass er nicht so genau weiß, was die Quadratur des Kreises ist.

[11]In der bei S. Fischer herausgekommenen Frankfurter Ausgabe der Gesammelten Werke Thomas Manns (1981) findet man die Stelle auf S. 582.

[12]Diese Stellen findet man auf S. 884–885.

als die Quadratur des Kreises.

Der entgleiste Beamte hatte sich im Lauf seiner Studien mit der Überzeugung durchdrungen, daß die Beweise, mit denen die Wissenschaft die Unmöglichkeit der Konstruktion erhärtet haben wollte, unstichhaltig seien und daß die planende Vorsehung ihn, Paravant, darum aus der unteren Welt der Lebendigen entfernt und hierher versetzt habe,weil sie ihn dazu ausersehen das transzendente Ziel in den Bereich irdisch genauer Erfüllung zu reißen. So stand es mit ihm. Er zirkelte und rechnete, wo er ging und stand, bedeckte Unmassen von Papier mit Figuren, Buchstaben, Zahlen, algebraischen Symbolen, und sein gebräuntes Gesicht, das Gesicht eines scheinbar urgesunden Mannes, trug den visionären und verbissenen Ausdruck der Manie. Sein Gespräch betraf ausschließlich und mit furchtbarer Eintönigkeit die Verhältniszahl pi, diesen verzweifelten Bruch, den das niedrige, Genie eines Kopfrechners namens Zacharias Dase eines Tages bis auf zweihundert Dezimalstellen berechnet hatte - und zwar rein luxuriöserweise, da auch mit zweitausend Stellen die Annäherungsmöglichkeiten an das Unerreichbar-Genaue so wenig erschöpft gewesen wären, daß man sie für unvermindert hätte erklären können. Alles floh den gequälten Denker, denn wen immer ihm an der Brust zu ergreifen gelang, der mußte glühende Redeströme über sich ergehen lassen, bestimmt, seine humane Empfindlichkeit zu wecken für die Schande der Verunreinigung des Menschengeistes durch die heillose Irrationalität dieses mystischen Verhältnisses. Die Fruchtlosigkeit ewiger Multiplikationen des Durchmessers mit pi, um den Umfang - des Quadrats über dem Halbmesser, um den Inhalt des Kreises zu finden, schuf dem Staatsanwalt Anfälle von Zweifeln, ob nicht die Menschheit sich die Lösung, des Problems seit Archimedes' Tagen viel zu schwer gemacht habe und ob diese Lösung nicht in Wahrheit die kindlich einfachste sei. ≫

9 Goldener Schnitt, DIN-Format

Von Johannes Kepler (1571–1630) stammt die Aussage:

- Die Geometrie birgt zwei große Schätze: der eine ist der Satz von Pythagoras, der andere der Goldene Schnitt. Den ersten können wir mit einem Scheffel Gold vergleichen, den zweiten können wir ein kostbares Juwel nennen.

Man sagt, eine Strecke sei nach dem *goldenen Schnitt* in zwei Teile geteilt, wenn sich die gesamte Strecke zum größeren Teil verhält wie dieser zum kleineren. Hat man also eine Strecke der Länge L, so bestimmt sich die Länge l der größeren nach dem goldenen Schnitt geteilten Strecke aus

$$\frac{L}{l} = \frac{l}{L-l}.$$

Hieraus erhält man für das "Goldene-Schnitt-Verhältnis" L/l die Gleichung

$$\frac{L}{l} = \frac{1}{L/l - 1} \qquad \text{bzw.} \qquad \left(\frac{L}{l}\right)^2 - \frac{L}{l} = 1.$$

Von den beiden Lösungen dieser quadratischen Gleichung ist natürlich nur die positive relevant, so dass

$$\frac{L}{l} = \frac{1+\sqrt{5}}{2} \qquad \text{bzw.} \qquad l = \frac{\sqrt{5}-1}{2}L.$$

Die Zahl[13]

$$\phi := \frac{1+\sqrt{5}}{2} = 1.6180339887498948482\cdots$$

heißt die *Goldene-Schnitt-Zahl*[14]. Dieses Verhältnis bzw. Proportion (goldener Schnitt, stetige Teilung, sectio aurea, divina proportione, section d'or, golden section, golden ratio) hat seit Jahrhunderten nicht nur Mathematiker und Mathematikerinnen fasziniert. In Abschnitt 11 gehen wir hierauf etwas genauer ein. Dort werden wir auch in Abbildung 5 die Konstruktion des goldenen Schnitts (mit Zirkel und Lineal) nach Euklid veranschaulichen. Einige weitere Konstruktionen des goldenen Schnitts findet man z. B. bei A. Beutelsbacher, B. Petri (2000), siehe auch `http://de.wikipedia.org/wiki/Goldener_Schnitt`.

Ein weiteres wichtiges Verhältnis ist das DIN-Format. Ein Rechteck mit Seitenlängen L und $l \le L$ hat DIN-Format, wenn

$$\frac{L}{l} = \frac{l}{L/2} \qquad \text{bzw.} \qquad \frac{L}{l} = \sqrt{2}.$$

Gleichbedeutend ist, dass das Halbieren einer Seite, welches DIN-Format hat, wieder ein DIN-Format ergibt. Ein Rechteck im DIN-Format mit einer Fläche von $1\,\mathrm{m}^2 = 10\,000\,\mathrm{cm}^2$, hat die Seitenlängen l (in cm) mit $\sqrt{2}l^2 = 100^2$, also $l = 100/2^{1/4} \approx 84.0896$ und $L = \sqrt{2}l \approx 118.9207$. Ein Rechteck mit diesen Maßen hat das Format DIN A0. Die ersten Formate sind daher

Format	hoch	breit
DIN A0	118.9	84.1
DIN A1	84.1	59.5
DIN A2	59.5	42.0
DIN A3	42.0	29.7
DIN A4	29.7	21.0

[13] Die Bezeichnung ϕ für die Goldene-Schnitt-Zahl soll den griechischen Bildhauer Phidias (er lebte etwa von 490/80 bis 430/20 v. Chr.) ehren, der in seinen Skulpturen, aber vor allem beim Bau des Parthenon (Beginn 449 v. Chr.) den goldenen Schnitt angewandt haben soll, siehe D. E. Knuth (1968, S. 81). Allerdings ist Phidias nicht Baumeister des Parthenon gewesen, auch seine Beteiligung bei der Ausgestaltung der Friese und Giebel ist nicht gesichert. Er könnte allerdings (hierfür gibt es nur Plutarch als Quelle) eine Art Oberaufseher über alle beim Bau des Parthenon beteiligten Künstler gewesen sein.

[14] Unter

`http://www.cs.arizona.edu/icon/oddsends/phi.htm`

findet man die ersten 5 000 Stellen von ϕ. Bei

`http://pi.lacim.uqam.ca/piDATA/golden.txt`

sollen es (wir haben es nicht nachgezählt) 10 Millionen Stellen sein.

10 Die Fibonacci-Zahlen

Leonardo Pisano Fibonacci (*1170, †1250)[15] kann als der erste große Mathematiker des christlichen Abendlandes angesehen werden. Weniger durch seine algebraischen und geometrischen Untersuchungen als vielmehr durch die nach ihm benannten Zahlen ist Fibonacci uns bekannt geworden. Es wird untersucht wie viele Nachkommen ein Kaninchenpaar bekommt. Es wird von den folgenden Annahmen ausgegangen.

- Jedes Kaninchenpaar wird im Alter von 2 Monaten gebärfähig.
- Jedes gebärfähige Paar bringt von da an jeden Monat ein neues Paar zur Welt.
- Kaninchen leben unendlich lange.

Im ersten Monat lebt ein Paar. Dieses wird im zweiten Monat gebärfähig und gebiert im dritten Monat ein neues Paar. Auch im vierten Monat bringt das erste Paar ein neues Paar zur Welt, während im fünften Monat beide Paar ein neues Paar zur Welt bringen. In der folgenden Tabelle bedeutet N ein neues Paar und G ein gebärfähiges Paar:

Monat	1	2	3	4	5	6	7
N, G	$1N$	$1G$	$1G+1N$	$2G+1N$	$3G+2N$	$5G+3N$	$8G+5N$
Σ	1	1	2	3	5	8	13

Bezeichnet f_k die Anzahl der Kaninchenpaare im k-ten Monat, so ist also

$$f_1 = f_2 = 1, \qquad f_{k+1} = f_k + f_{k-1} \quad (k = 2, 3, \ldots).$$

Die Folge $\{f_k\}$ nennt man die Fibonacci-Folge bzw. die Folge der Fibonacci-Zahlen[16]. Bei A. Beutelspacher, B. Petri (1996, S. 89) kommt die folgende hübsche Aufgabe vor:

- Ein Briefträger steigt täglich eine lange Treppe nach folgendem Muster empor: Die erste Stufe betritt er in jedem Fall. Von da an nimmt er jeweils nur eine Stufe oder aber zwei Stufen auf einmal.

  Auf wieviel verschiedene Arten kann der Briefträger die k-te Stufe erreichen?

Man rät richtig: Auf f_k Arten kann die k-te Stufe erreicht werden. Dies kann sehr leicht durch vollständige Induktion nach k bewiesen werden.

Von den vielen Zusammenhängen zwischen der Goldenen-Schnitt-Zahl ϕ und der Fibonacci-Folge seien hier nur die folgenden genannt.

- Es ist
$$\phi = \lim_{k\to\infty} \frac{f_{k+1}}{f_k}.$$

[15] Biographisches zu Fibonacci und einiges über seine wissenschaftlichen Leistungen kann man unter `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Fibonacci.html` nachlesen.

[16] Man kann natürlich auch $f_0 := 0$, $f_1 := 1$ und $f_k := f_{k-1} + f_{k-2}$ setzen.

Ferner:

- Es gilt die Formel von Binet:

$$f_k = \frac{1}{\sqrt{5}}(\phi^k - (-\phi)^{-k}), \qquad k = 0, 1, \ldots.$$

- Es ist

$$\phi^k = f_k\phi + f_{k-1}, \qquad k = 1, 2, \ldots.$$

Diese beiden Beziehungen beweist man leicht durch vollständige Induktion.

Weiter wollen wir auf Fibonacci-Zahlen nicht eingehen. Die Literatur hierzu ist außerordentlich reichhaltig.

11 Fra Luca Pacioli: Divina Proportione

In diesem Abschnitt wollen wir uns mit (Fra) Luca Pacioli (*1445, †1517) [17] und seinem Werk *De Divina proportione* (Venedig, 1509) mit Illustationen von Leonardo da Vinci[18]. beschäftigen. Ein wunderschönes Bild (Ritratto di Fra Luca Pacioli, 1495) von Jacopo de Barbari zeigt den Franziskanermönch Luca Pacioli mit einer Schiefertafel, auf der man geometrische Konstruktionen, wahrscheinlich zum goldenen Schnitt, sehen kann, sowie einem Polyeder. Man kann sich dieses Bild auch unter der Adresse

`http://www.georgehart.com/virtual-polyhedra/pacioli.html`

ansehen (oder man fährt nach Neapel).

Der Beginn der Divina Proportione ist wie folgt (A. Beutelspacher, B. Petri (1996) hat dieser Beginn offenbar so gut gefallen, dass sie mit diesem Zitat ihr Buch über den goldenen Schnitt beginnen), wobei wir den Text der Übersetzung von C. Winterberg angeben:

- Ein für alle klaren und wissbegierigen Geister nothwendiges Werk; wo jeder Studirende der Philosophie, Perspective, Malerei, Sculptur, Architektur, Musik und anderer mathematischer Fächer eine angenehme subtile und bewundernswerthe Gelehrsamkeit antreffen und sich mit verschiedensten Fragen der heiligsten Wissenschaft erfreuen wird.

Das Buch ist Fürst Ludovico Sforza, Herzog von Mailand, gewidmet. Aus Cap. II zitieren wir lediglich:

- ... Und es ist deswegen nicht zu verwundern, wenn es zu unsern Zeiten wenig gute Mathematiker gibt, weil die Seltenheit guter Lehrer schuld daran ist, zugleich mit dem Schlunde Schlaf, und müssigen Federn, und zum Theil der modernen Geister.

[17] Biographische Einzelheiten zu Pacioli findet man unter

`http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Pacioli.html`

[18] Ein Originalexemplar ist übrigens in der Niedersächsischen Staats- und Universitätsbibliothek Göttingen vorhanden. Man kann es sich ansehen, eine Ausleihe ist nicht möglich!

In Cap. V (Vom passenden Titel des vorliegenden Tractats) begründet Pacioli den Titel Divina Proportione damit, dass der goldene Schnitt Eigenschaften habe, die sonst nur Gott zukommen. Er gibt vier Eigenschaften an:

- Die erste ist, dass sie nur allein da sei und nicht mehr; und es ist nicht möglich, andere Species noch Abweichungen von ihr anzugeben, welche Einheit der theologischen wie auch philosophischen Lehre gemäß das höchste Beiwort Gottes selber ist. Die zweite Eigenschaft ist die der heil. Dreieinigkeit, d. h. wie in den Göttlichen ein und dieselbe Substanz zwischen drei Personen, Vater, Sohn und heil. Geist besteht, ebenso muss ein und dieselbe Proportion dieser Art stets zwischen drei Ausdrücken stattfinden, und kann sich nie weder bei mehr noch weniger (Ausdrücken) wiederfinden, was besprochen werden wird. Die dritte Eigenschaft ist, dass, wie Gott eigentlich nicht definirt noch durch Worte uns verständlich gemacht werden kann, ebensowenig diese unsere Proportion durch eine verständliche Zahl je bestimmt noch durch irgend eine rationale Grösse sich ausdrücken lässt, sondern stets verborgen und geheim bleibt, und daher von den Mathematikern irrational genannt wird. Die vierte Eigenschaft ist, dass ebenso wie Gott sich niemals ändern kann, und Alles in Allem, und Alles in jedem seiner Theile ist, so unsere vorliegende Proportion stets in jeder continuirlichen und discreten Grösse; mögen dieselben gross oder klein sein, ein und dieselbe und stets unveränderlich bleibt, und auf keine Art sich verändern, noch auch mit dem Verstande auf andere Art aufgefasst werden kann, wie unserer Fortgang zeigen wird. ...

In Cap. VI (von seiner würdigen Empfehlung) wird von der göttlichen Proportion gesagt:

- Diese unsere Proportion, erhabener Herzog, ist solchen Vorzugs und Auszeichnung wert, wie man es in Anbetracht ihrer unendlichen Macht nur irgend sagen kann, sofern als ohne ihre Kenntnis sehr viele der Bewunderung höchst würdige Dinge weder in der Philosophie noch in irgend einer anderen Wissenschaft jemals ans Licht gelangen könnten, ...

Bei Pacioli folgen dann dreizehn "Sätze" (dort "Wirkungen" genannt). Diese werden nicht bewiesen (es werden gelegentlich Zahlenbeispiele angegeben), sondern es wird auf Euklid[19] verwiesen. Wir wollen diese dreizehn "Wirkungen" durchgehen.

- Von der ersten Wirkung einer nach unserer Proportion getheilten Linie (Cap. VII).

  Wenn eine Linie nach der Proportion getheilt ist, die einen mittleren und zwei äussere Abschnitte hat (...), und hat man von ihrem größeren Abschnitt die Hälfte der ganzen Linie hinzugefügt, welche so proportional getheilt wurde, so wird mit Nothwendigkeit folgen, dass das Quadrat ihrer Summe stets das fünffache, d. h. fünfmal so viel als das Quadrat der genannten vollen Hälfte beträgt.

Die entsprechende Aussage bei Euklid (XIII, 1) lautet:

[19] Im Internet findet man eine englische Ausgabe der Elemente Euklids unter
`http://aleph0.clarku.edu/~djoyce/java/elements/elements.html`

- Teilt man eine Strecke stetig, so wird ihr größerer Abschnitt, wenn man die Hälfte der ganzen Strecke hinzufügt, quadriert fünfmal so groß wie das Quadrat über die Hälfte.

Für einen arithmetischen Beweis (L=Länge der gesamten Strecke, $l = L/\phi$=Länge der größeren Strecke, wobei $\phi := (1 + \sqrt{5})/2$) ist

$$\Big(\frac{L}{2} + \frac{L}{\phi}\Big)^2 = 5\Big(\frac{L}{2}\Big)^2$$

nachzuweisen, was einfach ist. Die Konstruktion bei Euklid zur Bestimmung des goldenen Schnitts , siehe Abbildung 5, verläuft folgendermaßen:

- Gegeben sei die Strecke AB, die nach dem goldenen Schnitt zu teilen ist.
- Konstruiere das Quadrat $ABDC$.
- Halbiere die Strecke AC im Punkte E.
- Schlage um E einen Kreis mit dem Radius $|EB|$ und finde den Punkt F als Schnittpunkt dieses Kreises mit der Verlängerung der Strecke CA über A hinaus.
- Schlage um A einen Kreis mit dem Radius $|AF|$ und finde den den Punkt P als Schnittpunkt mit der Strecke AB.
- Im Punkte P wird die Strecke AB nach dem goldenen Schnitt geteilt.

Ein geometrischer Beweis folgt aus der Konstruktion des goldenen Schnitts (II, 11), wie wir sie in Abbildung 5 durchgeführt haben. Nach Konstruktion ist nämlich $|AE| = \frac{1}{2}|AB|$, der Satz von Pythagoras, angewandt auf $\triangle EAB$, ergibt $|EB|^2 = \frac{5}{4}|AB|^2$. Wegen

$$|EB| = |EF| = |EA| + |AF| = \frac{|AB|}{2} + |AP|$$

ist daher

$$\Big(\frac{|AB|}{2} + |AP|\Big)^2 = 5\Big(\frac{|AB|}{2}\Big)^2,$$

das ist gerade die Behauptung.

- Von ihrer zweiten wesentlichen Wirkung (Cap. XI).

  Wenn eine Grösse in zwei Theile getheilt und zu der einen eine Grösse hinzugefügt wird, so dass das Quadrat dieser Summe das Fünffache des Quadrats der hinzugefügten Größe ist, so folgt mit Nothwendigkeit, dass die genannte zugefügte Größe die Hälfte der in die beiden Theile zerlegten ersten Grösse sei, und dass die, zu welcher sie hinzugefügt, ihr größerer Abschnitt, und dass sie die ganze in ihnen nach unserer Proportion getheilt sei.

Bei Euklid (XIII, 2) findet man eine etwas andere Formulierung:

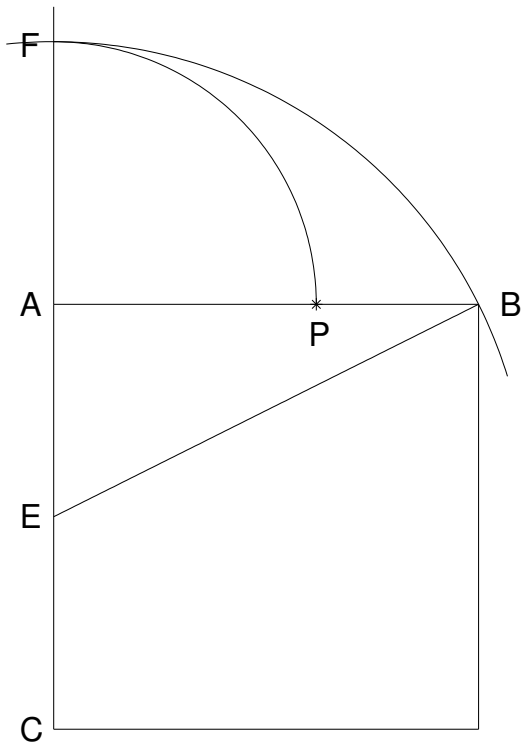


Abbildung 5: Konstruktion des goldenen Schnitts nach Euklid

- Wird quadriert eine Strecke fünfmal so groß wie das Quadrat eines Abschnittes von ihr, dann ist, wenn man das doppelte des genannten Abschnitts stetig teilt, der größere Abschnitt der Rest der ursprünglichen Strecke.

Ein analytischer Beweis der Aussage bei Euklid ist einfach. Ist nämlich $1 = \sqrt{5}\alpha$, so ist $2\alpha/\phi = 1 - \alpha$. Die Aussage bei Pacioli ist ebenfalls klar, wenn zu Beginn unter einer Teilung einer Größe eine stetige Teilung verstanden wird. Denn aus $\alpha + 1/\phi = \sqrt{5}\alpha$ folgt $\alpha = \frac{1}{2}$.

- Über ihre dritte besondere Wirkung (Cap. XII).

  Wenn eine Größe nach unserer Proportion getheilt ist, und wenn man dem kleineren Abschnitt die Hälfte des größeren hinzufügt, so wird alsdann stets das Quadrat der Summe das Fünffache des Quadrats der Hälfte des genannten größeren Abschnitts sein.

Fast wörtlich ist dies die Aussage bei Euklid (XIII, 3). Ein arithmetischer Beweis beruht auf

$$\Big(\Big(1 - \frac{1}{\phi}\Big) + \frac{1}{2\phi}\Big)^2 = 5\Big(\frac{1}{2\phi}\Big)^2,$$

was natürlich leicht zu zeigen ist. Wir wollen uns Euklids Beweis näher ansehen, siehe Abbildung 6. Die Strecke AB sei im Punkt C nach dem goldenen Schnitt geteilt, AC sei die größere der beiden Strecken. Die Strecke AC sei in D halbiert. Behauptet wird dann, dass $|DB|^2 = 5|DC|^2$. Wir bilden das Quadrat AE mit der Seitenlänge $|AB|$ (wir

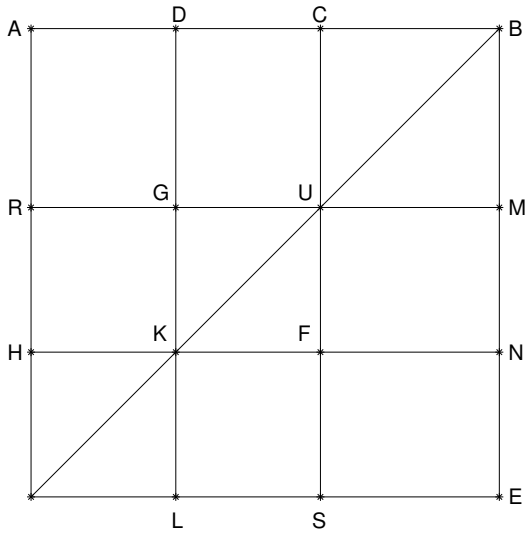


Abbildung 6: Beweis von Proposition 3 in Buch XIII

geben für Rechtecke und speziell Quadrate nur gegenüber liegende Ecken an). Auch die Seitenlängen dieses Quadrats seien entsprechend geteilt. Zu zeigen ist, dass der Flächeninhalt des Quadrats DN fünfmal so groß wie der des Quadrats GF ist. Sozusagen nach Definition des goldenen Schnitts (siehe Abbildung 5) ist $|AC|^2 = |AB|\,|CB|$. Wegen $|AC| = 2\,|DC|$ ist daher $|AB|\,|CB| = 4\,|DC|^2$. D. h. der Flächeninhalt des Rechtecks CE ist viermal so groß wie der des Quadrats GF. Der Flächeninhalt des Rechtecks DU ist aber offensichtlich gleich dem des Rechtecks FE. Hieraus folgt die Behauptung.

- Von ihrer vierten unsagbaren Wirkung (Cap. XIII).

  Wenn eine Grösse nach unserer göttlichen Proportion getheilt wird und man zu der ganzen Größe ihren größeren Abschnitt hinzufügt, so werden genannte Summe und genannter größerer Abschnitt Theile einer anderen ebenso getheilten Grösse sein. Und der größere Abschnitt dieser zweiten so getheilten Grösse wird immer die ganze zuerst genannte Grösse sein.

Bei Euklid (XIII, 5) heißt die entsprechende Stelle:

- Teilt man eine Strecke stetig und setzt ihr eine dem größeren Abschnitt gleiche an, dann ist die Summenstrecke stetig geteilt, und größerer Abschnitt ist die Ausgangsstrecke.

Ein analytischer Beweis basiert einfach auf der Gleichung

$$\frac{1}{\phi}\Big(1+\frac{1}{\phi}\Big) = 1.$$

Euklids Beweis wird in Abbildung 7 veranschaulicht. Die Strecke AB sei im Punkt C stetig geteilt, AC sei der größere Abschnitt und $|DA| = |AC|$. Zu zeigen ist, dass die Strecke DB in A stetig geteilt wird und die Ausgangsstrecke AB der größere Abschnitt

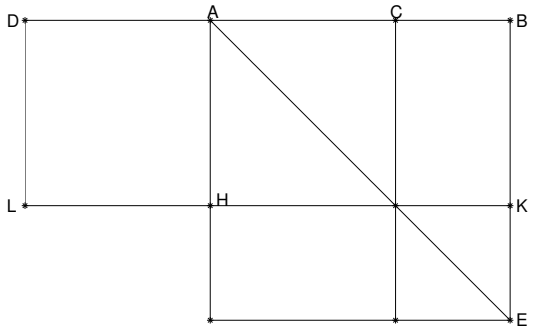


Abbildung 7: Beweis von Proposition 5 in Buch XIII

ist. Wie in Abbildung 7 angegeben, konstruiere man das Quadrat AE und das Rechteck DK. Man hat zu zeigen, dass ihre Flächeninhalte gleich sind. Da die Strecke AB in C nach dem goldenen Schnitt geteilt wird, ist der Flächeninhalt des Quadrates DH gleich dem des Rechtecks HE, woraus die Behaupung unmittelbar folgt.

- Von ihrer fünften wunderbaren Wirkung (Cap. XIV).

  Wenn eine Größe nach unserer genannten Proportion getheilt ist, so ist stets die Summe des Quadrats des kleineren Abschnittes und des Quadrats der ganzen Größe das dreifache des Quadrats des grösseren Abschnittes.

Genau diese Aussage findet man bei Euklid (XIII, 4). Die Aussage ist richtig, da

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

Bei Euklid basiert der Beweis auf Abbildung 8. Die Strecke AB sei in C nach dem

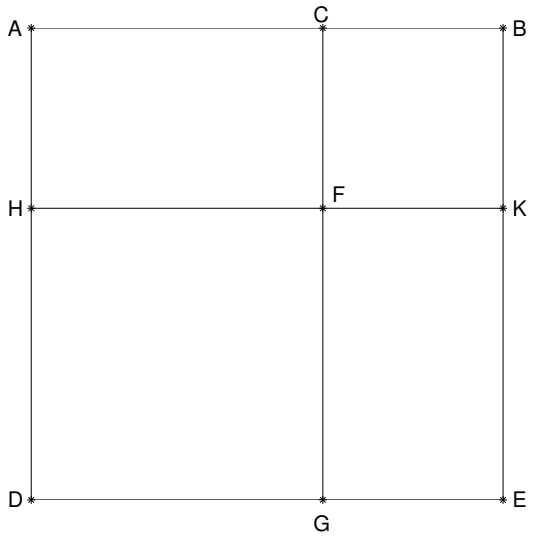


Abbildung 8: Beweis von Proposition 4 in Buch XIII

goldenen Schnitt geteilt, AC sei die größere Strecke. Man bilde das Quadrat AE mit

der Seitenlänge $|AB|$. Zu zeigen ist, dass der Flächeinhalt des Quadrats AE plus dem des Quadrats CK dreimal so viel wie der Flächeninhalt des Quadrats HG ist. Da C die Strecke AB nach dem goldenen Schnitt teilt, ist der Flächeninhalt des Rechtecks AK gleich dem des Quadrates HG. Der Rest des Beweises ist einfach und bleibt dem Leser überlassen.

- Von ihrer sechsten unnennbaren Wirkung (Cap. XV).

  Keine rationale Größe kann je nach unserer genannten Proportion so getheilt werden, ohne dass jeder ihrer Abschnitte irrational ... sei.

Siehe Euklid (XIII, 6). Gemeint ist ja wohl: Ist L rational, so sind L/ϕ und $L(1-1/\phi)$ irrational.

- Von ihrer siebenten unglaublichen Wirkung (Cap. XVI).

  Wenn man die Seite des gleichseitigen Sechsecks zu der Seite des gleichseitigen Zehnecks addirt, welche beide als in ein und demselben Kreis beschrieben sich verstehen, so wird ihre Summe immer eine nach unserer genannten Proportion getheilte Grösse sein. Und ihr grösserer Abschnitt wird die Sechseckseite sein.

Das ist genau die Aussage bei Euklid (XIII, 9). Ein analytischer Beweis der Aussage ist einfach. Sei nämlich s_n die Seitenlänge eines dem Einheitskreis eingeschriebenen regelmäßigen n-Ecks. Eine leichte Überlegung (siehe auch Abschnitt 31) zeigt, dass $s_n = 2\sin(\pi/n)$. Zu zeigen ist also, dass $s_6+s_{10} = \phi s_6$ bzw. $(\sin(\pi/6)+\sin(\pi/10))/\sin(\pi/6) = \phi$, was aber wegen $\sin(\pi/6) = 1/2$ und $\sin(\pi/10) = (\sqrt{5}-1)/4$ richtig ist.

- Von der umgekehrten Wirkung der vorhergehenden (Cap. XVII).

  Wenn eine Linie nach der Proportion getheilt ist, die einen mittleren und zwei äußere Abschnitte hat, so ist immer in dem Kreise wofür der größere Abschnitt die Seite des ihm einbeschriebenen Sechsecks ist, der kleinere die entsprechende Zehneckseite.

Es ist mir unklar, weshalb diese "Wirkung" von Pacioli aufgenommen wurde (allerdings fehlt auch ein steigerndes Adjektiv!), denn wegen der siebenten Wirkung ist diese Aussage trivial.

- Von ihrer neunten über die anderen hinausgehenden Wirkung (Cap. XVIII).

  Wenn man im Kreise das gleichseitige Fünfeck bildet, und über seine zwei benachbarten Ecken zwei gerade Linien von den Endpunkten seiner Seiten ausgehend spannt, so werden sich diese untereinander nothwendigerweise nach unserer Proportion theilen.

Bei Euklid (XIII, 8) heißt es etwas genauer:

- Diagonalen, die im gleichseitigen und gleichwinkligen Fünfeck zwei aufeinanderfolgenden Winkeln gegenüberliegen, teilen einander stetig; und ihre größeren Abschnitte sind der Fünfeckseite gleich.

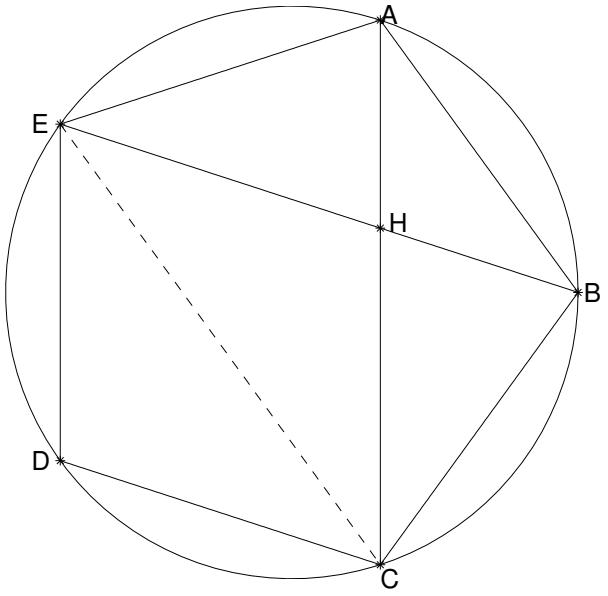


Abbildung 9: Proposition 8 in XIII bei Euklid

Ein analytischer Beweis ist nicht schwierig, wir verzichten aber darauf. Etwas verkürzt sieht der Beweis bei Euklid folgendermaßen aus, man vergleiche Abbildung 9 (wir benutzen im wesentlichen dieselben Bezeichnungen wie Eukid). Wir wollen zeigen, dass die Diagonalen AC und EB in H nach dem goldenen Schnitt getrennt werden.

1. Es ist $|EA| = |AB|$ und $\triangle ABE = \triangle ABC$, denn zwei Seiten und ein eingeschlossener Winkel sind gleich.

2. Es ist $\sphericalangle BAC = \sphericalangle ABE$ und $\sphericalangle AHE = 2\sphericalangle BAH$.

3. Es ist $\sphericalangle EAC = 2\sphericalangle BAC$.

4. Es ist $\sphericalangle HAE = \sphericalangle AHE$ und daher $|HE| = |EA| = |AB|$.

5. Es ist $\sphericalangle ABE = \sphericalangle AEB = \sphericalangle BAH$.

6. Die Dreiecke $\triangle ABE$ und $\triangle HAB$ sind gleichschenklig und winkelgleich. Daher ist bei beiden Dreiecken das Verhältnis der Länge der Grundseite zur Länge des Schenkels dasselbe. D. h. es ist

$$\frac{|EB|}{|AB|} = \frac{|AB|}{|BH|}.$$

Nun ist aber $|AB| = |EH|$, folglich

$$\frac{|EB|}{|EH|} = \frac{|EH|}{|BH|}.$$

Da $|EB| > |EH|$ ist $|EH| > |BH|$. Also ist schließlich bewiesen, dass die Stecke EB in H nach dem goldenen Schnitt getrennt wird.

- Über ihre 10. höchste Wirkung (Cap. XIX).

  Wenn eine Grösse nach der genannten Proportion getheilt ist, so gehen alle Wirkungen, welche aus ihr und ihren Abschnitten entspringen können, ihrer Beschaffenheit , Anzahl, Species und Gattung nach selbst aus irgend einer anderen ebenso getheilten Göße hervor.

Hier kann man wohl nur spekulieren, was gemeint sein könnte.

- Von ihrer 11. ausgezeichnetsten Wirkung (Cap. XX).

  Wenn man die Seite eines gleichseitigen Sechsecks nach unserer göttlichen Proportion theilen wird, so wird ihr größerer Abschnitt stets nothwendig die Seite des von demselben Kreise wie das Sechseck umschriebenen Zehnecks sein.

Bei Pacioli wird auf Buch XIV, 3 verwiesen (was eigentlich nicht zu den ursprünglichen Elementen gehört). Wegen der siebten Wirkung bzw. XII, 9 ist die Aussage aber eigentlich klar. Denn danach ist $(s_6 + s_{10})/s_6 = \phi$ (hierbei bedeutet s_n die Seitenlänge eines dem Einheitskreis eingeschriebenen regelmäßigen n-Ecks. Also ist $s_{10}/s_6 = \phi - 1 = 1/\phi$, womit die Behauptung bewiesen ist.

- Von ihrer zwölften fast unbegreiflichen Wirkung (Cap. XXI).

  Wenn eine Grösse nach unserer genannten Proportion getheilt wird, so verhält sich die Wurzel aus der Summe aus dem Quadrat der ganzen Grösse und dem Quadrat ihres grössern Abschnitts zur Wurzel der Summe aus dem Quadrat genannter Grösse und dem Quadrate ihres kleinern Abschnitts wie die Seite des Kubus zur Seite des Dreiecks des zwanzigflächigen Körpers.

Der "zwanzigflächige Körper" ist natürlich das Ikosaeder (20 Flächen (Dreiecke), 12 Ecken und 30 Kanten), siehe Abbildung 23. Der Durchmesser der Umkugel eines Würfels mit Kantenlänge l_6 ist $2r = l_6\sqrt{3}$ (das ist genau der Inhalt von Euklid XIII, 15). Die Kante eines Ikosaeders in einer Kugel vom Radius r ist

$$l_{20} = r\sqrt{2 - 2/\sqrt{5}} = \frac{r}{\sqrt{5}}\sqrt{10 - 2\sqrt{5}}$$

(siehe Euklid XIII, 16). Dann ist einerseits

$$\frac{\sqrt{1^2 + (1/\phi)^2}}{\sqrt{1^2 + (1 - 1/\phi)^2}} = \frac{\sqrt{\phi^2 + 1}}{\sqrt{\phi^2 + (\phi - 1)^2}} = \frac{\sqrt{\phi + 2}}{\sqrt{3}} = \sqrt{\frac{5 + \sqrt{5}}{6}}$$

und andererseits

$$\frac{l_6}{l_{20}} = \sqrt{\frac{5 + \sqrt{5}}{6}}.$$

Damit ist die Aussage bewiesen.

- Von ihrer dreizehnten werthesten Wirkung (Cap. XXII).

Diese “Wirkung” wollen wir nicht angeben. Sie ist ziemlich unklar formuliert und betrifft die Konstruktion des “edelsten von allen regelmäßigen Körpern, Dodekaeder genannt”. Die vier anderen platonischen Körper (Tetraeder, Hexaeder bzw. Würfel oder Kubus, Oktaeder und Ikosaeder) sind den vier Elementen (Erde, Feuer, Luft, Wasser) zugeordnet, während das Dodekaeder von Plato als Gestalt angesehen wurde, die das ganze Weltall umfasst.

Schließlich folgt Cap. XXIII (Wie aus Ehrfurcht vor unserem Heile die genannte Wirkungen endigen).

- Es scheint mir, erhabener Herzog, nicht angemessen, mich über noch mehr von ihren unendlichen Wirkungen für jetzt zu verbreiten, weil das Papier der Tinte nicht genügen würde, sie alle auszudrücken, sondern wir haben nur diese dreizehn unter den andern ausgewählt, aus Verehrung für die Schaar der Zwölf und ihres heiligsten Hauptes, unseres Erlösers Jesus Christus. Denn da wir ihr den göttlichen Namen auch der Zahl nach von 13 Artikeln mit Bezug auf unser Heil und zwar der zwölf Apostel mit unserm Erlöser beigelegt haben, so seien sie hiermit beendigt; ...

12 Was haben Goethe, Lichtenberg, Schopenhauer und Poe über Mathematiker gesagt?

Wir geben Aussagen von Johann Wolfgang von Goethe (1749–1832) und Georg Christoph Lichtenberg (1742–1799) über Mathematiker an. Zunächst Goethe, der sich im Gegensatz zur Mathematik und ihrer Bedeutung als Wissenschaft in der Mehrzahl der Fälle gegenüber den „Mathematikern“ negativ äußert. Eine positive Ausnahme ist sein Lob, wie Mathematiker ihre Argumente sorgfältig aneinanderreihen:

- »Diese Bedächtlichkeit, nur das Nächste ans Nächste zu reihen oder vielmehr das Nächste aus dem Nächsten zu folgern, haben wir von den Mathematikern zu lernen, und selbst da, wo wir uns keiner Rechnung bedienen, müssen wir immer so zu Werke gehen, als wenn wir dem strengsten Geometer Rechenschaft schuldig wären.«

Eine noch harmlose, sehr bekannte Aussage Goethes über Mathematiker ist:

- »Die Mathematiker sind eine Art Franzosen: Redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsbald etwas anderes.«

Schon weniger freundlich ist seine Äußerung vom 17. 5. 1829 in einem Brief an seinen Freund Zelter:

- »Daß aber ein Mathematiker, aus dem Hexengewirre seiner Formeln heraus, zur Anschauung der Natur käme und Sinn und Verstand, unabhängig, wie ein gesunder Mensch brauchte, werd ich wohl nicht erleben. «

Oder auch:

- »Die Mathematiker sind närrische Kerls und sind so weit entfernt, auch nur zu ahnen, worauf es ankommt, daß man ihnen ihren Dünkel nachsehen muß. Ich bin sehr neugierig auf den ersten, der die Sache einsieht und sich redlich dabei benimmt: denn sie haben doch nicht alle ein Brett vor dem Kopfe, und nicht alle haben bösen Willen. Übrigens wird mir dann doch bei dieser Gelegenheit immer deutlicher, was ich schon lange im Stillen weiß, daß diejenige Kultur, welche die Mathematik dem Geiste gibt, äußerst einseitig und beschränkt ist.«

Nun Lichtenberg:

- »Die sogenannten Mathematiker von Profession haben sich, auf die Unmündigkeit der übrigen Menschen gestützt, einen Kredit von Tiefsinn erworben, der viel Ähnlichkeit mit dem von Heiligkeit hat, den die Theologen für sich haben.«

- »Die Mathematik ist eine gar herrliche Wissenschaft, aber die Mathematiker taugen oft den Henker nicht. Es ist fast mit der Mathematik wie mit der Theologie. So wie die letzteren Beflissenen, zumal wenn sie in Ämtern stehen, Anspruch auf einen besonderen Kredit von Heiligkeit und eine nähere Verwandtschaft mit Gott machen, obgleich sehr viele darunter wahre Taugenichtse sind, so verlangt sehr oft der so genannte Mathematiker für einen tiefen Denker gehalten zu werden, ob es gleich darunter die größten Plunderköpfe gibt, die man finden kann, untauglich zu irgend einem Geschäft, das Nachdenken erfordert, wenn es nicht unmittelbar durch jene leichte Verbindung von Zeichen geschehen kann, die mehr das Werk der Routine, als des Denkens sind.«

Schließlich Aussagen von Schopenhauer (1788–1860):

- »Der einzige unmittelbare Nutzen, welcher der Mathematik gelassen wird, ist, daß sie unstäte und flatterhafte Köpfe gewöhnen kann, ihre Aufmerksamkeit zu fixiren.«

- »Daß die niedrigste Tätigkeit die arithmetische ist, wird dadurch belegt, daß sie die einzige ist, die auch durch eine Maschine ausgeführt werden kann. Nun läuft aber alle analysis finitorum et infinitorum im Grunde doch auf Rechnerei zurück. Danach bemesse man den mathematischen Tiefsinn.«

Thomas Mann beschreibt in *Königliche Hoheit*, welchen Eindruck mathematische Formeln auf ihn (seine Frau Katia hatte bekanntlich angefangen, Mathematik zu studieren; sein Schwiegervater Alfred Pringsheim war Professor der Mathematik in München) machten:

- «Was er sah, war sinnverwirrend. In einer krausen, kindlich dick aufgetragenenen Schrift, die Imma Spoelmanns besondere Federhaltung erkennen liess, bedeckte ein phantastischer Hokuspokus, ein Hexensabbat verschränkter Runen die Seiten. Griechische Schriftzeichen waren mit lateinischen und mit Ziffern in verschiedener Höhe verkoppelt, mit Kreuzen und Strichen durchsetzt, ober- und und unterhalb

waagrechter Linien bruchartig aufgereiht, durch andere Linien zeltartig überdacht, durch Doppelstricheichen gleichgewertet, durch runde Klammern zusammengefasst, durch eckige Klammern zu großen Formelmassen vereinigt. Einzelne Buchstaben, wie Schildwachen vorgeschoben, waren rechts oberhalb der umklammerten Gruppen ausgesetzt. Kabbalistische Male, vollständig unverständlich dem Laiensinn, umfassten mit ihren Armen Buchstaben und Zahlen, während Zahlenbrüche ihnen voranstanden und Zahlen und Buchstaben ihnen zu Häupten und Füßen schwebten. Sonderbare Silben, Abkürzungen geheimnisvoller Worte waren überall eingestreut, und zwischen den nekromantischen Kolonnen standen geschriebene Sätze und Bemerkungen in täglicher Sprache, deren Sinn gleichwohl so hoch über allen menschlichen Dingen war, dass man sie lesen konnte, ohne mehr davon zu verstehen als von einem Zaubergemurmel.»

In Thomas Manns Erzählung *Der kleine Herr Friedemann* findet sich folgende Passage:

- «Auch ein Student der Mathematik war anwesend, [...] der mit Eifer sprach. Er hatte die Behauptung aufgestellt, dass man durch einen Punkt mehr als eine Parallele zu einer Geraden ziehen könne, Frau Rechtsanwalt Hagenström hatte gerufen: ‚Dies ist unmöglich!' und nun bewies er es so schlagend, dass alle taten, als hätten sie es verstanden.»

In der Erzählung *Der entwendete Brief* von Edgar Allan Poe (1809–1849) findet sich der folgende Teil eines Gesprächs zwischen dem Ich-Erzähler und seinem Freund Auguste Dupin:

- »Aber ist denn dieser wirklich der Dichter?« fragte ich. »Es sind zwei Brüder, wie ich weiß, und beide haben als Schriftsteller einen Namen. Der Gesandte, glaube ich, hat eine gelehrte Abhandlung über Differentialrechnung geschrieben. Er ist Mathematiker und kein Dichter.«

  »Sie irren sich. Ich kenne ihn gut; er ist beides. Als Dichter und Mathematiker versteht er, schlau zu überlegen; als bloßer Mathematiker verstände er überhaupt nicht zu schlußfolgern und wäre sicherlich dem Präfekten in die Hände gefallen.«

  »Sie überraschen mich«, sagte ich. »Ihre Anschauung wird von der ganzen Welt Lügen gestraft. Sie werden doch wohl nicht eine seit Jahrhunderten festbegründete Ansicht umstoßen wollen? Die Vernunft des Mathematikers gilt seit langem als die Überlegungsfähigkeit par excellence.«

13 Das Buffonsche Nadelproblem

George Louis Leclerc, Comte de Buffon (1707–1788) stellte 1777 das folgende Problem und löste es: Eine Nadel der Länge L werde zufällig (jede Position ist gleich wahrscheinlich) auf eine horizontale Ebene geworfen, auf der sich parallele gerade Linien mit einem Abstand von $d > L$ befinden. Was ist die Wahrscheinlichkeit dafür, dass die Nadel eine der Linien berührt?

Die Position der Nadel ist durch $(\phi, x_0) \in [0, \pi] \times [0, d/2]$ (siehe Abbildung 10) bestimmt. Die Nadel trifft eine Linie genau dann, wenn $x_0 - (L/2)|\cos\phi| \leq 0$. Sei

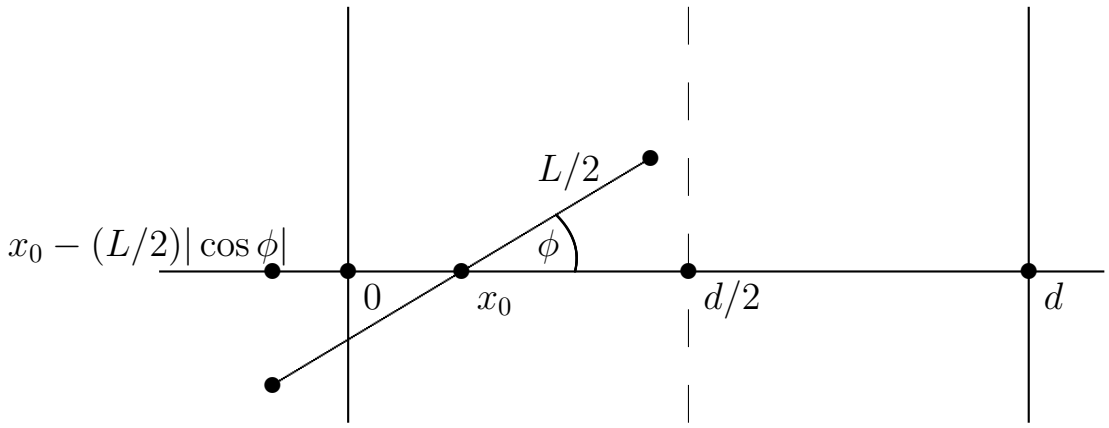


Abbildung 10: Das Buffonsche Nadelproblem

$$E := [0, \pi] \times [0, d/2)], \qquad T := \{(\phi, x_0) \in E : x_0 \leq (L/2)|\cos\phi|\}.$$

Die Wahrscheinlichkeit p, dass die Nadel eine Linie trifft, ist $p = \mu(T)/\mu(E)$. Hier ist $\mu(E)$, das Mass von E, natürlich $\mu(E) = (\pi d)/2$, während $\mu(T)$ durch

$$\mu(T) = \int_0^\pi \frac{L}{2}|\cos\phi|\, d\phi = \frac{L}{2} \cdot 2 = L$$

gegeben ist. Die gesuchte Wahrscheinlichkeit ist also

$$p = \frac{2L}{\pi d}.$$

Näherungsweise kann man π also folgendermaßen bestimmen: Man mache n Würfe mit mit der Nadel. Sei T_n die Anzahl der "Treffer" und $p_n := T_n/n$. Dann ist

$$\pi_n := \frac{2L}{p_n d} = \frac{2Ln}{T_n d}$$

eine Approximation für π. Die Konvergenz ist allerdings sehr schlecht. Im Netz findet man Applets zur Simulation dieses Experiments, z. B. unter `http://www.mathematik.ch/anwendungenmath/wkeit/buffon/Buffon_Nadelproblem.php` oder auch `http://mste.illinois.edu/reese/buffon/bufjava.html`.

14 Der Freundschaftssatz

Der Freundschaftssatz ist eine mathematische Fassung der folgenden Beobachtung:

- *In einer Gruppe von Personen mögen je zwei Personen genau einen gemeinsamen Freund haben. Dann gibt es in der Gruppe einen "Politiker", d. h. eine Person, die mit allen anderen Personen befreundet ist.*

Die mathematische Fassung ist

Satz *Sei $G = (V, E)$ ein Graph, in dem je zwei verschiedene Ecken genau einen gemeinsamen Nachbarn haben*[20]*. Dann gibt es in G eine Ecke, die zu allen anderen Ecken benachbart ist.*

Beweis: Man überzeugt sich leicht, dass es wirklich Graphen mit der angegebenen Eigenschaft gibt, nämlich die sogenannten "Windmühlengraphen", siehe Abbildung 11.

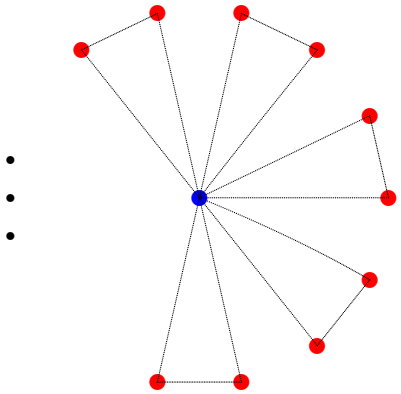


Abbildung 11: Ein Windmühlengraph

Im ersten Schritt zeigen wir:

- Zwei nicht benachbarte Ecken $u, v \in V$ haben denselben Grad, also dieselbe Anzahl von Nachbarn.

Denn: Sei $N(u) := \{w \in V : uw \in E\}$ die Menge der zu u benachbarten Ecken, entsprechend sei $N(v)$ definiert. Sei $T : N(u) \longrightarrow N(v)$ dadurch definiert, dass $T(w) \in N(v)$ bei gegebenem $w \in N(u)$ der eindeutige Nachbar von w und v ist. Es ist $T(w) \neq u$, da u und v nicht benachbart sind. Die Abbildung T ist eineindeutig, denn $w \in N(u)$ ist eindeutig als gemeinsamer Nachbar von u und $T(w) \in N(v)$ bestimmt. Aus Symmetriegründen ist T surjektiv und daher haben $N(u)$ und $N(v)$ gleich viele Elemente, womit die erste Zwischenbehauptung bewiesen ist.

Wir machen jetzt einen *Widerspruchsbeweis* und nehmen an, dass es keine Ecke gibt, die zu allen anderen Ecken benachbart ist.

- Gibt es eine Ecke mit dem Grad $k > 1$, so haben alle Ecken den Grad k.

Denn: Den Grad einer Ecke $v \in V$, also die Anzahl der Nachbarn von v, bezeichnen wir mit $d(v)$. Wir definieren die Mengen

$$A := \{u \in V : d(u) = k\}, \qquad B := \{v \in V : d(v) \neq k\}$$

und nehmen an, es sei $B \neq \emptyset$. Ecken aus A und B haben unterschiedlichen Grad und sind daher wegen der ersten Zwischenbehauptung benachbart. Die Mengen A und

[20] Hierbei sagen wir natürlich, zwei Ecken seien benachbart, wenn sie durch eine Kante miteinander verbunden sind.

B bestehen aus mehr als einer Ecke, denn andernfalls wäre diese eine Ecke aus A bzw. B ein "Politiker" bzw. eine Ecke, die zu allen anderen Ecken benachbart ist, was wir bei unserem Widerspruchsbeweis ausgeschlossen haben. Daher enthalten A und B jeweils mindestens zwei Elemente $u_1, u_2 \in A$ bzw. $v_1, v_2 \in B$. D. h. aber, dass die beiden Ecken u_1 und u_2 die zwei gemeinsamen Nachbarn v_1 und v_2 haben, was aber ausgeschlossen ist. Damit ist die Annahme $B \neq \emptyset$ zum Widerspruch geführt und die zweite Zwischenbehauptung bewiesen.

- Die Anzahl n der Ecken in G ist $n = k(k-1)+1$.

Denn: Wir zählen die Anzahl der Wege der Länge 2 in G auf zweierlei Art. Zu je zwei verschiedenen Ecken kann man genau einen Weg der Länge 2 bestimmen. Auf $\binom{n}{2}$ Weisen kann man aus n Ecken 2 Ecken auswählen, die Anzahl der Wege der Länge 2 ist also $\binom{n}{2}$. Andererseits gibt es zu jeder der n Ecken $v \in V$ genau k Nachbarn, und daher $\binom{k}{2}$ Wege der Länge 2, die v in der Mitte haben. Dies ergibt insgesamt $n\binom{k}{2}$ Wege der Länge 2. Aus

$$\frac{n(n-1)}{2} = \binom{n}{2} = n\binom{k}{2} = n\frac{k(k-1)}{2}$$

erhalten wir, wie behauptet, $n = k(k-1)+1$.

Nun kommt das Finale, bei welchem elementare Aussagen der linearen Algebra benutzt werden. Sei $V = \{v_1, \ldots, v_n\}$. Die *Adjazenzmatrix* $A = (a_{ij}) \in \{0,1\}^{n\times n}$ ist definiert durch

$$a_{ij} := \begin{cases} 1, & \text{falls } v_i v_j \in E, \\ 0, & \text{sonst.} \end{cases}$$

Die Adjazenzmatrix zu einem (einfachen) Graphen ist symmetrisch und besitzt verschwindende Diagonalelemente. Wegen der zweiten Zwischenbehauptung besitzt jede Ecke den Grad k, also enthält jede Zeile von A genau k Einsen. Ferner gibt es zu je zwei Zeilen genau eine Spalte, in der beide eine Eins haben. Es ist

$$(A^2)_{ij} = \sum_{l=1}^{n} a_{il}a_{lj} = \sum_{i=1}^{n} a_{il}a_{jl}.$$

Wir unterscheiden zwei Fälle: Ist $i = j$, so ist

$$(A^2)_{ii} = \sum_{l=1}^{n} a_{il}^2 = k.$$

Für $i \neq j$ gibt es genau ein l mit $a_{il} = a_{jl} = 1$, daher ist $(A^2)_{ij} = 1$ für $i \neq j$. Also ist

$$A^2 = \begin{pmatrix} k & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & k \end{pmatrix} = (k-1)I + ee^T,$$

wobei $e \in \mathbb{R}^n$ der Vektor ist, dessen Komponenten alle gleich 1 sind. Die Matrix A^2 hat die Eigenwerte $k-1$ (mit der Vielfachheit $n-1$ und Eigenvektoren aus span $\{e\}^\perp$) und

$k-1+n=k^2$ (mit der Vielfachheit 1 und dem Eigenvektor e). Wir wollen jetzt auf die Eigenwerte von A schließen. Die Eigenwerte von A^2 sind die Quadrate der Eigenwerte von A bzw. die Eigenwerte von A sind die Quadratwurzeln der Matrix A^2. Daher hat A genau $n-1$ Eigenwerte $\pm\sqrt{k-1}$, ferner den Eigenwert k mit Eigenvektor e, wie man aus

$$(Ae)_i = \sum_{j=1}^{n} a_{ij} = k = ke_i, \qquad i = 1, \ldots, n,$$

abliest. Wir nehmen an, dass r Eigenwerte von A gleich $\sqrt{k-1}$ und s Eigenwerte von A gleich $-\sqrt{k-1}$ sind. Da die Summe der Eigenwerte von A gleich der Spur von A bzw. gleich Null ist, ist

$$k + r\sqrt{k-1} - s\sqrt{k-1} = 0.$$

Insbesondere ist $r \neq s$ und

$$\sqrt{k-1} = \frac{k}{s-r}.$$

Dann ist aber $h := \sqrt{k-1} \in \mathbb{N}$, denn: Ist die Quadratwurzel einer natürlichen Zahl rational, so ist die Quadratwurzel selbst eine natürliche Zahl. Aus

$$h(s-r) = k = h^2 + 1$$

erkennt man, dass h sowohl h^2 als auch h^2+1 teilt, woraus $h=1$ bzw. $k=2$ folgt. Hieraus folgt aber, dass $G = K_3$ der vollständige Graph mit drei Ecken ist. Dies ist ein Widerspruch, denn wir hatten angenommen, dass es in G keine Ecke gibt, die mit allen anderen Ecken benachbart ist. Der Satz ist also bewiesen. □

Bemerkung: Der erste Beweis des Freundschaftssatzes (friendship theorem) stammt wohl von P. Erdős, A. Rényi, V. Sós (1966), man findet ihn auch bei M. Aigner, G. M. Ziegler (2002, S. 253 ff.). Für den ersten, kombinatorischen Teil des obigen Beweises haben wir einen Aufsatz von C. Huneke (2002) benutzt. □

Zum Schluss dieses Abschnitts wollen wir auf eine weitere Aussage eingehen, die gelegentlich ebenfalls als Freundschaftsatz (Theorem on friends and strangers) firmiert. Man stelle sich eine Party mit (bzw. Treffen von) sechs Leuten vor. Betrachten wir je zwei von ihnen. Entweder kannten sie sich schon vorher, dann nennen wir sie befreundet, oder sie kannten sich noch nicht, dann sagen wir, dass sie sich fremd sind. Der Freundschaftssatz sagt dann aus:

- In jeder Party mit sechs Leuten gibt es mindestens drei Leute, die sich fremd sind, oder mindestens drei Leute, die miteinander befreundet sind.

Nun formulieren wir den Freundschaftssatz in naheliegender Weise in graphentheoretischer Sprache. Siehe auch

`http://www.cut-the-knot.org/Curriculum/Combinatorics/ThreeOrThree.shtml`

mit einem hübschen Applet.

Satz *Man betrachte den vollständigen Graphen K_6, also einen Graphen, der sechs Ecken besitzt und bei dem jedes Paar von zwei Ecken eine Kante ist. Die 15 Kanten*

in K_6 seien mit den Farben rot oder blau gefärbt. Dann existiert in dem so gefärbten Graphen ein rotes oder ein blaues Dreieck.

Beweis: Man wähle eine der sechs Ecken und nenne sie P. Fünf Kanten verlassen P. Sie sind rot oder blau gefärbt. Mindestens drei von ihnen müssen dieselbe Farbe (rot oder blau) haben[21]. Seien dies PA, PB und PC, ihre Farbe sei etwa blau. Ist eine der Kanten AB, BC oder CA blau gefärbt, so bildet diese Kante zusammen mit den zwei Kanten von P zu den Ecken der Kante ein blaues Dreieck. Ist keine der Kanten AB, BC, CA blau, so sind sie alle drei rot und wir haben ein rotes Dreieck, nämlich $\triangle ABC$. Der Satz ist bewiesen. □

In einer Party von fünf Personen gibt es nicht notwendig drei Personen, die befreundet oder sich fremd sind. Als Beispiel betrachte man fünf Personen, die an einem runden Tisch sitzen, und jede Person nur mit ihren unmittelbaren Nachbarn zur Linken und Rechten befreundet ist. Andererseits gibt es auf jeder Party zwei Gäste, die gleich viele Freunde haben. Denn angenommen, die Party besteht aus n Gästen. Dann kann es nicht gleichzeitig jemanden geben, der mit niemandem befreundet ist, und jemand anders, der mit allen Gästen befreundet ist. Bei n Gästen gibt es also nur $n-1$ Freundschaften. Daher muss es zwei Gäste mit gleich vielen Freunden geben.

15 Das Königsberger Brückenproblem

Bei Wikipedia findet man: Das Königsberger Brückenproblem ist eine mathematische Fragestellung des frühen 18. Jahrhunderts, die anhand von sieben Brücken der Stadt Königsberg illustriert wurde. Das Problem bestand darin, zu klären, ob es einen Weg gibt, bei dem man alle sieben Brücken über den Pregel genau einmal überquert, und wenn ja, ob auch ein Rundweg möglich ist, bei dem man wieder zum Ausgangspunkt gelangt. Wie Leonhard Euler 1736 bewies, war ein solcher Weg bzw. "Eulerscher Weg" in Königsberg nicht möglich, da zu allen vier Ufergebieten bzw. Inseln eine ungerade Zahl von Brücken führte. Es dürfte maximal zwei Ufer (Ecken) mit einer ungeraden Zahl von angeschlossenen Brücken (Kanten) geben. Diese zwei Ufer könnten Ausgangs- bzw. Endpunkt sein. Die restlichen Ufer müssten eine gerade Anzahl von Brücken haben, um sie auch wieder verlassen zu können. In Abbildung 12 links (ebenfalls aus Wikipedia) wird der Pregel mit seinen sieben Brücken dargestellt. In Abbildung 12 rechts findet man den zugehörigen Graphen: Die vier Ufer sind durch vier Ecken gegeben, die je zwei Ufer verbindenden sieben Brücken durch Kanten. Zu den vier Ecken bzw. Ufern ist der zugehörige *Grad* angegegeben, d. h. die Anzahl der Kanten bzw. Brücken, die von dieser Ecke bzw. Ufer ausgehen. Die vom Pregel umflossene Insel ist z. B. eine Ecke mit dem Grad 5, die drei weiteren Ecken bzw. Ufer haben den Grad drei.

Das Lehrbuch *Einführung in die Kombinatorik* von H.Heise (Hanser, München 1976 und Akademie, Berlin 1977), offensichtlich noch mit einer Schreibmaschine mit eventuell nach unten verrutschenden Buchstaben geschrieben, wurde aus den Bibliotheken der DDR entfernt, als man in dem folgenden Abschnitt über das Königsberger Brückenproblem (Königsberg heißt jetzt Kaliningrad und gehörte 1976 zur Sowjetunion und

[21] Siehe die Socken-Fangfrage in Abschnitt 5.

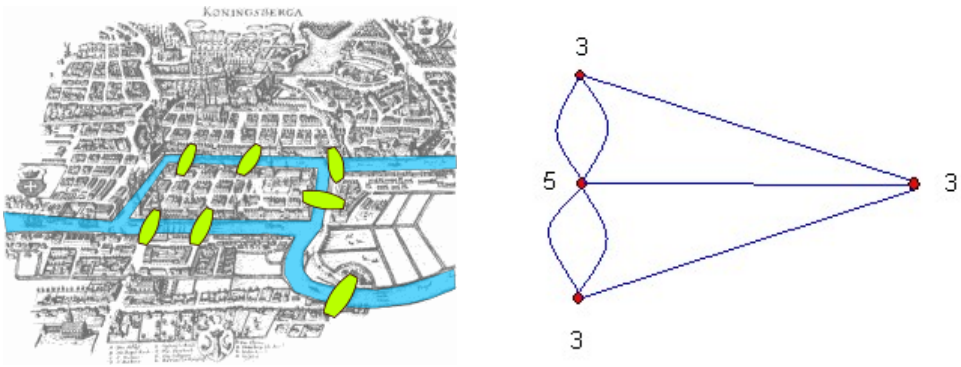


Abbildung 12: Das Königsberger Brückenproblem

jetzt zu Russland) die folgende Passage genauer las:

- I$_n$ Königsberg $_i$. Pr. gab$_e$lt sich $_{der}$ Pregel und u$_m$fließt eine Insel, d$_i$e Kneiphof heiß$_t$. In den $_d$reißiger Jahren d$_e$s achtzehnten Jahrhunderts wurde das Proble$_m$ gestellt, ob e$_s$ w$_o$hl möglich $_w$äre, in einem Spaziergang $_j$ede der sieben Königsberger Brück$_e$n genau einmal zu überschrei$_t$en. Daß ein solcher Spaz$_i$ergang unmöglich ist, war für L.EULER der Anlaß, $_m$it seiner anno 1735 der Akademie der Wissenschaften in St. Petersburg vorgelegten Abhandlung Solutio $_p$robl$_e$matis ad geomet$_{ria}$m situs pertinentis (Commentarii Academiae Petropo$_{li}$tanae 8 (1741) 128-140) einen der er$_s$ten Beiträge zur Topologie zu liefern. Das Proble$_m$ besteht darin, im nachfolgend gezeichneten Graphen einen einfachen Kantenzug zu finden, der alle Kanten enthält. Dabei repräsentiert die Ecke vom Grad 5 den Kneiphof $_u$nd die beiden Ecken vom Grad 2 die Krämerbrücke $_s$owie die Grüne Brücke.

Setzt man die verrutschten (wir haben dies hier nur etwas übertrieben nachvollziehen können) Buchstaben aneinander, so erhält man die Botschaft[22]: Nieder mit dem Sowjetimperialismus.

Nun wollen wir etwas genauer werden. Ein *Graph* $G = (V, E)$ besteht aus einer endlichen Menge $V = V(G)$, der Menge der *Ecken*, und einer Teilmenge $E = E(G)$ von (ungeordneten) Paaren aus V, der Menge der *Kanten*. Ein Graph $G = (V, E)$ heißt ein *Eulerscher Graph*, wenn es in ihm einen *Euler-Zug* gibt, d. h. einen geschlossenen Kantenzug[23], der jede Kante genau einmal enthält. Ein nicht notwendig geschlossener Kantenzug, der jede Kante im Graphen genau einmal enthält, heißt ein *offener Euler-Zug*. Ein Graph, in dem es einen offenen Euler-Zug gibt, heißt ein *semi-Eulerscher Graph*. Unter einem *Weg* in einem Graphen $G = (V, E)$ versteht man eine Folge von paarweise verschiedenen Ecken $x_1, \ldots, x_n \in V$, wobei $(x_i, x_{i+1}) \in E$, $i = 1, \ldots, n-1$. Dieser heißt ein geschlossener Weg oder ein *Kreis*, wenn darüberhinaus $(x_n, x_1) \in E$.

[22]Ein weiteres bekanntes Beispiel für eine versteckte Botschaft findet sich bei E. Isaacson, H. B. Keller (1966). Setzt man die Anfangsbuchstaben aufeinanderfolgender Sätze des Vorworts hintereinander, so erhält man die Botschaft: *Down with computers and their lackeys*.

[23]Ein *Kantenzug* ist eine Folge von paarweise verschiedenen Kanten, wobei aufeinanderfolgende Kanten eine gemeinsame Ecke haben.

Lemma *Die Kantenmenge eines Graphen kann genau dann in Kreise partitioniert werden, wenn jede Ecke geraden Grad besitzt. Insbesondere gilt: Hat jede Ecke in einem Graphen geraden Grad, so ist jede Ecke in einem Kreis enthalten.*

Beweis: Zunächst nehmen wir an, die Kantenmenge eines Graphen sei die disjunkte Vereinigung von Kreisen. Eine beliebige Ecke des Graphen ist entweder isoliert (dann hat sie den Grad 0) oder es führen Kanten zu ihr hin und von ihr fort. Ist die Ecke also in k Kreisen enthalten, so hat sie die Ordnung $2k$, und diese ist gerade.

Umgekehrt nehmen wir nun an, jede Ecke des Graphen habe geraden Grad. Außerdem können wir natürlich annehmen, dass nicht alle Ecken isoliert sind, es also überhaupt Kanten im Graphen gibt. Wie kann man einen Kreis im Graphen finden? Sei $x_0x_1 \dots x_l$ ein Weg maximaler Länge l (gleich Anzahl der Kanten) im Graphen. Da x_0 nicht isoliert ist und geraden Grad besitzt, hat x_0 außer x_1 noch einen weiteren Nachbarn, etwa y. Es ist $y = x_i$ für ein $i \in \{2, \dots, l\}$, denn andernfalls wäre $yx_0 \dots x_l$ ein Weg der Länge $l + 1$. Damit hat man wenigstens einen Kreis gefunden, nämlich $x_0x_1 \dots x_i$. Nun entferne man aus dem Graphen alle Kanten des gerade gefundenen Kreises und wende auf den so entstandenen Graphen (in diesem haben wieder alle Ecken geraden Grad) dieselbe Argumentation an. Nach endlich vielen Schritten ist die Behauptung bewiesen. □

Im folgenden Satz werden Eulersche Graphen charakterisiert.

Satz *Ein Graph $G = (V, E)$ mit keinen isolierten Ecken*[24] *und $E \neq \emptyset$ ist genau dann Eulersch, wenn G zusammenhängend (d. h. zu je zwei Ecken gibt es einen Kantenzug in G) ist und alle Ecken geraden Grad haben.*

Beweis: Sei G ein Eulerscher Graph ohne isolierte Ecken. Dann ist G zusammenhängend, denn in einem Euler-Zug kommt jede Kante und daher jede Ecke vor. Da der Euler-Zug jede Ecke über eine gewisse Kante erreicht und über eine andere wieder verlässt, hat jede Ecke geraden Grad.

Umgekehrt nehmen wir an, G sei zusammenhängend und jede Ecke habe geraden Grad. Nach dem vorigen Lemma gibt es in G wenigstens einen Kreis und damit einen geschlossenen Kantenzug. Sei C ein geschlossener Kantenzug mit einer maximalen Anzahl von Kanten, diese seien in $E(C) \subset E$ zusammengefasst. Ist $E = E(C)$, so ist man fertig, denn dann ist C ein (geschlossener) Euler-Zug. Andernfalls sei $G' := (V, E \setminus E(C))$. Da G zusammenhängend ist, gibt es eine Ecke u auf C, die mit einer Kante aus G' inzidiert. In G' hat wieder jede Ecke geraden Grad. Daher gibt es einen geschlossenen Kantenzug C' in G', der u enthält. Schiebt man C' beim Durchlaufen von C an der Stelle u ein, so erhält man einen Widerspruch zur Maximalität von C. □

Im nächsten Satz werden Bedingungen dafür angegeben, dass ein offener Euler-Zug von einer Ecke x zu einer Ecke $y \neq x$ existiert.

Satz *Sei $G = (V, E)$ ein zusammenhängender Graph. Dann existiert genau dann von der Ecke $x \in V$ ein offener Euler-Zug zu der Ecke $y \in V \setminus \{x\}$, wenn x und y die einzigen Ecken in G mit ungeradem Grad sind.*

Beweis: Seien x und y die beiden einzigen Ecken im Graphen mit ungeradem Grad. Man definiere $G^* := (V \cup \{u\}, E \cup \{ux\} \cup \{uy\})$, wobei $u \notin V$. Dann ist G^* zusam-

[24] Eine Ecke $x \in V$ heißt *isoliert*, wenn sie mit keiner Kante *inzidiert*, also $(x, y) \notin E$ für alle $y \in V$.

menhängend, ferner besitzen alle Ecken in G^* (also auch x, y und u) geraden Grad. Daher gibt es wegen des gerade ebenen bewiesenen Satzes in G^* einen Euler-Zug C^*. Lässt man in diesem Euler-Zug die Ecke u und die inzidierenden Kanten ux und uy fort, so erhält man einen offenen Euler-Zug von x nach y. Umgekehrt gebe es einen offenen Euler-Zug von x nach $y \neq x$. Da in jeder anderen Ecke außer x, y ebenso viele Kanten anfangen wie enden, ist der Grad einer jeden solchen Ecke gerade. Aus dem entsprechenden Grund ist der Grad von x und y ungerade. □

Beispiel: Schon als Kind (siehe auch `http://de.wikipedia.org/wiki/Haus_vom_Nikolaus`) versuchte man, "das Haus des Nikolaus" in einem Zug zu zeichnen, siehe Abbildung 13. Bekanntlich ist dies möglich, etwa durch den Kantenzug

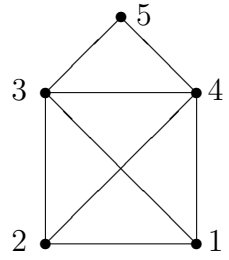


Abbildung 13: Das ist das Haus des Nikolaus

$$12, 23, 34, 45, 53, 31, 14, 42.$$

Klar ist jedenfalls, dass man in den Ecken 1 oder 2 mit dem Kantenzug beginnen muss, da dies die einzigen Ecken mit ungeradem Grad sind. Dagegen ist das in Abbildung 14 angegebene erweiterte Haus des Nikolaus ein Eulerscher Graph. Denn der Grad jeder

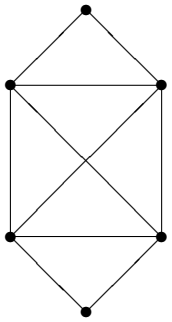


Abbildung 14: Das (erweiterte) Haus des Nikolaus

Ecke in dem Graphen ist 2 oder 4. □

16 Das Rösselsprungproblem und Hamilton-Kreise

Bei dem *Rösselsprungproblem* oder auch Springerproblem (engl.: knight's tour problem), siehe auch `http://de.wikipedia.org/wiki/Springerproblem` und `http://www.mayhematics.com/t/ktn.htm`, geht es um folgendes: Mit dem Springer sollen alle n^2 Felder eines $n \times n$-Bretts genau einmal in einem kontinuierlichen Zug erreicht

und zum Ausgangsfeld zurückgekehrt werden. Die n^2 Felder des Schachbretts seien die Ecken eines Graphen. Weiter sind zwei Ecken genau dann durch eine Kante verbunden, wenn zwischen den entsprechenden Feldern ein Rösselsprung möglich ist. Das Rösselsprungproblem ist dann offenbar äquivalent dazu, in dem definierten Graphen einen Hamilton-Kreis[25] zu bestimmen oder zu entscheiden, dass es so einen nicht gibt. Dies ist z. B. für $n = 4$ der Fall. Um dies einzusehen beachte man: Die Ecke "links oben" kann man nur von zwei Ecken aus erreichen. Die Ecke "rechts unten" kann nur von diesen zwei Ecken erreicht werden. Hierdurch hat man aber schon einen Kreis. Diese Ecken können nicht Teil eines Hamilton-Kreises sein, siehe Abbildung 15. Bei L. Volkmann (1991) ist eine auf Euler (1759) zurückgehende Lösung für $n := 8$ angegeben, die wir in Tabelle 1 links reproduzieren: In Abbildung 16 ist diese Lösung als Hamilton-

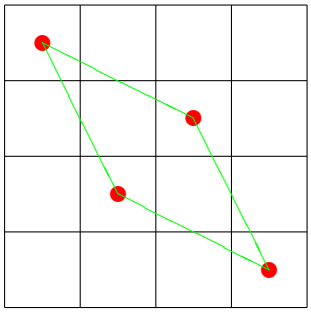


Abbildung 15: Das Rösselsprung-Problem für $n = 4$

58	43	60	37	52	41	62	35
49	46	57	42	61	36	53	40
44	59	48	51	38	55	34	63
47	50	45	56	33	64	39	54
22	7	32	1	24	13	18	15
31	2	23	6	19	16	27	12
8	21	4	29	10	25	14	17
3	30	9	20	5	28	11	26

63	22	15	40	1	42	59	18
14	39	64	21	60	17	2	43
37	62	23	16	41	4	19	58
24	13	38	61	20	57	44	3
11	36	25	52	29	46	5	56
26	51	12	33	8	55	30	45
35	10	49	28	53	32	47	6
50	27	34	9	48	7	54	31

Tabelle 1: Zwei Lösungen des Rösselsprungproblems

Kreis in einem Graphen veranschaulicht. Noch erstaunlicher ist ein Springerkreis, der gleichzeitig ein semi-magisches Quadrat[26] ist (alle Zeilen- und Spaltensummen ergeben 260, nicht aber die Diagonalsummen). Man findet ihn bei M. Löbbing, I. Wegener (1996). Er ist in Tabelle 1 rechts angegeben. Bei Löbbing-Wegener findet man weitere

[25] Unter einem *Weg* in einem Graphen $G = (V, E)$ versteht man eine Folge paarweise verschiedener Ecken $x_1, \ldots, x_n \in V$, wobei $(x_i, x_{i+1}) \in E$, $i = 1, \ldots, n-1$. Dieser heißt ein *geschlossener Weg* oder ein *Kreis*, wenn darüberhinaus $(x_n, x_1 \in E$. Ein Kreis, der jede Ecke in einem Graphen genau einmal enthält, heißt ein *Hamilton-Kreis*. Wenn in einem Graphen ein Hamilton-Kreis existiert, so heißt er ein Hamilton-Graph.

[26] Auf magische Quadrate gehen wir in Abschnitt 64 näher ein.

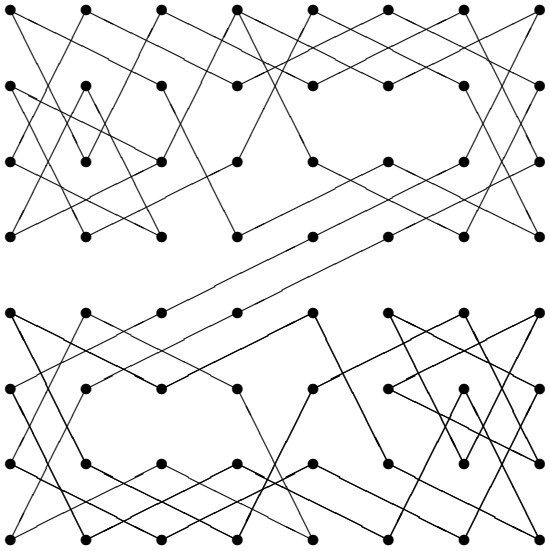


Abbildung 16: Eine Lösung des Rösselsprungproblems

interessante Bemerkungen zum Problem der Springerkreise, z. B. dass dieses in einer Folge von "Wetten, daß...?" schon eine Rolle spielte. Man kann dort nachlesen:

- Der Kandidat behauptete, dass er für ein beliebiges Feld den Springer so über das Schachbrett bewegen kann, dass jedes Feld genau einmal erreicht wird. Und der Kandidat hat die Wette gewonnen. Diese Leistung hat bei vielen Menschen Bewunderung hervorgerufen. Immerhin gibt es 64 Startfelder, und 64 Springerwege enthalten 4032 Springerzüge. Diese Vorgehensweise ist jedoch sträflich naiv. Es genügt doch, einen Springerkreis auswendig zu lernen, da jeder Springerkreis für jedes Startfeld einen Springerweg enthält. Diese Gedächtnisleisung wird noch geringer, wenn ein Springerkreis mit vielen Symmetrien gewählt wird.

Bemerkt sei, dass noch einmal am 22. Februar 2003 eine solche Wette in der Sendung "Wetten, dass..." vorkam und ein neunjähriger Junge mit verbundenen Augen von einer vorgegebenen Startecke einen Hamilton-Kreis angeben konnte.

Im Jahre 1859 stellte Sir William Hamilton das Problem, ob der in Abbildung 17 dargestellte Graph ein Hamilton-Graph ist (natürlich nannte er selber ihn so nicht). Wie man bei J. M. Aldous, R. J. Wilson (2000, S. 71) nachlesen kann, machte Hamilton aus dem Problem ein Spiel, in dem der Spieler einen Hamilton-Kreis zu finden hatte, wenn der Anfang des Kreises vorgegeben ist. Einen Hamilton-Kreis geben wir in Abbildung 18 an.

17 Witze über MathematikerInnen

Die folgende Geschichte, nicht eigentlich ein Witz, kenne ich von meinem Lehrer Lothar Collatz. Man stelle sich eine Tanzstunde vor, bei der die Herren Mathematiker

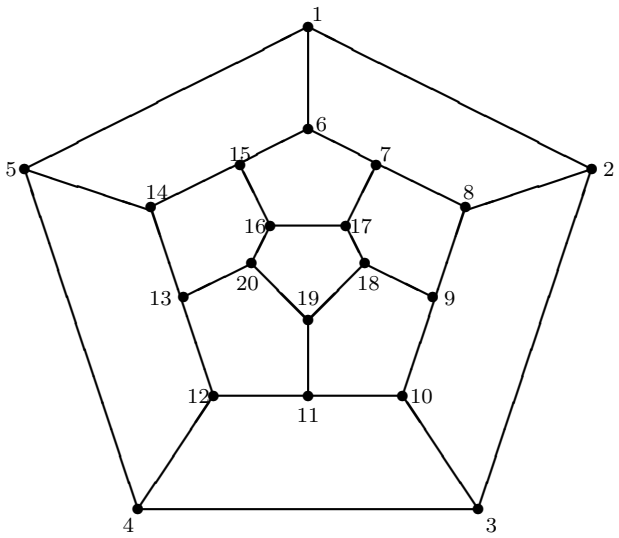


Abbildung 17: Ist dieser Graph Hamiltonsch?

sind, und zwar zur Hälfte reine, zur anderen Hälfte angewandte Mathematiker. Der Tanzlehrer bittet die Damen an die eine Seite des zehn Meter langen Saales, die Herren an die andere Seite. Er hat ein Tamburin in der Hand und fordert die Herren auf, bei jedem Schlag auf das Tamburin die Entfernung zu den Damen zu halbieren. Daraufhin verlassen die reinen Mathematiker die Tanzstunde, während die angewandten Mathematiker sich sagen: „Siebenmal auf das Tamburin geschlagen und das Ergebnis ist für (fast) alle praktischen Zwecke ausreichend."

Es gibt unzählige Witze über Mathematiker im Netz, darunter viele, über die zumindestens ich nicht lachen kann. Ich führe hier einige an, die ich ganz lustig finde. Man findet sie im Internet in vielen fast identischen Varianten. Daher gebe ich keine genauen Quellen an. Erwähnt seien nur eine Seite aus Kanada, nämlich `http://www.math.ualberta.ca/~runde/jokes.html`, sowie ein Aufsatz von P. Renteln, A. Dundes (2005), den man sich unter `http://www.ams.org/notices/200501/index.html` herunterladen kann.

- Ein Mathematiker, ein Physiker und ein Biologe sitzen im Zug und fahren durch Schottland. Während der Fahrt sehen sie auf einer Wiese ein schwarzes Schaf, worauf der Biologe meint: „Ah, ich sehe, dass die schottischen Schafe schwarz sind.'" Der Physiker sagt: „Du meinst wohl, dass manche schottischen Schafe schwarz sind." Darauf der Mathematiker: „Nein, wir wissen lediglich, dass es in Schottland mindestens ein Schaf gibt, und dass wenigstens eine Seite dieses Schafes schwarz ist."

- Zwei Mathematiker in einer Bar: Einer sagt zum anderen, dass der Durchschnittsbürger nur wenig Ahnung von Mathematik hat. Der zweite ist damit nicht einverstanden und meint, dass doch ein gewisses Grundwissen vorhanden ist. Als der erste mal kurz austreten muss, ruft der zweite die blonde Kellnerin, und meint, dass er sie in ein paar Minuten, wenn sein Freund zurück ist, etwas fragen wird, und sie möge doch bitte auf diese Frage mit 'ein Drittel x hoch drei' antwor-

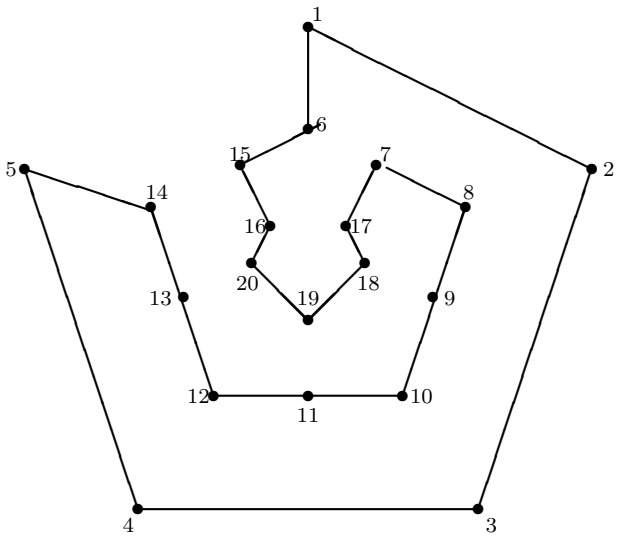


Abbildung 18: Ein Hamilton-Kreis zum Graphen in Abbildung 17

ten. Etwas unsicher bejaht die Kellnerin und wiederholt im Weggehen mehrmals: „Ein Drittel x hoch drei...“ Der Freund kommt zurück, und der andere meint: „Ich werde Dir zeigen, dass die meisten Menschen doch etwas von Mathematik verstehen. Ich frage jetzt die blonde Kellnerin da, was das Integral von x zum Quadrat ist.“ Der zweite lacht bloß und ist einverstanden. Also wird die Kellnerin gerufen und gefragt, was das Integral von x zum Quadrat sei. Diese antwortet: „Ein Drittel x hoch drei.“ Und im Weggehen dreht sie sich noch einmal um und meint: „Plus constant.“

- Ein Arzt, ein Rechtsanwalt und ein Mathematiker diskutieren darüber, was besser sei: Eine Freundin zu haben oder verheiratet zu sein. Der Arzt: Es ist besser verheiratet zu sein, um ein Gefühl der inneren Sicherheit zu haben. Das senkt den Blutdruck und ist somit gut für die Gesundheit! Der Anwalt: Es ist besser eine Freundin zu haben. Wenn man verheiratet ist und sie die Scheidung will, bringt das nur unnötige Schwierigkeiten! Der Mathematiker: Das beste ist, man hat beides! Denn wenn die Frau denkt, man sei bei der Freundin und die Freundin meint, man wäre bei der Frau, hat man genug Zeit für Mathematik[27].

- Zwei Männer fliegen in einem Heißluftballon. Es kommt Nebel auf, und sie können nichts mehr sehen. Nach einer Zeit lichtet sich der Nebel wieder, aber sie wissen

[27]Der Göttinger Mathematikprofessor Abraham Gotthelf Kästner hat 1756 in seiner Antrittsrede hierzu folgendes gesagt: „Die Beschäftgung mit der Mathematik ist in [...] höherem Maße dazu geeignet, heftigere Gemütsregungen zu besänftigen; denn sie kann nur bei ausgeglichenem Gemütszustand erfolgreich betrieben werden. [...] [Ich habe] immer geglaubt, daß jene reinste Lust, die bei wissenschaftlicher Arbeit entsteht, viel dazu beitrage, die sinnlichen Begierden wenigstens einzuschränken und die Seele von Lastern zu läutern. [...] Denjenigen [...], die um ihre Forschungen bemüht sind, bleibt keine Zeit übrig, in der sie ihren Lastern frönen können.“ Was Hofrat Behrens in Thomas Manns Zauberberg hierzu zu sagen weiß, haben wir auf Seite 19 angegeben. Andererseits gibt es auch das auf den sogenannten Irrenarzt Paul Möbius (1853–1907) zurückgehende Zitat „Die Mathematik ist dem Liebestrieb nicht abträglich“

nicht mehr, wo sie sind. Da sehen sie unter sich einen Mann, der im Garten arbeitet. Der eine ruft runter: „Wo sind wir hier?“ Es kommt keine Antwort. Endlich, als sie schon fast außer Hörweite sind, ruft der Mann hoch: „Sie sind in einer Gondel unterhalb eines einem Ballons.“ Daraufhin meint der eine Mann: „Das war bestimmt ein Mathematiker.“ Fragt der andere: „Wie kommst Du denn darauf?“ Antwortet der erste: „Das ist doch ganz klar. Erstens hat es ewig gedauert, bis eine Antwort kam. Zweitens war die Antwort richtig. Und drittens war sie zu überhaupt nichts zu gebrauchen.“

- Eine Gruppe von Ingenieuren und eine Gruppe von Mathematikern fahren mit dem Zug zu einer Tagung. Jeder der Ingenieure hat seine eigene Fahrkarte, aber die ganze Gruppe von Mathematikern hat nur eine einzige Karte. Plötzlich ruft einer der Mathematiker: „Der Schaffner kommt!“, worauf sich alle Mathematiker in eine der Toiletten zwängen. Der Schaffner kommt, kontrolliert die Ingenieure, sieht, dass das WC besetzt ist und klopft an die Tür: „Die Fahrkarte bitte!“. Einer der Mathematiker schiebt die Fahrkarte unter der Tür durch und der Schaffner zieht zufrieden wieder ab. Auf der Rückfahrt beschließen die Ingenieure, denselben Trick anzuwenden und sie kaufen nur eine Karte für die ganze Gruppe. Sie sind sehr verwundert als sie merken, dass die Mathematiker diesmal überhaupt keine Fahrkarte haben. Wieder ruft einer der Mathematiker : „Der Schaffner kommt!“. Sofort stürzen die Ingenieure auf das eine WC, die Mathematiker machen sich etwas gemächlicher auf den Weg zum anderen. Bevor der letzte Mathematiker die Toilette betritt, klopft er bei den Ingenieuren an: „Fahrkarte bitte!“

  Und die Moral der Geschichte: Man sollte keine Methoden anwenden, deren Sinn man nicht verstanden hat.

- Wie kann man einen extrovertierten Mathematiker von einem introvertierten unterscheiden? Der extrovertierte Mathematiker schaut auf *Deine* Füße, wenn er mit Dir spricht.

- Ein Mann ist mit einer Mathematikerin verheiratet. Er kommt nach Hause, schenkt seiner Frau einen großen Strauß Rosen und sagt: „Ich liebe Dich!“ Sie nimmt die Rosen, haut sie ihm um die Ohren, gibt ihm einen Tritt und wirft ihn aus der Wohnung. Was hat er falsch gemacht? Er hätte sagen müssen: „Ich liebe Dich und nur Dich!“

- Es gibt drei Arten von Mathematikern: Die, die bis drei zählen können und die, die es nicht können[28]

- Ein Mathematiker macht Urlaub in der Heide. An einem Tag trifft er einen Schäfer mit einer großen Schafherde. Er denkt, eines dieser kuscheligen Wolltiere würde für ihn und seine Familie ein wunderbares Haustier sein. Er fragt den Schäfer, was

[28] Ähnlich komisch ist der Ausspruch, der von Horst Hrubesch, dem früheren Kopfballungeheuer, stammen soll. „Ich sage nur ein Wort: Vielen Dank“. Auch die Aussprüche „Fremdwörter sind bei uns nicht usus“ und „Ruthchen, Du sollst nicht fluchen, verdammt noch mal!“ fallen in diese Kategorie.

eines der Schafe kosten würde. „Die Schafe sind nicht zu verkaufen“ antwortet der Schäfer. Der Mathematiker überlegt eine Weile und sagt dann: „Ich sage Ihnen die genaue Zahl der Schafe in Ihrer Herde ohne zu zählen. Erhalte ich dann ein Schaf?“ Nachdem der Schäfer das bejaht sagt der Mathematiker: „Die Zahl der Schafe ist 387.“ Nach einer Weile sagt der Schäfer: „Das ist richtig. Ich verliere nicht gerne eines meiner Schafe. Aber ich versprach es. Eines von ihnen gehört Ihnen!“ Der Mathematiker greift eines der Tiere und ist dabei wegzugehen, als der Schäfer sagt: „Warten Sie noch einen Moment! Ich sage Ihnen welchen Beruf Sie haben und wenn ich recht habe, erhalte ich das Tier zurück.“ Darauf der Mathematiker: „Ja, das ist fair.“ Der Schäfer: „Sie sind Mathematiker.“ Der Mathematiker ist sehr erstaunt: „Sie haben recht. Wie konnten Sie das wissen?“ Der Schäfer: „Das ist einfach. Sie sagten mir die genaue Zahl der Schafe ohne zu zählen und dann griffen Sie meinen Hund.“

Den in Abbildung 19 wiedergegebenen Cartoon habe ich bei

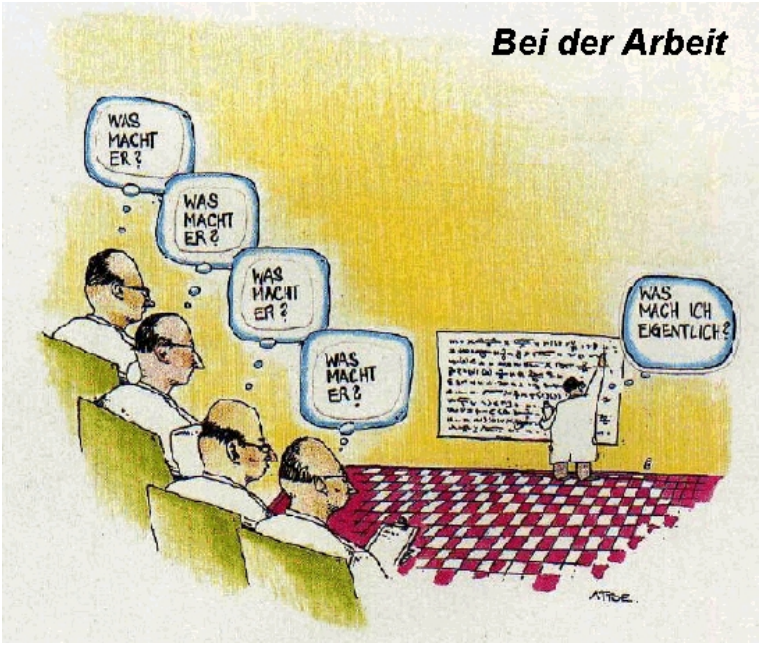


Abbildung 19: Wer kennt diese Situation?

`http://www.mathe.tu-freiberg.de/inst/theomath/Unterhaltsames.html`

gefunden.

18 Wahre Geschichten, über die nur Mathematiker-Innen lachen können

In dem Warteraum eines Arztes fand ich den folgenden Anschlag:

- Liebe Patienten!

  Aufgrund begrenzter Kapazität im Wartezimmer bitten wir Sie, erforderliche Begleitpersonen auf ein Minimum zu beschränken.

  Wir danken für Ihr Verständnis.

  Ihr Praxisteam.

Aus dem Rheinischen Merkur vom 10.06.2010:

- Schon seit geraumer Zeit ist in Berlin zu beobachten, dass das allgemeine Lächerlichmachen des politischen Gegners keine Grenzen mehr kennt. SPD-Parteichef Sigmar Gabriel hält es für nötig, öffentlich die Bundeskanzlerin als „Mutti“ zu verspotten, die in der Waschmaschine den Schongang für Vermögende einlegt. Selbst Koalitionäre diffamieren sich gegenseitig als „Wildsau“ und „Gurkentruppe“. Die Parlamentarier merken offenbar überhaupt nicht, wie sie durch die verbale Herabwürdigung des Mitbewerbers die Abneigung des Volkes gegenüber seinen Vertretern erst befeuern. Wer nicht ein **Minimum an Anstand**, Höflichkeit und Etikette wahrt, braucht sich nicht wundern, wenn politische Sitten insgesamt verlottern und der Graben zwischen Bürgern und Politik unüberbrückbar wird.

Gibt man bei Google “Minimum an Anstand” ein, so erhält man immerhin 183000 Ergebnisse. Darunter die Aussage in der Basler Zeitung:

- Wichtig ist aber auch, dass die politischen Auseinandersetzungen mit einem Minimum an Anstand geführt werden und dass auf die Argumente der Gegenseite eingegangen wird.

Weshalb eigentlich nicht mit einem Maximum an Anstand? Gelegentlich hört man (z.B. von einem Politiker über einen anderen) die Aussage: „Sie haben noch nicht einmal ein Minimum an Anstand!“ Mathematisch korrekter wäre: „Das Infimum an Anstand wird von Ihnen angenommen, Sie haben daher ein Minimum an Anstand!“ Es ist aber gut möglich, dass diese Aussage als ein Kompliment aufgefasst wird.

Das folgende Gespräch überhörte ich in einer Bäckerei.

- Eine Kundin sagt zu einer Verkäuferin vor einem wohlgefüllten Brotregal: „Ich hätte gerne das kleinste Brot, das Sie haben.“ Die Verkäuferin darauf: „Das tut mir leid, das habe ich gerade verkauft.“ Die Kundin antwortet: „Schade. Dann nehme ich zwei Brötchen.“

Ich habe es leider versäumt, einen Existenzbeweis in der Praxis zu führen.

19 Die Eulersche Polyederformel

Die Eulersche Polyederformel gibt den Zusammenhang zwischen der Anzahl der Ecken, der Anzahl der Flächen und der Anzahl der Kanten eines beschränkten, konvexen Polyeders im $\mathbb{R}^3$ bzw., und hierauf werden wir uns beschränken, eines *planaren* Graphen. Grob gesagt verstehen wir unter einem planaren Graphen einen Graphen, den man

ohne Kantenüberschneidungen in die Ebene einbetten (bzw. in der Ebene zeichnen) kann bzw. einen hierzu *isomorphen* Graphen. Um dies zu präzisieren benötigen wir einige Definitionen.

Definition Zwei Graphen $G = (V, E)$ und $G' = (V', E')$ heißen *isomorph*, wenn es eine umkehrbar eindeutige Abbildung ϕ von V auf V' gibt derart, dass $(x, y) \in E$ genau dann, wenn $(\phi(x), \phi(y)) \in E'$.

Beispiel: Man betrachte die beiden Graphen in Abbildung 20. Die Isomorphie vermit-

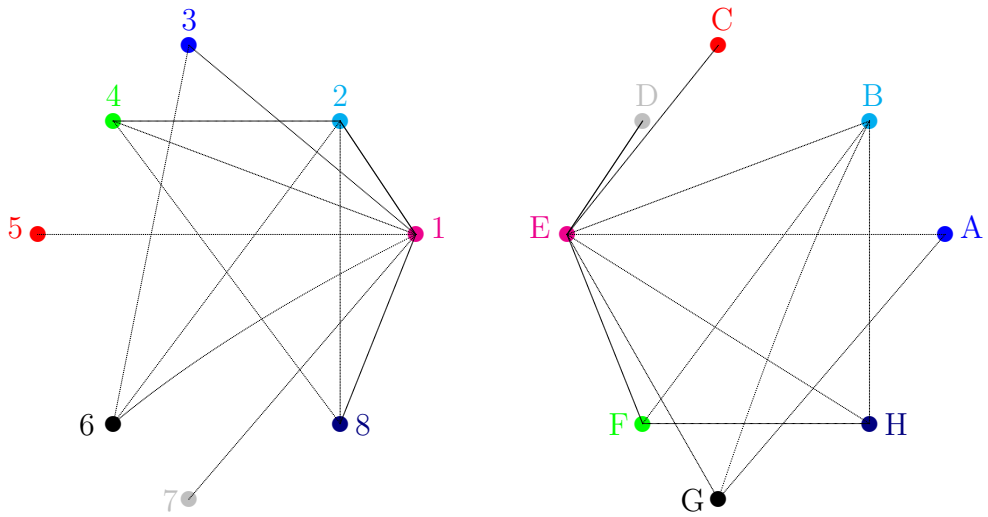


Abbildung 20: Zwei isomorphe Graphen

telnde Abbildung ϕ liest man aus

$$\phi : \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline E & B & A & F & C & G & D & H \end{array}$$

ab. □

Definition Eine stetige Abbildung $e\colon [0,1] \longrightarrow \mathbb{R}^2$ ist ein *Jordanbogen* im $\mathbb{R}^2$, wenn $e(t) \neq e(s)$ für alle $s, t \in [0,1]$ mit $s \neq t$. Ein Graph $G = (V, E)$ heißt ein *ebener Graph*, wenn folgende drei Bedingungen erfüllt sind:

1. Es ist $V = \{x_1, \ldots, x_n\} \subset \mathbb{R}^2$.
2. Es ist $E = \{e_1, \ldots, e_m\}$ mit Jordanbögen $e_1, \ldots, e_m$ im $\mathbb{R}^2$, wobei $e_i(0), e_i(1) \in V$ und $e_i(t) \notin V$ für alle $t \in (0,1)$, $i = 1, \ldots, m$. Die Ecke x_i und die Kante e_j heißen *inzident*, wenn $x_i \in \{e_j(0), e_j(1)\}$.
3. Die Kanten von G haben (außer Ecken) keine Schnittpunkte. Sind also e_i und e_j zwei verschiedene Kanten, so ist $e_i(s) \neq e_j(t)$ für alle $s, t \in (0,1)$. Zwei verschiedene Kanten heißen *inzident*, wenn sie mit einer gemeinsamen Ecke inzidieren.

Ist $G' = (V', E')$ ein Graph, der zu einem ebenen Graphen $G = (V, E)$ isomorph ist, so heißt G eine *Einbettung* von G' in den $\mathbb{R}^2$. Ein Graph, der zu einem ebenen Graphen isomorph ist, heißt ein *planarer Graph*.

Beispiel: Zeichnet man den vollständigen Graphen K_4 in der üblichen Weise, so findet eine Überschneidung der Kanten statt, siehe Abbildung 21 links. Rechts geben wir einen

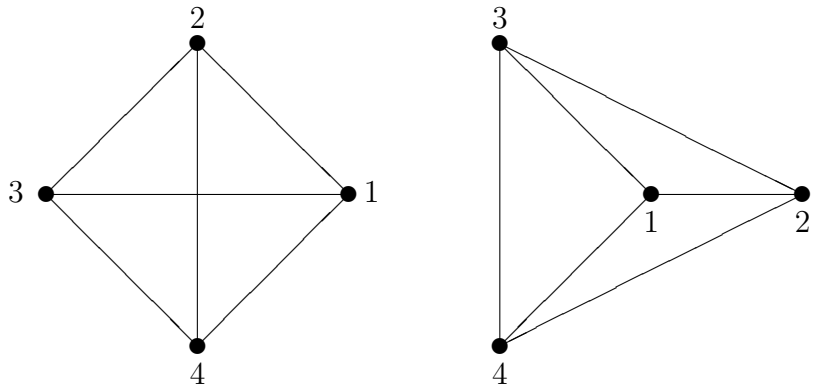


Abbildung 21: K_4 ist planar

isomorphen Graphen an, der K_4 ist also planar. Auch der vollständige bipartite Graph $K_{2,3}$ ist planar (Beweis?), nicht aber $K_{3,3}$ und K_5, wie wir später zeigen werden. Dass der $K_{3,3}$ nicht planar ist, kann folgendermaßen interpretiert werden. Angenommen, drei Nachbarn A, B und C wollen mit Gas, Strom und Wasser versorgt weden. Dann ist es nicht möglich, die Versorgungsleitungen kreuzungsfrei zu verlegen. Siehe Abbildung 22. Zwar kann man einige der Leitungen "außen herum" legen, aber nicht alle. □

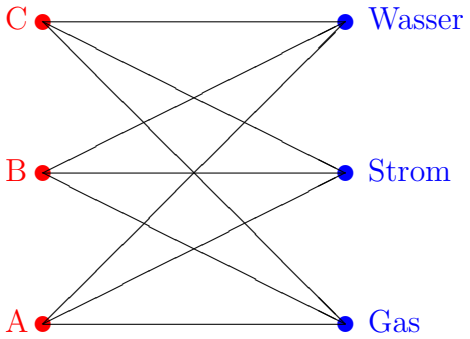


Abbildung 22: $K_{3,3}$ ist nicht planar

Sei $G = (V, E)$ ein zusammenhängender, ebener Graph. Fassen wir G als Vereinigung der Ecken und Kanten auf, so ist $\mathbb{R}^2 \setminus G$ disjunkte Vereinigung zusammenhängender Gebiete, von denen genau eines nicht beschränkt ist. Diese Zusammenhangskomponenten von $\mathbb{R}^2 \setminus G$ nennen wir *Länder*. Die Anzahl der Länder eines planaren Graphen ist natürlich die Anzahl der Länder einer ebenen Einbettung.

Nun definieren wir: Sei $G = (V, E)$ ein zusammenhängender Graph. Eine Kante $e \in E$ heißt eine *Brücke* in G, wenn $G' := (V, E \setminus \{e\})$ nicht zusammenhängend ist.

Die Aussage des nächsten Satzes wird die *Eulersche Polyederformel* genannt.

Eulersche Polyederformel *Sei $G = (V, E)$ ein zusammenhängender, planarer Graph mit n Ecken, m Kanten und f Ländern. Dann ist $n - m + f = 2$.*

Beweis: Wir beweisen die Behauptung durch vollständige Induktion nach m, der Anzahl der Kanten von G. Ist $m = 1$, so ist $n = 2$ und $f = 1$, die Aussage also richtig. Wir nehmen nun an, die Aussage sei für Graphen mit $\leq m - 1$ Kanten richtig. Angenommen, $e \in E$ sei keine Brücke. Nach Definition einer Brücke ist $G' := (V, E \setminus \{e\})$ zusammenhängend und natürlich planar. Dann hat G' genau so viel Ecken wie G, eine Kante und ein Land weniger als G, denn nach dem Entfernen der Grenze e wird aus zwei Ländern eines. Nach Induktionsvoraussetzung ist also

$$n - (m - 1) + (f - 1) = 2$$

bzw. $n - m + f = 2$. Zum Schluss müssen wir noch den Fall betrachten, dass jede Kante von G eine Brücke ist. Dann enthält G keinen Kreis[29] (denn wenn es einen Kreis geben würde, so wäre jede Kante in diesem Kreis keine Brücke). Dann ist aber[30] $m = n - 1$. Ferner ist $f = 1$, da ein zusammenhängender, kreisfreier Graph genau ein (nicht beschränktes) Land besitzt. Also gilt auch hier $n - m + f = 2$. Damit ist der Satz bewiesen. □

Denkt man sich aus einem beschränkten, konvexen Polyeder eine Seitenfläche entfernt und das Ergebnis in die Ebene entfaltet, so erhält man einen zusammenhängenden, planaren Graphen. Das unbeschränkte Land dieses Graphen entspricht der entfernten Fläche im Polyeder. Die Anzahl der Ecken und Kanten im Polyeder und im Graphen stimmen überein, die Anzahl der Flächen im Polyeder ist gleich der Anzahl der Länder im Graphen.

Beispiel: Es gibt bekanntlich fünf reguläre Polyeder: Tetraeder, Würfel, Oktaeder, Ikosaeder und Dodekaeder. Hierbei heißt ein (konvexer) Polyeder regulär, wenn alle seine Oberflächen aus demselben regelmäßigen Vieleck bestehen und in jeder Ecke gleich viele dieser Vielecke zusammenstoßen. In Abbildung 23 geben wir das Dodekaeder und das Ikosaeder wieder. Wir geben für diese fünf Körper die Anzahl der Ecken, Kanten und Flächen an:

	Ecken	Kanten	Flächen
Tetraeder	4	6	4
Würfel	8	12	6
Oktaeder	6	12	8
Ikosaeder	12	30	20
Dodekaeder	20	30	12

[29] Unter einem *Kreis* in einem Graphen verstehen wir einen *geschlossenen Weg*, also eine Folge von paarweise verschiedenen Ecken $x_1, \ldots, x_k \in V$, wobei $(x_i, x_{i+1} \in E$, $i = 1, \ldots, k-1$, und $(x_k, x_1) \in E$.

[30] **Satz** *Sei $G = (V, E)$ mit $n := |V|$ und $m := |E|$ ein zusammenhängender Graph, der keinen Kreis enthält. Dann ist $m = n - 1$.*

Beweis: Wir überlegen uns zunächst, dass es in G Ecken vom Grad 1 gibt. Ist nämlich $P = x, x_1, \ldots, y$ ein längster Weg in G, so ist haben x (und y) nur einen Nachbarn, d. h. es ist $d(x) = d(y) = 1$, denn andernfalls könnte der Weg verlängert werden, da G keinen Kreis enthält. Man setze $G_1 := (V_1, E_1)$ mit $V_1 := V \setminus \|x\}$, $E_1 := E \setminus \{(x, x_1)\}$. Dann ist auch G_1 ein zusammenhängender, kreisfreier Graph und $|V_1| - |E_1| = |V| - |E|$. Nach $n - 2$ Schritten erhalten wir einen zusammenhängenden, kreisfreien Graphen $G_{n-2} = (V_{n-2}, E_{n-2})$ mit $|V_{n-2}| = 2$ und $|E_{n-2}| = 1$. Folglich ist

$$n - m = |V| - |E| = |V_{n-2}| - |E_{n-2}| = 2 - 1 = 1.$$

Damit ist der Hilfssatz bewiesen. □

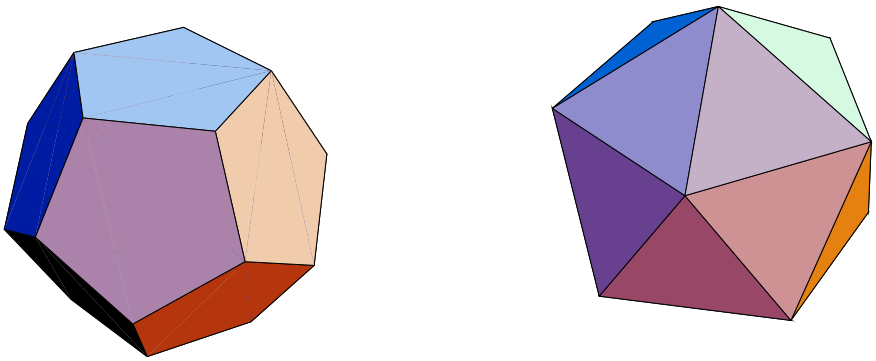


Abbildung 23: Das Dodekaeder und das Ikosaeder

Die Anzahl der Flächen bestimmt den Namen. □

20 Der Fünffarbensatz

Der Fünffarbensatz sagt aus, dass eine Landkarte bzw. ein zusammenhängender, planarer Graph mit fünf Farben so gefärbt werden kann, dass benachbarte Länder unterschiedlich gefärbt sind. Hierbei vereinbaren wir, dass zwei Länder, die nur einen Punkt gemeinsam haben, nicht als benachbart gelten, und Länder zusammenhängend sind, es also keine Exklaven gibt. Dass eine Färbung sogar mit nur vier Farben möglich ist, ist die Aussage des Vierfarbensatzes. Ein Beweis dieser Aussage, erst 1976 von K. Appel und W. Haken mit Hilfe eines Computers erbracht, ist weit außerhalb unserer Reichweite.

Beispiel: In Abbildung 24 geben wir eine Karte mit einem Teil Europas wieder. Genauer haben wir hier 10 Länder eingetragen. Bemerkt sei, dass wir (Entschuldigung!) Andorra, Liechtenstein, Monaco und San Marino fortgelassen haben. In Abbildung 25 geben wir diese Landkarte als einen ebenen Graphen an. Hier treffen sich Italien und Spanien nur in einer Ecke, sind also nicht benachbart. Ferner haben wir darauf geachtet, dass ein Grenzübertritt nur über genau eine Kante möglich ist, dass also keine unnötigen Ecken oder Kanten vorhanden sind[31]. □

Nun führen wir zum ebenen Graphen $G = (V, E)$ den dualen Graphen $G^* = (V^*, E^*)$ ein, und zwar folgendermaßen: Die Ecken von G^* sind die Länder in G, wobei auch das "Außenland" ein Land, also eine Ecke in G^* ist. Zwei Ecken in G^* (bzw. Länder in G)

[31] Dies gilt allerdings nicht für das Seefahrerland Portugal. Hier ist ein Übergang zur Außenwelt über zwei Kanten möglich. Wenn wir Mehrfachkanten zulassen würden (was durchaus möglich ist), hätten wir diese Ausnahmeregelung vermeiden können.

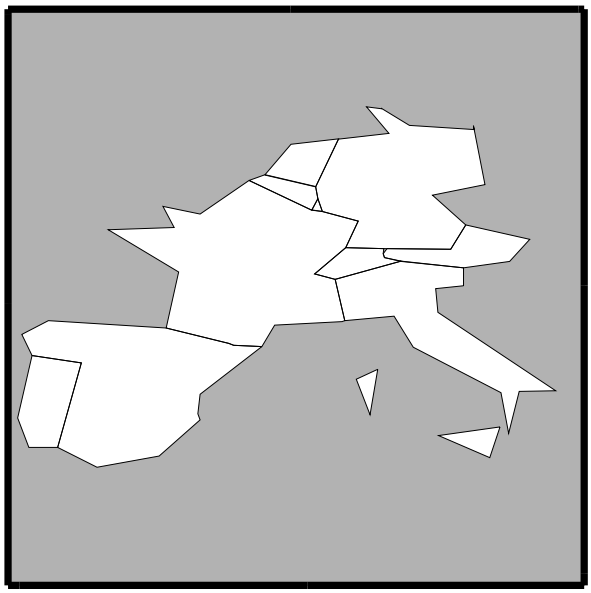


Abbildung 24: Einige Länder Europas

werden durch eine Kante verbunden, wenn die beiden Länder eine gemeinsame Grenze haben. Hierdurch wird das "Länderfärbungsproblem" in G auf ein "Eckenfärbungsproblem" in G^* zurückgeführt.

Beispiel: Wir geben in Abbildung 26 den dualen Graphen zu dem in Abbildung 25 angegebenen an. Hierzu haben wir die Länder mit ihrem Zentrum identifiziert, das sind die Ecken im dualen Graphen, und diese Zentren mit einer Kante verbunden, wenn die Länder eine gemeinsame Grenze haben. Schließlich wird der Außenwelt noch eine Ecke W zugeordnet, die noch mit allen Ländern außer den Binnenländern Luxemburg und Schweiz zu verbinden ist. Man erkennt, dass der duale Graph wieder zusammenhängend und eben ist. □

Dass mit G auch der duale Graph G^* zusammenhängend ist, ist nicht schwer einzusehen. Denn aus jedem beschränkten Gebiet von G gelangt man über Kanten benachbarter Gebiete zum äußeren Gebiet von G. Von jeder Ecke im dualen Graphen, die zu einem (beschränkten) Land gehört, ist also die Ecke, die die Außenwelt repräsentiert, erreichbar. Nicht ganz so einfach scheint die entsprechende Aussage über die Planarität des dualen Graphen zu sein. Wir werden dies ohne weiteren Beweis benutzen (und sind dann in guter Gesellschaft vieler Autoren).

Als Folgerung aus der Eulerschen Polyederformel erhalten wir eine Aussage, der wir entnehmen können, dass ein planarer Graph verhältnismäßig wenig Kanten besitzen kann, da deren Anzahl durch eine lineare Funktion der Eckenzahl beschränkt werden kann. Zuvor formulieren und beweisen wir aber das sogenannte Handshaking-Lemma.

Handshaking-Lemma *Sei $G = (V, E)$ ein (nicht notwendig planarer) Graph und $|E|$ die Anzahl der Kanten. Dann ist*

$$\sum_{x \in V} d(x) = 2|E|,$$

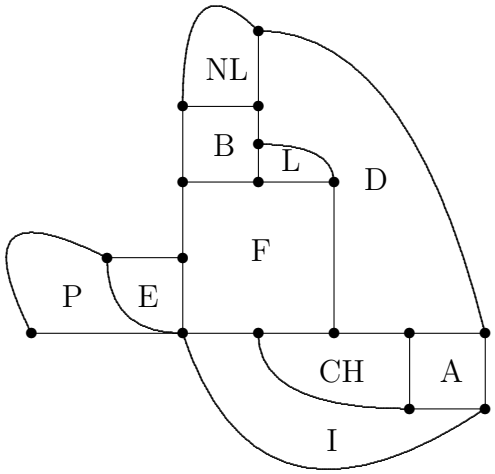


Abbildung 25: Länder Europas als ebener Graph

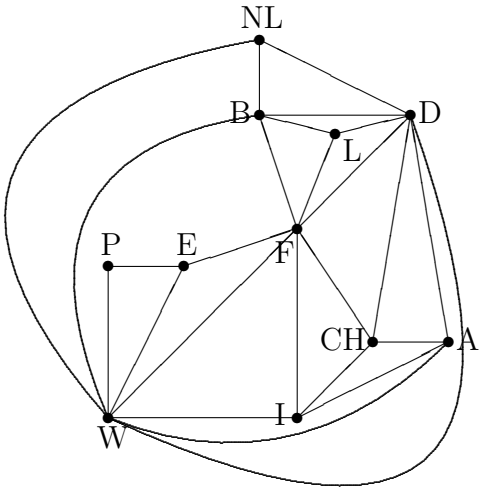


Abbildung 26: Der duale Graph zu dem Graphen in Abbildung 25

also die Summe der Grade aller Ecken gerade das Zweifache der Anzahl der Kanten.
Beweis: Wir betrachten die Menge aller Paare

$$P := \{(x, e) \in V \times E : x \in E\}$$

und berechnen die Anzahl ihrer Elemente auf zweierlei Weise. Da jede Kante $e \in E$ zwei Endecken besitzt, ist $|P| = 2|E|$. Andererseits gibt es zu jeder Ecke $x \in V$ genau $d(x)$ benachbarte Ecken, so dass $|P| = \sum_{x \in V} d(x)$. Damit ist der einfache Beweis schon abgeschlossen. □

Lemma *Sei $G = (V, E)$ ein zusammenhängender, planarer Graph mit $|V| \geq 3$. Dann gilt:*

1. *Es ist $|E| \leq 3|V| - 6$.*
2. *Enthält G kein Dreieck (das ist ein Kreis mit drei Ecken bzw. Kanten), so ist $|E| \leq 2|V| - 4$.*

3. Es existiert eine Ecke mit einem Grad ≤ 5.

Beweis: Natürlich können wir annehmen, dass G ein ebener Graph ist. Sei $n := |V|$ die Anzahl der Ecken, $m := |E|$ die Anzahl der Kanten und f die Anzahl der Länder. Für $m = 2$ ist die Aussage richtig, so dass $m \geq 3$ angenommen werden kann. Jedes Land wird von mindestens drei Kanten begrenzt, jede Kante gehört zum Rand von höchstens zwei Ländern[32]. Daher ist $3f \leq 2m$. Wegen der Eulerschen Polyederformel ist $3m - 3n + 6 \leq 2m$ bzw. $m \leq 3n - 6$, was schon die erste Behauptung ist. Enthält G kein Dreieck, so wird jedes Land von mindestens vier Kanten begrenzt, was $4f \leq 2m$ nach sich zieht. Einsetzen in die Eulersche Polyederformel liefert $m \leq 2n - 4$, die zweite Behauptung. Zum Beweis der dritten Behauptung bezeichnen wir mit $d(x)$ den Grad der Ecke $x \in V$. Angenommen, es sei $d(x) \geq 6$ für alle $x \in V$. Man definiere (wie beim Beweis des Handshaking-Lemmas) $P := \{(x, e) \in V \times E : x \in e\}$. Nach Annahme ist einerseits $|P| \geq 6n$, andererseits ist $|P| = 2m$, insbesondere also $m \geq 3n$. Dies ist ein Widerspruch zum ersten Teil des Korollars. □

Beispiel: Der K_5 hat $n = 5$ Ecken und $m = 10$ Kanten. Wegen der ersten der beiden obigen Aussagen ist K_5 nicht planar. Der vollständige bipartite Graph $K_{3,3}$ hat $n = 6$ Ecken und $m = 9$ Kanten, weiter enthält $K_{3,3}$ natürlich kein Dreieck. Aus der zweiten der beiden obigen Aussagen folgt, dass $K_{3,3}$ nicht planar ist. □

Es geht jetzt um das Färben planarer Graphen mit möglichst wenig Farben. Dass dazu i. Allg. mindestens vier Farben nötig sind, erkennt man an an dem Graphen in Abbildung 27 links, wobei man daran denken muss, dass auch das unbeschränkte Land

Abbildung 27: Vier Farben sind notwendig

gefärbt werden muss. Rechts haben wir den dualen Graphen angegeben. In diesem gibt es eine Ecke mit drei Nachbarn. Also sind mindestens vier Farben notwendig.

Nun wollen wir den *Sechsfarbensatz* formulieren und beweisen. Wir folgen hier und auch beim entsprechenden Fünffarbensatz der Darstellung bei J. M. Aldous, R. J. Wilson (2000, S. 284 ff.).

Sechsfarbensatz *Ein zusammenhängender, ebener Graph bzw. die zugehörige Landkarte lässt sich durch sechs Farben färben.*

Beweis: Wir zeigen, dass eine Eckenfärbung eines zusammenhängenden, planaren Graphen $G = (V, E)$ mit sechs Farben möglich ist (wobei natürlich zwei benachbarte Ecken unterschiedlich gefärbt sind). Eine Anwendung dieses Resultats auf den dualen Graphen

[32] Es folgt eine Art Handshaking-Argument. Man betrachte die Menge von Paaren, bestehend aus einem Land und einer dieses Land begrenzenden Kante. Die Anzahl der Elemente dieser Menge ist mindestens $3f$ und höchstens $2m$.

liefert dann die Behauptung[33]. Wir beweisen die Aussage durch vollständige Induktion nach der Anzahl n der Ecken von G. Die Aussage ist trivialerweise richtig für Graphen mit $n \leq 6$ Ecken. Wir nehmen an, sie sei richtig für Graphen mit weniger als n Ecken und zeigen, dass eine Eckenfärbung eines zusammenhängenden, planaren Graphen mit sechs (oder weniger) Farben möglich ist. Wegen des dritten Teils des obigen Lemmas gibt es eine Ecke v in G mit $d(v) \leq 5$, die also höchstens fünf Nachbarn hat. Diese Ecke mitsamt der Kanten, auf denen sie liegt, entferne man aus dem Graphen G und erhalte so den (nach wie vor planaren, möglicherweise aber nicht zusammenhängenden) Graphen H mit $n-1$ Ecken. Diesen Prozess veranschaulichen wir uns in Abbildung 28, wobei wir davon ausgehen, dass die Ecke v die fünf Nachbarn a, b, c, d, e besitzt. Nach

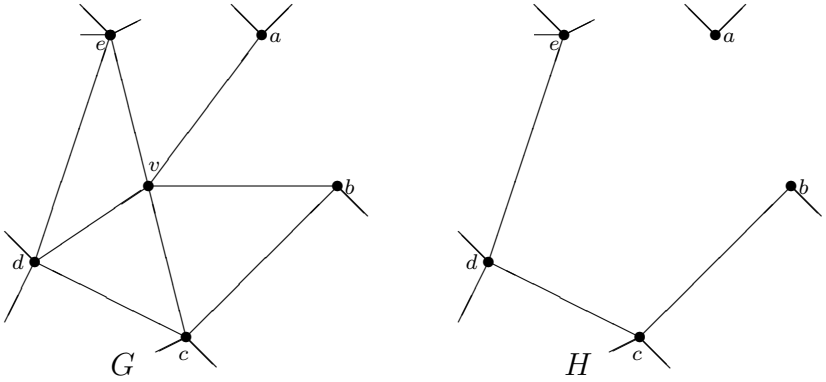


Abbildung 28: Gewinne H aus G durch Entfernen der Ecke v

Induktionsannahme können die Ecken von H (bzw. der Komponenten von H, wenn H nicht zusammenhängend ist) mit sechs Farben so gefärbt werden, dass benachbarte Ecken unterschiedlich gefärbt sind. Die höchstens fünf Nachbarn von v sind dadurch gefärbt, hierzu wurden natürlich höchstens fünf Farben benötigt. Zur Färbung von v steht noch eine Farbe zur Verfügung, so dasseine Eckenfärbung von G mit höchstens sechs Farben gelungen ist. Der Beweis ist damit abgeschlossen. □

Mit etwas mehr Aufwand können wir auch den *Fünffarbensatz* beweisen.

Fünffarbensatz *Ein zusammenhängender, ebener Graph bzw. die zugehörige Landkarte lässt sich durch fünf Farben färben.*

Beweis: Der Beginn des Beweises ist wie beim Sechsfarbensatz. Entsprechend zeigen wir, dass die Ecken eines zusammenhängenden, planaren Graphen G mit fünf Farben so gefärbt werden können, dass benachbarte Ecken unterschiedliche Farben haben. Dies geschieht wieder durch vollständige Induktion nach n, der Anzahl der Ecken von G. Der Induktionsanfang kann für $n = 1, 2, 3, 4, 5$ gemacht werden, die Annahme ist, die Aussage sei für zusammenhängende, planare Graphen mit weniger als n Ecken richtig. Auch der Beginn des Induktionsbeweises ist wie beim Beweis des Sechsfarbensatz. Es gibt eine Ecke v in G mit höchstens fünf Nachbarn. Diese Ecke mitsamt der Kanten, auf denen sie liegt, entferne man wieder aus dem Graphen G und erhalte so den Graphen

[33]Hierbei benutzen wir, dass der zu einem zusammenhängenden, planaren Graphen duale Graph ebenfalls zusammenhängend und planar ist.

H mit $n-1$ Ecken (siehe Abbildung 28). Nach Induktionsvoraussetzung können die Ecken von H (bzw. der Komponenten von H, wenn H nicht zusammenhängend ist) mit fünf Farben so gefärbt werden, dass benachbarte Ecken unterschiedlich gefärbt sind. Wir können annehmen, dass v fünf Nachbarn besitzt, die mit fünf verschiedenen Farben gefärbt sind, da man andernfalls noch eine Farbe zur Färbung von v übrig hätte. Die fünf Farben seien etwa blau (b), grün (g), magenta (m), rot (r) und türkis (t), siehe Abbildung 29. Wir greifen uns zwei beliebige Nachbarn von v heraus, etwa

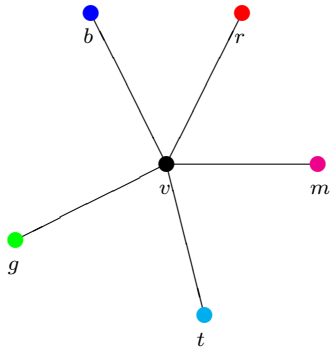


Abbildung 29: Die Ecke v und ihre fünf Nachbarn

die rot (r) und türkis (t) gefärbten Nachbarn. Wir betrachten Wege, die nur aus roten (r) oder türkisfarbenen (t) Ecken bestehen, beginnend bei den Nachbarn r bzw. t und machen eine Fallunterscheidung. Im ersten Fall sind alle r- und t-Ecken, die vom r-Nachbarn erreicht werden können, verschieden von denen, die vom t-Nachbarn aus erreicht werden können. Insbesondere gibt es dann keinen Weg vom r-Nachbarn zum t-Nachbarn, der aus (natürlich abwechselnd) roten und türkisfarbenen Nachbarn besteht. In diesem Fall vertausche man einfach die Farben derjenigen Ecken, die vom r-Nachbarn aus auf einem r-t-Weg erreicht werden kann und färbe anschließend v rot. Diesen Fall veranschaulichen wir in Abbildung 30. Im zweiten Fall gibt es einen r-t-Weg vom r-

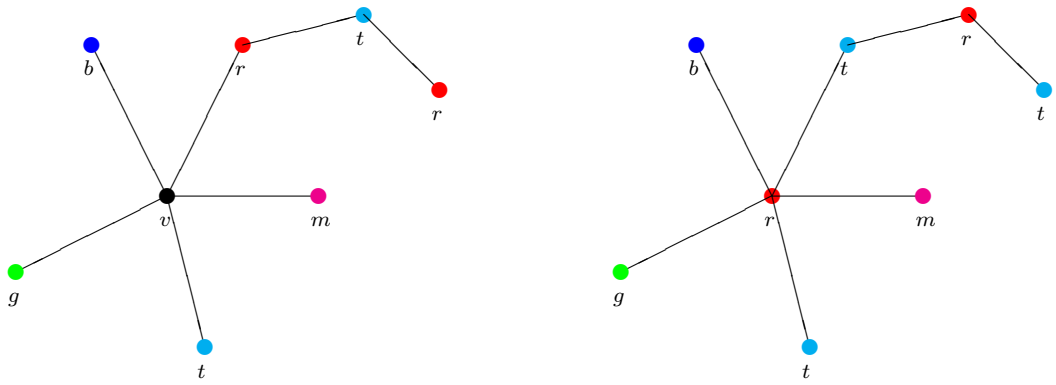


Abbildung 30: Vertausche Farben r und t für vom r-Nachbarn ausgehende r-t-Wege

Nachbarn zum t-Nachbarn. Ein Vertauschen der Farben würde nichts nützen, da v dann nach wie vor einen r- und einen t-Nachbarn hätte. Andererseits kann es keinen Weg

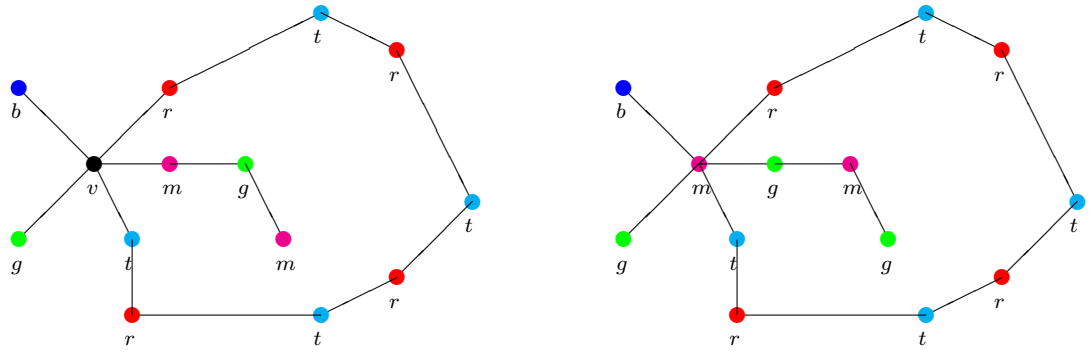


Abbildung 31: Vertausche grün (g) und magenta (m), färbe v mit magenta

aus magentafarbenen (m) und grünen (g) Ecken vom m-Nachbarn zum g-Nachbarn geben, da sonst der r-t-Weg gekreuzt werden müsste, was bei einem planaren Graphen natürlich nicht möglich ist. Daher kann man im m-g-Weg von v über den m-Nachbarn die Farben magenta (m) und grün (g) vertauschen und anschließend die Ecke v mit der Farbe magenta (m) färben, man hat den zweiten Fall also auf den ersten bezüglich eines anderen Farbpaares zurückgeführt. Dies ist in Abbildung 31 veranschaulicht. Insgesamt ist der Satz bewiesen. □

Beispiel: Wir kommen zurück auf Abbildung 26 und färben die dort angegebenen Ecken (bzw. die Hauptstädte entsprechender Länder Westeuropas) mit vier Farben so, dass benachbarte Länder unterschiedlich gefärbt sind, siehe Abbildung 32. □

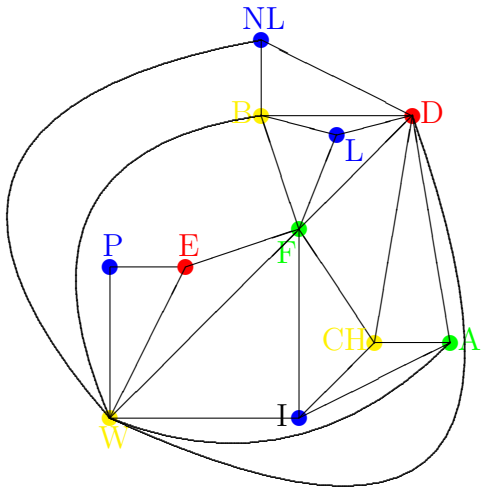


Abbildung 32: Färbung einer Landkarte mit vier Farben

Beispiel: Ein[34] Unternehmen muss gewisse Chemikalien, nämlich genau sieben verschiedene Arten, lagern. Gewisse Chemikalien müssen von anderen getrennt werden, können aber durchaus mit anderen zusammen gelagert werden. In der folgenden Tabelle ist durch $*$ gekennzeichnet, dass entsprechende Chemikalien nur getrennt voneinander

[34] Auch hier folgen wir J. M. Aldous, R. J. Wilson (2000, S. 276).

gelagert werden können.

	A	B	C	D	E	F	G
A	–	$*$	$*$	$*$	–	–	$*$
B	$*$	–	$*$	$*$	$*$	–	$*$
C	$*$	$*$	–	$*$	–	$*$	–
D	$*$	$*$	$*$	–	–	$*$	–
E	–	$*$	–	–	–	–	–
F	–	–	$*$	$*$	–	–	$*$
G	$*$	$*$	–	–	–	$*$	–

Die Frage ist, wie man die Chemikalien lagern sollte, wobei man mit möglichst wenig Sonderlagern auskommen möchte. Wir definieren einen Graphen, dessen Ecken die sieben Chemikalien sind. Diese Ecken werden durch eine Kante verbunden, wenn die entsprechenden Chemikalien getrennt gelagert werden müssen, siehe Abbildung 33.

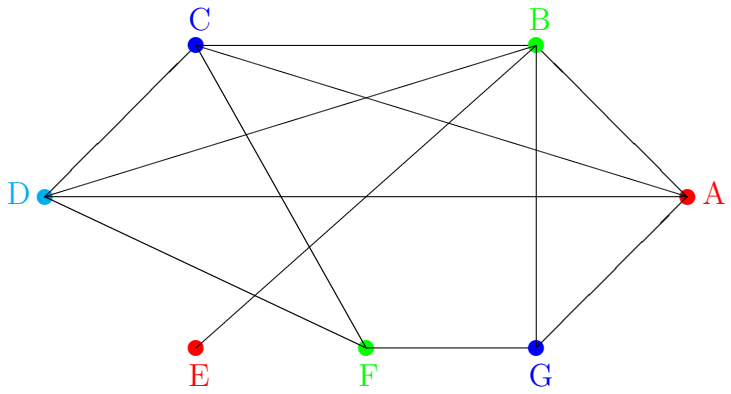


Abbildung 33: Unverträgliche Chemikalien: Mögliche Eckenfärbung

Dies ist kein planarer Graph. Denn dieser Graph enthält den bipartiten, vollständigen Graphen $K_{3,3}$ mit den Ecken $\{A, B, F\}$ und $\{C, D, G\}$ und dieser ist bekanntlich nicht planar. Mögliche Aufteilungen sind $\{A, E\}$, $\{B, F\}$, $\{C, G\}$, $\{D\}$ oder $\{A, E\}$, $\{B, F\}$, $\{C\}$, $\{D, G\}$. Hier haben wir also einen nichtplanaren Graphen, dessen Ecken durch vier Farben so gefärbt werden können, dass benachbarte Ecken unterschiedliche Farben haben. Da eine entsprechende Färbung mit drei Farben nicht möglich ist, sagt man, der Graph habe die chromatische Zahl 4. Die Aussage des *Vierfarbensatzes* ist es, dass jeder zusammenhängende, planare Graph eine chromatische Zahl ≤ 4 besitzt. Für nichtplanare Graphen ist das natürlich i. Allg. nicht richtig, denn z. B. hat der vollständige Graph K_n die chromatische Zahl n. Die Bestimmung der chromatischen Zahl eines gegebenen Graphen ist ein schwieriges Problem, auf das wir nicht eingehen können. □

21 Irrationale Zahlen

Eine positive reelle Zahl heißt *rational*, wenn sie sich als Quotient zweier natürlicher Zahlen darstellen lässt, andernfalls *irrational*. Als kleine Fingerübung wollen wir

zunächst die folgende, eigentlich überhaupt nicht merkwürdige und schon Euklid bekannte Aussage beweisen:

- $\sqrt{2}$ ist irrational.

Denn: Wir nehmen an, die Quadratwurzel aus 2 sei rational. Dann existieren zwei teilerfremde natürliche Zahlen p und q mit $\sqrt{2} = p/q$. Dann ist $p^2 = 2q^2$. Da die rechte Seite der Gleichung gerade ist, ist auch die linke Seite p^2 gerade. Daraus folgt, dass auch p gerade ist und folglich $r := p/2$ eine natürliche Zahl ist. Dann ist

$$2q^2 = p^2 = (2r)^2 = 4r^2.$$

Nach Division durch 2 folgt $q^2 = 2r^2$. Also sind q^2 und damit auch q gerade Zahlen. Dass aber p und q beide durch 2 teilbar sind, steht im Widerspruch dazu, dass p und q teilerfremd sind. Damit ist die Behauptung bewiesen. □

Von Johann Heinrich Lambert (1728–1777) stammt aus dem Jahre 1766 der erste Beweis, dass π irrational ist. Legrende bewies 1794, dass π^2 irrational ist, woraus trivialerweise folgt, dass auch π irrational ist. Wir wollen einen Beweis für die Irrationalität von π angeben, der auf I. Niven (1947) zurückgeht. Ein wirklich genialer Beweis, bei dem benutzt wird, dass π die erste positive Nullstelle des Sinus ist. Zu Recht ist er in das BUCH der Beweise von M. Aigner, G. M. Ziegler (2002, S. 33 ff.) aufgenommen worden.

Satz *π ist irrational.*

Beweis: Angenommen, es sei $\pi = p/q$ mit $p, q \in \mathbb{N}$. Wir definieren

$$P_n(x) := \frac{x^n(p - qx)^n}{n!}.$$

Dann ist

$$n!P_n(x) = \sum_{k=n}^{2n} c_k x^k$$

mit $c_n, \ldots, c_{2n} \in \mathbb{Z}$. Wir wollen uns überlegen, dass

$$P_n^{(j)}(0) \in \mathbb{Z}, \qquad j = 0, 1, \ldots,$$

und

$$P_n^{(j)}(\pi) \in \mathbb{Z}, \qquad j = 0, 1, \ldots.$$

Denn: Zunächst ist trivialerweise $P_n^{(j)}(0) = 0$, $j = 0, \ldots, n-1$, und $j > 2n$. Weiter ist $P_n^{(j)}(0) = j!c_j/n! \in \mathbb{Z}$, $j = n, \ldots, 2n$. Damit ist die erste Behauptung schon bewiesen. Es ist $P_n(x) = P_n(p/q - x) = P_n(\pi - x)$ und folglich $P_n^{(j)}(x) = (-1)^j P_n^{(j)}(\pi - x)$. Setzt man hier $x = 0$, so erhält man aus dem ersten Teil den zweiten.

Nun definiere man

$$Q_n(x) := \sum_{j=0}^{n} (-1)^j P_n^{(2j)}(x).$$

Dann ist

$$\begin{aligned}\frac{d}{dx}[Q_n'(x)\sin x - Q_n(x)\cos x] &= Q_n''(x)\sin x + Q_n(x)\sin x \\ &= [Q_n''(x) + Q_n(x)]\sin x \\ &= P_n(x)\sin x.\end{aligned}$$

Folglich ist

$$\int_0^\pi P_n(x)\sin x\,dx = [Q_n'(x)\sin x - Q_n(x)\cos x]_0^\pi = Q_n(\pi) + Q_n(0).$$

Nun ist $Q_n(\pi) + Q_n(0) \in \mathbb{Z}$, da $P_n^{(j)}(\pi), P_n^{(j)}(0) \in \mathbb{Z}$. Für $x \in (0,\pi)$ ist andererseits

$$0 < P_n(x)\sin x < P_n(x) = \frac{x^n p^n (1-x/\pi)^n}{n!} \le \frac{\pi^n p^n}{n!}.$$

Daher ist $\int_0^\pi P_n(x)\sin x\,dx$ positiv und beliebig klein für hinreichend großes n. Das ist ein Widerspruch dazu, dass $\int_0^\pi P_n(x)\sin x\,dx$ für jedes n eine ganze Zahl ist und damit zu der Ausgangsannahme, dass π rational sei. □

Ganz ähnlich kann bewiesen werden, dass π^2 irrational ist, was ein stärkeres Ergebnis als das gerade eben bewiesene ist.

Satz *π^2 ist irrational.*

Beweis: Wir machen einen Widerspruchsbeweis und nehmen an, es sei $\pi^2 = p/q$ mit $p, q \in \mathbb{N}$. Diesmal definieren wir

$$P_n(x) := \frac{x^n(1-x)^n}{n!}.$$

Dann ist

$$n!P_n(x) = \sum_{k=n}^{2n} c_k x^k$$

mit $c_n, \ldots, c_{2n} \in \mathbb{Z}$. Wir wollen uns überlegen, dass

$$P_n^{(j)}(0) \in \mathbb{Z}, \qquad j = 0, 1, \ldots,$$

und

$$P_n^{(j)}(1) \in \mathbb{Z}, \qquad j = 0, 1, \ldots.$$

Denn: Zunächst ist trivialerweise $P_n^{(j)}(0) = 0$, $j = 0, \ldots, n-1$, und $j > 2n$. Weiter ist $P_n^{(j)}(0) = j!c_j/n! \in \mathbb{Z}$, $j = n, \ldots, 2n$. Damit ist die erste Behauptung schon bewiesen. Es ist $P_n(x) = P_n(1-x)$ und folglich $P_n^{(j)}(x) = (-1)^j P_n^{(j)}(1-x)$. Setzt man hier $x = 0$, so erhält man aus dem ersten Teil den zweiten.

Nun definiere man

$$Q_n(x) := q^n[\pi^{2n}P_n(x) - \pi^{2n-2}P_n''(x) + \cdots + (-1)^n\pi^0 P_n^{(2n)}(x)].$$

Wegen des gerade eben bewiesenen Teils ist $Q_n(0), Q_n(1) \in \mathbb{Z}$. Wegen $P_n^{(2n+2)}(x) = 0$ ist ferner

$$\begin{aligned}\frac{d}{dx}[Q_n'(x)\sin(\pi x) - \pi Q_n(x)\cos(\pi x)] &= [Q_n''(x) + \pi 2 Q_n(x)]\sin(\pi x) \\ &= q^n \pi^{2n+2} P_n(x)\sin(\pi x) \\ &= \pi 2 p^n P_n(x)\sin(\pi x).\end{aligned}$$

Folglich ist

$$\begin{aligned}\pi q^n \int_0^1 P_n(x)\sin(\pi x)\,dx &= \Big[\frac{Q_n'(x)\sin(\pi x)}{\pi} - Q_n(x)\cos(\pi x)\Big]_0^1 \\ &= Q_n(0) + Q_n(1) \\ &\in \mathbb{Z}.\end{aligned}$$

Andererseits ist

$$0 < \pi q^n \int_0^1 P_n(x)\sin(\pi x)\,dx \le \frac{\pi q^n}{n!}.$$

Wie im letzten Beweis erhält man einen Widerspruch. □

22 Das Fermat-Weber Problem

Das folgende Problem scheint 1629 zum ersten Mal von Fermat formuliert worden zu sein:

- Gegeben seien drei Punkte in der Ebene. Man finde einen Punkt in der Ebene derart, dass die Summe der Abstände dieses Punktes zu den drei vorgegebenen Punkten minimal ist.

Die Verallgemeinerung auf m Punkte im $\mathbb{R}^n$ heißt das Fermat-Weber[35]-Problem:

- Gegeben seien $m \ge 3$ paarweise verschiedene Punkte $a_1, \ldots, a_m \in \mathbb{R}^n$ und positive reelle Zahlen $w_1, \ldots, w_m$. Man bestimme eine Lösung $x^* \in \mathbb{R}^n$ von

  $$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^m w_i \,\|x - a_i\|_2 \quad \text{auf} \quad M := \mathbb{R}^n,$$

  wobei $\|\cdot\|_2$ die euklidische Norm auf dem $\mathbb{R}^n$ bedeutet.

[35] Alfred Weber (1868–1958) war Nationalökonom, Soziologe und Kulturphilosoph und begründete die volkswirtschaftliche Standorttheorie. Er lehrte 1904–1907 in Prag, danach in Heidelberg. In Prag promovierte (mit der schlechtesten, zum Bestehen noch ausreichenden Note) 1906 Franz Kafka bei ihm. Ein anderer bekannter Doktorand Alfred Webers ist Erich Fromm, der 1922 in Heidelberg in Soziologie bei ihm promovierte. Als überzeugter Gegner des Nationalsozialismus wurde Alfred Weber bei der Bundespräsidentenwahl 1954 ohne seine Zustimmung von der KPD für das Amt des Bundespräsidenten vorgeschlagen. Alfred Weber ist Bruder (des noch berühmteren) Max Weber (1864–1920), Jurist, Soziologe und Nationalökonom.

Die ökonomische Interpretation (man spricht in denWirtschaftswissenschaften auch von dem "Standortproblem") könnte die folgende sein: Eine Warenhauskette mit Filialen in $a_1, \ldots, a_k$ und Zulieferern in $a_{k+1}, \ldots, a_m$ will den Standort eines zusätzlichen Lagers bestimmen. Dieser soll so gewählt werden, dass eine gewichtete Summe der Abstände vom Lager zu den Filialen und von den Zulieferern zum Lager minimal wird.

Wir wollen auf das Fermat-Weber-Problem nicht weiter eingehen, sondern nur einen hübschen geometrischen Beweis dafür angeben, dass beim eingangs genannten Fermat-Problem der gesuchte Punkt (auch Fermat- oder Torricelli-Punkt genannt) derjenige ist, von dem die drei Seiten des (spitzwinkligen) Dreiecks unter einem Winkel von 120° gesehen werden.

Gegeben sei ein spitzwinkliges Dreieck in der Ebene mit den Ecken A, B und C. In diesem Dreieck wähle man sich einen beliebigen Punkt P und verbinde ihn mit den Ecken. Das innere Dreieck $\triangle APB$ drehe man um 60° um B und erhalte das Dreieck $\triangle C'P'B$. In Abbildung 34 ist die Konstruktion angegeben. Dann sind $\triangle ABC'$ grün

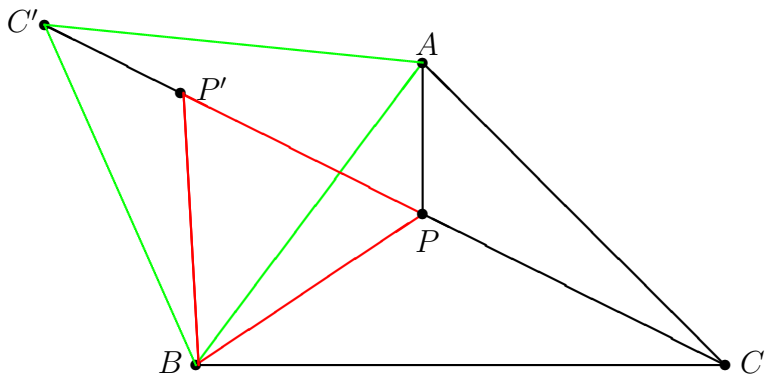


Abbildung 34: Konstruktion zum Fermat-Problem

und $\triangle PBP'$ rot gleichseitig, die Winkel in diesen Dreiecken also jeweils 60°. Daher ist

$$|AP| + |BP| + |CP| = |C'P'| + |P'P| + |PC|,$$

und die rechtsstehende Summe ist die Länge eines i. Allg. gebrochenen Streckenzuges. Dieser ist minimal, wenn er ein Geradensegment ist. In diesem Falle ist

$$\sphericalangle BPC = 180° - \sphericalangle BPP' = 120°$$

und

$$\sphericalangle APB = \sphericalangle C'P'B = 180° - \sphericalangle PP'B = 120°.$$

Der gesuchte Punkt P, für den $|AP|+|BP|+|CP|$ minimal ist, ist also derjenige Punkt P, für den

$$\sphericalangle APB = \sphericalangle BPC = \sphericalangle CPA = 120°.$$

Diese Lösung des Fermat-Problems kann man bei H. S. M. Coxeter (1969, S. 21) nachlesen. In Abbildung 35 geben wir eine mögliche Konstruktion des Fermat-Torricelli-

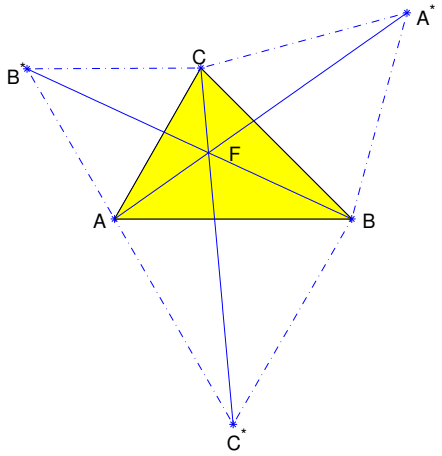


Abbildung 35: Konstruktion des Fermat-Torricelli-Punktes

Punktes an: Über den Seiten des gegebenen Dreiecks $\triangle ABC$ konstruiere man drei gleichseitige Dreiecke und gewinne dadurch Punkte A^*, B^* und C^*. Verbindet man diese Punkte mit den gegenüberliegenden Ecken des Dreiecks, also mit A, B und C, so schneiden sich diese drei Verbindungsstrecken in einem Punkt F, dem Fermat-Torricelli-Punkt.

23 In welchen Jahren fällt der Himmelfahrtstag auf den 1. Mai?

Im Sommersemester 2008 hielt ich noch einmal eine Vorlesung über Angewandte Mathematik für Lehramtskandidaten. Diese Vorlesung fand jeweils an einem Mittwoch und einem Donnerstag statt. Da der Himmelfahrtstag stets ein Donnerstag ist, war klar, dass mindestens eine Vorlesung ausfällt. Da auch der 1. Mai ein Feiertag ist, war zu befürchten, dass eine weitere Vorlesung ausfällt. Ein Blick in den Kalender zeigte aber, dass Himmelfahrtstag und der 1. Mai im Jahr 2008 zusammenfallen. Naheliegenderweise trat daher die Frage auf, in welchen Jahren der Himmelfahrtstag auf den 1. Mai fällt. Da Himmelfahrt genau 40 Tage nach dem Ostersonntag ist, ist das genau dann der Fall, wenn der Ostersonntag auf den 23. März fällt. Bei Wikipedia, siehe *Gaußsche Osterformel*, kann man nachlesen: Seit den Beschlüssen des ersten Konzils von Nicäa 325 n. Chr. und auf Grund der im Jahr 525 n. Chr. im Auftrag von Papst Johannes I. begonnenen Arbeiten durch Exiguus wird das Osterfest am ersten Sonntag (dem Ostersonntag) nach dem Frühlings-Vollmond gefeiert. Tag des Frühlingsanfangs ist nach Beschluss der 21. März. Ein am 21. März stattfindender Vollmond gilt bereits als frühestmöglicher Frühlings-Vollmond. Der 22. März ist deshalb der früheste Kalendertag, auf den Ostern fallen kann. Im Julianischen Kalender fällt der letzte mögliche Ostersonntag auf den 25. April. Diese Begrenzung wurde im Gregorianischen Kalender in einer Zusatzbestimmung beibehalten. Somit gibt es in

beiden Kalendern insgesamt 35 verschiedene Ostertermine. Weitere historische Bemerkungen und ein Verfahren zur Berechnung des Osterdatums findet man bei D. KNUTH (1968, S. 155–156). Eine hierauf basierende MATLAB-Funktion `easter.m` findet man unter `http://www.mathworks.de/moler/exm/chapters.html` (Individual Files).

Wir geben jetzt eine MATLAB-Funktion `Ostern` an, die zu einem gegebenen Jahr `X` das Datum des Ostersonntags berechnet. Diese benutzt die beiden MATLAB-Funktionen `floor` und `mod`. Bei gegebenem reellen x ergibt `floor(x)` die größte ganze Zahl, die kleiner oder gleich x ist. Für natürliche Zahlen $m \geq n$ ist `mod(m,n)` die nichtnegative ganze Zahl, die sich als Rest bei der Division von m durch n ergibt. Z. B. ist `mod(13,5)` gleich 3.

```
function oster_datum=Ostern(X);
%****************************************************************
%Fuer das Jahr X wird das Datum des Ostersonntags berechnet.
%Ist oster_datum<=31, so faellt der Ostersonntag auf den
%oster_datum. Maerz. Ist dagegen oster_datum>31,
%so ist Ostersonntag der oster_datum-31. April.
%****************************************************************
a=mod(X,19);    b=mod(X,4);   c=mod(X,7);
d=floor((floor(X/100)*8+13)/25)-2;  e=floor(X/100)-floor(X/400)-2;
f=mod(15+e-d,30), g=mod(6+e,7); h=mod(19*a+f,30);
i=h;
if (h==29)
      i=28;
end;
if (h==28)&(a>10)
      i=27;
end;
j=mod(2*b+4*c+6*i+g,7);    oster_datum=i+j+22;
```

Beispiel Wir wollen das Datum für den Ostersonntag im Jahre $X = 2013$ berechnen. Der Reihe nach erhalten wir:

$$a = 18, \quad b = 1, \quad c = 4, \quad d = 4, \quad e = 13, \quad f = 24, \quad g = 5$$

und weiter

$$h = 6, \quad i = 6, \quad j = 3.$$

Wegen $i + j + 22 = 31$ fällt im Jahre 2013 der Ostersonntag auf den 31. März.

Mit dem obigen Verfahren kann man ausrechnen, dass nach 2008 erst im Jahre 2160 wieder der Himmelfahrtstag auf den 1. Mai fällt.

24 Das allgemeine Dreieck

Wenn man mit Google "Das allgemeine Dreieck" sucht, so wird man auf einen (lustigen) Aufsatz von Bernhard Tergan[36] stoßen, den man z. B. auch als Anhang 2 im Buch von

[36] siehe `http://www.holger-lang.de/haupt/mathe/dreieck/dreieck.html`

F. Wille (1984) findet. Und zwar verstehen wir unter einem *allgemeinen Dreieck* ein spitzwinkliges Dreieck (alle Winkel im Dreieck sind also $\leq 90°$), welches sich "so viel wie möglich" von einem rechtwinkligen und einem gleichschenkligen Dreieck (also einem mit zwei gleichen Winkeln) unterscheidet. Ein allgemeines Dreieck ist also eines, das "leicht erkennbar" weder rechtwinklig noch gleichseitig ist. Genauer betrachten wir die folgende Aufgabe:

- Seien α, β, γ Winkel in einem (spitzwinkligen) Dreieck mit $90° \geq \alpha \geq \beta \geq \gamma \geq 0$ und natürlich $\alpha + \beta + \gamma = 180°$. Unter allen solchen Winkeln bestimme man diejenigen, für die

  $$f(\alpha, \beta, \gamma) := \min(90° - \alpha, \alpha - \beta, \beta - \gamma, \gamma)$$

  maximal ist.

Die Lösung ist einfach zu erhalten, wenn man folgendes beachtet: Für ein Tripel (α, β, γ) mit $90° \geq \alpha \geq \beta \geq \gamma \geq 0$ und $\alpha + \beta + \gamma = 180°$ ist

$$f(\alpha, \beta, \gamma) \leq \frac{3(90° - \alpha) + 2(\alpha - \beta) + (\beta - \gamma)}{6} = \frac{270° - (\alpha + \beta + \gamma)}{6} = 15°.$$

Mit $(\alpha^*, \beta^*, \gamma^*) := (75°, 60°, 45°)$ ist $f(\alpha^*, \beta^*, \gamma^*) = 15°$ und daher ist $(\alpha^*, \beta^*, \gamma^*)$ eine Lösung der gestellten Aufgabe. Dies ist aber auch die einzige Lösung. Denn angenommen, auch $(\alpha^{**}, \beta^{**}, \gamma^{**})$ sei eine Lösung, d. h. es sei

$$90° \geq \alpha^{**} \geq \beta^{**} \geq \gamma^{**} \geq 0, \qquad \alpha^{**} + \beta^{**} + \gamma^{**} = 180°$$

und $f(\alpha^{**}, \beta^{**}, \gamma^{**}) = 15°$. Hieraus folgt $15° \leq 90° - \alpha^{**}$, $15° \leq \alpha^{**} - \beta^{**}$, $15° \leq \beta^{**} - \gamma^{**}$ und hieraus der Reihe nach $\alpha^{**} \leq 75°$, $\beta^{**} \leq 45°$ und $\gamma^{**} \leq 45°$. Würde in den letzten drei Ungleichungen nur einmal das strikte Ungleichheitszeichen gelten, hätte man einen Widerspruch zu $\alpha^{**} + \beta^{**} + \gamma^{**} = 180°$. Also ist $(\alpha^{**}, \beta^{**}, \gamma^{**}) = (\alpha^*, \beta^*, \gamma^*)$, die Eindeutigkeit einer Lösung ist bewiesen.

Jetzt betrachten wir noch die entsprechende Aufgabe für ein stumpfwinkliges Dreieck, untersuchen also die folgende Aufgabe:

- Seien α, β, γ Winkel in einem (stumpfwinkligen) Dreieck mit $\alpha \geq 90° \geq \beta \geq \gamma \geq 0$ und natürlich $\alpha + \beta + \gamma = 180°$. Unter allen solchen Winkeln bestimme man diejenigen, für die

  $$f(\alpha, \beta, \gamma) := \min(\alpha - 90°, 90° - \beta, \beta - \gamma, \gamma)$$

  maximal ist.

Die Lösung kann ähnlich wie im spitzwinkligen Fall erhalten werden, wobei $f(\cdot)$ lediglich durch ein etwas anderes gewichtetes arithmetisches Mittel abgeschätzt wird. Für ein Tripel mit $\alpha \geq 90° \geq \beta \geq \gamma \geq 0$ und $\alpha + \beta + \gamma = 180°$ ist

$$f(\alpha, \beta, \gamma) \leq \frac{(\alpha - 90°) + (\beta - \gamma) + 2\gamma}{4} = \frac{(\alpha + \beta + \gamma) - 90°}{4} = 22.5°.$$

Mit $(\alpha^*, \beta^*, \gamma^*) := (112.5°, 45°, 22.5°)$ ist $f(\alpha^*, \beta^*, \gamma^*) = 22.5°$ und daher ist $(\alpha^*, \beta^*, \gamma^*)$ eine Lösung der gestellten Aufgabe. Ganz ähnlich wie im spitzwinkligen Fall zeigt man, dass dies die einzige Lösung ist.

25 Das Napoleon-Dreieck, Napoleonische Probleme

Die Aussage des folgenden Satzes ist als Satz von Napoleon bekannt. Er wird dem französischen Feldherrn und Kaiser Napoleon Bonaparte zugeordnet, von dem gesagt wird, er sei ein begabter Amateurmathematiker gewesen.

Satz *Über den Seiten eines Dreiecks $\triangle ABC$ werden nach außen gleichseitige Dreiecke errichtet, also etwa die Dreiecke $\triangle ADC$, $\triangle CEB$ und $\triangle BFA$. Dann bilden die Mittelpunkte (bzw. geometrischen Schwerpunkte, Flächenschwerpunkte) dieser drei gleichseitigen Dreiecke ein gleichseitiges Dreieck, das sogenannte Napoleon-Dreieck.*

Zunächst veranschaulichen wir uns den Satz von Napoleon in Abbildung 36. Bevor wir

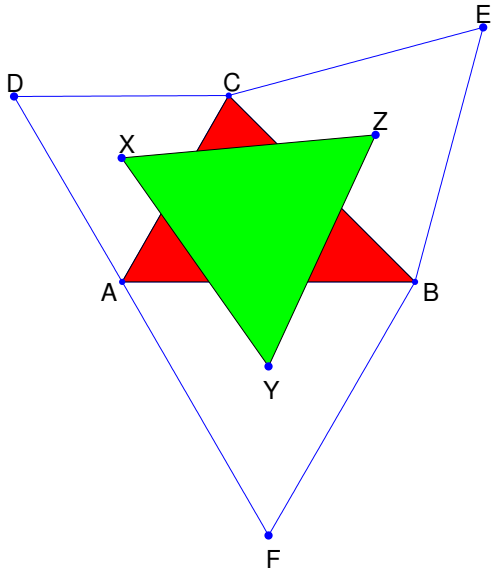


Abbildung 36: Der Satz von Napoleon

ihn beweisen, formulieren und beweisen wir das folgende

Lemma *Sei $\eta := e^{2\pi i/3}$ und $\triangle UVW$ ein Dreieck in der komplexen Zahlenebene $\mathbb{C}$. Dann ist $\triangle UVW$ genau dann ein gleichschenkliges Dreieck mit Basis UV und Winkel an der Spitze $\sphericalangle UWV = 120°$, wenn*

$$\eta U - V + (1-\eta)W = 0 \qquad \textit{bzw.} \qquad (\eta-1)U + (\eta^2-1)V + 3W = 0.$$

Beweis: Offenbar ist $\triangle UVW$ das beschriebene Dreieck genau dann, wenn eine Drehung von U um W mit dem Winkel $2\pi/3 = 120°$ nach V führt bzw. $\eta(U-W) = V-W$ oder $\eta U - V + (1-\eta)W = 0$ gilt. Dies wiederum ist gleichwertig zu

$$(1-\eta^2)\eta U + (\eta^2-1)V + (1-\eta^2)(1-\eta)W = 0,$$

was wegen $\eta^3 = 1$ und $(1-\eta^2)(1-\eta) = 3$ gleichwertig mit

$$(\eta-1)U + (\eta^2-1)V + 3W = 0$$

ist. □

Beweis des Satzes von Napoleon: Wir wenden auf die Dreiecke $\triangle ACX$, $\triangle BAY$ und $\triangle CBZ$ (siehe Abbildung 36) das eben bewiesene Lemma an. Hiernach ist

$$\begin{aligned}(\eta-1)A+(\eta^2-1)C+3X &= 0,\\ (\eta-1)B+(\eta^2-1)A+3Y &= 0,\\ (\eta-1)C+(\eta^2-1)B+3Z &= 0.\end{aligned}$$

Multipliziert man die erste Gleichung mit 1, die zweite mit η, die dritte mit η^2 und addiert anschließend die so erhaltenen Gleichungen, so erhält man

$$[(\eta-1)+\eta(\eta^2-1)]A+[\eta(\eta-1)+\eta^2(\eta^2-1)]B+[(\eta^2-1)+\eta^2(\eta-1)]C+3[X+\eta Y+\eta^2 Z]=0.$$

Unter Benutzung von $\eta^3=1$ erhält man, dass die Koeffizienten von A, B und C verschwinden und daher $X+\eta Y+\eta^2 Z=0$ gilt. Wegen eines Lemmas im Abschnitt 6 über den Satz von Morley ist dies gleichbedeutend damit, dass $\triangle XYZ$ ein gleichseitiges, positiv orientiertes Dreieck ist. Damit ist der Satz von Napoleon bewiesen. □

Es gibt nicht nur einen Satz von Napoleon, sondern auch mindestens zwei *Probleme von Napoleon* bzw. *Napoleonische Probleme*. Bei dem ersten, dem leichteren der beiden Probleme, sind auf einem gegebenen Kreis um den Punkt O alleine mit Hilfe eines Zirkels vier Punkte A, B, C, D zu finden derart, dass $ABCD$ ein Quadrat bilden. Für die **Lösung** mache man die folgende Konstruktion, siehe Abbildung 37.

- Man wähle einen beliebigen Punkt A auf dem Kreis. Trage mit dem Radius des Kreises von A aus die Punkte F, G und C ab.
- Sei E Schnittpunkt der Kreise um A und C jeweils mit dem Radius $|AG|$.
- Es ist $|OE|$ die Seitenlänge des gesuchten Quadrates. Als Schnittpunkte des gegebenen Kreises und des Kreies um A mit dem Radius $|OE|$ gewinne man also die beiden übrigen Punkte B und D.

Nun muss die Konstruktion gerechtfertigt werden. Hierzu muss gezeigt werden, dass $|OE|=\sqrt{2}|AO|$. Es ist $|AG|=\sqrt{3}|AO|$. Wegen des Satzes von Pythagoras (angewandt auf $\triangle COE$) ist

$$|OE|=\sqrt{|EC|^2-|OC|^2}=\sqrt{|AC|^2-|AO|^2}=\sqrt{2}|AO|.$$

Das war zu zeigen.

Bei dem zweiten, dem schwereren der beiden Napoleonischen Probleme, ist ein Kreis gegeben. Alleine mit einem Zirkel ist der Mittelpunkt des Kreises zu bestimmen. Als **Lösung** zu diesem Problem geben wir die folgende Konstruktion an. Hierbei sei $\mathcal{C}$ der (schwarze) Kreis, dessen Mittelpunkt zu bestimmen ist.

- Sei A ein Punkt auf $\mathcal{C}$. Bilde einen (roten) Kreis $\mathcal{C}_1$ mit A als Mittelpunkt (und einem Radius, der gewissen, später zu präzisierenden Bedingungen genügt), der $\mathcal{C}$ in den Punkten B_1 und B_2 trifft.

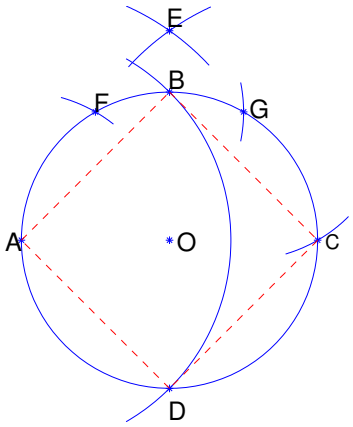


Abbildung 37: Lösung des ersten Napoleonischen Problems

- Um B_1 und B_2 schlage man jeweils einen (grünen) Kreis mit dem Radius $|AB_1| = |AB_2|$. Der von A verschiedene Schnittpunkt dieser beiden Kreise sei mit C bezeichnet.
- Mit C als Mittelpunkt schlage man einen (magentafarbenen) Kreis $\mathcal{C}_2$ mit dem Radius $|AC|$. Dieser schneide den (roten) Kreis $\mathcal{C}_1$ in den Punkten D_1 und D_2.
- Mit D_1 und D_2 als Mittelpunkt schlage man jeweils einen (blauen) Kreis mit dem Radius $|AD_1| = |AD_2|$. Diese schneiden sich in A und in O, dem gesuchten Zentrum des gegebenen Kreises.

Damit die Konstruktion durchführbar ist, darf der Radius des Kreises $\mathcal{C}_1$ weder zu klein noch zu groß sein. Ist r der Radius des gegebenen Kreises $\mathcal{C}$, so sollte der Radius r_1 von $\mathcal{C}_1$ aus $(r/2, 2r)$ sein. Denn wäre $r_1 \geq 2r$, so würde $\mathcal{C}_1$ den Kreis $\mathcal{C}$ entweder nicht schneiden ($r_1 > 2r$) oder die beiden Schnittpunkte B_1 und B_2 fallen zusammen ($r_1 = 2r$), so dass es nur einen grünen Kreis gibt. Ist $r_1 \leq r/2$, so schneidet entweder $\mathcal{C}_2$ nicht den Kreis $\mathcal{C}$ (für $r_1 < r/2$) oder die Schnittpunkte D_1 und D_2 des roten und des magentafarbenen Kreises fallen zusammen, so dass es nur einen blauen Kreis gibt.

In Abbildung 38 veranschaulichen wir uns die Konstruktion. Die Farbe, in der wir die bei der Konstruktion auftretenden Kreise zeichnen, haben wir oben schon angegeben. Jetzt haben wir zu rechtfertigen, dass die obige Konstruktion zum Ziel führt. Hierzu geben wir einen analytischen Beweis an und nehmen o. B. d. A. an, der Ausgangskreis sei durch

$$\mathcal{C} := \{(x, y) : x^2 + y^2 = 1\}$$

gegeben. Ferner sei, ebenfalls ohne Einschränkung der Allgemeinheit, $A := (0, -1)$ der vorgegebene Punkt auf $\mathcal{C}$. Mit $r_1 \in [\frac{1}{2}, 2]$ ist dann

$$\mathcal{C}_1 := \{(x, y) : x^2 + (y + 1)^2 = r_1^2\}.$$

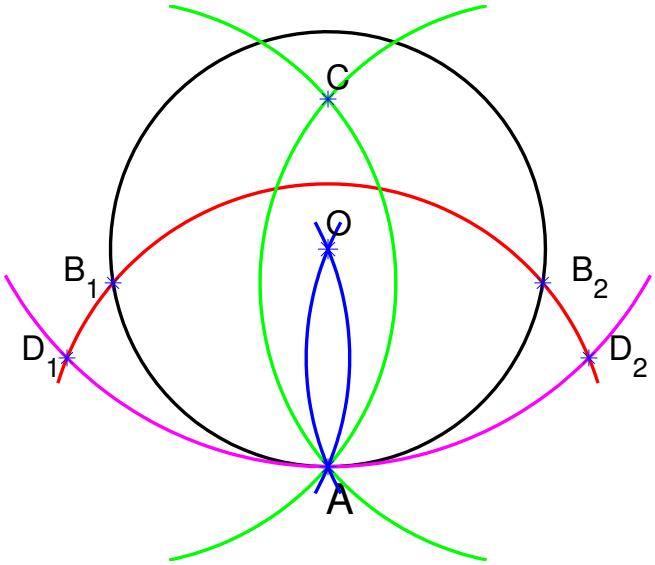


Abbildung 38: Lösung des zweiten Napoleonischen Problems

Die Schnittpunkte B_1 und B_2 der Kreise $\mathcal{C}$ und $\mathcal{C}_1$ sind dann mit

$$y_B := \frac{r_1^2}{2} - 1$$

gegeben durch

$$B_1 := (-\sqrt{1-y_B^2}, y_B), \qquad B_2 := (\sqrt{1-y_B^2}, y_B).$$

Der Radius $q_1 := |AB_1| = |AB_2|$ der beiden (grünen) Kreise um B_1 bzw. B_2 ist

$$q_1 = \sqrt{(1-y_B^2) + (y_B+1)^2} = \sqrt{2}\sqrt{y_B+1} = r_1.$$

Die Schnittpunkte der beiden Kreise sind daher die Lösungen der Gleichungen

$$(x + \sqrt{1-y_B^2})^2 + (y-y_B)^2 = r_1^2, \qquad (x - \sqrt{1-y_B^2})^2 + (y-y_B)^2 = r_1^2$$

und das sind $A = (0,-1)$ und $C := (0, y_C)$ mit $y_C := 2y_B + 1 = r_1^2 - 1$. Der Radius $r_2 := |AC|$ des (magentafarbenen) Kreises $\mathcal{C}_2$ ist

$$r_2 = y_C + 1 = 2(y_B+1) = r_1^2.$$

Die Schnittpunkte D_1 und D_2 des (roten) Kreises $\mathcal{C}_1$ und des (magentafarbenen) Kreises $\mathcal{C}_2$ sind die Lösungen der Gleichungen

$$x^2 + (y+1)^2 = r_1^2, \qquad x^2 + (y-y_C)^2 = r_2^2.$$

Dies Lösungen sind gegeben durch

$$D_1 := \Big(-\sqrt{r_1^2 - \frac{1}{4}}, -\frac{1}{2}\Big), \qquad D_2 := \Big(\sqrt{r_1^2 - \frac{1}{4}}, -\frac{1}{2}\Big).$$

Der Radius $q_2 := |AD_1| = |AD_2|$ der beiden (blauen) Kreise um D_1 bzw. D_2 ist $q_2 = r_1$. Die Schnittpunkte der beiden (blauen) Kreise sind daher Lösungen der Gleichungen

$$\Big(x + \sqrt{r_1^2 - \frac{1}{4}}\Big)^2 + \Big(y + \frac{1}{2}\Big)^2 = r_1^2, \qquad \Big(x - \sqrt{r_1^2 - \frac{1}{4}}\Big)^2 + \Big(y + \frac{1}{2}\Big)^2 = r_1^2.$$

Die Lösungen sind $A = (0, -1)$ und $O := (0, 0)$. Wir haben also den Mittelpunkt O des Kreises $\mathcal{C}$ alleine mit Zirkel-Konstruktionen gewonnen. In Abbildung 39 geben wir die beiden Extremsituationen wieder, nämlich einmal (links), dass der Radius von $\mathcal{C}_1$ die Hälfte des Radius von $\mathcal{C}$ ist, zum anderen (rechts), dass der Radius von $\mathcal{C}_1$ genau das Doppelte des Radius von $\mathcal{C}$ ist. Im ersten Fall gibt es nur einen blauen, im zweiten

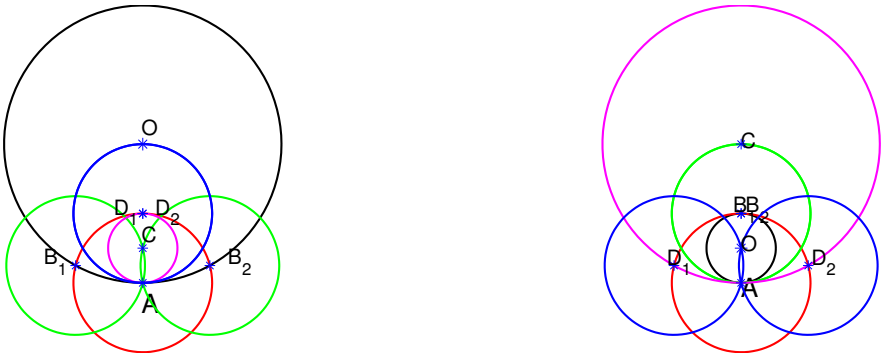


Abbildung 39: Extremsituationen bei der Lösung des zweiten Napoleonischen Problems

Fall nur einen grünen Kreis.

26 Das arithmetisch-geometrische Mittel

Das *arithmetische Mittel* zweier nichtnegativer Zahlen a und b ist $(a + b)/2$, das zugehörige *geometrische Mittel* ist $\sqrt{ab}$. Es gilt die Ungleichung vom geometrisch-arithmetischen Mittel:

$$\sqrt{ab} \le \frac{1}{2}(a + b),$$

deren Beweis man sofort aus

$$\Big(\frac{a+b}{2}\Big)^2 = ab + \Big(\frac{a-b}{2}\Big)^2 \ge ab$$

erhält. Eine geometrische Veranschaulichung haben wir in Abbildung 40 dargestellt.

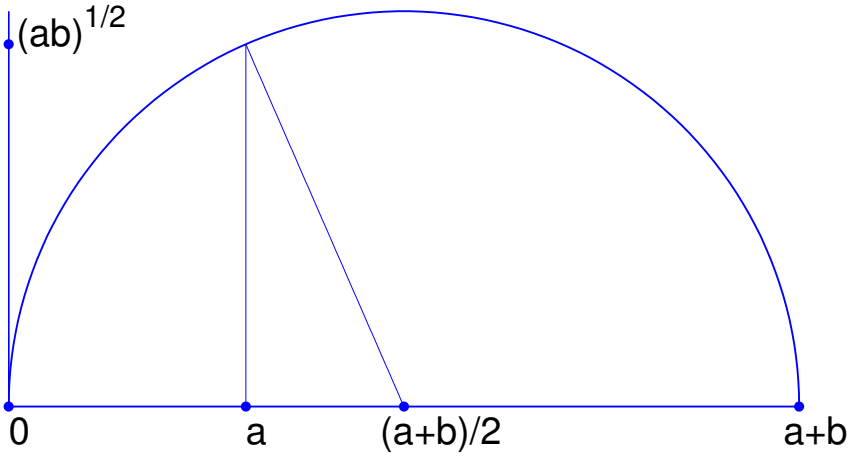


Abbildung 40: Das arithmetische und das geometrische Mittel

Das *arithmetisch-geometrische Mittel* $M(a,b)$ zweier nichtnegativer reeller Zahlen a, b ist eine Zahl, die zwischen dem geometrischen und dem arithmetischen Mittel von a, b liegt.

Definition Seien a, b zwei nichtnegative reelle Zahlen. Man setze $a_0 := a$, $b_0 := b$ und definiere die Folgen $\{a_k\}$ und $\{b_k\}$ durch

$$a_{k+1} := \frac{a_k + b_k}{2}, \qquad b_{k+1} := \sqrt{a_k b_k} \qquad (k = 0, 1, \ldots).$$

Der gemeinsame Grenzwert der beiden Folgen $\{a_k\}$ und $\{b_k\}$ heißt das *arithmetisch-geometrische Mittel* von a und b und wird mit $M(a,b)$ bezeichnet.

Bemerkung In der obigen Definition sind zwei Behauptungen enthalten, nämlich dass die Folgen $\{a_k\}$ und $\{b_k\}$ beide konvergent sind und dass die beiden Limiten übereinstimmen. Diese Behauptungen sind richtig. Denn wegen obiger Ungleichung vom geometrisch-arithmetischen Mittel ist $b_k \le a_k$, $k = 1, 2, \ldots$. Weiter ist

$$a_k - a_{k+1} = a_k - \frac{a_k + b_k}{2} = \frac{1}{2}(a_k - b_k) \ge 0, \qquad k = 1, 2, \ldots$$

und

$$b_{k+1} - b_k = \sqrt{a_k b_k} - b_k = \sqrt{b_k}(\sqrt{a_k} - \sqrt{b_k}) \ge 0, \qquad k = 1, 2, \ldots.$$

Also ist

$$b_1 \le b_2 \le \cdots \le b_k \le b_{k+1} \le \cdots \le a_{k+1} \le a_k \le \cdots \le a_2 \le a_1.$$

Damit sind $\{a_k\}$ bzw. $\{b_k\}$ monoton nicht wachsende bzw. nicht fallende, nach unten bzw. nach oben beschränkte Folgen und damit Konvergent. Die Limiten seien mit α bzw. β bezeichnet. Aus $a_{k+1} = (a_k + b_k)/2$ folgt mit $k \to \infty$, dass $\alpha = (\alpha + \beta)/2$ und damit $\alpha = \beta$. Damit ist nachgewiesen, dass das arithmetisch-geometrischen Mittel wohldefiniert ist.

Beispiel Wir berechnen $M(1, \sqrt{2})$ mit Hilfe von MuPAD, wobei wir `DIGITS:=30` setzen, und erhalten die folgenden Werte:

k	a_k	b_k
0	1.00000000000000000000000000000	1.41421356237309504880168872421
1	1.20710678118654752440084436210	1.18920711500272106671749997056
2	1.19815694809463429555917216633	1.19812352149312012260658557182
3	1.19814023479387720908287886908	1.19814023467730720579838378819
4	1.19814023473559220744063132863	1.19814023473559220744063132863
5	1.19814023473559220743992249228	1.19814023473559220743992249228
6	1.19814023473559220743992249228	1.19814023473559220743992249228

Also ist

$$M(1, \sqrt{2}) \approx 1.19814023473559220743992249228.$$

Die Konvergenz ist, wie man sieht, sehr gut. Allgemein liegt das daran, dass mit $c_k := \sqrt{a_k^2 - b_k^2}$ die Gleichungs-Ungleichungskette

$$c_{k+1} = \sqrt{a_{k+1}^2 - b_{k+1}^2} = \frac{1}{2}(a_k - b_k) = \frac{c_k^2}{4a_{k+1}} \leq \frac{c_k^2}{4M(a,b)}$$

gilt. Man spricht hier von *quadratischer Konvergenz*, ohne dass wir dies hier näher ausführen wollen.

Im Alter von 22 Jahren entdeckte Carl Friedrich Gauß die Beziehung

$$\frac{\pi}{2M(1, \sqrt{2})} = \int_0^1 \frac{dt}{\sqrt{1 - t^4}},$$

wobei Gauß durch numerische Berechnungen beobachtete, dass beide Ausdrück bis auf 11 Dezimalen übereinstimmen. Die Gültigkeit dieser Identität wird eine Folgerung aus dem nächsten Satz sein.

Satz Für positive a, b gilt

$$T(a,b) := \frac{2}{\pi} \int_0^{\pi/2} \frac{d\theta}{\sqrt{a^2 \cos^2\theta + b^2 \sin^2\theta}} = \frac{1}{M(a,b)}.$$

Beweis: Der Trick beim Beweis besteht im Nachweis[37] von

$$(*) \qquad T(a,b) = T\Big(\frac{1}{2}(a+b), \sqrt{ab}\Big)$$

[37]Einen alternativen Beweis werden wir in einer Fußnote in Abschnitt 32 bringen.

(sogenannte *Landen-Transformation*). Durch wiederholte Anwendung dieser Beziehung und Grenzübergang erhält man

$$T(a,b) = T(M(a,b), M(a,b)) = \frac{1}{M(a,b)}$$

und das ist die Behauptung.

Zum Beweis von $(*)$ machen wir die Variablentransformation $t := b\tan\theta$. Dann ist

$$\cos^2\theta = \frac{b^2}{b^2+t^2}, \qquad \sin^2\theta = \frac{t^2}{b^2+t^2}, \qquad d\theta = \frac{b}{b^2+t^2}dt.$$

Folglich ist

$$T(a,b) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}}.$$

Nun ist

$$\int_{-\infty}^{+\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} = \int_{-\infty}^{0} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} + \int_{0}^{+\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}}.$$

Beim ersten Integral machen wir die Variablentransformation $t = x - C(x)$, beim zweiten die Transformation $t = x + C(x)$, wobei wir zur Abkürzung $C(x) := \sqrt{ab+x^2}$ setzen. Dann ist

$$\begin{aligned}
T(a,b) &= \frac{1}{\pi}\int_{-\infty}^{+\infty} \Big\{ \frac{1 - x/C(x)}{\sqrt{[a^2+(x-C(x))^2][b^2+(x-C(x))^2]}} \\
&\qquad + \frac{1 + x/C(x)}{\sqrt{[a^2+(x+C(x))^2][b^2+(x+C(x))^2]}} \Big\}\, dx \\
&= \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{1}{C(x)} \Big\{ \frac{C(x) - x}{\sqrt{[a^2+(C(x)-x)^2][b^2+(C(x)-x)^2]}} \\
&\qquad + \frac{C(x) + x}{\sqrt{[a^2+(C(x)+x)^2][b^2+(C(x)+x)^2]}} \Big\}\, dx \\
&= \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{1}{C(x)} \frac{2}{\sqrt{2(C(x)^2+x^2)+(a^2+b^2)}}\, dx \\
&\qquad \text{(Erweitere mit } C(x)+x \text{ bzw. } C(x)-x\text{)} \\
&= \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{dx}{\sqrt{[((a+b)/2)^2+x^2][\sqrt{ab}^{\,2}+x^2]}} \\
&= T\Big(\frac{1}{2}(a+b), \sqrt{ab}\Big),
\end{aligned}$$

womit die obige Beziehung $(*)$ und damit auch der Satz bewiesen ist. □

Bemerkung Aus dem letzten Satz erhält man mit dessen Bezeichnungen, dass für $x \in (0,1]$ gilt

$$\frac{1}{M(1,x)} = T(1,x)$$

$$
\begin{aligned}
&= \frac{2}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{\cos^2\theta + x^2\sin^2\theta}} \\
&= \frac{2}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{1-(1-x^2)\sin^2\theta}} \\
&= \frac{2}{\pi}K(\sqrt{1-x^2}),
\end{aligned}
$$

wobei

$$
K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}}
$$

das sogenannte *vollständige elliptische Integral erster Art zum Modul* $k \in (0,1]$ ist.

Bemerkung Hat man umgekehrt bei gegebenem $k \in [0,1)$ das vollständige elliptische Integral $K(k)$ zu berechnen, so kann man

$$
K(k) = \frac{\pi}{2M(1,\sqrt{1-k^2})}
$$

ausnutzen.

Bemerkung Mit Hilfe der positiven Homogenität (d. h. für positive reelle Zahlen a, b und $\lambda > 0$ ist $M(\lambda a, \lambda b) = \lambda M(a,b)$) sowie der Symmetrie (d. h. $M(a,b) = M(b,a)$ für positive reelle Zahlen a, b, was z.B. aus $M(a,b) = M((a+b)/2, \sqrt{ab})$ folgt) erhalten wir einen Beweis für die von Gauß numerisch beobachtete Identität:

$$
\begin{aligned}
\frac{1}{M(1,\sqrt{2})} &= \frac{1}{M(\sqrt{2},1)} \\
&= \frac{1}{\sqrt{2}M(1,1/\sqrt{2})} \\
&= \frac{\sqrt{2}}{\pi}K(1/\sqrt{2}) \\
&= \frac{\sqrt{2}}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{1-\frac{1}{2}\sin^2\theta}} \\
&= \frac{\sqrt{2}}{\pi}\int_0^{1} \frac{dt}{\sqrt{(1-t^2)(1-\frac{1}{2}t^2)}} \\
&\qquad \text{(Substitution } t = \sin\theta\text{)} \\
&= \frac{2}{\pi}\int_0^{1} \frac{dt}{\sqrt{(1-t^2)(2-t^2)}} \\
&= \frac{2}{\pi}\int_0^{1} \frac{dx}{\sqrt{1-x^4}}.
\end{aligned}
$$

Beim letzten Schritt haben wir die Substitution $x^2 = t^2/(2-t^2)$ bzw. $t^2 = 2x^2/(1+x^2)$ gemacht. Dann ist

$$
(1-t^2)(2-t^2) = \Big(1 - \frac{2x^2}{1+x^2}\Big)\Big(2 - \frac{2x^2}{1+x^2}\Big) = \frac{2(1-x^2)}{(1+x^2)^2}
$$

und

$$2t\,dt = \frac{4x}{(1+x^2)^2}\,dx$$

bzw.

$$dt = \frac{\sqrt{2}}{(1+x^2)^{3/2}}\,dx.$$

Also ist, wie behauptet,

$$\int_0^1 \frac{dt}{\sqrt{(1-t^2)(2-t^2)}} = \int_0^1 \frac{\sqrt{2}}{(1+x^2)^{3/2}} \cdot \frac{1+x^2}{\sqrt{2}\sqrt{1-x^2}}\,dx = \int_0^1 \frac{dx}{\sqrt{1-x^4}}.$$

Bemerkung Die Auslenkung $\phi(t)$ eines mathematischen Pendels der Länge l, welches zur Anfangszeit $t_0 = 0$ im Ruhezustand ist und eine Anfangsauslenkung $\phi_0 > 0$ besitzt, genügt der Anfangswertaufgabe (die einfache bzw. zweifache Ableitung von ϕ nach der Zeit t wird durch $\dot{\phi}$ bzw. $\ddot{\phi}$ bezeichnet)

$$\ddot{\phi} + \omega_0^2 \sin\phi = 0, \qquad \phi(0) = \phi_0, \quad \dot{\phi}(0) = 0$$

mit $\omega_0 := \sqrt{g/l}$, wobei g die Erdbeschleunigung ist. Man kann zeigen, dass $\phi(\cdot)$ periodisch mit einer Periode $T = T(\phi_0, \omega_0)$ ist, welche gegeben ist durch

$$T(\phi_0, \omega_0) := \frac{4}{\omega_0} \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}} = \frac{4}{\omega_0} K(k),$$

wobei $k := \sin\frac{1}{2}\phi_0$ und $K(k)$ wieder das vollständige elliptische Integral erster Art ist.

27 Die Ungleichung vom geometrisch-arithmetischen Mittel

Die Ungleichung vom geometrisch-arithmetischen Mittel wurde zu Beginn des Abschnitts 26 (für $n = 2$) schon erwähnt. Allgemein gilt

Ungleichung vom geometrisch-arithmetischen Mittel *Seien $a_1, \ldots, a_n$ positive reelle Zahlen. Dann gilt*

$$\Big(\prod_{k=1}^n a_k\Big)^{1/n} \le \frac{1}{n}\sum_{k=1}^n a_k,$$

die Ungleichung vom geometrisch-arithmetischen Mittel. Hierbei gilt Gleichheit genau dann, wenn alle a_k gleich sind.

Beweis: Wir folgen einem sehr schönen Induktionsbeweis, der von Cauchy stammen soll und den man in dem BUCH der Beweise von M. Aigner, G. M. Ziegler (2002, S. 128) findet. Wir nennen $P(n)$ die Aussage, dass die Ungleichung vom geometrisch-arithmetischen Mittel mit $n \in \mathbb{N}$ Termen gültig ist bzw. die Ungleichung

$$\prod_{k=1}^n a_k \le \Big(\frac{1}{n}\sum_{k=1}^n a_k\Big)^n$$

gilt, ferner hier Gleichheit genau dann gilt, wenn alle a_k gleich sind. Ein üblicher Induktionsbeweis zeigt nach der Induktionsverankerung bei $n = 1$ oder $n = 2$, dass aus $P(n)$ auch $P(n+1)$ folgt. Hier funktioniert der Induktionsbeweis anders. Die Aussage $P(2)$ ist richtig, da die Ungleichung vom geometrisch-arithmetischen Mittel für zwei positive, reelle Zahlen a_1, a_2 äquivalent zu $(a_1 - a_2)^2 \geq 0$ ist, Gleichheit gilt hier genau dann, wenn $a_1 = a_2$. Nun zeigen wir:

(A) Aus $P(n)$ folgt $P(n-1)$,

(B) Aus $P(n)$ und $P(2)$ folgt $P(2n)$.

Aus diesen beiden Aussagen erhält man offenbar $P(n)$ für alle $n \in \mathbb{N}$. Zum Nachweis von (A) setzen wir

$$A := \frac{1}{n-1} \sum_{k=1}^{n-1} a_k.$$

Wegen $P(n)$ ist

$$\begin{aligned} \Big(\prod_{k=1}^{n-1} a_k\Big) A &\leq \Big[\frac{1}{n}\Big(\sum_{k=1}^{n-1} a_k + A\Big)\Big]^n \\ &= \Big(\frac{(n-1)A + A}{n}\Big)^n \\ &= A^n \end{aligned}$$

und daher

$$\prod_{k=1}^{n-1} a_k \leq A^{n-1} = \Big(\frac{1}{n-1} \sum_{k=1}^{n-1} a_k\Big)^{n-1}.$$

Gilt hier Gleichheit, so gilt wegen $P(n)$, dass $a_k = A$, $k = 1, \ldots, n-1$. Dies ist $P(n-1)$. Zum Nachweis von (B) beachten wir, dass

$$\begin{aligned} \prod_{k=1}^{2n} a_k &= \Big(\prod_{k=1}^{n} a_k\Big)\Big(\prod_{k=n+1}^{2n} a_k\Big) \\ &\leq \Big(\frac{1}{n}\sum_{k=1}^{n} a_k\Big)^n \Big(\frac{1}{n}\sum_{k=n+1}^{2n} a_k\Big)^n \\ &= \Big[\Big(\frac{1}{n}\sum_{k=1}^{n} a_k\Big)\Big(\frac{1}{n}\sum_{k=n+1}^{2n} a_k\Big)\Big]^n \\ &\leq \Big(\frac{1}{2n}\sum_{k=1}^{2n} a_k\Big)^{2n}, \end{aligned}$$

wobei wir zunächst $P(n)$ und anschließend $P(2)$ benutzt haben. Gilt hier Gleichheit, so folgt wegen $P(n)$ zunächst, dass $a_1 = \cdots = a_n$ und $a_{n+1} = \cdots = a_{2n}$. Wegen $P(2)$ erhält man, dass alle a_k, $k = 1, \ldots, 2n$ gleich sind. Insgesamt ist auch (B) und damit der Satz bewiesen. □

Bemerkung: Es gibt viele Beweise der Ungleichung vom geometrisch-arithmetischen Mittel. Ein Beweis von George Polya benutzt lediglich die elementare Tatsache, dass $\exp(a) \geq 1 + a$ für alle $a \in \mathbb{R}$. Wir bezeichnen mit $\overline{a}_{\text{arithm}}$ das arithmetische Mittel der nichtnegativen $a_1, \ldots, a_n$ und nehmen $\overline{a}_{\text{arithm}} > 0$ an. Multipliziert man die n Ungleichungen $\exp(a_k/\overline{a}_{\text{arithm}} - 1) \geq a_k/\overline{a}_{\text{arithm}}$, $k = 1, \ldots, n$, miteinander, so erhält man

$$\underbrace{\exp\Big(\sum_{k=1}^{n} a_k/\overline{a}_{\text{arithm}} - n\Big)}_{=1} \geq \prod_{k=1}^{n} (a_k/\overline{a}_{\text{arithm}})$$

und hieraus folgt sofort die Ungleichung vom geometrisch-arithmetischen Mittel. □

28 Die Keplerschen Gesetze und ihre Herleitung

Das Newtonsche Gravitationsgesetz sagt aus, dass zwei (punktförmige) Objekte aufeinander eine Kraft ausüben, deren Länge (die Kraft ist ein Vektor) linear jeweils von ihrer Masse und umgekehrt proportional vom Quadrat ihrer Entfernung abhängt. Im Folgenden sei das eine Objekt die Sonne mit der Masse M, das andere Objekt ein gegebener Planet mit der Masse m. Der Planet bewegt sich in einer Ebene, als dessen Nullpunkt die (unbewegliche) Sonne genommen wird. Die Bewegungsgleichungen sind dann

$$m\ddot{x} = -\frac{\gamma m M}{\|x\|^3} x.$$

Hierbei ist $x(t) = (x_1(t), x_2(t))$ der Ort des Planeten zur Zeit t und γ die Gravitationskonstante. Die Bewegungsgleichung ist also eigentlich ein System von zwei Differentialgleichungen zweiter Ordnung, nämlich

$$(*) \qquad \begin{aligned} \ddot{x}_1 &= -\frac{\gamma M}{(x_1^2 + x_2^2)^{3/2}} x_1, \\ \ddot{x}_2 &= -\frac{\gamma M}{(x_1^2 + x_2^2)^{3/2}} x_2. \end{aligned}$$

Das Zweikörperproblem besteht darin, die Bahn des Planeten zu beschreiben. Die Lösung dieses Problems durch J. Kepler (1571-1630) gehört sicherlich zu den größten Leistungen in der Geschichte der Menschheit. Aus den Bewegungsgleichungen $(*)$ wollen wir die drei Keplerschen Gesetze[38] der Planetenbewegung ableiten, wobei wir W. Walter (1990) folgen. Dies sind bekanntlich:

(K1) Die Bahnen der Planeten sind Ellipsen, in deren einem Brennpunkt die Sonne steht.

(K2) Der von der Sonne zu einem Planeten weisende Radiusvektor überstreicht in gleichen Zeiten gleiche Flächen.

[38] Es gibt im Internet einige Seiten, auf denen animierte Erläuterungen der Keplerschen Gesetze zu sehen sind. Man gebe z. B. bei Google *Keplersche Gesetze Animation* ein.

(K3) Das Verhältnis zwischen dem Quadrat der Umlaufzeit und dem Kubus der großen Achse (der Bahnellipse) ist für alle Planeten des Sonnensystems konstant.

Satz *Für Planeten mit der Bewegungsgleichung* $(*)$ *gelten die drei Keplerschen Gesetze.*

Beweis: Wir suchen eine Lösung der Anfangswertaufgabe

$$\begin{aligned} \ddot{x}_1 &= -\frac{\gamma M}{(x_1^2+x_2^2)^{3/2}}x_1, \qquad x_1(0) = R, \qquad \dot{x}_1(0) = v_1, \\ \ddot{x}_2 &= -\frac{\gamma M}{(x_1^2+x_2^2)^{3/2}}x_2, \qquad x_2(0) = 0, \qquad \dot{x}_2(0) = v_2, \end{aligned}$$

die sich darstellen lässt in der Form

$$x_1(t) = r(t)\cos\phi(t), \qquad x_2(t) = r(t)\sin\phi(t),$$

wobei wir voraussetzen, dass $R > 0$ der Abstand des Planeten von der Sonne zur Zeit $t_0 = 0$ ist. Ferner wird $v_2 \neq 0$ vorausgesetzt (andernfalls wäre $x_2(t) = 0$, die Bewegung des Planeten würde auf einer Geraden erfolgen). Benutzt man die komplexe Schreibweise $z(t) = r(t)e^{i\phi(t)} = x_1(t) + ix_2(t)$, so ist

$$\dot{z} = (\dot{r} + ir\dot{\phi})e^{i\phi}, \qquad \ddot{z} = (\ddot{r} + 2i\dot{r}\dot{\phi} + ir\ddot{\phi} - r\dot{\phi}^2)e^{i\phi}.$$

Die Bewegungsgleichungen sind also äquivalent zu

$$\ddot{r} + 2i\dot{r}\dot{\phi} + ir\ddot{\phi} - r\dot{\phi}^2 = -\frac{\gamma M}{r^2}.$$

Zerlegt man in Real- und Imaginärteil, so ergeben sich die beiden folgenden, mit $(*)$ äquivalenten Gleichungen:

$$(1) \qquad \ddot{r} - r\dot{\phi}^2 + \frac{\gamma M}{r^2} = 0, \qquad 2\dot{r}\dot{\phi} + r\ddot{\phi} = 0.$$

Die Anfangsbedingungen sind gegeben durch

$$(2) \qquad r(0) = R, \qquad \phi(0) = 0, \qquad \dot{r}(0) = v_1, \qquad \dot{\phi}(0) = \frac{v_2}{R}.$$

Also ist die gegebene Anfangswertaufgabe zu (1), (2) äquivalent.

Zum Nachweis des ersten Keplerschen Gesetzes gehen wir folgendermaßen vor. Zunächst definieren wir mit noch unbekannten Konstanten $\epsilon \geq 0$, $p > 0$ und $0 \leq \alpha < 2\pi$ die Funktion f durch

$$f(\phi) := \frac{p}{1 + \epsilon\cos(\phi - \alpha)}.$$

Anschließend sei

$$t(\phi) := \frac{1}{A}\int_0^\phi f^2(\phi)\,d\phi \qquad \text{mit} \quad A := Rv_2,$$

also

$$\frac{dt(\phi)}{d\phi} = \frac{1}{A}f^2(\phi), \qquad t(0) = 0.$$

Mit $\phi = \phi(t)$ sei die Umkehrfunktion zu $t = t(\phi)$ bezeichnet und $r(t) := f(\phi(t))$ gesetzt. Aus

$$t = t(\phi(t)) = \frac{1}{A}\int_0^{\phi(t)} f^2(\phi)\, d\phi$$

erhält man durch Differentiation nach t, dass

$$\dot{\phi}(t) = \frac{A}{f^2(\phi(t))}.$$

Wir wollen uns überlegen, dass bei geeigneter Wahl der noch freien Konstanten ϵ, p, α durch (r, ϕ) eine Lösung von (1) und (2) gegeben ist. Am einfachsten ist die zweite Gleichung in (1) einzusehen. Es ist nämlich

$$r^2(t)\dot{\phi}(t) = f^2(\phi(t))\frac{A}{f^2(\phi(t))} = A,$$

insbesondere also auch

$$0 = \frac{d}{dt}[r^2(t)\dot{\phi}(t)] = r(t)[2\dot{r}(t)\dot{\phi}(t) + r(t)\ddot{\phi}(t)].$$

Wegen $r(t) > 0$ ist die zweite Gleichung in (1) erfüllt. Von den Anfangsbedingungen in (2) ist $\phi(0) = 0$ schon erfüllt (wegen $t(0) = 0$). Nun kommen wir zu der ersten Gleichung in (1). Zunächst folgt aus $r(t) := f(\phi(t))$, dass

$$\begin{aligned}
\dot{r}(t) &= f'(\phi(t))\dot{\phi}(t) \\
&= f'(\phi(t))\frac{A}{f^2(\phi(t))} \\
&= \frac{p\epsilon\sin(\phi(t)-\alpha)}{(1+\epsilon\cos(\phi(t)-\alpha))^2}\cdot\frac{A(1+\epsilon\cos(\phi(t)-\alpha))^2}{p^2} \\
&= \frac{\epsilon A}{p}\sin(\phi(t)-\alpha),
\end{aligned}$$

anschließend

$$\ddot{r}(t) = \frac{\epsilon A}{p}\cos(\phi(t)-\alpha)\dot{\phi}(t).$$

Folglich ist

$$\begin{aligned}
r^2(t)\Big[\ddot{r}(t) - r(t)\dot{\phi}^2(t) + \frac{\gamma M}{r^2(t)}\Big] &= r^2(t)\Big[\frac{\epsilon A}{p}\cos(\phi(t)-\alpha)\dot{\phi}(t) - \frac{A^2}{r^3(t)} + \frac{\gamma M}{r^2(t)}\Big] \\
&= \frac{\epsilon A^2}{p}\cos(\phi(t)-\alpha) - \frac{A^2}{r(t)} + \gamma M \\
&= \frac{\epsilon A^2}{p}\cos(\phi(t)-\alpha) - \frac{A^2}{p}[1+\epsilon\cos(\phi(t)-\alpha)] + \gamma M \\
&= \gamma M - \frac{A^2}{p}.
\end{aligned}$$

Setzt man also

$$p := \frac{A^2}{\gamma M} = \frac{R^2 v_2^2}{\gamma M},$$

so erfüllt (r, ϕ) auch die erste Gleichung in (1). Die Konstanten ϵ und α werden durch die Anfangsbedingungen $r(0) = R$ und $\dot{r}(0) = v_1$ festgelegt. Dies führt auf die Gleichungen

$$R = \frac{p}{1 + \epsilon \cos \alpha}, \qquad v_1 = -\frac{\epsilon A}{p} \sin \alpha$$

bzw. nach Einsetzen von p auf

$$\epsilon \cos \alpha = \frac{R v_2^2 - \gamma M}{\gamma M}, \qquad \epsilon \sin \alpha = -\frac{R v_1 v_2}{\gamma M}.$$

Diese beiden Gleichungen sind lösbar, da $\epsilon \geq 0$ und $\alpha \in [0, 2\pi)$ als Polarkoordinaten des Punktes

$$Q := \frac{1}{\gamma M}(R v_2^2 - \gamma M, -R v_1 v_2)$$

bestimmt werden können. Für $Q \neq 0$ sind $\epsilon \geq 0$ und $\alpha \in [0, 2\pi)$ sogar eindeutig festgelegt. Die letzte Anfangsbedingung ist ebenfalls erfüllt:

$$\dot{\phi}(0) = \frac{A}{f^2(\phi(0))} = \frac{A}{r^2(0)} = \frac{R v_2}{R^2} = \frac{v_2}{R}.$$

Damit ist gezeigt: Die Anfangswertaufgabe (1), (2) besitzt eine Lösung (r, ϕ) mit

$$r(t) = \frac{p}{1 + \epsilon \cos(\phi(t) - \alpha)},$$

wobei $p > 0$, $\epsilon \geq 0$ und $\alpha \in [0, 2\pi)$ geeignete Konstanten sind.

Durch

$$K := \{(f(\phi) \cos \phi, f(\phi) \sin \phi) : \phi \in [0, 2\pi]\}$$

mit

$$f(\phi) := \frac{p}{1 + \epsilon \cos(\phi - \alpha)}$$

ist ein Kegelschnitt gegeben und zwar eine Ellipse ($\epsilon < 1$), eine Parabel ($\epsilon = 1$) oder eine Hyperbel ($\epsilon > 1$). Wir gehen jetzt davon aus, dass die Anfangsdaten so vernünftig sind, dass es sich bei der Planetenbahn um eine Ellipse handelt, da sie ja geschlossen ist. Damit ist das erste Keplersche Gesetz bewiesen. Die Ellipse habe die Halbachsen $a \geq b$. Diese lassen sich aus ϵ und p berechnen und man erhält

$$a = \frac{p}{1 - \epsilon^2}, \qquad b = \frac{p}{\sqrt{1 - \epsilon^2}}.$$

Nun zum zweiten Keplerschen Gesetz. Man bezeichne mit $F(t_1, t_2)$ die Größe der vom Fahrstrahl für $t_1 \leq t \leq t_2$ überstrichenen Fläche, also den Flächeninhalt des von den Strahlen $\phi = \phi(t_1)$, $\phi = \phi(t_2)$ und der Kurve $f(\phi)e^{i\phi}$, $\phi(t_1) \leq \phi \leq \phi(t_2)$, begrenzten Gebietes

$$S_{1,2} := \{(r \cos \phi, r \sin \phi) : 0 \leq r \leq f(\phi),\ \phi(t_1) \leq \phi \leq \phi(t_2)\}.$$

Dann ist

$$F(t_1,t_2)=\frac{1}{2}\int_{\phi(t_1)}^{\phi(t_2)} f^2(\phi)\,d\phi=\frac{1}{2}\int_{t_1}^{t_2} f^2(\phi(t))\dot{\phi}(t)\,dt=\frac{1}{2}\int_{t_1}^{t_2} r^2(t)\dot{\phi}(t)\,dt=\frac{A}{2}(t_2-t_1).$$

Hierbei erhält man die erste Gleichung (Leibnizsche Sektorformel), siehe W. Walter (1990, S. 251), z. B. aus der Transformationsformel für mehrfache Integrale. Damit ist auch das zweite Keplersche Gesetz bewiesen.

Nun sei T die Umlaufzeit des Planeten, also $\phi(T)=2\pi$. Dann ist $F(0,T)=\frac{1}{2}AT$ die Fläche der Ellipse, die andererseits bekanntlich πab beträgt. Folglich ist $T=2\pi ab/A$ und daher

$$T^2=\frac{4\pi^2a^2b^2}{A^2}=\frac{4\pi^2a^2b^2}{p\gamma M}=\frac{4\pi^2a^2b^2}{(b^2/a)\gamma M}=\frac{4\pi^2}{\gamma M}a^3.$$

Damit ist auch das dritte Keplersche Gesetz bewiesen. □

Beispiel Wir wollen die Ellipsenbahn berechnen, auf der die Lösung $(x_1(t),x_2(t))$ der Anfangswertaufgabe

$$(*)\qquad \begin{array}{rclcrclcrcl} \ddot{x}_1 &=& -\dfrac{10}{(x_1^2+x_2^2)^{3/2}}x_1, & & x_1(0) &=& 3, & & \dot{x}_1(0) &=& 1,\\ \ddot{x}_2 &=& -\dfrac{10}{(x_1^2+x_2^2)^{3/2}}x_2, & & x_2(0) &=& 0, & & \dot{x}_2(0) &=& 1 \end{array}$$

liegt. In der Darstellung der Ellipse

$$K:=\{(f(\phi)\cos\phi,f(\phi)\sin\phi):0\le\phi\le 2\pi\},\qquad f(\phi):=\frac{p}{1+\epsilon\cos(\phi-\alpha)},$$

erhalten wir (siehe den Beweis der Keplerschen Gesetze) $p=0.9$, während $\epsilon\ge 0$ und $\alpha\in[0,2\pi)$ aus

$$\epsilon\cos\alpha=-0.7,\qquad \epsilon\sin\alpha=-0.3$$

zu berechnen sind. Hieraus erhält man $\epsilon=\sqrt{0.58}\approx 0.7615773106$ und danach α aus $\cos\alpha=-0.7/\epsilon\approx -0.9$ und $\sin\alpha=-0.3/\epsilon\approx -0.4$, was $\alpha=2\pi-\arccos(-0.7/\epsilon)\approx 3.546484440$ ergibt. In Abbildung 41 geben wir die so berechnete Ellipse K an, auf der der Planet, welcher der Anfangswertaufgabe $(*)$ genügt, sich bewegt. Man kann natürlich die Bewegungsgleichungen $(*)$ auch, z. B. mit einem System wie MATLAB oder Maple, numerisch lösen und die erhaltene Lösung in einer (x_1,x_2)-Ebene plotten. Tut man dies, so wird man Übereinstimmung mit dem in Abbildung 41 gefundenen Ergebnis feststellen.

29 Der Heiratssatz

Wir stellen das folgende Problem, das durch den sogenannten *Heiratssatz* gelöst wird:

- Gegeben sei eine Menge $U=\{u_1,\dots,u_m\}$ von Damen und eine Menge $W=\{w_1,\dots,w_n\}$ von Herren. Wir sagen, ein Paar $(u,w)\in U\times W$ (ein Paar besteht also ganz konventionell aus einer Dame und einem Herrn) sei *befreundet*, wenn

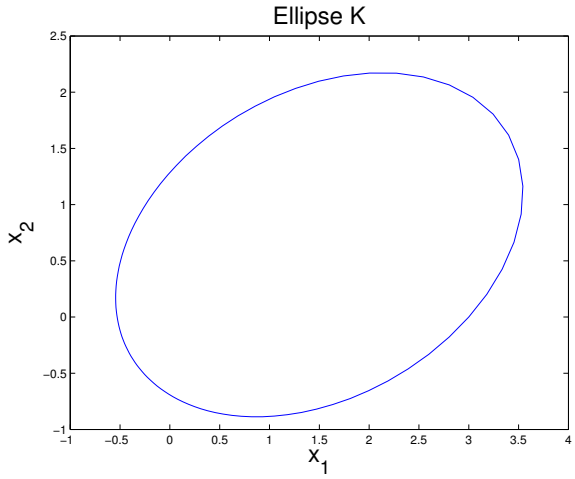


Abbildung 41: Exakte Planetenbahn beim Zweikörperproblem

beide einer gegenseitigen langfristigen Beziehung, z. B. einer Heirat, zustimmen würden[39]. Unter welchen Bedingungen gibt es zu jeder Dame $u \in U$ einen Herren $w \in W$ derart, dass das Paar (u, w) befreundet ist? Hierbei soll natürlich Bigamie ausgeschlossen werden, d. h. jeder Herr darf höchstens eine Dame als Partnerin und jede Dame höchstens einen Herrn als Partner erhalten.

Beispiel: In Abbildung 42 geben wir für fünf Damen und sechs Herren die zugehörigen

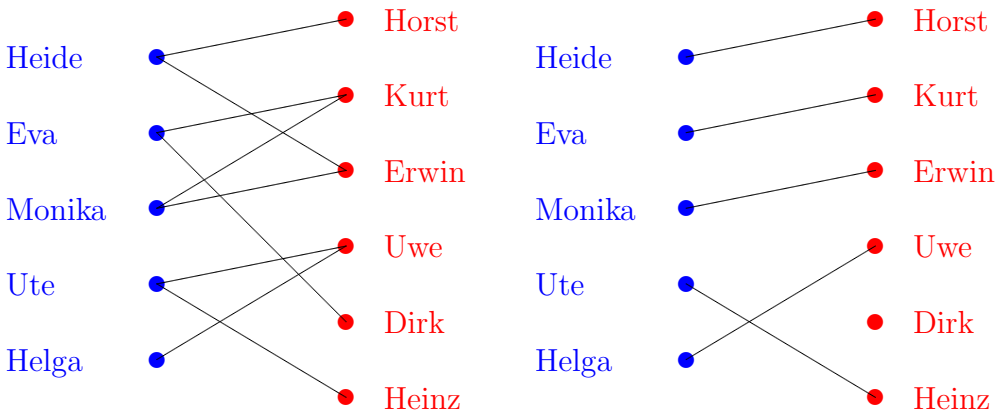


Abbildung 42: Befreundete Paare und ein zulässiger Heiratsplan

Freundschaftsbeziehungen sowie einen zulässigen Heiratsplan an. □

Ein nur scheinbar anderes Problem ist das folgende:

- Einem Arbeitsamt liegen Stellenangebote und Stellungsgesuche vor. Dabei haben einige Arbeitssuchende mehrere Berufe angegeben, für die sie qualifiziert sind. Es

[39] Hierbei kann es zu einer Dame $u \in U$ durchaus mehr als einen Herren geben, mit dem sie eine langfristige Beziehung eingehen könnte. Entsprechendes gilt natürlich auch für die Herren.

sollen möglichst viele Jobs an Arbeitssuchende vermittelt werden und bestimmt werden, unter welchen Voraussetzungen jeder Arbeitssuchende einen Job erhalten kann. Natürlich kann für jedes Stellenangebot nur ein Bewerber angenommen werden, und jeder Arbeitssuchende kann nur einen Job ausführen.

Eine Formulierung dieser Probleme als eine graphentheoretische Aufgabe ist einfach. Hierzu führen wir einen *bipartiten* Graphen $G = (V, E)$ mit der Eckenmenge $V := U \cup W$ der Damen und Herren bzw. der Arbeitssuchenden und der Jobs ein. Wenn ein Paar $(u, w) \in U \times W$ befreundet bzw. der Arbeitssuchende u für den Job w qualifiziert ist, so sei $e = uw$ eine Kante im Graphen[40]. Ein „zulässiger Heiratsplan“ bzw. ein „zulässiger Vermittlungsplan“ (gleich werden wir dies ein *Matching* nennen) ist eine Teilmenge $F \subset E$ der Kantenmenge mit der Eigenschaft, dass zwei verschiedene Elemente von F keine Ecke gemeinsam haben, es also zu keiner Bigamie kommt bzw. jedem Arbeitssuchendem höchstens ein Job vermittelt wird und jedem Job höchstens ein Arbeitssuchender zugeteilt wird. Gesucht ist ein zulässiger Plan mit einer maximalen Anzahl von Kanten, also mit möglichst vielen Heiraten bzw. Vermittlungen.

Allgemein, d. h. nicht nur für bipartite Graphen, definieren wir:

Definition Ein *Matching* in einem Graphen $G = (V, E)$ ist eine Kantenmenge $F \subset E$ mit der Eigenschaft, dass zwei verschiedene Elemente von F keine Ecke gemeinsam haben. Die Anzahl der Kanten in einem maximal großen Matching, also die Zahl

$$m(G) := \max_{F \subset E \text{ Matching}} |F|,$$

heißt *Matching-Zahl* von G. Ein Matching F heißt ein *Maximum-Matching*, wenn $|F| = m(G)$.

Bemerkung: Ein Maximum-Matching sollte nicht mit einem *maximalen Matching*, einem Matching, das durch Hinzufügen einer Kante kein Matching mehr ist, verwechselt werden. In Abbildung 43 geben wir einen Graphen, in Abbildung 44 links ist ein

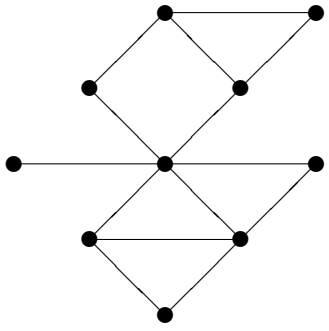


Abbildung 43: Ein Graph

zugehöriges maximales Matching und in Abbildung 44 rechts ein Maximum-Matching zu sehen. □

[40]Bei einem bipartiten Graphen ist die Eckenmenge in zwei nichtleere, disjunkte Eckenmengen partitioniert und jede Kante verbindet zwei Ecken aus unterschiedlichen Eckenmengen.

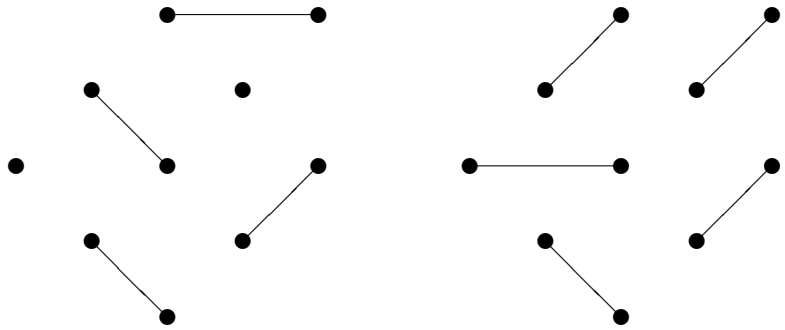


Abbildung 44: Ein maximales Matching und ein Maximum-Matching

Den auf Philip Hall (1935) zurückgehenden *Heiratssatz* geben wir in zwei offensichtlich äquivalenten Versionen an. Einmal in einer graphentheoretischen Version, zum anderen einer auf H. Weyl zurückgehenden Version, für die sich der Name Heiratssatz unmittelbar erschließt.

Satz *Gegeben sei ein bipartiter Graph $G = (U \cup W, E)$. Für $T \subset U$ sei*

$$N(T) := \{w \in W : \text{ Es existiert ein } u \in T \text{ mit } uw \in E\}.$$

(Die Menge $N(T)$ kann als Menge der Nachbarn von T bzw. der Menge der Heiratskandidaten der Damen aus T aufgefasst werden.) Dann ist $m(G) = |U|$ genau dann, wenn $|T| \leq |N(T)|$ für alle $T \subset U$.

Heiratssatz *Gegeben sei eine Menge U von m Damen sowie eine Menge V von n Herren. Von jeder beliebigen Dame und jedem beliebigen Herrn sei bekannt, ob sie miteinander befreundet sind. Eine Verheiratung aller m Damen mit befreundeten Herren (und zwar so, dass keine Bigamie eintritt) ist genau dann möglich, wenn die sogenannte Partybedingung erfüllt ist, d. h. je k Damen aus U mit mindestens k Herren aus V befreundet sind, $k = 1, \ldots, m$, also bei jeder Party mit k Damen kein Mangel an befreundeten Herren auftritt.*

Beweis: Der folgende Beweis findet sich bei P. R. Halmos, H. E. Vaughan (1950), siehe auch W. Maak (1950, S. 234) oder das schon angegebene Buch von F. Wille (1982, S. 12–13). Die Notwendigkeit der Partybedingung ist klar. Denn gibt es k Damen (mit $k \in \{1, \ldots, m\}$), die mit weniger als k Herren befreundet sind, so können diese Damen nicht alle verheiratet werden. Zum Beweis, dass die Partybedingung auch hinreichend für die Verheiratung aller m Damen ist, wenden wir vollständige Induktion nach m, der Anzahl der Damen an. Für $m = 1$ ist die Aussage trivial. Die Aussage des Heiratssatzes sei für $m - 1$ Damen richtig, das ist die Induktionsannahme. Nun seien m Damen zu verheiraten. Wenn je k Damen, $k \in \{1, \ldots, m-1\}$, mehr als k (weniger als k können es wegen der Partybedingung nicht sein), also mindestens $k+1$ Freunde haben, so verheirate man irgendeine von ihnen mit einem ihrer Freunde. Es bleiben $m-1$ Damen übrig und je k von ihnen haben mindestens k Freunde (der schon vergebene Herr zählt natürlich nicht mehr als Freund). Daher können diese $m-1$ Damen nach Induktionsvoraussetzung verheiratet werden. Es bleibt der Fall zu betrachten, dass für ein gewisses $k \in \{1, \ldots, m-1\}$ es k Damen gibt, die mit genau

k Herren befreundet sind. Nach Induktionsvoraussetzung können diese k Damen mit ihren Freunden verheiratet werden. Die noch übrigen $m - k$ ledigen Damen erfüllen wieder die Partybedingung. Denn wären gewisse h von ihnen nur mit weniger als h unverheirateten Männern befreundet, so wären diese h ledigen Damen zusammen mit den k (inzwischen) verheirateten (oder wenigstens versprochenen) Damen, also insgesamt $k + h$ Damen, mit weiger als $k + h$ Herren befreundet, was der vorausgesetzten Partybedingung des Satzes widerspricht. Deshalb können die restlichen $m - k$ ledigen Damen auch noch verheiratet werden und der Heiratssatz ist bewiesen. □

Beispiel: In Abbildung 45 links geben wir einen bipartiten Graphen, daneben ein

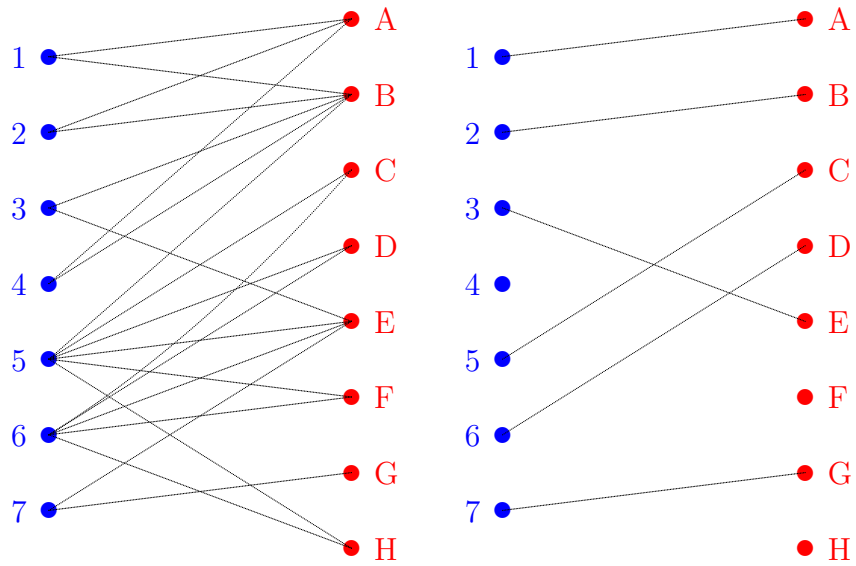


Abbildung 45: Ein bipartiter Graph und ein Maximum-Matching

zugehöriges Maximum-Matching an. Es ist kein Wunder, dass nicht alle sieben Damen verheiratet werden können. Denn die Voraussetzung des Heiratssatzes ist nicht erfüllt, da die drei Damen 1, 2 und 4 nur mit den zwei Herren A und B befreundet sind. □

In einem Korollar zum Heiratssatz wird die Matching-Zahl in einem bipartiten Graphen, also die maximale Anzahl der Damen, die verheiratet werden können, angegeben.

Korollar *Sei der bipartite Graph $G = (U \cup W, E)$ gegeben. Für $T \subset U$ sei wieder $N(T) \subset W$ die Menge der Nachbarn von T bzw. die Menge der Freunde der Damen aus T. Dann ist die Matching-Zahl gegeben durch*

$$m(G) = |U| - \max_{T \subset U}(|T| - |N(T)|).$$

Beweis: Sei $\delta^* := \max_{T \subset U}(|T| - |N(T)|)$. Es ist $\delta^* \geq 0$, da $\delta^* \geq |\emptyset| - |N(\emptyset)| = 0$. Ist $\delta^* = 0$, so ist die Partybedingung für die Menge U der Damen erfüllt, wegen des Heiratssatzes ist die Aussage des Korollars dann richtig. Daher nehmen wir im Folgenden an, es sei $\delta^* \geq 1$. Sei T^* eine Teilmenge von U mit $\delta^* = |T^*| - |N(T^*)|$. In T^* gibt es $\delta^* = |T^*| - |N(T^*)|$ Damen ohne geeigneten Partner, d. h. es ist $m(G) \leq |U| - \delta^*$.

Wir haben zu zeigen, dass hier Gleichheit gilt. Hierzu erweitern wir die Menge W durch δ^* weitere Herren, die wir uns in der Menge D zusammengefasst denken. Und zwar seien dies Herren, die für alle Damen in U als Heiratskandidaten in Frage kommen. Sei also D eine beliebige Menge mit $|D| = \delta^*$ und $W \cap D = \emptyset$. Man definiere den bipartiten Graphen $G^* := (U \cup (W \cup D), E^*)$, wobei $E^* := E \cup \{ud : u \in U, d \in D\}$. Für eine beliebige Teilmenge $T \subset U$ von Damen ist $N^*(T) = N(T) \cup D$ die Menge der Nachbarn von T im Graphen G^*, d. h. Freunde der Gruppe T von Damen bezüglich der erweiterten Gruppe von Herren sind alle Freunde aus W und die gesamte hinzugekommene Gruppe D von Herren. Für ein beliebiges $T \subset U$ ist daher $|N^*(T)| = |N(T)| + \delta^* \geq |T|$ nach Definition von δ^*, d. h. in G^* ist die Partybedingung erfüll. Daher gibt es wegen des Heiratssatzes ein Matching $F^* \subset E^*$ mit $|F^*| = |U|$ bzw. können alle Damen aus U mit Herren aus $W \cup D$ verheiratet werden. Aus dem Matching F^* in G^* erhalte man ein Matching F, indem man alle Kanten, die nach D führen, entfernt bzw. alle geplanten Hochzeiten mit Herren aus D annulliert. Deren Anzahl ist höchstens gleich der Anzahl der Elemente in D, also δ^*, und folglich $|F| \geq |U| - \delta^*$. Folglich ist $m(G) \geq |U| - \delta^*$ und insgesamt $m(G) = |U| - \delta^*$. Damit ist das Korollar zum Heiratssatz bewiesen. □

29.1 Der Satz von Birkhoff-von Neumann

Im Zusammenhang mit der Vervollständigung partieller lateinische Rechtecke bzw. Quadrate werden wir eine schöne Anwendung des Heiratssatzes in Unterabschnitt 65.2 kennenlernen. Hier wollen wir den Heiratssatz beim Beweis des Satzes von Birkhoff-von Neumann anwenden. Hierzu müssen zwei Begriffe eingeführt werden.

Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ mit nichtnegativen Einträgen $a_{ij} \geq 0$, $i, j = 1, \ldots, n$, heißt *doppelt stochastisch*, wenn $\sum_{j=1}^{n} a_{ij} = 1$, $i = 1, \ldots, n$, und $\sum_{i=1}^{n} a_{ij} = 1$, $j = 1, \ldots, n$, wenn also alle Zeilen- und alle Spaltensummen gleich 1 sind. Eine *Permutationsmatrix* ist eine doppelt stochastische Matrix mit Einträgen 0 und 1. Eine Permutationsmatrix ist also eine $n \times n$-Matrix, die in jeder Zeile und in jeder Spalte genau eine 1 als Eintrag enthält und sonst nur Nullen.

Satz (Birkhoff-von Neumann) *Eine Matrix $A \in \mathbb{R}^{n \times n}$ ist genau dann doppelt stochastisch, wenn sie eine Konvexkombination von Permutationsmatrizen ist, wenn es also $l \in \mathbb{N}$ und Permutationsmatrizen $P_1, \ldots, P_l$ sowie nichtnegative Koeffizienten $\lambda_1, \ldots, \lambda_l$ mit $\sum_{i=1}^{l} \lambda_i = 1$ und $A = \sum_{i=1}^{l} \lambda_i P_i$ gibt.*

Beweis: Eine Konvexkombination von Permutationsmatrizen ist doppelt stochastisch, wie man durch einfaches Nachrechnen bestätigt. Mathematisch interessant ist also nur die Umkehrung. Hierzu beweisen wir ein geringfügig allgemeineres Resultat:

- Eine nichtnegative Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, deren Zeilen- und Spaltensummen sämtlich gleich einer positiven Zahl γ sind, lässt sich darstellen in der Form

$$A = \sum_{i=1}^{l} \lambda_i P_i$$

mit Permutationsmatrizen $P_1, \ldots, P_l$ sowie $\lambda_1, \ldots, \lambda_l \geq 0$ mit $\sum_{i=1}^{l} \lambda_i = \gamma$.

Ist hier $\gamma = 1$, so hat man die Aussage des Satzes von Birkhoff-von Neumann. Der Beweis erfolgt durch vollständige Induktion nach der Anzahl m der von Null verschiedenen Einträge von A. Es ist $m \geq n$, da es wegen $\gamma > 0$ in jeder Zeile und jeder Spalte mindestens einen von Null verschiedenen Eintrag gibt. Ist $m = n$, so ist $A = \gamma P$ mit einer gewissen Permutationsmatrix P und die Behauptung ist in diesem Falle richtig. Wir nehmen an, die Aussage sei für Matrizen mit weniger als m von Null verschiedenen Einträgen richtig. Wir werden den Heiratssatz anwenden und sagen, dass eine Dame (Zeile) $i \in \{1, \ldots, n\}$ mit einem Herrn (Spalte) $j \in \{1, \ldots, n\}$ befreundet sei, wenn $a_{ij} > 0$. Wir überlegen uns, dass die Partybedingung erfüllt ist, also k Damen mit mindestens k Herren, $k = 1, \ldots, n$, befreundet sind. Denn wäre dies nicht der Fall, so würde es ein $k \in \{1, \ldots, n\}$ geben derart, dass in den zugehörigen k Zeilen von A alle von Null verschiedenen Einträge in höchstens $k-1$ Spalten von A auftreten. Die Summe der k Zeilensummen ist einerseits $k\gamma$, während eine Summation über die Spalten höchstens $(k-1)\gamma$ und damit einen Widerspruch ergibt. Folglich ist eine (zulässige) Verheiratung aller n Damen mit den n Herren möglich, d. h. es gibt paarweise verschiedene $j(1), \ldots, j(n) \in \{1, \ldots, n\}$ mit $a_{ij(i)} > 0$, $i = 1, \ldots, n$. Wir definieren die Permutationsmatrix $P_1 := (p_{ij})$ mit

$$p_{ij} := \begin{cases} 1, & j = j(i), \\ 0, & j \neq j(i), \end{cases}$$

und setzen

$$\lambda_1 := \min_{i=1,\ldots,n} a_{ij(i)}, \qquad A_1 := A - \lambda_1 P_1.$$

Dann ist $\lambda_1 > 0$ und A_1 eine nichtnegative Matrix, deren Zeilen- und Spaltensummen sämtlich gleich $\gamma_1 := \gamma - \lambda_1$ sind. Es ist $\gamma_1 \geq 0$ und $\gamma_1 = 0$ genau dann, wenn $A_1 = 0$ bzw. $A = \lambda_1 P_1$. O. B. A. ist also $\gamma_1 > 0$. Da A_1 weniger als m von Null verschiedene Einträge besitzt, können wir die Induktionsvoraussetzung anwenden. Hiernach lässt sich A_1 in der Form $A_1 = \sum_{i=2}^{l} \lambda_i P_i$ mit Permutationsmatrizen $P_2, \ldots, P_l$ und nichtnegativen $\lambda_2, \ldots, \lambda_l$ mit $\sum_{i=2}^{l} \lambda_i = \gamma_1$ darstellen. Wegen $A = \sum_{i=1} \lambda_i P_i$ und $\sum_{i=1}^{l} \lambda_i = \gamma$ ist der Induktionsschluss vollzogen und der Satz bewiesen. □

Die Menge

$$\mathcal{B}_n := \{A \in \mathbb{R}^{n \times n} : \ A \text{ ist doppelt stochastisch}\}$$

heißt das *Birkhoff-Polytop*. Dieses ist offensichtlich eine konvexe Menge. Unter einer *Ecke* von $\mathcal{B}_n$ verstehen wir ein Element $A \in \mathcal{B}_n$, welches sich nicht als echte Konvexkombination von zwei verschiedenen Elementen aus $\mathcal{B}_n$ darstellen lässt, bzw. für welches die Implikation

$$A_1, A_2 \in \mathcal{B}_n, \quad t \in (0,1), \quad A = (1-t)A_1 + tA_2 \Longrightarrow A = A_1 = A_2$$

gilt. Als Korollar zum Satz von Birkhoff-von Neumann formulieren wir:

Korollar *Ein $A \in \mathcal{B}_n$ ist genau dann eine Ecke von $\mathcal{B}_n$, wenn A eine Permutationsmatrix ist.*

Beweis: Sei $A \in \mathcal{B}_n$ eine Ecke von $\mathcal{B}_n$. Wegen des Satzes von Birkhoff-von Neumann lässt sich A in der Form $A = \sum_{i=1}^{l} \lambda_i P_i$ mit Permutationsmatrizen $P_1, \ldots, P_l$ und nichtnegativen $\lambda_1, \ldots, \lambda_l$ mit $\sum_{i=1}^{l} \lambda_i = 1$ darstellen. Wir können annehmen, dass $\lambda_l \in (0,1)$.

Denn ist z. B. $\lambda_l = 1$, so ist $A = P_l$ und A wie behauptet eine Permutationsmatrix, während ein Index l mit $\lambda_l = 0$ keinen Beitrag zur Darstellung von A liefert, also von vornherein hätte weggelassen werden können. Dann ist aber

$$A = (1 - \lambda_l) \underbrace{\sum_{i=1}^{l-1} \frac{\lambda_i}{1 - \lambda_l} P_i}_{\in \mathcal{B}_n} + \lambda_l P_l.$$

Da A eine Ecke des Birkhoff-Polytops $\mathcal{B}_n$ ist, ist $A = P_l$ eine Permutationsmatrix.

Sei nun umgekehrt $A \in \mathcal{B}_n$ eine Permutationsmatrix und $A = (1-t)A_1 + tA_2$ mit $A_1, A_2 \in \mathcal{B}_n$ und $t \in (0,1)$. Aus $(A)_{ij} = 0$ folgt, dass auch $(A_1)_{ij} = (A_2)_{ij} = 0$. Da A eine Permutationsmatrix ist und daher in jeder Zeile und jeder Spalte genau eine 1 als Eintrag und sonst nur Nullen besitzt, ist $A = A_1 = A_2$ und folglich A eine Ecke des Birkhoff-Polytops $\mathcal{B}_n$. Das Korollar zum Satz von Birkhoff-von Neumann ist hiermit bewiesen. □

Bemerkung: Mehr zum Birkhoff-Polytop findet man z. B. bei `http://en.wikipedia.org/wiki/Birkhoff_polytope`. Für $n \leq 10$ ist das Volumen des Birkhoff-Polytops $\mathcal{B}_n$ bekannt, siehe `http://oeis.org/A037302`. Für $n = 10$ ist das Ergebnis z. B. in `http://www.math.binghamton.edu/dennis/Birkhoff/` angegeben. □

30 Das Zuordnungsproblem

Beim *Zuordnungsproblem* sind zwei endliche Mengen mit gleich vielen Elementen gegeben, die in eineindeutiger (also bijektiver) Weise einander zugeordnet werden sollen und zwar so, dass dabei ein Gesamtgewinn maximiert bzw. die Gesamtkosten minimiert werden. Wir geben zunächst ein typisches Beispiel an:

- Es sollen n Arbeitern n Jobs zugeordnet werden, wobei nichtnegative Kosten c_{ij} anfallen, wenn der i-te Arbeiter den j-ten Job übernimmt. Unter einer *Zuordnung* verstehen wir, dass jeder Arbeiter genau einen Job erhält und jeder Job von genau einem Arbeiter übernommen wird. Gesucht ist eine Zuordnung, bei der die Gesamtkosten minimal sind.

Für eine mathematische Formulierung definieren wir die Variablen

$$x_{ij} := \begin{cases} 1, & \text{falls der } i\text{-te Arbeiter den } j\text{-ten Job erhält,} \\ 0, & \text{sonst.} \end{cases}$$

Durch die $n \times n$-Matrix $X = (x_{ij})$ ist aber nur dann eine Zuordnung gegeben, wenn X eine *Permutationsmatrix* ist, d. h. alle Einträge in X entweder Einsen oder Nullen sind und in jeder Zeile und jeder Spalte von X genau eine 1 und sonst nur Nullen stehen. Dies kann man auch durch

$$x_{ij} \in \{0, 1\} \qquad (i, j = 1, \ldots, n)$$

und

$$\sum_{j=1}^{n} x_{ij} = 1 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} x_{ij} = 1 \quad (j = 1, \ldots, n)$$

ausdrücken. Unter diesen Nebenbedingungen sind die Gesamtkosten

$$f(X) := \sum_{i,j=1}^{n} c_{ij} x_{ij}$$

zu minimieren.

Wir geben ein konkretes Beispiel an:

- Ein Beduine hat drei Töchter, für die es drei heiratswillige Bewerber gibt. Der erste Bewerber verlangt als Mitgift für die älteste Tochter 10 Kamele, für die mittlere 5 Kamele und für die jüngste 8 Kamele. Der zweite Bewerber verlangt als Mitgift: Für die älteste Tochter 5 Kamele, für die mittlere 6 und für die jüngste 7 Kamele. Schließlich verlangt der dritte Bewerber 5 Kamele für die älteste, 3 Kamele für die mittlere und 6 Kamele für die jüngste Tochter. Der Beduine will die Mitgift minimieren. Welchem Bewerber sollte er welche Tochter geben?

Es gibt aber auch Zuordnungsprobleme, bei denen eine Zielfunktion zu *maximieren* ist.

- Gegeben seien n heiratswillige Frauen und ebenso viele Männer. Die Frauen F_i, $i = 1, \ldots, n$, haben ein Sympathiemaß d_{ij} für eine Verbindung mit den Männern M_j, $j = 1, \ldots, n$, ermittelt. Gesucht ist eine Zuordnung mit einer maximalen „Sympathiesumme“, also eine Permutationsmatrix $X = (x_{ij})$ derart, dass

  $$g(X) := \sum_{i,j=1}^{n} d_{ij} x_{ij}$$

  maximal ist.

Ein solches Maximierungsproblem kann man auf ein Minimierungsproblem mit nichtnegativen Kosten zurückführen. Hierzu definiere man $D := \max_{1 \le i,j \le n} d_j$ und anschließend $c_{ij} := D - d_{ij}$, $1 \le i, j \le n$. Dann sind einerseits die Einträge c_{ij} nichtnegativ, andererseits ist mit einer $n \times n$-Permutationsmatrix $X = (x_{ij})$, also einer Zuordnung,

$$f(X) := \sum_{i,j=1}^{n} c_{ij} x_{ij} = \sum_{i,j=1}^{n} (D - d_{ij}) = D - \sum_{i,j=1}^{n} d_{ij} x_{ij} = D - g(X).$$

Dies bedeutet, dass das Maximieren von g bzw. das Minimieren von $-g$ gleichwertig zum Minimieren von f ist.

Unser Ziel besteht darin, die sogenannte *ungarische*[41] *Methode* zur Lösung des Zuordnungsproblems, $f(X) := \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}$ über alle $n \times n$-Permutationsmatrizen zu minimieren, darzustellen. Eine der Grundlagen des Verfahrens ist, dass sich die optimale Zuordnung bei gewissen Änderungen der Matrix $C = (c_{ij})$ nicht verändert. Genauer wird durch den folgenden Initialisierungsschritt zwar die Matrix C, nicht aber die optimale Zuordnung verändert:

[41] Die ungarischen Mathematiker Denés König und Jenő Egerváry haben wichtige Vorarbeiten geleistet.

- Subtrahiere $u_i := \min_{j=1,\ldots,n} c_{ij}$ von der i-ten Zeile in (c_{ij}) für $i = 1, \ldots, n$. Die neue Matrix werde wieder mit (c_{ij}) bezeichnet.
- Subtrahiere $v_j := \min_{i=1,\ldots,n} c_{ij}$ von der j-ten Spalte in (c_{ij}) für $j = 1, \ldots, n$. Die neue Matrix werde wieder mit (c_{ij}) bezeichnet.

Für eine $n \times n$-Permutationsmatrix $X = (x_{ij})$ ist

$$\sum_{i,j=1}^{n} (c_{ij} - u_i - v_j) x_{ij} = \sum_{i,j=1}^{n} c_{ij} x_{ij} - \sum_{i=1}^{n} u_i - \sum_{j=1}^{n} v_j,$$

daher bleibt durch die obige Modifikation der Kostenmatrix C die optimale Zuordnung unverändert.

Beispiel: Wie betrachten noch einmal das obige Beispiel, bei dem ein Beduine die zu zahlende Mitgift minimieren will. Als Modifikationen der Kostenmatrix erhalten wir

$$C := \begin{pmatrix} 10 & 5 & 8 \\ 5 & 6 & 7 \\ 5 & 3 & 6 \end{pmatrix} \longrightarrow \begin{pmatrix} 5 & 0 & 3 \\ 0 & 1 & 2 \\ 2 & 0 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 5 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}.$$

Wir kommen auf dieses Beispiel später wieder zurück. □

Wesentliches Hilfsmittel bei der Formulierung der ungarischen Methode ist die folgende Aussage, die sich als Korollar zu einem Korollar des *Heiratssatzes* (siehe Abschnitt 29) herausstellen wird. Hierbei verstehen wir unter einer *Linie* in einer Matrix eine Zeile oder eine Spalte in dieser Matrix.

Satz *Sei $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ eine Matrix mit nichtnegativen Einträgen c_{ij}, $1 \leq i, j \leq n$. Dann ist die Minimalzahl der alle Nullen enthaltenden Linien (wir sprechen von einer minimalen Bedeckung der Nullen) gleich der Maximalzahl der Nullen, von denen nicht zwei auf einer Linie liegen.*

Beispiel: In der oben angegebenen Matrix

$$C := \begin{pmatrix} 5 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}$$

sind alle Nullen in der zweiten Zeile und der zweiten Spalte enthalten. Die vier Nullen in C kann man offenbar nicht durch weniger als zwei Linien bedecken. Weiter liegen die Nullen in den Positionen $(1, 2)$ und $(2, 1)$ nicht auf einer Linie, mit jeder weiteren Null würden zwei Nullen auf einer Linie liegen. □

Beweis des Satzes: Wir definieren einen bipartiten Graphen $G = (U \cup W, E)$ durch die Eckenmengen $U := \{1, \ldots, n\}$, $W := \{1, \ldots, n\}$ sowie die Kantenmenge $E := \{(i, j) \in U \times W : c_{ij} = 0\}$. Hierbei wird U als Menge der Zeilen, W als Menge der Spalten in C interpretiert, so dass $U \cup W$ die Menge der Linien ist und Kanten die Positionen in der Matrix sind, in denen Nullen stehen. Ein Matching $F \subset E$ in G ist dann eine Teilmenge der Positionen von Nullen in C mit der Eigenschaft, dass zwei

verschiedene Positionen $(i,j) \neq (k,l)$ nicht auf einer Linie liegen, also $i \neq k$ (nicht auf einer Zeile) oder $j \neq l$ (nicht auf einer Spalte) gilt. Sind in $D \subset U \cup W$, also einer Teilmenge der Linien, alle Nullen in C enthalten, so ist trivialerweise $|F| \leq |D|$. Daher ist die Matching-Zahl $m(G)$ bzw. die Maximalzahl der Nullen, von denen nicht zwei auf einer Linie liegen, kleiner oder gleich der Minimalzahl von Linien, die alle Nullen in C enthalten bzw. bedecken. Andererseits existiert wegen des Korollars zum Heiratssatz in Abschnitt 29 eine Teilmenge $T \subset U$ mit

$$m(G) = |U| - (T - N(T)) = |U \setminus T| + N(T).$$

Hierbei ist $N(T)$ die Menge der Nachbarn von T, also

$$N(T) := \{w \in W : \text{ Es existiert } u \in T \text{ mit } uw \in E\}.$$

Man definiere die Menge der Linien $D := (U \setminus T) \cup N(T)$ und beachte, dass $|D| = |U| - |T| + |N(T)| = m(G)$. Wir überlegen uns, dass D alle Nullen in C bedeckt. Nehmen wir an, es sei $c_{ij} = 0$. Ist $i \in U \setminus T$, so wird die Null in der Position (i,j) durch die Zeile $i \in D$ bedeckt. Ist dagegen $i \in T$, so ist $j \in N(T) \subset D$ und die Null in Position (i,j) wird durch die Spalte $j \in D$ bedeckt. Daher ist die Matching-Zahl $m(G)$ größer oder gleich der Minimalzahl von Linien, die alle Nullen in C bedecken. Der Satz ist damit bewiesen. □

Nach Schilderung des Initialisierungsschrittes fahren wir mit der Beschreibung der ungarischen Methode fort. Hierbei nennen wir eine Menge von m Nulleinträgen in C, von denen nicht zwei auf einer Linie liegen, eine 0-*Diagonale der Länge* m.

- Man bestimme in C eine minimale Bedeckung der Nullen. Falls die minimale Bedeckung die Länge n hat, so ist durch eine maximale 0-Diagonale der Länge n, also n Null-Einträgen in C, die gesuchte Permutationsmatrix bzw. das Matching mit minimalen Kosten bestimmt.

Wir gehen daher jetzt davon aus, die minimale Bedeckung bestehe aus weniger als n Linien, etwa aus den Zeilen Z mit $z := |Z|$ und den Spalten S mit $s := |S|$, und es sei $z + s < n$. Sei d das Minimum der unbedeckten Einträge in C, also $d := \min_{i \notin Z, j \notin S} c_{ij}$. Es ist $d > 0$, denn nach Definition einer Bedeckung müssen *alle* Null-Einträge bedeckt sein. Wir ziehen zunächst d von den $n - z$ Zeilen von C ab, die nicht bedeckt sind. Anschließend addieren wir d zu den s Spalten der Bedeckung. Die neue Matrix nennen wir vorübergehend C'. Wir setzen also

$$c'_{ij} := \begin{cases} c_{ij} - d, & \text{falls } i \notin Z \text{ und } j \notin S, \\ c_{ij}, & \text{falls } i \in Z \text{ und } j \notin S \text{ oder } i \notin Z \text{ und } j \in S, \\ c_{ij} + d, & \text{falls } i \in Z \text{ und } j \in S. \end{cases}$$

Insbesondere ist auch (c'_{ij}) eine nichtnegative $n \times n$-Matrix. Für eine beliebige $n \times n$-Permutationsmatrix $X = (x_{ij})$ ist dann

$$\sum_{i,j=1}^{n} c'_{ij} x_{ij} \;=\; \sum_{i \in Z, j \in S} c'_{ij} x_{ij} + \sum_{i \notin Z, j \in S} c'_{ij} x_{ij}$$

$$
\begin{aligned}
&\quad + \sum_{i\in Z, j\notin S} c'_{ij}x_{ij} + \sum_{i\notin Z, j\notin S} c'_{ij}x_{ij} \\
&= \sum_{i\in Z, j\in S} (c_{ij}+d)x_{ij} + \sum_{i\notin Z, j\in S} c_{ij}x_{ij} \\
&\quad + \sum_{i\in Z, j\notin S} c_{ij}x_{ij} + \sum_{i\notin Z, j\notin S} (c_{ij}-d)x_{ij} \\
&= \sum_{i,j=1}^{n} c_{ij}x_{ij} + d\sum_{i\in Z, j\in S} x_{ij} - d\sum_{i\notin Z, j\notin S} x_{ij} \\
&= \sum_{i,j=1}^{n} c_{ij}x_{ij} + d\Big[\sum_{i\in Z, j\in S} x_{ij} - \sum_{i\notin Z}\Big(1 - \sum_{j\in S} x_{ij}\Big)\Big] \\
&= \sum_{i,j=1}^{n} c_{ij}x_{ij} + d\Big(\sum_{i=1}^{n}\sum_{j\in S} x_{ij} - (n-z)\Big) \\
&= \sum_{i,j=1}^{n} c_{ij}x_{ij} + d(s+z-n).
\end{aligned}
$$

Daher ist die Lösungsmenge des Zuordnungsproblems zur Kostenmatrix C die gleiche wie für die Matrix C'. Weiter ist sz die Anzahl der doppelt bedeckten (also durch eine Zeile *und* eine Spalte) Einträge, $(n-z)(n-s)$ die der unbedeckten Einträge. Nun vergleichen wir die Summe aller Einträge in den Matrizen C' und C. Wir erhalten

$$
\begin{aligned}
\sum_{i,j=1}^{n} c'_{ij} - \sum_{i,j=1}^{n} c_{ij} &= \sum_{i,j=1}^{n} (c'_{ij} - c_{ij}) \\
&= -d(n-z)(n-s) + dzs \\
&= dn(z+s-n) \\
&< 0.
\end{aligned}
$$

Die neue Matrix C', die wieder nichtnegative Einträge besitzt, hat also echt kleinere Gesamtkosten[42]. Insgesamt lautet dieser Schritt also:

- Sei d der kleinste unbedeckte Eintrag in C. Ziehe in C von allen unbedeckten Einträgen d ab und addiere d zu allen doppelt bedeckten Einträgen. Die entstehende Matrix werde wieder mit C bezeichnet. Hiermit gehe man zu dem vorigen Schritt zurück.

Beispiel: Sei

$$
C := \begin{pmatrix} 5 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}.
$$

[42]Ist die Ausgangsmatrix C *ganzzahlig*, so ist daher gesichert, dass die ungarische Methode nach endlich vielen Schritten abbricht.

Eine minimale Bedeckung der Nullen in C ist durch $Z = \{2\}$ und $S = \{2\}$ gegeben. In C sind die entsprechenden Zeilen und Spalten durchgestrichen.

$$C := \left(\begin{array}{ccc} 5 & 0 & 1 \\ \cancel{0} & \cancel{1} & \cancel{0} \\ 2 & 0 & 1 \end{array}\right).$$

Der minimale nicht durchgestrichene Eintrag ist $d := 1$. Als neue Matrix erhält man

$$C := \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Als Lösungen des zu C gehörigen Zuordnungsproblems erhält man

$$X_1 := \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad X_2 := \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Mit der zu Beginn gegebenen Matrix

$$C := \begin{pmatrix} 10 & 5 & 8 \\ 5 & 6 & 7 \\ 5 & 3 & 6 \end{pmatrix}$$

erhalten wir, dass der Beduine im für ihn günstigsten Fall 16 Kamele abzugeben hat. Die für ihn günstigsten Heiratspläne sind nicht eindeutig und können an den Permutationsmatrizen X_1 und X_2 abgelesen werden. □

Wir geben ein weiteres Beispiel an.

Beispiel: Eine Baufirma hat in 5 Städten A_i Baustellenkräne stehen. Diese werden in 5 anderen Städten B_j benötigt. Die Entfernung (in 100 km) c_{ij} von A_i nach B_j ist in der Matrix

$$C := \begin{pmatrix} 8 & 3 & 11 & 13 & 16 \\ 2 & 8 & 17 & 2 & 7 \\ 12 & 9 & 4 & 4 & 6 \\ 5 & 11 & 9 & 7 & 14 \\ 6 & 8 & 9 & 3 & 13 \end{pmatrix}$$

eingetragen. Die gesamte Wegstrecke soll minimiert werden. Als Ergebnis des Initialisierungsschrittes erhalten wir:

$$C \longrightarrow \begin{pmatrix} 5 & 0 & 8 & 10 & 13 \\ 0 & 6 & 15 & 0 & 5 \\ 8 & 5 & 0 & 0 & 2 \\ 0 & 6 & 4 & 2 & 9 \\ 3 & 5 & 6 & 0 & 10 \end{pmatrix} \longrightarrow \left(\begin{array}{ccccc} 5 & 0^* & 8 & 10 & 11 \\ 0^* & 6 & 15 & 0 & 3 \\ \cancel{8} & \cancel{5} & \cancel{0^*} & \cancel{0} & \cancel{0} \\ 0 & 6 & 4 & 2 & 7 \\ 3 & 5 & 6 & 0^* & 8 \end{array}\right).$$

In der letzten Matrix haben wir vier Nullen, von denen je zwei nicht auf einer Linie liegen, durch einen Stern gekennzeichnet. Außerdem haben wir durch Streichen einer

Zeile und von drei Spalten eine minimale Bedeckung der Nullen angegeben. Als nächste Matrix C bzw. Lösung X erhalten wir

$$C := \begin{pmatrix} 5 & 0^* & 5 & 10 & 8 \\ 0 & 6 & 12 & 0 & 0^* \\ 11 & 8 & 0^* & 3 & 0 \\ 0^* & 6 & 1 & 2 & 4 \\ 3 & 5 & 3 & 0^* & 5 \end{pmatrix}, \qquad X := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Dies ist eine Lösung, da wir in C eine 0-Diagonale der Länge 5 entdeckt haben. □

31 Die archimedische Methode zur Berechnung der Zahl π

Die Zahl π ist Quotient aus Umfang und Durchmesser eines Kreises. Von Archimedes ($\approx$ 287 v. Chr.–212 v. Chr.) stammt die Methode, durch einem Kreis ein- und umschriebene regelmäßige Vielecke und Berechnung derer Umfänge eine monotone Einschließung der Zahl π zu erhalten.

Das wichtigste Ergebnis seiner Arbeit "Kreismessung" ($KYKΛOY\ METRHΣIΣ$) ist die Aussage III. bzw. Proposition 3 (in Übersetzungen findet man die Arbeit von Archimedes bei F. Rudio (1892, S. 73–81) sowie L. Berggren et al. (1997, S. 7–14)):

- Der Umfang eines jeden Kreises ist dreimal so groß als der Durchmesser und noch um etwas größer, nämlich um weniger als ein Siebentel, aber um mehr als zehn Einundsiebenzigstel des Durchmessers.

Also ist $3\frac{10}{71} < \pi < 3\frac{1}{7}$. Wir wollen den archimedischen Algorithmus jetzt schildern, wobei wir die Darstellung von G. Miel (1983) zum Teil benutzen.

Gegeben sei ein Kreis mit dem Radius $r = \frac{1}{2}$ bzw. dem Durchmesser 1. Sein Umfang ist also $2\pi r = \pi$. Ein einbeschriebenes regelmäßiges 6-Eck (siehe Abbildung 46 links) hat den Umfang $6r = 3$. Daher ist $3 < \pi$. In Abbildung 46 rechts haben wir ein

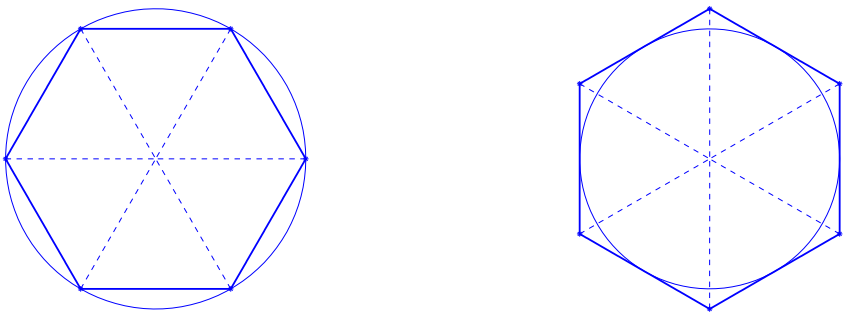


Abbildung 46: Ein einbeschriebenes bzw. umschriebenes regelmäßiges 6-Eck

dem Kreis mit Radius $r = \frac{1}{2}$ umschriebenes regelmäßiges 6-Eck angegeben. Bezeichnet man die Seitenlänge des umschriebenen regelmäßigen 6-Ecks mit s, so erhält man mit Hilfe des Satzes von Pythagoras, dass $r^2 + (s/2)^2 = s^2$ bzw. $s = 2r/\sqrt{3} = \sqrt{3}/3$. Das umschriebene regelmäßige 6-Eck hat also den Umfang $6s = 2\sqrt{3}$. Daher ist $\pi < 2\sqrt{3} \approx 3.464102$.

Der Umfang E_n bzw. U_n eines einem Kreis vom Durchmesser 1 einbe- bzw. umschriebenen regelmäßigen n-Ecks kann leicht mit Mitteln der Trigonometrie, die allerdings Archimedes nicht zur Verfügung standen, angegeben werden. Und zwar ist

$$E_n = n \sin\Big(\frac{\pi}{n}\Big), \qquad U_n = n \tan\Big(\frac{\pi}{n}\Big).$$

Ersteres erkennt man aus Abbildung 47 links. Hier ist $|AO| = |BO| = \frac{1}{2}$ und $\sphericalangle AOB =$

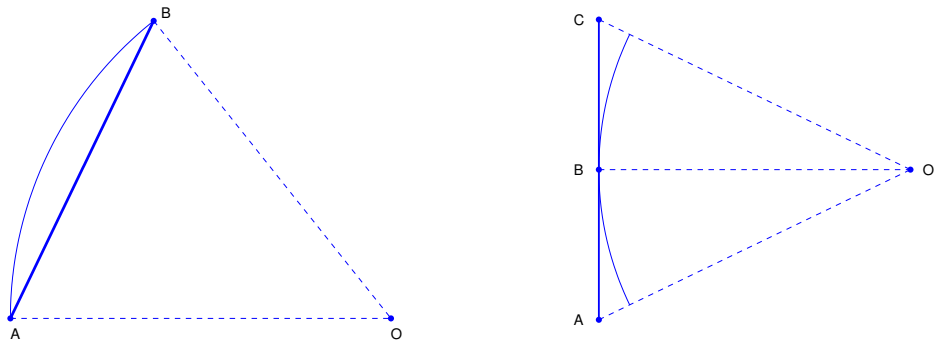


Abbildung 47: Der Umfang eines einbe- bzw. umschriebenen regelmäßigen n-Ecks

$2\pi/n$, folglich $\sphericalangle OAB = \sphericalangle ABO = \pi/2 - \pi/n$. Wegen des Sinussatzes und $\sin(\pi/2 - \pi/n) = \cos(\pi/n)$ ist

$$\frac{|AO|}{\cos(\pi/n)} = \frac{|AB|}{\sin(2\pi/n)}$$

und daher

$$E_n = n\,|AB| = n\frac{\sin(2\pi/n)}{\cos(\pi/n)}\,|AO| = n \sin\Big(\frac{\pi}{n}\Big).$$

Die entsprechende Beziehung für den Umfang des umschriebenen regelmäßigen n-Ecks erkennt man mit Hilfe von Abbildung 47 rechts. Hier ist $|BO| = \frac{1}{2}$, ferner $\sphericalangle AOC = 2\pi/n$ und $\sphericalangle AOB = \sphericalangle BOC = \pi/n$. Daher ist

$$U_n = n|AC| = 2n|AB| = 2n|BO| \tan\Big(\frac{\pi}{n}\Big) = n \tan\Big(\frac{\pi}{n}\Big),$$

wie oben behauptet.

Den folgenden Satz beweisen wir mit trigonometrischen Hilfsmitteln. Einen elementargeometrischen Beweis findet man bei G. Miel (1983).

Satz *Mit E_n bzw. U_n werde der Umfang eines einem Kreis vom Durchmesser 1 einbe- bzw. umschriebenen regelmäßigen n-Ecks bezeichnet. Dann ist*

$$\frac{1}{U_{2n}} = \frac{1}{2}\Big(\frac{1}{U_n} + \frac{1}{E_n}\Big), \qquad \frac{1}{E_{2n}} = \sqrt{\frac{1}{U_{2n}} \cdot \frac{1}{E_n}}$$

bzw.

$$U_{2n} = 2\frac{U_n E_n}{U_n + E_n}, \qquad E_{2n} = \sqrt{U_{2n} E_n}.$$

Beweis: Wir haben oben mit trigonometrischen Hilfsmitteln nachgewiesen, dass

$$E_n = n \sin\Big(\frac{\pi}{n}\Big), \qquad U_n = n \tan\Big(\frac{\pi}{n}\Big).$$

Zur Abkürzung setzen wir $\alpha := \pi/(2n)$. Dann ist

$$\begin{aligned}
\frac{1}{2}\Big(\frac{1}{U_n} + \frac{1}{E_n}\Big) &= \frac{1}{2n}\Big(\frac{1}{\tan(2\alpha)} + \frac{1}{\sin(2\alpha)}\Big) \\
&= \frac{1}{2n}\Big(\frac{\cos(2\alpha)+1}{\sin(2\alpha)}\Big) \\
&= \frac{1}{2n} \cdot \frac{\cos\alpha}{\sin\alpha} \\
&= \frac{1}{2n\tan(\pi/(2n))} \\
&= \frac{1}{U_{2n}}.
\end{aligned}$$

Entsprechend einfach ist der Beweis der zweiten Beziehung. □

Wegen $U_6 = 2\sqrt{3}$ und $E_6 = 3$ können mit der eben angegebenen Rekursionsformel $U_{2^k 6}$ und $E_{2^k 6}$ leicht berechnet werden. Um einen Eindruck von der "Konvergenzgüte" zu gewinnen, geben wir an (nach wie vor natürlich bezogen auf einen Kreis mit dem Durchmesser 1):

k	$E_{2^k 6}$	$U_{2^k 6}$
0	3.00000000000000	3.46410161513775
1	3.10582854123025	3.21539030917347
2	3.13262861328124	3.15965994209750
3	3.13935020304687	3.14608621513143
4	3.14103195089051	3.14271459964537

Die Konvergenz ist also nicht besonders gut, so dass ein stures Fortsetzen dieser Idee nur langsam zu besseren Ergebnissen führt. Trotzdem wurde die Methode von Archimedes von seinen Nachfolgern weitergetrieben. Wir geben nur eine kleine Übersicht (siehe hierzu auch J. Arndt, C. Haenel (2000)).

	Archimedes $\approx$ 230 v. Chr.	Aryabhata 499 n. Chr.	Vietà 1593	Roomen 1593	Van Ceulen 1610
Seitenzahl	$6 \cdot 2^4$	$6 \cdot 2^6$	$6 \cdot 2^{16}$	2^{30}	2^{62}
Genauigkeit	10^{-3}	10^{-4}	10^{-9}	10^{-15}	10^{-35}

Ludolph van Ceulen (1539–1610) soll einen großen Teil seines Lebens dafür verwandt haben, π auf 35 Stellen genau zu berechnen. Er hatte anscheinend eine ausnehmend verständnisvolle Frau. Sie gab die letzten 3 Ziffern posthum heraus und ließ auf seinem

(inzwischen verlorenen) Grabstein π auf 35 Stellen einmeißeln. Zumindest in Deutschland hieß π lange die Ludolphsche Zahl.

Nun kommen wir zu einer wesentlichen Verbesserung der archimedischen Methode, die 1621 von W. Snellius (1580–1626) gefunden und von C. Huygens (1629–1695) bewiesen wurde[43]. Im Lehrsatz VII seiner Arbeit "De circuli magnitudine inventa" (Über die gefundene Größe des Kreises) bewies Huygens 1654 mit geometrischen Methoden:

- Der Umfang eines jeden Kreises ist größer als der Umfang eines ihm eingeschriebenen gleichseitigen Polygones, vermehrt um den dritten Teil des Überschusses, um welchen dieser Umfang den Umfang eines andern eingeschriebenen Polygones von halb so viel Seiten übertrifft.

Im Klartext: Gegeben sei ein Kreis vom Durchmesser 1, also dem Umfang π. Mit E_n wird der Umfang eines eingeschriebenen regelmäßigen n-Ecks bezeichnet. Dann sagt der Lehrsatz VII bei Huygens aus, dass

$$\pi > E_n + \frac{1}{3}(E_n - E_{n/2}) = \frac{4E_n - E_{n/2}}{3}$$

bzw.

$$\pi > \frac{4E_{2n} - E_n}{3}, \qquad n = 3, 4, \ldots.$$

Der *geometrische* Beweis bei Huygens ist nicht ganz einfach. Wir geben einen analytischen Beweis an, wobei wir ausnutzen, dass $E_n = n\sin(\pi/n)$. Mit $x = \pi/n$ hat man zu zeigen, dass

$$f(x) := x - \frac{8\sin(x/2) - \sin x}{3} > 0 \qquad \text{für alle } x \in (0, \pi/3].$$

Es ist

$$f'(x) = 1 - \frac{4\cos(x/2) - \cos x}{3}$$

und

$$f''(x) = \frac{2\sin(x/2) - \sin x}{3} = \frac{2}{3}\sin(x/2)[1 - \cos(x/2)].$$

Da $f(0) = f'(0) = 0$ und $f''(x) > 0$ für $x \in (0, \pi/2)$, ist $f(x) > 0$ für $x \in (0, \pi/3]$ und damit der Lehrsatz von Huygens bewiesen. Die Güte der Approximation kann man mit Hilfe der Reihendarstellung des Sinus erkennen. Hiernach ist

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots.$$

Hieraus erhält man leicht, dass

$$0 < \pi - \frac{4E_{2n} - E_n}{3} < \frac{\pi^5}{480} \cdot \frac{1}{n^4}.$$

[43]Eine deutsche Übersetzung findet man bei F. Rudio (1892, S. 85–131), Kopien eines Teils der lateinisch geschriebenen Arbeit bei L. Berggren et al. (1997, S. 81–86).

Für $n = 2^{29}$ erhält man z. B. einen maximalen Fehler von $7.7 \cdot 10^{-36}$. Huygens brauchte also "nur" $E_{2^{29}}$ und $E_{2^{30}}$ zu berechnen, um dieselbe Genauigkeit wie van Ceulen mit $E_{2^{62}}$ zu erhalten.

Lehrsatz IX bei Huygens lautet:

- Der Umfang eines jeden Kreises ist kleiner als zwei Drittel des Umfangs eines ihm eingeschriebenen gleichseitigen Polygones, vermehrt um ein Drittel des dazu ähnlichen umgeschriebenen Polygones.

Bezeichnet man mit U_n den Umfang eines einem Kreis vom Umfang 1 umschriebenen regelmäßigen n-Ecks, so sagt dieser Lehrsatz aus, dass

$$\pi < \frac{2E_n + U_n}{3}.$$

Berücksichtigt man, dass $E_n = n \sin(\pi/n)$ und $U_n = n \tan(\pi/n)$, so ist also

$$\pi < \frac{2n \sin(\pi/n) + n \tan(\pi/n)}{3}, \qquad n = 3, 4, \ldots$$

zu zeigen. Wir setzen $x = \pi/n$ und haben zu zeigen, dass

$$f(x) := \frac{2 \sin x + \tan x}{3} - x > 0 \qquad \text{für } x \in (0, \pi/3].$$

Nun ist $f(0) = 0$ und

$$f'(x) = \frac{1}{3}\Big(2 \cos x + \frac{1}{\cos^2 x}\Big) - 1, \qquad f''(x) = \frac{2}{3} \sin x \Big(\frac{1}{\cos^3 x} - 1\Big).$$

Wegen $f'(0) = 0$ und $f''(x) > 0$ für $x \in (0, \pi/2)$ ist $f(x) > 0$ für $x \in (0, \pi/2)$, womit der Lehrsatz von Huygens mit analytischen Mitteln bewiesen ist. Berücksichtigt man, dass

$$\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \cdots,$$

so erhält man

$$\frac{2E_n + U_n}{3} = \pi + \frac{\pi^5}{20\, n^4} + \cdots.$$

Während E_n so schnell (oder langsam) mit $n \to \infty$ gegen π konvergiert wie $1/n^2$ gegen Null, so konvergiert $\frac{1}{3}(2E_n + U_n)$ so schnell gegen π wie $1/n^4$ gegen Null. Wie bei anderen Extrapolationsmethoden kann dieser Prozess fortgesetzt werden, worauf wir aber nicht mehr eingehen wollen. Zum Schluss dieses Abschnitts geben wir in der folgenden Tabelle einige Werte von E_n, $E_n^{(1)} := \frac{1}{3}(4E_n - E_{n/2})$, $U_n^{(1)} := \frac{1}{3}(2E_n + U_n)$ und U_n an.

n	E_n	$E_n^{(1)}$	$U_n^{(1)}$	U_n
6	3.00000000000		3.15470053838	3.46410161514
12	3.10582854123	3.14110472164	3.14234913054	3.21539030917
24	3.13262861328	3.14156197063	3.14163905622	3.15965994210
48	3.13935020305	3.14159073297	3.14159554041	3.14608621513
96	3.14103195089	3.14159253351	3.14159283381	3.14271459965

32 Der Brent-Salamin-Algorithmus

Der Brent-Salamin-Algorithmus (mit einigem Recht könnte man auch vom Gauß-Legendre-Algorithmus sprechen) ist ein verblüffend einfach zu formulierender Algorithmus zur schnellen Berechnung der Zahl π. Grundlage sind Aussagen über elliptische Integrale und das arithmetisch-geometrische Mittel, welche schon Gauß und Legendre bekannt waren. Der Algorithmus sieht folgendermaßen aus:

- Setze
$$a_0 := 1, \quad b_0 := \frac{1}{\sqrt{2}}, \quad t_0 := \frac{1}{4}, \quad p_0 := 1.$$
- Für $k = 0, 1, \ldots$:

  Berechne
$$\pi_k := \frac{(a_k + b_k)^2}{4t_k}$$
  und anschließend
$$a_{k+1} := \frac{a_k + b_k}{2}, \quad b_{k+1} := \sqrt{a_k b_k}, \quad t_{k+1} := t_k - p_k(a_k - a_{k+1})^2, \quad p_{k+1} := 2p_k.$$

Bemerkung: Wir wissen, dass $\{a_k\}$ und $\{b_k\}$ monoton fallend bzw. wachsend gegen den gemeinsamen Grenzwert $M(1, 1/\sqrt{2})$, das arithmetisch-geometrische Mittel von 1 und $1/\sqrt{2}$, konvergieren, siehe Abschnitt 26. Offenbar ist $p_k = 2^k$, wegen $a_k - a_{k+1} = \frac{1}{2}(a_k - b_k)$ ist ferner

$$t_k = t_0 - \frac{1}{4}\sum_{j=0}^{k-1} 2^j (a_j - b_j)^2 = \frac{1}{4}\Big(1 - \sum_{j=0}^{k-1} 2^j (a_j - b_j)^2\Big),$$

wie man leicht durch vollständige Induktion nach k beweist. Daher ist

$$\pi_k = \frac{(a_k + b_k)^2}{4t_k} = \frac{4a_{k+1}^2}{1 - \sum_{j=0}^{k-1} 2^j (a_j - b_j)^2}.$$

Der Zähler konvergiert gegen $4M(1, 1/\sqrt{2})^2$ der Nenner gegen $1 - \sum_{j=0}^{\infty} 2^j (a_j - b_j)^2$. Denn die Reihe $\sum_{j=0}^{\infty} 2^j (a_j - b_j)^2$ ist konvergent. Dies kann man folgendermaßen einsehen. Es ist leicht zu zeigen, dass

$$a_j - b_j = \frac{(a_{j-1} - b_{j-1})^2}{8a_{j+1}} \le \frac{(a_{j-1} - b_{j-1})^2}{8M(1, 1/\sqrt{2})}.$$

Da wir schon wissen, dass $\{a_j\}$ und $\{b_j\}$ den gemeinsamen Limes $M(1, 1/\sqrt{2})$ besitzen, kann hieraus auf die Existenz von Konstanten $c_0 > 0$ und $q \in (0, 1)$ mit $2^j (a_j - b_j)^2 \le c_0 q^j$, $j = 0, 1, \ldots$, geschlossen werden und dies sichert die Konvergenz obiger Reihe, da sie durch eine geometrische Reihe majorisiert wird. Daher konvergiert die Folge $\{\pi_k\}$ gegen π, falls

$$\pi = \frac{4M(1, 1/\sqrt{2})^2}{1 - \sum_{j=0}^{\infty} 2^j (a_j - b_j)^2}.$$

Dass diese Aussage richtig ist, ist von R. P. Brent (1975) und E. Salamin (1976, Theorem 1b) unabhängig voneinander bewiesen worden. Wir werden einen Beweis im Anschluss bringen. □

Mit Hilfe von Mupad und `DIGITS:=60` erhalten wir die folgenden Ergebnisse:

k	π_k
0	2.914213562373095048801688724209698078569671875376948073176 68
1	3.140579250522168248311331268975823311773440237512948335643 49
2	3.141592646213542282149344431982695774314437223345602794559 54
3	3.141592653589793238279512774801863974381225504835446935787 33
4	3.141592653589793238462643383279502884197114678283648921556 62
5	3.141592653589793238462643383279502884197169399375105820974 94
6	3.141592653589793238462643383279502884197169399375105820974 94

Zumindest die ersten 60 Ziffern von π_5 stimmen mit den entsprechenden Ziffern von π überein.

Den angekündigten Satz beweisen wir mit Hilfe von drei Hilfssätzen.

Hilfssatz 1 (Landen) *Für positive reelle Zahlen a, b sei*

$$I(a,b) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{a^2\cos^2\theta + b^2\sin^2\theta}}, \qquad J(a,b) := \int_0^{\pi/2} \sqrt{a^2\cos^2\theta + b^2\sin^2\theta}\, d\theta.$$

Dann gilt:

$$(*) \qquad I(a,b) = I\Big(\frac{a+b}{2}, \sqrt{ab}\Big), \qquad I(a,b) = \frac{\pi/2}{M(a,b)}$$

sowie

$$(**) \qquad J(a,b) = 2J\Big(\frac{a+b}{2}, \sqrt{ab}) - abI(a,b).$$

Beweis: Mit etwas anderen Bezeichnungen ist $(*)$ schon in Abschnitt 26 bewiesen worden. Zum Beweis von $(**)$ machen wir die Variablentransformation $t := \tan\theta$. Dann ist

$$\cos^2\theta = \frac{1}{1+t^2}, \quad \sin^2\theta = \frac{t^2}{1+t^2}, \quad d\theta = \frac{dt}{1+t^2}$$

und folglich

$$I(a,b) = \int_0^{\infty} \frac{dt}{\sqrt{(1+t^2)(a^2+b^2t^2)}} \qquad J(a,b) = \int_0^{\infty} \frac{\sqrt{a^2+b^2t^2}}{(1+t^2)^{3/2}}\, dt.$$

Jetzt definieren wir

$$L(s) := \frac{(a+b)s}{a-bs^2}, \qquad c := \sqrt{a/b}$$

und beachten, dass $L(\cdot)$ wegen

$$L'(s) = \frac{(a+b)(a+bs^2)}{(a-bs^2)^2} > 0$$

auf $[0,c)$ und (c,∞) monoton wachsend ist. Ferner ist

$$L(0)=0, \qquad \lim_{s\to c-} L(s)=\infty$$

sowie

$$\lim_{s\to c+}(-L)(s)=\infty, \qquad \lim_{s\to\infty}(-L)(s)=0.$$

Jetzt machen wir die Variablentransformationen $t=L(s)$ bzw. $t=-L(s)$ und erhalten, dass

$$J\Big(\frac{a+b}{2},\sqrt{ab}\Big)=\int_0^c \frac{L'(s)\sqrt{(a+b)/2)^2+abL^2(s)}}{(1+L^2(s))^{3/2}}\,ds$$

und

$$J\Big(\frac{a+b}{2},\sqrt{ab}\Big)=\int_c^\infty \frac{L'(s)\sqrt{(a+b)/2)^2+abL^2(s)}}{(1+L^2(s))^{3/2}}\,ds.$$

Insgesamt ist

$$J\Big(\frac{a+b}{2},\sqrt{ab}\Big)=\frac{1}{2}\int_0^\infty \frac{L'(s)\sqrt{(a+b)/2)^2+abL^2(s)}}{(1+L^2(s))^{3/2}}\,ds.$$

Zur Abkürzung setzen wir

$$A(s):=a-bs^2, \quad B(s):=a+bs^2, \quad C(s):=a^2-b^2s^4, \quad D(s):=a^2+b^2s^2.$$

Durch Nachrechnen bestätigt man, dass

$$\frac{(a+b)^2}{2}B^2(s)-ab(1+s^2)D(s)=D^2(s)-\frac{a^2-b^2}{2}C(s).$$

Ferner ist

$$1+L^2(s)=\frac{(1+s^2)D(s)}{A^2(s)}, \qquad L'(s)=\frac{(a+b)B(s)}{A^2(s)}$$

sowie

$$\Big(\frac{a+b}{2}\Big)^2+abL^2(s)=\frac{(a+b)^2B^2(s)}{4A^2(s)}.$$

Folglich ist nach einfacher Rechnung

$$\begin{aligned}
2J\Big(\frac{a+b}{2},\sqrt{ab}\Big)-abI(a,b)) &= \frac{(a+b)^2}{2}\int_0^\infty \frac{B^2(s)}{((1+s^2)D(s))^{3/2}}\,ds \\
&\quad - ab\int_0^\infty \frac{ds}{\sqrt{(1+s^2)D(s)}} \\
&= \int_0^\infty \frac{(a+b)^2B^2(s)/2-ab(1+s^2)D(s)}{((1+s^2)D(s))^{3/2}}\,ds \\
&= \int_0^\infty \frac{D^2(s)}{((1+s^2)D(s))^{3/2}}ds \\
&\quad - \frac{a^2-b^2}{2}\int_0^\infty \frac{C(s)}{((1+s^2)D(s))^{3/2}}ds
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \frac{\sqrt{D(s)}}{(1+s^2)^{3/2}}\,ds \\
&\qquad - \frac{a^2-b^2}{2}\int_0^\infty \frac{C(s)}{((1+s^2)D(s))^{3/2}}ds \\
&= I(a,b) - \frac{a^2-b^2}{2}\int_0^\infty \frac{d}{ds}\Big(\frac{s}{\sqrt{(1+s^2)D(s)}}\Big)\,ds \\
&= I(a,b) - \frac{a^2-b^2}{2}\cdot\frac{s}{\sqrt{(1+s^2)D(s)}}\,\Big|_0^\infty \\
&= I(a,b).
\end{aligned}
$$

Damit ist der Satz bewiesen[44]. □

Hilfssatz 2 *Seien a_0, b_0 positive reelle Zahlen. Sei*

$$a_{k+1} := \frac{a_k + b_k}{2}, \qquad b_{k+1} := \sqrt{a_k b_k}$$

für $k = 0, 1, \ldots$. Mit den Bezeichnungen von Hilfssatz 1 ist dann

$$J(a_0, b_0) = \Big(a_0^2 - \frac{1}{2}\sum_{j=0}^{\infty} 2^j(a_j^2 - b_j^2)\Big) I(a_0, b_0).$$

Beweis: Aus $(**)$ und $(*)$ in Hilfssatz 1 erhalten wir

$$2J(a_{j+1}, b_{j+1}) - J(a_j, b_j) = a_j b_j I(a_j, b_j) = a_j b_j I(a_0, b_0).$$

Folglich ist

$$
\begin{aligned}
& 2^{j+1}[J(a_{j+1}, b_{j+1}) - a_{j+1}^2 I(a_0, b_0)] - 2^j[J(a_j, b_j) - a_j^2 I(a_0, b_0)] \\
= \;& 2^j[2J(a_{j+1}, b_{j+1}) - J(a_j, b_j)] - 2^j[2a_{j+1}^2 - a_j^2] I(a_0, b_0) \\
= \;& 2^j[a_j b_j + a_j^2 - 2a_{j+1}^2] I(a_0, b_0) \\
= \;& \frac{1}{2} 2^j (a_j^2 - b_j^2) I(a_0, b_0).
\end{aligned}
$$

[44] Die Aussage $(*)$ kann im Prinzip genau so bewiesen werden, womit wir einen alternativen Beweis zu dem aus Abschnitt 26 erhalten. Denn mit den obigen Bezeichnungen ist

$$
\begin{aligned}
I\Big(\frac{a+b}{2}\sqrt{ab}\Big) &= \frac{1}{2}\int_0^\infty \frac{L'(s)}{\sqrt{(1+L^2(s))((a+b)/2)^2 + abL^2(s)}}\,ds \\
&= \frac{1}{2}\int_0^\infty \frac{(a+b)B(s)}{A^2(s)} \cdot \frac{A(s)}{\sqrt{(1+s^2)D(s)}} \cdot \frac{2A(s)}{(a+b)B(s)}\,ds \\
&= \int_0^\infty \frac{ds}{\sqrt{(1+s^2)D(s)}}\,ds \\
&= I(a,b).
\end{aligned}
$$

Durch Aufsummieren von $j = 0$ bis $j = k-1$ erhalten wir, dass

$$\frac{1}{2}\Big(\sum_{j=0}^{k-1} 2^j(a_j^2 - b_j^2)\Big) I(a_0, b_0) = 2^k[J(a_k, b_k) - a_k^2 I(a_0, b_0)] - [J(a_0, b_0) - a_0^2 I(a_0, b_0)]$$

bzw.

$$(*) \qquad J(a_0, b_0) = \Big(a_0^2 - \frac{1}{2}\sum_{j=0}^{k-1} 2^j(a_j^2 - b_j^2)\Big) I(a_0, b_0) - 2^k[a_k^2 I(a_0, b_0) - J(a_k, b_k)].$$

Nun beachten wir, dass die Reihe $\sum_{j=0}^{\infty} 2^j(a_j^2 - b_j^2)$ konvergiert, da $\{a_j^2 - b_j^2\}$ quadratisch gegen Null konvergiert, und

$$\begin{aligned}
2^k[a_k^2 I(a_0, b_0) - J(a_k, b_k)] &= 2^k[a_k^2 I(a_k, b_k) - J(a_k, b_k)] \\
&= 2^k \int_0^{\pi/2} \frac{a_k^2 - (a_k^2 \cos^2\theta + b_k^2 \sin^2\theta)}{\sqrt{a_k^2 \cos^2\theta + b_k^2 \sin^2\theta}}\, d\theta \\
&= 2^k (a_k^2 - b_k^2) \int_0^{\pi/2} \frac{\sin^2\theta}{\sqrt{a_k^2 \cos^2\theta + b_k^2 \sin^2\theta}}\, d\theta \\
&\leq 2^k (a_k^2 - b_k^2) I(a_k, b_k) \\
&= 2^k (a_k^2 - b_k^2) I(a_0, b_0) \\
&\to 0,
\end{aligned}$$

da $\{a_k^2 - b_k^2\}$ quadratisch gegen Null konvergiert. Mit $k \to \infty$ in $(*)$ haben wir den Hilfssatz bewiesen. □

Hilfssatz 3 (Legendre) *Sei $k \in (0, 1)$ und $k' := \sqrt{1 - k^2}$. Dann ist*

$$K(k)E(k') + K(k')E(k) - K(k)K(k') = \frac{\pi}{2}.$$

Hierbei ist

$$K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}}$$

das vollständige elliptische Integral erster Art und

$$E(k) := \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2\theta}\, d\theta$$

das vollständige elliptische Integral zweiter Art.

Beweis: Den Beweis übernehmen wir von G. Almkvist, B. Berndt (1988). Sei $c := k^2$. Wir fassen K und E als Funktion in c auf und definieren

$$L(c) := K(k)E(k') + K(k')E(k) - K(k)K(k').$$

Wir werden zeigen, dass die Ableitung von $L(\cdot)$ nach c verschwindet, $L(\cdot)$ also konstant ist und die Konstante durch Berechnung von $\lim_{c\to 0+} L(c)$ erhalten. Zunächst ist

$$\begin{aligned}
\frac{d}{dc}[E(k)-K(k)] &= \frac{d}{dc}\int_0^{\pi/2}\Big(\sqrt{1-c\sin^2\theta}-\frac{1}{\sqrt{1-c\sin^2\theta}}\Big)\,d\theta\\
&= -\frac{d}{dc}\int_0^{\pi/2}\frac{c\sin^2\theta}{\sqrt{1-c\sin^2\theta}}\,d\theta\\
&= \int_0^{\pi/2}\frac{(c/2)\sin^4\theta-\sin^2\theta}{(1-c\sin^2\theta)^{3/2}}\,d\theta\\
&= \frac{1}{2c}\int_0^{\pi/2}\frac{(1-c\sin^2\theta)^2-1}{(1-c\sin^2\theta)^{3/2}}\,d\theta\\
&= \frac{1}{2c}\int_0^{\pi/2}\sqrt{1-c\sin^2\theta}\,d\theta-\frac{1}{2c}\int_0^{\pi/2}\frac{d\theta}{(1-c\sin^2\theta)^{3/2}}\\
&= \frac{1}{2c}E(k)-\frac{1}{2c}\int_0^{\pi/2}\frac{d\theta}{(1-c\sin^2\theta)^{3/2}}.
\end{aligned}$$

Da

$$\begin{aligned}
\frac{d}{d\theta}\Big(\frac{\sin\theta\cos\theta}{\sqrt{1-c\sin^2\theta}}\Big) &= \frac{(\cos^2\theta-\sin^2\theta)(1-c\sin^2\theta)-c\sin^2\theta\cos^2\theta}{(1-c\sin^2\theta)^{3/2}}\\
&= \frac{(1-2\sin^2\theta)(1-c\sin^2\theta)+c\sin^2\theta(1-\sin^2\theta)}{(1-c\sin^2\theta)^{3/2}}\\
&= \frac{(1-c\sin^2\theta)^2-(1-c)}{c(1-c\sin^2\theta)^{3/2}}\\
&= \frac{1}{c}\sqrt{1-c\sin^2\theta}-\frac{1-c}{c}\cdot\frac{1}{(1-c\sin^2\theta)^{3/2}},
\end{aligned}$$

ist

$$\begin{aligned}
\frac{d}{dc}[E(k)-K(k)] &= \frac{1}{2c}E(k)-\frac{1}{2c}\int_0^{\pi/2}\frac{d\theta}{(1-\sin^2\theta)^{3/2}}\\
&= \frac{1}{2c}E(k)-\frac{1}{2(1-c)}\cdot\frac{1-c}{c}\int_0^{\pi/2}\frac{d\theta}{(1-c\sin^2\theta)^{3/2}}\\
&= \frac{1}{2c}E(k)+\frac{1}{2(1-c)}\Big[\underbrace{\int_0^{\pi/2}\frac{d}{d\theta}\Big(\frac{\sin\theta\cos\theta}{\sqrt{1-c\sin^2\theta}}\Big)\,d\theta}_{=0}\\
&\qquad -\frac{1}{c}\int_0^{\pi/2}\sqrt{1-c\sin^2\theta}\,d\theta\Big]\\
&= \frac{1}{2c}\Big(1-\frac{1}{1-c}\Big)E(k)\\
&= -\frac{1}{2(1-c)}E(k).
\end{aligned}$$

Hieraus schließen wir, dass

$$\begin{aligned}\frac{d}{dc}[E(k')-K(k')] &= \frac{d}{dc}\int_0^{\pi/2}\Big(\sqrt{1-(1-c)\sin^2\theta}-\frac{1}{\sqrt{1-(1-c)\sin^2\theta}}\Big)\,d\theta\\ &= \frac{1}{2c}E(k').\end{aligned}$$

Weiter ist

$$\begin{aligned}\frac{d}{dc}E(k) &= \frac{d}{dc}\int_0^{\pi/2}\sqrt{1-c\sin^2\theta}\,d\theta\\ &= -\int_0^{\pi/2}\frac{\sin^2\theta}{2\sqrt{1-c\sin^2\theta}}\,d\theta\\ &= \frac{1}{2c}\int_0^{\pi/2}\frac{(1-c\sin^2\theta)-1}{\sqrt{1-c\sin^2\theta}}\,d\theta\\ &= \frac{1}{2c}[E(k)-K(k)]\end{aligned}$$

und entsprechend

$$\frac{d}{dc}E(k')=\frac{d}{dc}\int_0^{\pi/2}\sqrt{1-(1-c)\sin^2\theta}\,d\theta=-\frac{1}{2(1-c)}[E(k')-K(k')].$$

Mit

$$\begin{aligned}L(c) &:= K(k)E(k')+K(k')E(k)-K(k)K(k')\\ &= E(k)E(k')-[E(k)-K(k)][E(k')-K(k')]\end{aligned}$$

ist daher

$$\begin{aligned}\frac{d}{dc}L(c) &= \Big(\frac{d}{dc}E(k)\Big)E(k')+E(k)\Big(\frac{d}{dc}E(k')\Big)\\ &\quad-\Big(\frac{d}{dc}[E(k)-K(k)]\Big)[E(k')-K(k')]\\ &\quad-[E(k)-K(k)]\Big(\frac{d}{dc}[E(k')-K(k')]\Big)\\ &= \frac{1}{2c}[E(k)-K(k)]E(k')-\frac{1}{2(1-c)}E(k)[E(k')-K(k')]\\ &\quad+\frac{1}{2(1-c)}E(k)[E(k')-K(k')]-\frac{1}{2c}[E(k)-K(k)]E(k')\\ &= 0\end{aligned}$$

und folglich $L(\cdot)$ konstant. Wir zeigen, dass diese Konstante $\pi/2$ ist, indem wir

$$\lim_{c\to 0+}L(c)=\frac{\pi}{2}$$

nachweisen. Zunächst ist

$$E(k) - K(k) = -c \int_0^{\pi/2} \frac{\sin^2\theta}{\sqrt{1 - c\sin^2\theta}}\, d\theta$$

und daher $E(k)-K(k) = O(c)$ bzw. $|E(k)-K(k)| \le \alpha_1 c$ mit einer Konstanten $\alpha_1 > 0$ für alle hinreichend kleinen $c > 0$. Weiter ist

$$K(k') = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - (1-c)\sin^2\theta}} \le \frac{\pi}{2}\sqrt{c}$$

bzw. $(0 \le) K(k') \le \alpha_2/\sqrt{c}$ mit der Konstanten $\alpha_2 := \pi/2$ für alle $c \in (0,1)$. Daher ist

$$\lim_{c\to 0+} [E(k) - K(k)]K(k') = 0.$$

Daher ist

$$\lim_{c\to 0+} L(c) = \underbrace{\lim_{c\to 0+} [E(k) - K(k)]K(k')}_{=0} + \underbrace{\lim_{c\to 0} E(k')}_{=1} \cdot \underbrace{\lim_{c\to 0+} K(k)}_{=\pi/2} = \frac{\pi}{2}.$$

Damit ist der Hilfssatz 3 schließlich bewiesen. □

Satz (Brent-Salamin) *Sei $a_0 := 1$, $b_0 := 1/\sqrt{2}$. Die Folgen $\{a_k\}$, $\{b_k\}$ seien durch*

$$a_{k+1} := \frac{a_k + b_k}{2}, \qquad b_{k+1} := \sqrt{a_k b_k} \qquad (k = 0, 1, \ldots)$$

definiert, $M(1, 1/\sqrt{2})$ sei ihr (gemeinsamer) Limes. Dann ist

$$\pi = \frac{4M(1, 1/\sqrt{2})^2}{1 - \sum_{j=0}^{\infty} 2^j (a_j - b_j)^2}.$$

Beweis: In Hilfssatz 3 setzen wir $k := 1/\sqrt{2}$. Dann ist auch $k' := \sqrt{1-k^2} = 1/\sqrt{2}$ und aus Hilfssatz 3 erhalten wir

$$(*) \qquad 2K\Big(\frac{1}{\sqrt{2}}\Big)E\Big(\frac{1}{\sqrt{2}}\Big) - K^2\Big(\frac{1}{\sqrt{2}}\Big) = \frac{\pi}{2}.$$

In Hilfssatz 2 setze man $a_0 := 1$, $b_0 := 1/\sqrt{2}$. Unter Berücksichtigung von $I(1, 1/\sqrt{2}) = K(1/\sqrt{2})$ und $J(1, 1/\sqrt{2}) = E(1/\sqrt{2})$ erhalten wir aus diesem Hilfssatz die Beziehung

$$(**) \qquad E\Big(\frac{1}{\sqrt{2}}\Big) = \Big(1 - \frac{1}{2}\sum_{j=0}^{\infty} 2^j (a_j^2 - b_j^2)\Big) K\Big(\frac{1}{\sqrt{2}}\Big).$$

Wegen Hilfssatz 1 ist

$$(***) \qquad M(1, 1/\sqrt{2}) = \frac{\pi}{2K(1/\sqrt{2})}.$$

Nun gehen wir von der Legendre-Beziehung $(*)$, ersetzen hierin $E(1/\sqrt{2})$ mit Hilfe von $(**)$ und anschließend $K(1/\sqrt{2})$ mit $(***)$. Wir erhalten

$$\begin{aligned}
\frac{\pi}{2} &= 2K\Big(\frac{1}{\sqrt{2}}\Big)E\Big(\frac{1}{\sqrt{2}}\Big) - K^2\Big(\frac{1}{\sqrt{2}}\Big) \\
&= 2K\Big(\frac{1}{\sqrt{2}}\Big)\Big(1 - \frac{1}{2}\sum_{j=0}^{\infty} 2^j(a_j^2 - b_j^2)\Big)K\Big(\frac{1}{\sqrt{2}}\Big) - K^2\Big(\frac{1}{\sqrt{2}}\Big) \\
&= K^2\Big(\frac{1}{\sqrt{2}}\Big)\Big(1 - \sum_{j=0}^{\infty} 2^j(a_j^2 - b_j^2)\Big) \\
&= \frac{\pi^2}{4M^2(1,1/\sqrt{2})}\Big(1 - \sum_{j=0}^{\infty} 2^j(a_j^2 - b_j^2)\Big) \\
&= \frac{\pi^2}{4M^2(1,1/\sqrt{2})}\Big(\frac{1}{2} - \sum_{j=1}^{\infty} 2^j(a_j^2 - b_j^2)\Big) \\
&= \frac{\pi^2}{8M^2(1,1/\sqrt{2})}\Big(1 - \sum_{j=1}^{\infty} 2^{j+1}(a_j^2 - b_j^2)\Big) \\
&= \frac{\pi^2}{8M^2(1,1/\sqrt{2})}\Big(1 - 4\sum_{j=0}^{\infty} 2^j(a_{j+1}^2 - b_{j+1}^2)\Big) \\
&= \frac{\pi^2}{8M^2(1,1/\sqrt{2})}\Big(1 - \sum_{j=0}^{\infty} 2^j(a_j - b_j)^2\Big)
\end{aligned}$$

und hieraus die Behauptung. □

33 Das Geburtstagsparadoxon

Wir zitieren einmal wieder Wikipedia: Das Ergebnis der Frage:

> *Wie groß ist die Wahrscheinlichkeit, dass bei 23 Personen zwei von ihnen am selben Tag Geburtstag haben (ohne Beachtung des Jahrganges)?*

ist für viele verblüffend und wird deshalb als paradox wahrgenommen. So schätzen die meisten Menschen die Wahrscheinlichkeit um eine Zehnerpotenz falsch ein. Sie liegt nicht zwischen 1 und 5 % (wie zumeist geschätzt), sondern über 50 %; bei 50 Personen sogar bei über 97 %.

Bei der folgenden Herleitung wird der 29. Februar vernachlässigt und daher angenommen, dass es genau 365 verschiedene Geburtstage gibt. Weiter wird angenommen, dass alle Geburtstage gleich wahrscheinlich sind (was wohl in der Wirklichkeit nicht der Fall ist, da im Sommer mehr Kinder als im Winter geboren werden). Unter diesen Annahmen fragen wir:

> *Wie groß ist die Wahrscheinlichkeit $p(n)$, dass bei n Personen zwei von ihnen am selben Tag Geburtstag haben (ohne Beachtung des Jahrganges)?*,

Sei $q(n) := 1 - p(n)$. Dann ist $q(n)$ die Wahrscheinlichkeit dafür, dass bei n Personen alle an verschiedenen Tagen Geburtstag haben. Offenbar ist

$$q(1) = \frac{365}{365}, \qquad q(2) = \frac{364}{365},$$

allgemein

$$q(n) = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}$$

und daher

$$p(n) = 1 - \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n} = 1 - \frac{364 \cdots (365 - n + 1)}{365^{n-1}}.$$

In Abbildung 48 haben wir $p(\cdot)$ über $[1, 100]$ aufgetragen. Z. B. ist $p(23) \approx 0.5073$ und

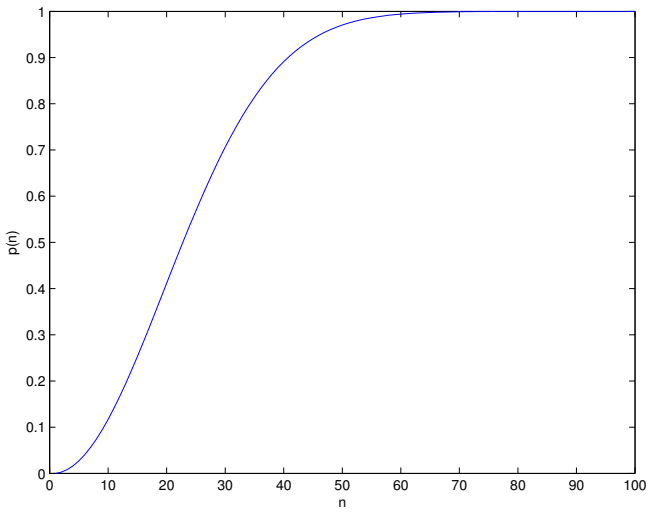


Abbildung 48: Die Wahrscheinlichkeit $p(n)$ beim Geburtstagsparadoxon

$p(75) \approx 0.9997$.

Entsprechend ist die Wahrscheinlichkeit p dafür, dass bei einer siebenstelligen Zahl (z.B. einer Telefonnummer) zwei Ziffern übereinstimmen, gegeben durch

$$p = 1 - \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4}{10^7} \approx 0.9395.$$

Bemerkungen: Eine mögliche Verallgemeinerung besteht darin, dass man danach fragt, wie groß die Wahrscheinlichkeit ist, dass in einer Gruppe von n Personen genau $m < n$ Personen am selben Tag Geburtstag haben. Dies ist eine wesentlich schwerer zu beantwortende Fragestellung. Wir verweisen lediglich auf `http://mathworld.wolfram.com/BirthdayProblem.html`.

Natürlich kann man auch danach fragen, wie groß die Wahrscheinlichkeit dafür ist, dass unter n Personen eine Person an einem *bestimmten* Tag, etwa dem 31. Dezember,

Geburtstag hat, wobei wir weiter den 29. Februar außer Acht lassen. Die Wahrscheinlichkeit für eine Person an einem bestimmten Tag Geburtstag zu haben beträgt $\frac{1}{365}$. Daher ist die Wahrscheinlichkeit für eine Person an einem bestimmten Tag *nicht* Geburtstag zu haben gleich $q := 1 - \frac{1}{365}$. Dann ist q^2 die Wahrscheinlichkeit dafür, dass keine von zwei Personen an dem bestimmten Tag Geburtstag hat und dementsprechend $1-q^2$ die Wahrscheinlich, dass unter zwei Personen wenigstens eine an dem bestimmten Tag Geburtstag hat. Allgemein ist $1 - q^n$ die Wahrscheinlichkeit dafür, dass unter n Personen wenigstens eine an einem bestimmten Tag Geburtstag hat. Z. B. ist $1-q^n \geq \frac{1}{2}$ bzw. $q^n \leq \frac{1}{2}$, wenn $n \geq \ln(\frac{1}{2})/\ln(1 - \frac{1}{365})$ bzw. $n \geq 253$. □

34 Das Ziegenproblem

Gibt man *Ziegenproblem*, siehe auch G. von Randow (2010), bei Google ein, so erhält man über 21 000 Web-Seiten. In der Formulierung wie bei Wikipedia handelt es sich um das folgende Problem:

> „Nehmen Sie an, Sie wären in einer Spielshow und hätten die Wahl zwischen drei Toren. Hinter einem der Tore ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, sagen wir, Tor Nummer 1, und der Showmaster, der weiß, was hinter den Toren ist, öffnet ein anderes Tor, sagen wir, Nummer 3, hinter dem eine Ziege steht. Er fragt Sie nun: 'Möchten Sie das Tor Nummer Zwei?' Ist es von Vorteil, die Wahl des Tores zu ändern?“

Eine etwas genauere Beschreibung (ebenfalls bei Wikipedia zu finden) ist die folgende:

> „Angenommen Sie befinden sich in einer Spielshow und haben die Wahl zwischen drei Toren. Hinter einem Tor ist ein Auto, hinter den anderen sind Ziegen. Das Auto und die Ziegen sind vor der Show zufällig hinter die Tore verteilt worden. Die Regeln der Spielshow sind folgende: Nachdem Sie ein Tor gewählt haben bleibt dieses zunächst geschlossen. Der Showmaster, der weiß, was sich hinter den Toren befindet, muss nun eines der beiden verbleibenden Tore öffnen, und hinter dem von ihm geöffneten Tor muss sich eine Ziege befinden. Wenn hinter beiden verbleibenden Toren jeweils eine Ziege steht, öffnet er eines der beiden Tore zufällig. Nachdem der Showmaster ein Tor mit einer Ziege geöffnet hat fragt er Sie, ob Sie bei Ihrer ersten Wahl bleiben oder zum letzten verbleibenden Tor wechseln möchten. Nehmen Sie an Sie wählen Tor 1 und der Showmaster öffnet Tor 3 mit einer Ziege. Er fragt Sie dann: „Möchten Sie zu Tor 2 wechseln?“ Ist es zu Ihrem Vorteil, Ihre Wahl zu ändern?“

Die folgenden beiden Argumente, von denen das erste wesentlich öfter genannt wird, kommen zu unterschiedlichen Ergebnissen.

- Es hat keinen Zweck zu wechseln, denn das Auto steht hinter Tor 1 oder Tor 2. Beides ist mit der Wahrscheinlichkeit $\frac{1}{2}$ gleich wahrscheinlich, so dass es keinen Sinn macht zu wechseln.

- Man sollte wechseln und das Tor 2 wählen. Denn mit der Wahrscheinlichkeit $\frac{1}{3}$ ist das Auto hinter dem gewählten Tor 1 und damit mit der Wahrscheinlichkeit $\frac{2}{3}$ hinter einem der beiden Tore 2 oder 3. Da es aber nicht hinter Tor 3 steht, steht es mit Wahrscheinlichkeit $\frac{2}{3}$ hinter Tor 2.

Was ist die richtige Antwort? Die richtige Antwort ist die zweite, der Kandidat der Spielshow sollte also seine Wahl ändern. Wir geben eine erste Begründung hierfür an. Angenommen, zwei Kandidaten A und B verständigen sich darauf, dasselbe Tor zu wählen, etwa das erste. Der Kandidat A bleibt bei seiner Wahl (unabhängig von der durch den Showmaster gegebenen Information), während Kandidat B seine Entscheidung ändert. Die Wahrscheinlichkeit eines Auto-Gewinns für A ist offenbar $\frac{1}{3}$. Immer wenn A gewinnt, verliert B, und umgekehrt. Daher ist die Wahrscheinlichkeit eines Gewinns für B durch $1 - \frac{1}{3} = \frac{2}{3}$ gegeben. Es folgt eine zweite Begründung, bei der wir alle möglichen Fälle auflisten und abzählen, in welchen Fällen der Kandidat durch Wechseln gewinnt bzw. verliert. Ohne Einschränkung gehen wir davon aus, dass der Kandidat Tor 1 wählt.

1. Das Auto steht hinter Tor 1.

   Beim Wechseln verliert der Kandidat.

2. Das Auto steht hinter Tor 2.

   Dann muss der Showmaster Tor 3 öffnen. Durch Wechseln von Tor 1 zu Tor 2 gewinnt der Kandidat.

3. Das Auto steht hinter Tor 3.

   Dann muss der Showmaster Tor 2 öffnen. Durch Wechseln von Tor 1 zu Tor 3 gewinnt der Kandidat.

In zwei von drei Fällen gewinnt der Kandidat also durch Wechseln. Eigentlich kann man nach diesen beiden Argumenten nicht verstehen, weshalb das Ziegenproblem einen solchen Wirbel verursacht hat.

Unter `http://www.nytimes.com/2008/04/08/science/08monty.html` findet man eine hübsche Visualisierung zum Ziegenproblem (gelegentlich auch Monty Hall Problem nach einem amerikanischen Showmaster benannt).

Eine Variante des Ziegenproblems ist das *Gefangenenparadoxon*, im englischen *Three Prisoners Problem* (nicht zu verwechseln mit dem *Gefangenenproblem*, siehe 35). Wir zitieren einmal wieder nach Wikipedia:

> „In einem Gefängnis sitzen drei zum Tode verurteilte Gefangene: Anton, Brigitte und Clemens. Genau einer von ihnen soll begnadigt werden. Dazu wird ein Los gezogen, das allen die gleiche Chance gibt, begnadigt zu werden. Der Gefangene Anton, der also eine Überlebenswahrscheinlichkeit von 1/3 hat, bittet den Wärter, der das Ergebnis des Losentscheids kennt, ihm einen seiner Leidensgenossen Brigitte oder Clemens zu nennen, der oder die sterben muss. Der Wärter antwortet ‚Brigitte' und lügt nicht. Wie hoch ist nun Antons Überlebenswahrscheinlichkeit?"

Die Lösung ist: die Wahrscheinlichkeit bleibt $\frac{1}{3}$. Es ist auf den ersten Blick überraschend, dass Antons Überlebenschance immer noch $\frac{1}{3}$ ist, obwohl nur noch er und Clemens Todeskandidaten sind. Die Überlebenswahrscheinlichkeit von Clemens ist auf $\frac{2}{3}$ gestiegen. Woran liegt das? Man kann das Gefangenenparadoxon auf das Ziegenproblem zurückführen. Die drei Tore im Ziegenproblem entsprechen den drei Gefangenen. Die Existenz eines Gewinnes hinter einem bestimmten Tor entspricht der Begnadigung eines Gefangenen, weiter das Öffnen eines Tores der Nennung eines Opfers und der Showmaster im Ziegenproblem dem Wärter im Gefangenenparadoxon.

35 Das Gefangenenproblem

In der Literatur gibt es viele unterschiedliche Gefangenenprobleme, so dass man eigentlich nicht von *dem* Gefangenenproblem sprechen kann. Wir wollen uns hier hauptsächlich mit dem folgenden Problem beschäftigen:

> Es gibt n Gefangene, wobei n eine gerade Zahl ist. Diese Gefangenen haben alle Ausweise, die in n Fächer derart verteilt sind, dass in jedem Fach genau ein Ausweis liegt. Jeder Gefangene darf alleine zu den Fächern gehen und $n/2$ beliebige Fächer öffnen, um zu sehen, ob sein Ausweis in dem Fach liegt. Es findet keine Kommunikation zwischen den Gefangenen statt und die Ausweise bleiben in dem jeweiligen Fach. Findet *jeder* Gefangene seinen Ausweis, so werden *alle* freigelassen. Findet aber ein einziger Gefangener seinen Ausweis nicht, so bleiben alle gefangen.

Die Wahrscheinlichkeit, dass ein einzelner Gefangener seinen Ausweis findet, ist $\frac{1}{2}$, da er ja die Hälfte aller Fächer öffnen darf. Die Wahrscheinlichkeit, dass alle n Gefangenen ihren Ausweis finden, ist daher $(\frac{1}{2})^n$, schon für moderat großes n eine kleine Zahl. Um eine bessere Strategie zu erhalten und beschreiben können, denken wir uns die Gefangenen durchnummeriert, die Menge der Gefangenen sei also $\{1,\dots,n\}$. Ebenso denken wir uns die Fächer durchnummeriert. Wir nehmen an, dass im i-ten Fach der Ausweis des $\pi(i)$-ten Gefangenen, $i = 1,\dots,n$, liegt. Dann ist $\{\pi(1),\dots,\pi(n)\}$ eine *Permutation* von $\{1,\dots,n\}$, d. h. die Abbildung $\pi : \{1,\dots,n\} \longrightarrow \{1,\dots,n\}$ ist bijektiv (also injektiv und surjektiv). Wir schreiben diese auch in der Form

$$\pi = \begin{pmatrix} 1 & 2 & \cdots & n \\ \pi(1) & \pi(2) & \cdots & \pi(n) \end{pmatrix}.$$

Jede Permutation kann in der Zykelschreibweise dargestellt werden. Hierzu wähle man $k \in \{1,\dots,n\}$ und nennt $\begin{pmatrix} k & \pi(k) & \pi^2(k) & \cdots & \pi^{|k|-1} \end{pmatrix}$ einen *Zykel*, wobei $|k|$, die *Länge* des Zykel, minimal mit der Eigenschaft $\pi^{|k|}(k) = k$ ist, ferner naheliegenderweise die Potenzen von π induktiv durch $\pi^1(k) := \pi(k)$ und $\pi^{j+1}(k) := \pi(\pi^j(k))$ definiert sind. Ist z. B. die Permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 5 & 3 & 2 & 8 & 9 & 1 & 4 & 6 & 7 & 11 & 10 & 12 \end{pmatrix}$$

gegeben, so lässt sich π in der Zykelschreibweise darstellen als

$$\pi = \begin{pmatrix} 1 & 5 & 9 & 7 & 4 & 8 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 \end{pmatrix} \begin{pmatrix} 10 & 11 \end{pmatrix} \begin{pmatrix} 12 \end{pmatrix}.$$

Natürlich ist die Permutation π den Gefangenen nicht bekannt, sie legt aber für den k-ten Gefangenen, $k = 1, \ldots, n$, die folgende Strategie nahe: Der k-te Gefangene geht zum Fach k. Liegt dort sein Ausweis, so ist $\pi(k) = k$ und er ist fertig. Andernfalls geht er zum Fach $\pi(k)$ und schaut nach, ob dort sei Ausweis liegt bzw. $\pi^2(k) = k$ gilt. Ist dies der Fall, so ist er fertig. Andernfalls gehe er zum Fach mit der Nummer $\pi^2(k)$ usw. Der k-te Gefangene erreicht also seinen Ausweis rechtzeitig, also durch das Öffnen von höchstens $n/2$ Fächern, wenn der mit k beginnende Zykel der Permutation π eine Länge $|k| \leq n/2$ besitzt. Im obigen Beispiel einer Permutation mit 12 Elementen gibt es einen Zykel der Länge 7, die entsprechenden Gefangenen mit den Nummern 1, 5, 9, 7, 4, g und 6 würden mit obiger Strategie ihren Ausweis nicht rechtzeitig erhalten. *Alle* Gefangenen erhalten ihren Ausweis mit der obigen Strategie rechtzeitig, wenn es in der Permutation π *keinen* Zykel mit einer Länge $> n/2$ gibt. Die Wahrscheinlichkeit $p(n)$ dafür, dass alle n Gefangenen frei kommen, ist daher der Quotient aus der Anzahl der Permutationen mit n Elementen, die keinen Zykel der Länge $> n/2$ besitzen, und der Anzahl aller Permutationen von n Elementen. Da letztere Anzahl bekanntlich $n!$ ist, ist

$$p(n) = 1 - \frac{q(n)}{n!},$$

wobei $q(n)$ die Anzahl der Permutationen von n Elementen ist, die einen Zykel der Länge $> \frac{1}{2}n$ besitzen. Es ist $q(n) = \sum_{i=n/2+1}^{n} q_i(n)$, wobei $q_i(n)$ die Anzahl der Permutationen einer n-elementigen Menge ist, die einen Zykel der Länge[45] i besitzen. Aus einer n-elementigen Menge kann man auf $\binom{n}{i}$ verschiedene Weise eine i-elementige Teilmenge auswählen. In einer i-elementigen Menge halte man ein Element fest. Die übrigen $(i-1)$ Elemente lassen sich auf $(i-1)!$ verschiedene Weise anordnen. Daher gibt es in einer n-elementigen Menge $\binom{n}{i}(i-1)!$ verschiedene Zykel der Länge i. Die übrigen $(n-i)$ Elemente lassen sich auf $(n-i)!$ Arten anordnen. Daher ist

$$q_i(n) = \binom{n}{i}(i-1)!(n-i)!$$

und folglich

$$p(n) = 1 - \sum_{i=n/2+1}^{n} \frac{n!\,(i-1)!\,(n-i)!}{i!\,(n-i)!\,n!} = 1 - \sum_{i=n/2+1}^{n} \frac{1}{i}.$$

Nun ist nach einfacher Argumentation

$$\int_{n/2}^{n} \frac{dx}{x} \geq \sum_{i=n/2+1}^{n} \frac{1}{i}$$

und daher

$$p(n) = 1 - \sum_{i=n/2+1}^{n} \frac{1}{i} \geq 1 - \int_{n/2}^{n} \frac{dx}{x} = 1 - \ln n + \ln(n/2) = 1 - \ln 2 \geq 0.3068.$$

[45] Wegen $i > n/2$ ist dies automatisch ein Zykel maximaler Länge.

Mit der angegebenen Strategie ist die Wahrscheinlichkeit, dass alle Gefangenen freikommen, also mindestens 30%.

Als Gefangenenproblem wird gelegentlich aber auch die folgende Aufgabe verstanden:

> Sie sitzen in einer Zelle mit zwei Türen. Vor jeder steht ein Wächter. Einer von beiden, Sie wissen aber nicht welcher, sagt immer die Wahrheit, der andere nie. Sie können die Zelle durch eine der beiden Türen verlassen. Die eine führt zum Henker, die andere in die Freiheit. Bevor Sie entscheiden, welche Tür Sie wählen, dürfen Sie eine Frage stellen. Wie lautet die Frage, die in die Freiheit führt?

Die Lösung besteht darin, einen der beiden Wärter zu fragen: „Wenn ich den anderen Wärter frage, welche Tür in die Freiheit führt, welche Tür zeigt er mir?“ Durch die andere Tür sollte ich die Zelle verlassen. Weshalb? Im Prinzip liegt das daran, dass $(-1) \cdot (+1) = -1$ ist. Denn frage ich den Lügner, so zeigt er mir die Tür zum Henker. Das gleiche gilt, wenn ich den frage, der immer die Wahrheit sagt.

Und noch ein Gefangenenproblem:

> Drei (intelligente) Gefangene bekommen die Chance, freigesprochen zu werden, sofern einer von ihnen die folgende Aufgabe lösen. Ihnen werden die Augen verbunden und jeder von ihnen bekommt einen von fünf Hüten auf den Kopf. Es sind drei schwarze und zwei weiße Hüte. Die Augenbinden werden abgenommen, so dass jeder von ihnen sehen kann, welche Farbe die Hüte der anderen beiden Gefangenen haben, nicht aber, welche Farbe der Hut auf dem eigenen Kopf hat. Die Aufgabe besteht darin, die Farbe des eigenen Hutes zu bestimmen. Außerdem gilt die Regel: Wenn man eine falsche Aussage macht, so wird die Haftstrafe verdoppelt, es kann aber die Aussage verweigert werden. Der erste Gefangene sagt nichts. Der zweite Gefangene sagt auch nichts. Der dritte jedoch sagt: „Ich weiß, welche Farbe mein Hut hat.“ Noch erstaunlicher ist, dass der dritte Gefangene blind ist!

Die richtige Antwort ist: Der Hut ist schwarz. Da die Argumentation relativ einfach ist und mit Mathematik kaum etwas zu tun hat, wollen wir auf eine Begründung verzichten.

Bei `http://www.php-resource.de/forum/showthread/t-44276.html` findet man eines der unzähligen Probleme, in denen Gefangene und Hüte vorkommen:

> In einem Gefangenenlager befinden sich n Gefangene. Der Kommandant stellt den Gefangen eine Aufgabe. Jeder der diese Aufgabe lösen kann, kommt frei. Die Aufgabe: Die n Gefangenen stellen sich in einer Reihe auf, so dass jeder seine Vordermänner sehen kann. Nun bekommt jeder Gefangene einen Hut auf. Jeder Hut hat eine der vier Farben: rot, gelb, grün oder blau. Jeder Gefangene kann die Farbe seines Hutes nicht sehen. Der Kommandant fängt von hinten an zu fragen: „Welche Farbe hat dein Hut?“ Die Gefangenen können die Antworten der vorherigen Gefangenen hören. Die Gefangenen bekommen 5 Minuten Zeit, sich vorher eine Strategie auszudenken und abzusprechen. Wie viele Gefangene können freikommen und welche Strategie haben sie? Hierbei darf nur eine der vier Farben genannt werden.

Die Antwort: Die ersten $n-1$ Gefangenen können freikommen, der letzte (ihn beißen bekanntlich die Hunde) kommt nur durch Glück frei. Hierzu einigen die Gefangenen sich auf eine Codierung der Farben. Das könnte folgendermaßen aussehen:

Farbe	Punkte
rot	1
gelb	2
grün	3
blau	4

Der Hut des i-ten Gefangenen habe den Punktwert p_i. Der i-te Gefangene kann die Hüte der $i-1$ Gefangenen vor ihm sehen und daher die Summe $q_i := \sum_{k=1}^{i-1} p_k$, $i = 1, \ldots, n$, berechnen. Die leere Summe ist gleich Null, d. h. $q_1 = 0$. Jeder Gefangene rechnet den Rest von q_i bei Teilung durch 4 aus, also $r_i := q_i \bmod 4$, wobei $r_i \in \{0, 1, 2, 3\}$. Der letzte Gefangene teilt dem Kommandanten (und damit den übrigen Gefangenen) die Farbe mit, die zum Rest r_n gehört: blau für 0 (bzw. 4), rot für 1, gelb für 2 und grün für 3. Der $(n-1)$-te Gefangene kennt seinen Rest r_{n-1} und den Rest r_n des n-ten Gefangenen. Da $p_{n-1} \equiv (r_n - r_{n-1}) \bmod 4$ (d. h. die Differenz $p_{n-1} - (r_n - r_{n-1})$ ist durch 4 teilbar) kennt der vorletzte Gefange die Farbe seines Hutes. Es ist lediglich $r_n - r_{n-1}$ durch eventuelle Addition von 4 auf eine zwischen 1 und 4 zurückzuführen und die Codierungstafel zu benützen. So kann offenbar fortgefahren werden und sukzessive können die ersten $n-1$ Gefangenen die Farbe ihres Hutes bestimmen. Wir wollen das Verfahren durch ein Beispiel mit 10 Gefangenen illustrieren.

i	Farbe	p_i	q_i	r_i	$r_{i+1} - r_i$		Farbe
1	rot	1	0	0	1	1	rot
2	rot	1	1	1	1	1	rot
3	blau	2	2	2	−2	2	blau
4	grün	3	4	0	3	3	grün
5	gelb	4	7	3	0	4	gelb
6	grün	3	11	3	−1	3	grün
7	rot	1	14	2	1	1	rot
8	gelb	4	15	3	0	4	gelb
9	rot	1	19	3	−3	1	rot
10	blau	2	20	0		4	gelb

Alle haben ihre Hutfarben richtig berechnet, nur der letzte hat Pech gehabt.

Eine sehr einfache (deshalb geben wir auch keine Lösung an) Aufgabe mit Gefangenen und Hüten findet man bei `http://www.logisch-gedacht.de/logikraetsel/bergwerk/`:

> In einem fernen Land wurden von einem Stamm der Ureinwohner 50 Gefangene gehalten. Diese mussten für den Stammesführer in einem Bergwerk arbeiten. Als eines Tages der Stammesführer 50 Jahre alt wurde, entschloss er sich, seinen Gefangenen die Chance zu geben, ihre Freiheit zurückzuerlangen. Er brachte sie alle in die Dunkelheit des Bergwerks und setzte jedem einen Hut auf. Dann

sagte er zu ihnen: „Ich habe jedem von euch einen Hut aufgesetzt - entweder einen schwarzen oder einen weißen. Gleich werdet ihr einer nach dem anderen aus dem Bergwerk gelassen. Eure Aufgabe ist es, euch in einer Reihe nach Farben sortiert aufzustellen: die mit den schwarzen Hüten links, die mit den weißen Hüten rechts. Ihr dürft euch weder unter einander verständigen noch euren eigenen Hut ansehen. Gelingt euch das, so werdet ihr die Freiheit erhalten.“ Wie konnten die Gefangenen diese Aufgabe bewältigen?

36 Der Satz von Stone-Weierstraß

Der berühmte Weierstraßsche Approximationssatz sagt aus, dass sich jede auf einem kompakten Intervall stetige Funktion auf diesem Intervall gleichmäßig durch eine Folge von Polynomen approximieren lässt. Eine weitreichende Verallgemeinerung stammt von M. H. Stone (1937, 1948). Dieser Satz von Stone-Weierstraß ist für mich einer der erstaunlichsten Sätze der Mathematik. Ich hörte als Student das erste Mal von diesem Satz in einer Vorlesung von Hel Braun über Topologie an der Universität Hamburg. In dieser speziellen Vorlesung wurde Hel Braun krankheitshalber von Emil Artin vertreten!

Satz (Stone-Weierstraß) *Sei $B \subset \mathbb{R}^n$ kompakt und $C(B)$ die Menge der auf B definierten, stetigen und reellwertigen Funktionen. Auf $C(B)$ definieren wir die Maximumnorm $\|\cdot\|_\infty$ durch $\|f\|_\infty := \max_{x\in B} |f(x)|$. Sei $\mathcal{A} \subset C(B)$ eine Teilalgebra, d. h. ein linearer Teilraum mit der Eigenschaft, dass mit $f, g \in \mathcal{A}$ auch[46] $f \cdot g \in \mathcal{A}$. Es gelte:*

1. *Es ist $1 \in \mathcal{A}$, wobei $1(x) := 1$ für alle $x \in B$, d. h. $\mathcal{A}$ enthält die konstanten Funktionen.*
2. *Zu $x, y \in B$ mit $x \neq y$ existiert ein $g \in \mathcal{A}$ mit $g(x) \neq g(y)$, d. h. $\mathcal{A}$ trennt Punkte von B.*

Dann existiert zu jedem $f \in C(B)$ eine Folge $\{g_k\} \subset \mathcal{A}$ die auf B gleichmäßig gegen f konvergiert, für die also $\lim_{k\to\infty} \|f - g_k\|_\infty = 0$.

Beweis: Mit $\mathrm{cl}(\mathcal{A})$ bezeichnen wir den Abschluss von $\mathcal{A}$ in $C(B)$ bezüglich gleichmäßiger Konvergenz auf B. D. h. es sei

$$\mathrm{cl}(\mathcal{A}) := \{f \in C(B) : \text{ Es existiert } \{g_k\} \subset \mathcal{A} \text{ mit } \lim_{k\to\infty} \|f - g_k\|_\infty = 0.\}$$

Wir haben zu zeigen, dass $\mathrm{cl}(\mathcal{A}) = C(B)$. Der Beweis erfolgt in mehreren Schritten.

(a) $\mathrm{cl}(\mathcal{A})$ ist eine Teilalgebra von $C(B)$, d. h. mit $f_1, f_2 \in \mathrm{cl}(\mathcal{A})$ sowie $\alpha \in \mathbb{R}$, sind $f_1 + f_2$, $f_1 \cdot f_2$ sowie $\alpha f_1 \in \mathrm{cl}(\mathcal{A})$.

Der Beweis hierfür ist offensichtlich.

(b) Sei $a > 0$ gegeben. Dann gibt es eine Folge $\{p_k\}$ von Polynomen mit

$$\lim_{k\to\infty} \max_{t\in[-a,a]} | \, |t| - p_k(t)| = 0,$$

welche also gleichmäßig auf $[-a, a]$ gegen die Betragsfunktion konvergiert.

[46] Für $f, g \in C(B)$ ist natürlich $f \cdot g \in C(B)$ durch $(f \cdot g)(x) := f(x)g(x)$ definiert.

O. B. d. A. ist $a = 1$. Definiere die Folge $\{p_k\}$ von Polynomen durch $p_0(t) := 0$ sowie $p_{k+1}(t) := p_k(t) + \frac{1}{2}(t^2 - p_k(t)^2)$, $k = 0, 1, \ldots$. In Abbildung 49 haben wir die Betragsfunktion sowie p_1, p_2 und p_3 dargestellt. Durch vollständige Induktion nach k zeigt

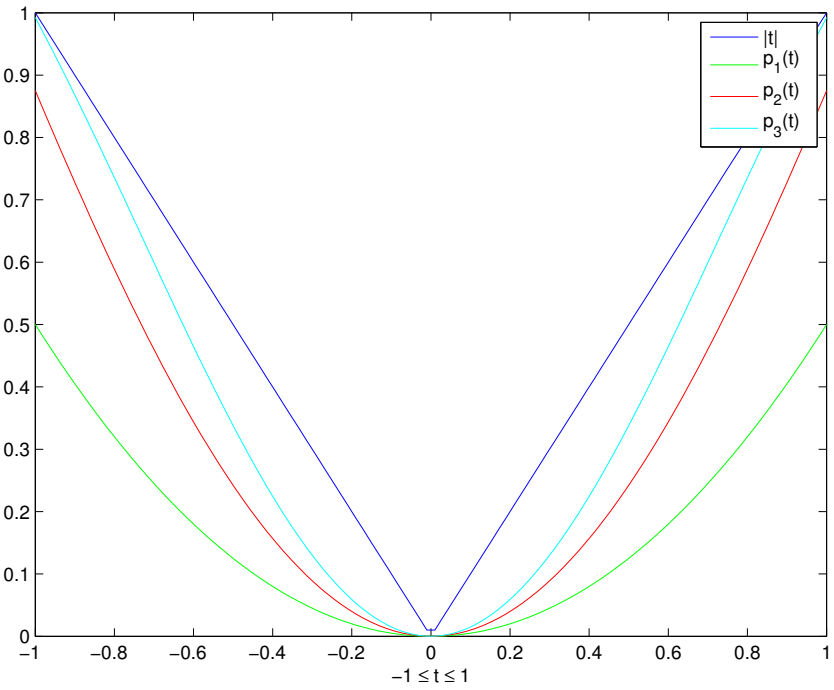


Abbildung 49: Approximation der Betragsfunktion durch Polynome

man, dass $p_k(t) \leq |t|$ und $0 \leq p_k(t) \leq p_{k+1}(t)$ für alle $t \in [-1, 1]$. Denn für $k = 0$ sind diese beiden Aussagen offenbar richtig. Sei dies auch für k der Fall. Wegen

$$|t| - p_{k+1}(t) = \underbrace{(|t| - p_k(t))}_{\geq 0}\underbrace{(1 - \tfrac{1}{2}(|t| + p_k(t)))}_{\geq 0} \geq 0$$

ist die erste Aussage auch für $k + 1$ richtig, was für die zweite Aussage wegen der Rekursionsformel evident ist. Also konvergiert $\{p_k\}$ auf $[-1, 1]$ punktweise und ist monoton nicht fallend. Der Limes ist die Betragsfunktion, insbesondere ist diese stetig. Aus dem Satz von Dini[47] folgt die auf $[-1, 1]$ gleichmäßige Konvergenz von $\{p_k\}$ gegen die Betragsfunktion.

(c) Ist $f \in \mathrm{cl}(\mathcal{A})$, so ist auch $|f| \in \mathrm{cl}(\mathcal{A})$, wobei $|f|(x) := |f(x)|$ für alle $x \in B$.

Wegen (b) existiert eine Folge $\{p_k\}$ von Polynomen mit

$$\lim_{k\to\infty} \max_{|s| \leq \|f\|_\infty} ||s| - p_k(s)| = 0.$$

[47]Der Satz von Dini sagt aus: Eine auf einem kompakten Intervall I monoton nicht fallende oder monoton nicht wachsende, punktweise konvergente Funktionenfolge mit stetiger Grenzfunktion ist auf I sogar gleichmäßig konvergent.

Da $\mathrm{cl}(\mathcal{A})$ nach (a) eine Teilalgebra von $C(B)$ ist und die konstanten Funktionen enthält, ist $p_k \circ f \in \mathrm{cl}(\mathcal{A})$ (wobei natürlich $(p_k \circ f)(x) := p_k(f(x))$). Wegen

$$\begin{aligned} \| \, |f| - p_k \circ f\|_\infty &= \max_{x \in B} | \, |f(x)| - p_k(f(x))| \\ &\leq \max_{|s| \leq \|f\|_\infty} | \, |s| - p_k(s)| \\ &\to 0 \quad \text{mit } k \to \infty \end{aligned}$$

ist $|f| \in \mathrm{cl}(\mathcal{A})$.

(d) Mit $f, g \in \mathrm{cl}(\mathcal{A})$ sind auch $\max(f,g), \min(f,g) \in \mathrm{cl}(\mathcal{A})$, wobei

$$\max(f,g)(x) := \max(f(x), g(x)), \quad \min(f,g)(x) := \min(f(x), g(x)).$$

Dies folgt sofort aus

$$\max(f,g) = \frac{1}{2}(f + g + |f - g|), \quad \min(f,g) = \frac{1}{2}(f + g - |f - g|)$$

sowie (a) und (c).

(e) Zu $x, y \in B$ mit $x \neq y$ und $\alpha, \beta \in \mathbb{R}$ existiert ein $g_{xy} \in \mathcal{A}$ mit $g_{xy}(x) = \alpha$ und $g_{xy}(y) = \beta$.

Denn: Da $\mathcal{A}$ nach Voraussetzung Punkte trennt, gibt es ein $g \in \mathcal{A}$ mit $g(x) \neq g(y)$. Dann setze man

$$g_{xy} := \frac{\alpha - \beta}{g(x) - g(y)} g + \frac{\beta g(x) - \alpha g(y)}{g(x) - g(y)} 1.$$

Nach diesen Vorbereitungen kommt jetzt der eigentliche Beweis des Satzes. Wir geben uns $f \in C(B)$ und ein $\epsilon > 0$ beliebig vor und zeigen die Existenz eines $g \in \mathrm{cl}(\mathcal{A})$ mit $\|f - g\|_\infty < \epsilon/2$, woraus offenbar die Behauptung folgt, da $g \in \mathrm{cl}(\mathcal{A})$ beliebig genau durch Elemente aus $\mathcal{A}$ approximiert werden kann.

(f) Bei vorgegebenen $f \in C(B)$ und $\epsilon > 0$ existiert zu jedem $x \in B$ ein $g_x \in \mathrm{cl}(\mathcal{A})$ mit $g_x(x) = f(x)$ und $g_x(z) < f(z) + \epsilon/2$ für alle $z \in B$.

Zu jedem Paar $(x, y) \in B \times B$ existiert nach (e) ein $g_{xy} \in \mathcal{A}$ mit $g_{xy}(x) = f(x)$ und $g_{xy}(y) = f(y)$ (setze $g_{xx} := f(x) \cdot 1$). Nun definiere man

$$O_{xy} := \{z \in B : g_{xy}(z) < f(z) + \epsilon/2\}.$$

Dann ist O_{xy} eine offene Teilmenge von B und $y \in O_{xy}$, so dass $B = \bigcup_{y \in B} O_{xy}$ für beliebiges $x \in B$ eine offene Überdeckung von B ist. Da B kompakt ist, existiert eine endliche Teilüberdeckung $B = \bigcup_{i=1}^{M} O_{xy_i}$ mit $\{y_1, \ldots, y_M\} \subset B$. Hierbei sind M und die y_i i. Allg. von x abhängig. Man definiere

$$g_x := \min(g_{xy_1}, \ldots, g_{xy_M}).$$

Es ist $g_x \in \mathrm{cl}(\mathcal{A})$ (wegen (d)). Wegen $g_{xy_i}(x) = f(x)$ ist auch $g_x(x) = f(x)$. Ein beliebiges $z \in B$ ist in O_{xy_i} mit einem gewissen $i \in \{1, \ldots, M\}$ enthalten, so dass $g_x(z) \leq g_{xy_i}(z) < f(z) + \epsilon/2$. Damit ist auch der Beweisschritt (f) beendet.

(g) Es existiert ein $g \in \mathrm{cl}(\mathcal{A})$ mit $f(z) - \epsilon/2 < g(z) < f(z) + \epsilon/2$ für alle $z \in B$ bzw. mit $\|f - g\| < \epsilon/2$.

Zum Nachweis definiere man

$$O_x := \{z \in B : g_x(z) > f(z) - \epsilon/2\}.$$

Für jedes $x \in B$ ist O_x offen und $x \in O_x$ wegen $g_x(x) = f(x)$ (siehe (f)). Dann ist auch $B = \bigcup_{x \in B} O_x$ eine offene Überdeckung von B. Wiederum kann eine endliche Teilüberdeckung $B = \bigcup_{j=1}^{N} O_{x_j}$ mit $\{x_1, \ldots, x_N\} \subset B$ ausgewählt werden. Nun setze man

$$g := \max(g_{x_1}, \ldots, g_{x_N}).$$

Bei gegebenem $z \in B$ ist einerseits $g(z) = g_{x_j}(z)$ für ein gewisses $j \in \{1, \ldots, N\}$ und daher $g(z) < f(z) + \epsilon/2$. Andererseits ist $z \in O_{x_k}$ mit einem gewissen $k \in \{1, \ldots, N\}$. Daher ist $g(z) \geq g_{x_k}(z) > f(z) - \epsilon/2$. Auch der Beweisschritt (g) ist beendet und damit der Satz von Stone-Weierstraß bewiesen. □

Folgerung Sei $B \subset \mathbb{R}^n$ kompakt und $\mathcal{P}$ die Menge der reellwertigen Polynome in n Variablen. Dann ist $\mathcal{P}$ eine Teilalgebra von $C(B)$, die die konstanten Funktionen enthält und Punkte von B trennt. Wegen des Satzes von Stone-Weierstraß ist $\mathrm{cl}(\mathcal{P}) = C(B)$, jede auf B stetige, reellwertige Funktion kann also auf B gleichmäßig durch Polynome approximiert werden. Dies gilt insbesondere für $B := [a, b] \subset \mathbb{R}$, womit auch der klassische Weierstraßsche Approximationssatz bewiesen ist.

37 Der Brouwersche Fixpunktsatz

Der Brouwersche Fixpunktsatz ist einer der schönsten Sätze der Mathematik, er hat viele Anwendungen und ist nicht leicht zu beweisen. Dies ist die Aussage:

Brouwerscher Fixpunktsatz *Sei $K \subset \mathbb{R}^n$ nichtleer, konvex und kompakt, ferner $f\colon K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ eine stetige Abbildung mit $f(K) \subset K$. Dann besitzt f mindestens einen Fixpunkt in K, d. h. es existiert ein $x^* \in K$ mit $f(x^*) = x^*$.*

Bei J. Franklin (1980) kann man nachlesen: The Brouwer fixed-point theorem is one of the most important results in modern mathematics. It is easy to state but hard to prove. The statement is easy enough to be understood by anyone, even by someone who can't add or subtract. But its proof has usually been so difficult that it has been taught only in graduate courses on topology.... This is what the Brouwer theorem says in everyday terms. Sit down with a cup of coffee. Gently and continuously swirl the coffee about in the cup. Put the cup down, and let the motion subside. When the coffee is still, Brouwer says there is at least one point in the coffee that has returned to the exact spot in the cup where it was when you first sat down.

Bemerkung: Es genügt, den oben formulierten Brouwerschen Fixpunktsatz für $K := B^n[0; 1]$, die (euklidische) Einheitskugel im $\mathbb{R}^n$, zu beweisen. Denn sei allgemein $K \subset \mathbb{R}^n$ nichtleer, konvex und kompakt. Als beschränkte Menge ist K in einer hinreichend großen (euklidischen) Kugel $B^n[0; r]$ um den Nullpunkt mit dem Radius $r > 0$ enthalten. Da K nichtleer, konvex und abgeschlossen ist, existiert zu jedem $x \in \mathbb{R}^n$ genau ein

$P_K(x) \in K$ (die sogenannte Projektion von x auf K) mit $\|P_k(x) - x\|_2 \leq \|y - x\|_2$ für alle $y \in K$. Die Projektion $P_K(x)$ von x auf K ist charakterisiert durch

$$(x - P_K(x))^T(y - P_K(x)) \leq 0 \quad \text{für alle } y \in K.$$

Hieraus folgt leicht $\|P_K(u) - P_K(v)\|_2 \leq \|u - v\|_2$ für alle $u, v \in \mathbb{R}^n$, insbesondere ist die Abbildung $P_K : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig. Diese Aussagen nennt man auch den *Projektionssatz für konvexe Mengen im* $\mathbb{R}^n$[48]. Eine Veranschaulichung findet man in Abbildung 50. Nun definiere man $g\colon B^n[0;1] \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch $g(z) := (1/r)f(P_K(rz))$. Dann ist g als Komposition stetiger Abbildungen offensichtlich selbst stetig, nach Konstruktion ist weiter $g(B^n[0;1]) \subset B^n[0;1]$. Ist der Brouwersche Fixpunktsatz für die Einheitskugel richtig, so existiert also ein $z^* \in B^n[0;1]$ mit $g(z^*) = z^*$ bzw. $f(P_K(rz^*)) = rz^*$. Als Projektion von rz^* auf K ist $P_K(rz^*) \in K$, wegen $f(K) \subset K$ ist $rz^* \in K$ und daher $P_K(rz^*) = rz^*$. Folglich ist $x^* := rz^*$ ein Fixpunkt von f in K. □

Es gibt viele Beweise des Brouwerschen Fixpunktsatzes. Im oben genannten Buch von Joel Franklin wird ein Beweis von J. Milnor (1980) reproduziert. Kurze Zeit später erschien eine Vereinfachung von C. A. Rogers (1980). Wir folgen im wesentlichen R. Howard (2004). Mit $\|\cdot\|$ (wir lassen den oft üblichen Index $_2$ weg) bezeichnen wir die euklidische Norm auf dem $\mathbb{R}^n$, mit $B^n := \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ die abgeschlossene Einheitskugel im $\mathbb{R}^n$, mit $B_0^n := \{x \in \mathbb{R}^n : \|x\| < 1\}$ die offene Einheitskugel im $\mathbb{R}^n$ und mit $S^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$ die Einheitssphäre im $\mathbb{R}^n$ (jeweils bezüglich

[48] **Projektionssatz für konvexe Mengen im** $\mathbb{R}^n$ *Sei $K \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex sowie $x \in \mathbb{R}^n$. Dann existiert genau ein $P_K(x) \in K$, die Projektion von x auf K, mit*

$$\|P_k(x) - x\|_2 \leq \|y - x\|_2 \quad \textit{für alle } y \in K.$$

Ein $y^* \in K$ ist genau dann die Projektion von x auf K, wenn $(x - y^*)^T(y - y^*) \leq 0$ für alle $y \in K$.
Beweis: Wir betrachten die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(y) := \frac{1}{2}\|y - x\|_2^2, \quad y \in K,$$

und überlegen uns, dass (P) genau eine Lösung besitzt. Mit einem beliebigen $y_0 \in K$ betrachte man hierzu die Aufgabe, die Zielfunktion $f(\cdot)$ von (P) auf $K_0 := K \cap \{y \in K : f(y) \leq f(y_0)\}$ zu minimieren. Die Menge $K_0 \subset \mathbb{R}^n$ ist beschränkt und abgeschlossen, also kompakt, und $f(\cdot)$ stetig. Daher nimmt $f(\cdot)$ auf K_0 sein Minimum an, und dieses ist auch eine Lösung von (P). Damit ist die Existenz einer Lösung $y^* = P_K(x)$ von (P) geklärt. Für ein beliebiges $y \in K$ und $t \in (0,1)$ ist

$$0 \leq f(y^* + t(y - y^*)) - f(y^*) = t(y^* - x)^T(y - y^*) + \frac{t^2}{2}\|y - y^*\|_2^2.$$

Nach Division durch t und anschließendem Grenzübergang $t \to 0+$ erhält man $(x - y^*)^T(y - y^*) \leq 0$ für jede Lösung y^* von (P) und alle $y \in K$. Sind $y_1^*, y_2^* \in K$ Lösungen von (P), so ist $(x - y_1^*)^T(y_2^* - y_1^*) \leq 0$ und $(x - y_2^*)^T(y_1^* - y_2^*) \leq 0$. Addition dieser beiden Ungleichungen liefert $(y_2^* - y_1^*)^T(y_2^* - y_1^*) \leq 0$ und damit $y_1^* = y_2^*$. Folglich existiert genau eine Lösung von (P). Ist schließlich $y^* \in K$ und $(x - y^*)^T(y - y^*) \leq 0$ für jedes $y \in K$, so ist

$$f(y) - f(y^*) = \underbrace{(y^* - x)^T(y - y^*)}_{\geq 0} + \frac{1}{2}\|y - y^*\|_2^2 \geq 0,$$

also y^* eine Lösung von (P). Damit ist der Projektionssatz für konvexe Mengen im $\mathbb{R}^n$ bewiesen. □

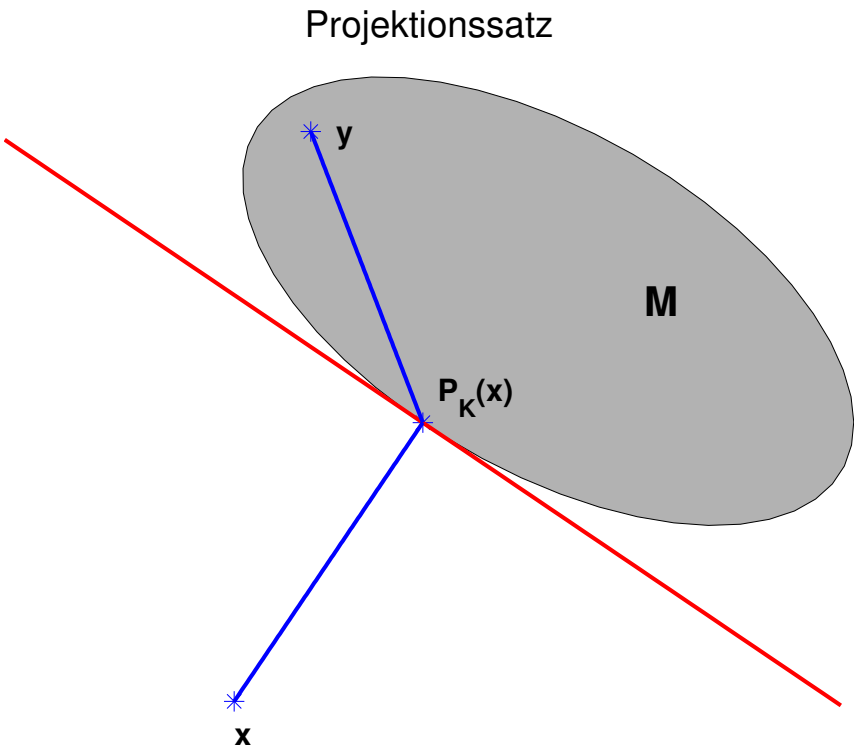


Abbildung 50: Der Projektionssatz

der euklidischen Norm). Wir beginnen mit (Lemma 2 bei C. A. Rogers (1980) und Lemma 1 bei R. Howard (2004)):

Lemma Es gibt keine stetig differenzierbare Abbildung $f\colon B^n \longrightarrow S^{n-1}$ mit $f(x) = x$ für alle $x \in S^{n-1}$.

Beweis: Der Beweis erfolgt durch Widerspruch. Wir nehmen also an, dass eine stetig differenzierbare Abbildung $f\colon B^n \longrightarrow S^{n-1}$ mit $f(x) = x$ für alle $x \in S^{n-1}$ existiert. Für $t \in [0,1]$ definieren wir

$$f_t(x) := (1-t)x + tf(x) = x + tg(x),$$

wobei $g(x) := f(x) - x$. Für $x \in B^n$ ist

$$\|f_t(x)\| \leq (1-t)\underbrace{\|x\|}_{\leq 1} + t\underbrace{\|f(x)\|}_{=1} \leq (1-t) + t = 1$$

und daher ist $f_t : B^n \longrightarrow B^n$. Für alle $x \in S^{n-1}$ ist

$$f_t(x) = (1-t)x + tf(x) = (1-t)x + tf(x) = x,$$

d. h. jeder Punkt von S^{n-1} ist ein Fixpunkt von f_t. Da f stetig differenzierbar ist, ist es auch g. Insbesondere ist g lipschitzstetig auf B^n, es exitiert also eine Konstante $C > 0$ mit

$$\|g(x_2) - g(x_1)\| \leq C\,\|x_2 - x_1\| \quad \text{für alle } x_1, x_2 \in B^n.$$

Nun nehmen wir an, es seien $x_1, x_2 \in B^n$ zwei verschiedene Punkte mit $f_t(x_1) = f_t(x_2)$. Dann ist $x_2 - x_1 = t(g(x_1) - g(x_2))$ und daher

$$\|x_2 - x_1\| = t\,\|g(x_1) - g(x_2)\| \leq Ct\,\|x_2 - x_1\|.$$

Wegen $x_1 \neq x_2$ impliziert dies $Ct \geq 1$. Ist also $t < 1/C$, so ist $f_t\colon B^n \longrightarrow B^n$ injektiv. Wir definieren $G_t := f_t(B_0^n) \subset B^n$ als Bild der offenen Einheitskugel B_0^n unter f_t. Mit $f'(x) = ((\partial f_i/\partial x_j)(x)) \in \mathbb{R}^{n\times n}$ bezeichnen wir die Funktionalmatrix (oder auch Jacobi-Matrix) von f in x. Offenbar ist

$$f_t'(x) = I + tg'(x),$$

wobei I die Einheitsmatrix im $\mathbb{R}^n$ ist. Da $\det(f_t'(x))$ stetig von t und x abhängt und $\det(f_0'(x)) = 1$ gilt, existiert ein $t_0 > 0$ mit $\det(f_t'(x)) > 0$ für alle $t \in [0, t_0]$ und alle $x \in B^n$. Wegen des Satzes über die inverse Funktion[49] ist G_t für alle $t \in [0, t_0]$ eine offene Menge, insbesondere ist $G_t \subset B_0^n$. Indem man t_0 notfalls noch kleiner macht, können wir annehmen, dass f_t für alle $t \in [0, t_0]$ injektiv ist.

Wir behaupten nun, dass $G_t = B_0^n$ für alle $t \in [0, t_0]$. Angenommen, dies sei für ein $t \in [0, t_0]$ nicht der Fall. Dann gibt es einen Punkt $y_0 \in \partial G_t \cap B_0^n$, wobei $\partial G_t = \mathrm{cl}(G_t)\backslash G_t$ den Rand von G_t bedeutet. Denn ist G_t eine echte Teilmenge von B_0^n, so gibt es ein $v_0 \in B_0^n$ mit $v_0 \notin G_t$. Man wähle $u_0 \in G_t$ beliebig. Dann ist $\{(1-\lambda)u_0 + \lambda v_0 : \lambda \in [0,1]\} \subset B_0^n$ und man kann ein $\lambda \in (0,1]$ mit $y_0 := (1-\lambda)u_0 + \lambda v_0 \in \partial G_t \cap B_0^n$ bestimmen. Da $y_0 \in \partial G_t$, existiert eine Folge $\{f_t(x_k)\} \subset G_t$ mit $\lim_{k\to\infty} f_t(x_k) = y_0$. Wegen der Kompaktheit von B^n kann man von $\{x_k\} \subset B_0^n \subset B_n$ zu einer konvergenten Teilfolge übergehen und annehmen, dass $\lim_{k\to\infty} x_k = x_0$ mit einem $x_0 \in B^n$. Wegen der Stetigkeit von f und damit von f_t ist $f_t(x_0) = y_0$. Wegen $y_0 \notin G_t$ ist $x_0 \in B^n \setminus B_0^n = S^{n-1}$. Dann ist aber x_0 ein Fixpunkt von f_t, also $y_0 = f_t(x_0) = x_0 \in S^{n-1}$, ein Widerspruch zu $y_0 \in B_0^{n-1}$. Damit ist $G_t = f_t(B_0^n) = B_0^n$ bewiesen. Insgesamt ist $f_t\colon B_0^n \longrightarrow B_0^n$ für jedes $t \in [0, t_0]$ eine Bijektion.

Definiere die Funktion $F\colon [0,1] \longrightarrow \mathbb{R}$ durch

$$F(t) := \int_{B_0^n} \det(f_t'(x))\,dx = \int_{B_0^n} \det(I + tg'(x))\,dx.$$

Offensichtlich ist F ein Polynom in t. Wie wir gerade gezeigt haben, ist $f_t\colon B_0^n \longrightarrow B_0^n$ für jedes $t \in [0, t_0]$ eine Bijektion. Wegen der Transformationsformel für mehrfache Integrale ist $F(t)$ das Volumen des Bildes $f_t(B_0^n) = B_0^n$, also $F(t) = \mathrm{vol}(B_0^n)$ für jedes $t \in [0, t_0]$. Ist aber ein Polynom auf einem Intervall konstant, so ist es überall konstant. Also gilt $F(t) = \mathrm{vol}(B_0^n)$ für alle $t \in [0,1]$. Insbesondere ist $F(1) = \mathrm{vol}(B_0^n) > 0$. Nun überlegen wir uns, dass $\det(f_1'(x)) = 0$ für jedes $x \in B_0^n$, woraus $F(1) = 0$ folgt und damit der gewünschte Widerspruch erreicht ist. Seien hierzu $x \in B_0^n$ und $v \in \mathbb{R}^n$

[49] Der Satz über die inverse Funktion (oder auch Satz von der Umkehrfunktion) sagt aus: Sei $U \subset \mathbb{R}^n$ offen und $F\colon U \longrightarrow \mathbb{R}^n$ stetig differenzierbar. Sei $u \in U$, $v := F(u)$ und die Funktionalmatrix $F'(u)$ invertierbar bzw. nichtsingulär. Dann gibt es eine offene Umgebung $U_u \subset U$ von u und eine offene Umgebung V_v von v, so dass F die Menge U_u bijektiv auf V_v abbildet und die inverse Funktion (bzw. Umkehrfunktion) $G := F^{-1} : V_v \longrightarrow U_u$ stetig differenzierbar ist. Es gilt $G'(F(u)) = (F'(u))^{-1}$.

beliebig. Für alle hinreichend kleinen $t > 0$ ist $x + tv \in B_0^n$ und daher $f_1(x + tv) = f(x + tv) \in S^{n-1}$. Daher ist

$$0 = \frac{d}{dt}\|f_1(x+tv)\|^2\Big|_{t=0} = \frac{d}{dt}f_1(x+tv)^T f_1(x+tv)\Big|_{t=0} = 2v^T f_1'(x)^T f_1(x).$$

Da dies für alle $v \in \mathbb{R}^n$ gilt, ist $f_1'(x)^T f_1(x) = 0$. Da $f_1(x) \neq 0$ als Element von S^{n-1}, ist $f_1'(x)^T$ und dann auch $f_1'(x)$ singulär bzw. $\det(f_1'(x)) = 0$. Damit ist das Lemma schließlich bewiesen. □

Nun ist der Beweis des Brouwerschen Fixpunktsatzes nicht mehr schwer.

Beweis des Brouwerschen Fixpunktsatzes: Es genügt zu zeigen, dass eine stetige Abbildung f, die die abgeschlossene Einheitskugel B^n in sich abbildet, einen Fixpunkt in B^n besitzt. Wegen des Satzes von Stone-Weierstraß (siehe 36) gibt es es eine Folge $\{p_k\}_{k\in\mathbb{N}}$ stetig differenzierbarer Funktionen $p_k\colon B^n \longrightarrow \mathbb{R}^n$ (diese können sogar als Polynome gewählt werden) mit $\|f(x) - p_k(x)\| \leq 1/k$ für alle $x \in B^n$, welche also insbesondere auf B^n gleichmäßig gegen f konvergiert. Dann ist

$$\|p_k(x)\| \leq \underbrace{\|f(x)\|}_{\leq 1} + \|f(x) - p_k(x)\| \leq 1 + \frac{1}{k} \quad \text{für alle } x \in B^n.$$

Definiert man $q_k := (1+1/k)^{-1}p_k$, so ist $q_k\colon B^n \longrightarrow B^n$ und $\{q_k\}$ konvergiert gleichmäßig gegen f. Wir behaupten, dass jedes q_k einen Fixpunkt in B^n besitzt. Ist dies für ein k nicht richtig, so definiere man die Abbildung $f_k\colon B^n \longrightarrow S^{n-1}$ dadurch, dass $f_k(x)$ für $x \in B^n$ der Punkt auf dem von $q_k(x)$ ausgehenden Strahl durch x ist, der auf S^{n-1} liegt. Genauer ist $f_k(x) = q_k(x) + \lambda_k(x)(x - q_k(x))$, wobei $\lambda_k(x) \geq 1$ durch $\|f_k(x)\|^2 = 1$ bestimmt ist. Dies führt auf die quadratische Gleichung

$$\lambda_k(x)^2 \|x - q_k(x)\|^2 + 2\lambda_k(x)(x - q_k(x))^T q_k(x) + (\|q_k(x)\|^2 - 1) = 0.$$

Ist die Diskriminante

$$\Delta_k(x) := 4[(x - q_k(x))^T q_k(x)]^2 + 4\|x - q_k(x)\|^2 (1 - \|q_k(x)\|^2)$$

auf B^n positiv (sie ist jedenfalls nichtnegativ), so ist λ_k und damit auch q_k auf B^n stetig differenzierbar. Wegen $\|x - q_k(x)\|^2 > 0$ (wir hatten angenommen, dass q_k keinen Fixpunkt in B^n besitzt) verschwindet die Diskriminante genau dann in $x \in B^n$, wenn $(x - q_k(x))^T q_k(x) = 0$ und $\|q_k(x)\|^2 = 1$. Dann ist aber

$$\|x - q_k(x)\|^2 = \|x\|^2 - 2x^T q_k(x) + \|q_k(x)\|^2 = \|x\|^2 - 2 + 1 \leq 0,$$

also $x \in B^n$ ein Fixpunkt von q_k, was gerade ausgeschlossen war. Also ist $f_k\colon B^n \longrightarrow S^{n-1}$ eine stetig differenzierbare Abbildung. Ist $x \in S^{n-1}$ bzw. $\|x\| = 1$, so ist $\lambda_k(x) = 1$ bzw. $f_k(x) = x$. Dies ist ein Widerspruch zum vorangegangenen Lemma und die Existenz eines Fixpunktes $x_k \in B^n$ von q_k ist für jedes $k \in \mathbb{N}$ bewiesen. Da B^n kompakt ist, besitzt $\{x_k\}$ eine gegen ein $x^* \in B^n$ konvergente Teilfolge, die wir wieder mit $\{x_k\}$

bezeichnen. Da $\{q_k\}$ gleichmäßig gegen f konvergiert, ist

$$\begin{aligned}\|f(x^*) - q_k(x_k)\| &\leq \|f(x^*) - f(x_k)\| + \|f(x_k) - q_k(x_k)\| \\ &\leq \underbrace{\|f(x^*) - f(x_k)\|}_{\to 0} + \underbrace{\|f - q_k\|_\infty}_{\to 0} \\ &\to 0\end{aligned}$$

und daher

$$f(x^*) = \lim_{k\to\infty} q_k(x_k) = \lim_{k\to\infty} x_k = x^*.$$

Daher ist $x^* \in B^n$ Fixpunkt von f und der Brouwersche Fixpunktsatz ist bewiesen. □

38 Der Fundamentalsatz der Algebra

Der Fundamentalsatz der Algebra ist zum ersten Mal in der Dissertation von C. F. Gauss (1799) bewiesen worden. Es gibt viele Beweise, wir geben einen besonders schönen an, der den Brouwerschen Fixpunktsatz benutzt.

Fundamentalsatz der Algebra *Seien $a_0, \ldots, a_n \in \mathbb{C}$ mit $n \geq 1$ und $a_n \neq 0$ die Koeffizienten des komplexen Polynoms $p(z) := a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$. Dann besitzt p mindestens eine komplexe Nullstelle.*

Beweis: Man definiere das normierte Polynom $q(z) := p(z)/a_n$ mit den Koeffizienten $b_i := a_i/a_n$, $i = 0, \ldots, n$. Wir zeigen, dass q (und damit auch p) eine Nullstelle besitzt. Man definiere

$$D := \{z \in \mathbb{C} : |z| \leq R\},$$

wobei $R := 2 + |b_0| + |b_1| + \cdots + |b_{n-1}|$. Die Abbildung $g : D \subset \mathbb{C} \longrightarrow \mathbb{C}$ definiere man durch

$$g(z) := \begin{cases} z - \dfrac{q(z)}{R(z/|z|)^{n-1}}, & |z| \leq 1, \\ z - \dfrac{q(z)}{Rz^{n-1}}, & \text{sonst.} \end{cases}$$

Dann ist g auf D stetig. Wir zeigen, dass $g(D) \subset D$. Für $z \in D$ mit $|z| \leq 1$ ist

$$\begin{aligned}|g(z)| &= \left|z - \frac{q(z)}{R(z/|z|)^{n-1}}\right| \\ &\leq |z| + \left|\frac{q(z)}{R(z/|z|)^{n-1}}\right| \\ &= |z| + \frac{|q(z)|}{R} \\ &\leq 1 + \frac{|b_0| + |b_1| + \cdots + |b_{n-1}| + 1}{R} \\ &\leq 1 + 1 \\ &\leq R,\end{aligned}$$

für $1 < |z| \leq R$ ist

$$\begin{aligned}
|g(z)| &= \Big|z - \frac{q(z)}{Rz^{n-1}}\Big| \\
&= \Big|z - \frac{b_0 + b_1 z + \cdots + b_{n-1} z^{n-1}}{Rz^{n-1}} - \frac{z}{R}\Big| \\
&\leq \Big(1 - \frac{1}{R}\Big)|z| + \frac{|b_0| + |b_1|\,|z| + \cdots + |b_{n-1}|\,|z|^{n-1}}{R|z|^{n-1}} \\
&\leq \Big(1 - \frac{1}{R}\Big)R + \frac{|b_0| + |b_1| + \cdots + |b\text{–}n - 1}{R} \\
&= R - 1 + \frac{R-2}{R} \\
&< R.
\end{aligned}$$

Folglich ist $g(D) \subset D$. Der Brouwersche Fixpunktsatz (man identifiziere die komplexe Ebene $\mathbb{C}$ mit dem $\mathbb{R}^2$) liefert die Existenz eines Fixpunktes $z^* \in D$. Aus der Definition von g folgt $q(z^*) = 0$ und hieraus $p(z^*) = 0$. Damit ist der Fundamentalsatz der Algebra bewiesen. □

39 Das Gefangenendilemma und andere Zwei-Personen-Spiele

Man stelle sich die folgende Situation vor (wieder wörtlich aus Wikipedia übernommen):

> Zwei Gefangene werden verdächtigt, gemeinsam eine Straftat begangen zu haben. Beide Gefangene werden in getrennten Räumen verhört und haben keine Möglichkeit, sich zu beraten. Die Höchststrafe für das Verbrechen beträgt sechs Jahre. Wenn die Gefangenen sich entscheiden zu schweigen, werden beide wegen kleinerer Delikte zu je zwei Jahren Haft verurteilt. Gestehen jedoch beide die Tat, erwartet beide eine Gefängnisstrafe, wegen der Zusammenarbeit mit den Ermittlungsbehörden jedoch nicht die Höchststrafe, sondern lediglich von vier Jahren. Gesteht nur einer und der andere schweigt, bekommt der erste als Kronzeuge eine symbolische einjährige Bewährungsstrafe und der andere bekommt die Höchststrafe von sechs Jahren. Falls der andere aber schweigt, so wird seine Strafe von zwei Jahren auf ein Jahr reduziert.

Wir nennen die beiden Gefangenen *Spieler* und bezeichnen sie mit A und B. Beide haben zwei *Strategien* zur Verfügung, nämlich zu schweigen oder zu gestehen. In einer Auszahlungsmatrix eingetragen erhält man das folgende Bild.

	B schweigt	B gesteht
A schweigt	A: -2, B: -2	A: -6, B: -1
A gesteht	A: -1, B: -6	A: -4, B: -4

Hierbei sind die Einträge mit einem Minuszeichen versehen, weil wir standardmäßig von einer Maximierungsaufgabe ausgehen. Maximieren des Negativen einer Haftstrafe bedeutet Minimieren der angedrohten Haftstrafe. Wollen die beiden Gefangenen ihre Gesamtstrafe minimieren, so ist es für sie am günstigsten zu schweigen. Denn dann werden beide zu je zwei Jahren Gefängnis verurteilt, was insgesamt vier Jahre ergibt. Würden die beiden Gefangenen also als Kollektiv agieren und kooperieren können, was aber nicht möglich ist, so wäre es für sie am besten zu schweigen. Für jeden einzelnen der beiden Gefangenen, etwa für A, also individuell gesehen, scheint es aber vorteilhafter zu sein auszusagen bzw. zu gestehen. Denn wenn auch B gesteht, reduziert A durch sein Geständnis seine Strafe von sechs auf vier Jahre. Falls B aber schweigt, so wird die Strafe von A von zwei Jahren auf ein Jahr reduziert. Das Dilemma besteht darin, dass das für beide zusammen günstigste Verhalten für den Einzelnen ungünstig sein kann.

Allgemein spricht man von einem (strategischen) *Zwei-Personen-Spiel* (oder auch *Bi-Matrix-Spiel*), wenn zwei Spieler A und B jeweils eine endliche Menge von (reinen) Strategien $\{s_1, \ldots, s_m\}$ bzw. $\{t_1, \ldots, t_n\}$ besitzen. Wählt der Spieler A die i-te Strategie s_i und der Spieler B die j-te Strategie t_j, so entstehen A bzw. B ein Gewinn in Höhe von a_{ij} bzw. b_{ij} Einheiten. Hiermit sind die den Spielern A und B bekannten Matrizen $A, B \in \mathbb{R}^{m\times n}$ gegeben. Positive Einträge in A bzw. B werden als Einnahmen bzw. Gewinn, negative Einträge als Abgaben bzw. Verluste interpretiert. Neben den reinen Strategien spielen sogenannte *gemischte Strategien* in der Spieltheorie eine wichtige Rolle. Hierzu definiere man

$$\Sigma_m := \{p \in \mathbb{R}^m : p \geq 0,\, e^T p = 1\}, \qquad \Sigma_n := \{q \in \mathbb{R}^n : q \geq 0,\, e^T q = 1\}$$

als Menge der gemischten Strategien für Spieler A bzw. B. Hierbei bezeichnet e den Vektor im $\mathbb{R}^m$ bzw. R^n (das geht aus dem Zusammenhang hervor), dessen Komponenten alle gleich 1 sind. Ein Vektor $p = (p_i) \in \Sigma_m$ gibt an, dass der Spieler A die Strategie s_i mit Wahrscheinlichkeit p_i spielt, $i = 1, \ldots, m$. Entsprechendes gilt für einen Vektor $q = (q_j) \in \Sigma_n$. Der reinen Strategie s_i für den Spieler A entspricht somit der i-te Einheitsvektor e_i im $\mathbb{R}^m$, entsprechend der reinen Strategie t_j für den Spieler B der j-te Einheitsvektor im $\mathbb{R}^n$.

Sehr wichtig in der sogenannten *Spieltheorie* (wir beschränken uns auf Zwei-Personen-Spiele) ist die folgende Definition.

Definition Ein Paar $(p^*, q^*) \in \Sigma_m \times \Sigma_n$ gemischter Strategien heißt ein *Nash*[50]*-Gleichgewichtspunkt* für ein Spiel mit den Auszahlungs- bzw. Kostenmatrizen $A, B \in \mathbb{R}^{m\times n}$, wenn

$$(p^*)^T A q^* \geq p^T A q^* \quad \text{für alle } p \in \Sigma_m$$

und

$$(p^*)^T B q^* \geq (p^*)^T B q \quad \text{für alle } q \in \Sigma_n.$$

Bei Wahl eines Paares $(p, q) \in \Sigma_m \times \Sigma_n$ gemischter Strategien haben A bzw. B einen Gewinn von $p^T A q$ bzw. $p^T B q$ zu erwarten. Ein Paar $(p^*, q^*) \in \Sigma_m \times \Sigma_n$ ist daher

[50]Benannt nach dem Mathematiker John Forbes Nash Jr., der 1994 den Nobelpreis für Wirtschaftswissenschaften erhielt und vor allem durch den Spielfilm *A beautiful mind* einem breiten Publikum bekannt wurde.

ein Nash-Gleichgewichtspunkt, wenn A durch keine andere Wahl als p^* einen größeren Gewinn erwarten kann, falls Spieler B die (gemischte) Strategie q^* wählt. Ähnliches gilt für den Spieler B.

Beispiel: Beim obigen Gefangenendilemma haben beide Spieler dieselben Strategien, nämlich zu schweigen oder zu gestehen, und die Auszahlungsmatrizen sind gegeben durch

$$A := \begin{pmatrix} -2 & -6 \\ -1 & -4 \end{pmatrix}, \qquad B := \begin{pmatrix} -2 & -1 \\ -6 & -4 \end{pmatrix}.$$

Wir wollen uns überlegen, dass das reine Strategiepaar, bei dem beide Gefangenen gestehen, ein Nash-Gleichgewichtspunkt ist. Hierzu haben wir zu zeigen, dass

$$-4 = e_2^T A e_2 \geq p^T A e_2 = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}^T \begin{pmatrix} -6 \\ -4 \end{pmatrix} \quad \text{für alle } p \in \Sigma_2$$

und entsprechend

$$-4 = e_2^T B e_2 \geq e_2^T B q = \begin{pmatrix} -6 & -4 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \quad \text{für alle } q \in \Sigma_2.$$

Beides ist offensichtlich richtig. □

Beispiel: Beim sogenannten *Kampf der Geschlechter* handelt es sich (zumindest in der Spieltheorie) um das folgende Problem. Ein Mann (Spieler A) und eine Frau (Spieler B) wollen einen Abend gemeinsam verbringen und entweder zu einem Fußballspiel oder in ein Konzert gehen. Der Mann geht wesentlich lieber zum Fußball als ins Konzert, bei der Frau ist es genau umgekehrt. Wichtig für beide ist aber, dass sie den Abend miteinander verbringen. Es sei die folgende Auszahlungsmatrix gegeben:

	B geht zum Fußball	B geht ins Konzert
A geht zum Fußball	A: 3, B: 1	A: 0, B: 0
A geht ins Konzert	A: 0, B: 0	A: 1, B: 3

Die beiden Matrizen A und B sind in diesem Fall also durch

$$A := \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B := \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

gegeben. Die Strategiemenge besteht für beide darin, zum Fuball oder ins Konzert zu fehen. Die reinen Strategiepaare (e_1, e_1) bzw. (e_2, e_2), dass also A und B zum Fußball bzw. A und B zum Konzert gehen, sind offenbar Nash-Gleichgewichtspunkte. Denn: Es ist

$$3 = e_1^T A e_1 \geq p^T A e_1 = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}^T \begin{pmatrix} 3 \\ 0 \end{pmatrix} \quad \text{für alle } p \in \Sigma_2$$

und entsprechend

$$1 = e_1^T B e_1 \geq e_1^T B q = \begin{pmatrix} 1 & 0 \end{pmatrix}^T \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \quad \text{für alle } q \in \Sigma_2,$$

also ist (e_1, e_1) ein Nash-Gleichgewichtspunkt. Entsprechend zeigt man, dass auch (e_2, e_2) ein Nash-Gleichgewichtspunkt ist. Mit

$$p^* := \begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}, \qquad q^* := \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix}$$

ist aber auch ein Gleichgewichtspunkt (p^*, q^*) aus gemischten Strategien gegeben, wie man leicht nachweist. Das bedeutet, dass in 25% aller Fälle Mann und Frau das Lieblingsereignis des Partners aufsuchen sollten. □

Beispiel: Beim „Knobelspiel“ Stein-Schere-Papier spielen zwei Spieler A und B gegeneinander. Die Strategiemenge für beide besteht darin, Stein, Schere oder Papier zu ziehen. Hierbei schlägt Stein Schere, Schere schlägt Papier und Papier schlägt Stein. Wir erhalten die folgende Auszahlungsmatrix stammt

		B		
		Stein	Schere	Papier
A	Stein	A: 0, B: 0	A: 1, B: -1	A: -1, B: 1
	Schere	A: -1, B:1	A: 0, B: 0	A: 1, B: -1
	Papier	A: 1, B: -1	A: -1, B: 1	A: 0, B: 0

bzw. die Matrizen

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}.$$

Man erkennt, dass $B = -A$. Alles was Spieler A gewinnt, verliert B und umgekehrt. Solche Spiele heißen *Zwei-Personen-Nullsummen-Spiele*. Durch (p^*, q^*) mit

$$p^* := \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}, \qquad q^* := \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$$

ist offenbar ein Nash-Gleichgewichtspunkt gegeben. □

Der folgende Satz (in einer allgemeineren Version für nichtkooperative n-Personen-Spiele) stammt aus der Dissertation von John F. Nash (1950).

Satz *Jedes Zwei-Personen-Spiel besitzt einen Nash-Gleichgewichtspunkt.*

Beweis: Die Idee des Beweises besteht darin, eine gewisse stetige Abbildung $F\colon \Sigma_m \times \Sigma_n \subset \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^m \times \mathbb{R}^n$ mit $F(\Sigma_m \times \Sigma_n) \subset \Sigma_m \times \Sigma_n$ zu definieren, aus dem Brouwerschen Fixpunktsatz auf die Existenz eines die Existenz eines Fixpunkts (p^*, q^*) von F in $\Sigma_m \times \Sigma_n$ zu schließen und nachzuweisen, dass dieser Fixpunkt ein Nash-Gleichgewichtspunkt ist.

Sei $(p,q) \in \Sigma_m \times \Sigma_n$. Man definiere

$$c_i(p,q) := \max(e_i^T Aq - p^T Aq, 0), \quad i = 1, \ldots, m,$$

sowie

$$d_j(p,q) := \max(x^T Be_j - p^T Bq, 0), \quad j = 1, \ldots, n,$$

und anschließend

$$c(p,q) := (c_i(p,q)) \in \mathbb{R}^m, \qquad d(p,q) := (d_j(p,q)) \in \mathbb{R}^n$$

sowie

$$F(p,q) := \Big(\frac{p + c(p,q)}{1 + e^T c(p,q)}, \frac{q + d(p,q)}{1 + e^T d(p,q)}\Big).$$

Offenbar ist mit (p,q) auch $F(p,q)$ ein Paar gemischter Strategien bzw. $F(\Sigma_m \times \Sigma_n) \subset \Sigma_m \times \Sigma_n$. Da $\Sigma_m \times \Sigma_n$ nichtleer, konvex und kompakt ist, liefert der Brouwersche Fixpunktsatz die Existenz eines Fixpunkts (p^*, q^*) von F in $\Sigma_m \times \Sigma_n$. Angenommen, (p^*, q^*) sei kein Nash-Gleichgewichtspunkt. Das impliziert die Existenz eines $p \in \Sigma_m$ mit $p^T Aq^* > (p^*)^T Aq^*$ oder eines $q \in \Sigma_n$ mit $(p^*)^T Aq > (p^*)^T Aq^*$. Nehmen wir die erste der beiden Möglichkeiten (der Beweis für den zweiten Fall verläuft analog). Dann ist also

$$0 < p^T Aq^* - (p^*)^T Aq^* = \sum_{i=1}^m p_i [e_i^T Aq^* - (p^*)^T Aq^*]$$

und folglich $e_k^T Aq^* - (p^*)^T Aq^* > 0$ und damit $c_k(p^*, q^*) > 0$ für wenigstens ein $k \in \{1, \ldots, m\}$. Da alle Komponenten von $c(p^*, q^*)$ nichtnegativ sind, ist $e^T c(p^*, q^*) > 0$. Weiter ist $(p^*)^T Aq^* = \sum_{i=1}^m p_i^* e_i^T Aq^*$ ein gewichtetes Mittel der Ausdrücke $e_i^T Aq^*$, $i = 1, \ldots, m$ bzw.

$$0 = \sum_{i=1}^m p_i^* [(p^*)^T Aq^* - e_i^T Aq^*] = \sum_{i: p_i^* > 0} p_i^* [(p^*)^T Aq^* - e_i^T Aq^*].$$

Daher existiert ein $i \in \{1, \ldots, m\}$ mit $p_i^* > 0$ und $e_i^T Aq^* \le (p^*)^T Aq^*$. Für dieses i gilt $c_i(p^*, q^*) = 0$ und daher

$$\frac{p_i^* + c_i(p^*, q^*)}{1 + e^T c(p^*, q^*)} = \frac{p_i^*}{1 + e^T c(p^*, q^*)} < p_i^*,$$

ein Widerspruch dazu, dass (p^*, q^*) ein Fixpunkt von F ist. Der Satz ist bewiesen. □

Spezialisiert man den gerade eben bewiesenen Satz über die Existenz eines Nash-Gleichgewichtspunktes bei Zwei-Personen-Spielen auf Zwei-Personen-Nullsummenspiele, so erhält man das auf JOHN VON NEUMANN (1928) zurückgehende Min-Max-Theorem (auch Hauptsatz der Theorie der Matrixspiele genannt). Wir formulieren dieses Ergebnis als

Min-Max-Theorem *Sei $A \in \mathbb{R}^{m \times n}$ und (wie oben)*

$$\Sigma_m := \{p \in \mathbb{R}^m : p \ge 0,\, e^T p = 1\}, \qquad \Sigma_n := \{q \in \mathbb{R}^n : q \ge 0,\, e^T q = 1\}.$$

Dann ist

$$\max_{p\in\Sigma_m}\min_{q\in\Sigma_n} p^T Aq = \min_{q\in\Sigma_n}\max_{p\in\Sigma_m} p^T Aq.$$

Beweis: Man betrachte ein Zwei-Personen-Nullsummenspiel mit der Auszahlungsmatrix $A \in \mathbb{R}^{m\times n}$ für den Spieler A und der Auszahlungsmatrix $B := -A$ für den Spieler B. Wie wir gerade eben bewiesen haben, besitzt das zugehörige Spiel einen Nash-Gleichgewichtspunkt $(p^*, q^*) \in \Sigma_m \times \Sigma_n$, d. h. unter Berücksichtigung von $B = -A$ gilt

$$(p^*)^T Aq^* \geq p^T Aq^* \quad \text{für alle } p \in \Sigma_m$$

und

$$(p^*)^T Aq^* \leq (p^*)^T Aq \quad \text{für alle } q \in \Sigma_n.$$

Man betrachte die beiden Optimierungsaufgaben

$$\text{(A)} \qquad \text{Maximiere} \quad \phi(p) := \min_{q\in\Sigma_n} p^T Aq, \quad p \in \Sigma_m$$

und

$$\text{(B)} \qquad \text{Minimiere} \quad \psi(q) := \max_{p\in\Sigma_m} p^T Aq, \quad q \in \Sigma_n.$$

Da (p^*, q^*) ein Nash-Gleichgewichtspunkt ist, ist

$$(p^*)^T Aq^* \leq \psi(q^*) \leq (p^*)^T Aq^* \leq \phi(p^*) \leq (p^*)^T Aq^*$$

und folglich $\psi(q^*) = (p^*)^T Aq^* = \phi(p^*)$. Hieraus folgt aber, dass $p^* \in \Sigma_m$ eine Lösung von (A) und $q^* \in \Sigma_n$ eine Lösung von (B) ist. Denn für ein beliebiges $p \in \Sigma_m$ ist

$$\phi(p^*) = \psi(q^*) \geq p^T Aq^* \geq \phi(p),$$

also $p^* \in \Sigma_m$ eine Lösung von (A). Für ein beliebiges $q \in \Sigma_n$ ist entsprechend

$$\psi(q^*) = \phi(p^*) \leq (p^*)^T Aq \leq \psi(q),$$

also $q^* \in \Sigma_n$ eine Lösung von (B), weiter stimmen die Optimalwerte $\phi(p^*)$ von (A) und $\psi(q^*)$ von (B) überein, d. h. es ist

$$\phi(p^*) = \max_{p\in\Sigma_m}\min_{q\in\Sigma_n} p^T Aq = \min_{q\in\Sigma_n}\max_{p\in\Sigma_m} p^T Aq = \psi(q^*).$$

Dies ist die Aussage des Min-Max-Theorems. □

Bemerkungen: Wir benutzen die Bezeichnungen des Beweises zum Min-Max-Theorem. Wählt Spieler A die gemischte Strategie $p \in \Sigma_m$, so ist $\phi(p)$ sein zu erwartender Mindestgewinn. Diesen wird er versuchen zu maximieren. Spieler A hat also die Optimierungsaufgabe (A) zu lösen. Entsprechend ist $\psi(q)$ bei gegebenem $q \in \Sigma_n$ der zu erwartende Maximalverlust für Spieler B, diesen wird er versuchen zu minimieren. Spieler B hat also die Optimierungsaufgabe (B) zu lösen. Das Min-Max-Theorem sagt aus, dass der von Spieler A zu erwartende maximale Mindestgewinn gleich dem minimalen

Maximalverlust für Spieler B ist. Wir nennen ein Zwei-Personen-Nullsummenspiel *fair*, wenn dieser Wert gleich Null ist, andernfalls *unfair*.

Der Beweis des Min-Max-Theorems zeigt eine Möglichkeit auf, die Komponenten p^* und q^* eines Nash-Gleichgewichtspunktes (p^*, q^*) zu einem Zwei-Personen-Nullsummenspiel als Lösung der Optimierungsaufgaben

$$\text{(A)} \qquad \text{Maximiere} \quad \phi(p) := \min_{q \in \Sigma_n} p^T A q, \quad p \in \Sigma_m$$

bzw.

$$\text{(B)} \qquad \text{Minimiere} \quad \psi(q) := \max_{p \in \Sigma_m} p^T A q, \quad q \in \Sigma_n$$

zu erhalten. Hierzu ist es zweckmäßig, die sogenannten *Zielfunktionen* $\phi(\cdot)$ bzw. $\psi(\cdot)$ von (A) bzw. (B) noch etwas umzuformulieren. Wir bezeichnen mit $a_{ij} := e_i^T A e_j$ den Eintrag von A in der i-ten Zeile und der j-ten Spalte. Dann ist

$$\phi(p) = \min_{j=1,\ldots,n} p^T A e_j, \qquad \psi(q) = \max_{i=1,\ldots,m} e_i^T A q.$$

Denn: Für ein beliebiges $q \in \Sigma_n$ und gegebenem $p \in \Sigma_m$ ist

$$p^T A q = p^T A \Big(\sum_{j=1}^n q_j e_j \Big) = \sum_{j=1}^n q_j p^T A e_j \geq \Big(\min_{j=1,\ldots,n} p^T A e_j \Big) \sum_{j=1}^n q_j = \min_{j=1,\ldots,n} p^T A e_j$$

und folglich

$$\phi(p) = \min_{q \in \Sigma_n} p^T A q \geq \min_{j=1,\ldots,n} p^T A e_j.$$

Andererseits ist $e_j \in \Sigma_n$, $j = 1, \ldots, n$, und daher $\phi(p) \leq p^T A e_j$, $j = 1, \ldots, n$ und folglich $\phi(p) \leq \min_{j=1,\ldots,n} p^T A e_j$. Insgesamt ist $\phi(p) = \min_{j=1,\ldots,n} p^T A e_j$. Entsprechend beweist man die Darstellung $\psi(q) = \max_{i=1,\ldots,m} e_i^T A q$. Die Aufgabe (A) für Spieler A bzw. die Aufgabe (B) für Spieler B kann daher geschrieben werden als

$$\text{(A)} \qquad \text{Maximiere} \quad \phi(p) = \min_{j=1,\ldots,n} p^T A e_j, \quad p \in \Sigma_m$$

bzw.

$$\text{(B)} \qquad \text{Minimiere} \quad \psi(q) = \max_{i=1,\ldots,m} e_i^T A q, \quad q \in \Sigma_n.$$

Es ist $\alpha \leq \min_{j=1,\ldots,n} p^T e_j$ bzw. $\alpha \leq p^T A e_j = e_j^T A^T p$, $j = 1, \ldots, n$, genau dann, wenn $\alpha e \leq A^T p$, wobei e wieder der Vektor ist, dessen Komponenten alle gleich 1 sind. Daher ist (A) äquivalent zu ($\leq$- oder $\geq$-Beziehungen zwischen Vektoren sind komponentenweise zu verstehen)

$$\text{(A)} \qquad \begin{cases} \text{Maximiere} \quad \alpha = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \alpha \end{pmatrix} \quad \text{auf} \\ M := \{(p, \alpha) \in \mathbb{R}^m \times \mathbb{R} : p \geq 0,\ \alpha e \leq A^T p,\ e^T p = 1\}. \end{cases}$$

Entsprechend ist (B) äquivalent

$$\text{(B)} \qquad \begin{cases} \text{Minimiere} \quad \beta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} q \\ \beta \end{pmatrix} \quad \text{auf} \\ N := \{(q,\beta) \in \mathbb{R}^n \times \mathbb{R} : q \geq 0,\ Aq \leq \beta e,\ e^T q = 1\}. \end{cases}$$

Hierbei haben wir die Bezeichnungen der Probleme nicht geändert, Spieler A hat also (A), Spieler B die Aufgabe (B) zu lösen. Die beiden Aufgaben (A) und (B) sind *lineare Optimierungsaufgaben*, welche mit Standardmethoden (z. B. Simplexverfahren) oder Computer-Software gelöst werden können. □

Beispiel: Das folgende Beispiel findet man bei L. Collatz, W. Wetterling (1971). Die Spieler A und B haben je 3 Karten auf der Hand, und zwar A die Karten Pik As, Karo As und Karo Zwei, B die Karten Pik As, Karo As und Pik Zwei. Beide Spieler legen jeweils zugleich eine ihrer Karten auf den Tisch. A gewinnt, wenn die hingelegten Karten die gleiche Farbe haben, andernfalls B. Ein As hat den Wert 1, eine Zwei den Wert 2. Die Höhe des Gewinnes ist gleich dem Wert derjenigen Karte, die der Gewinner hingelegt hat. Das Spiel hat also die Auszahlungsmatrix

A \ B	♢	♠	♠♠
♢	1	−1	−2
♠	−1	1	1
♢♢	2	−1	−2

Man hat den Eindruck, das Spiel sei unfair, weil die Auszahlungsmatrix

$$A := \begin{pmatrix} 1 & -1 & -2 \\ -1 & 1 & 1 \\ 2 & -1 & -2 \end{pmatrix}$$

fünf negative Elemente gegenüber vier positiven enthält. Das gibt Anlass zur Formulierung der

Zusatzregel: Wenn beide Spieler ihre Zweierkarte hinlegen, so soll keiner an den anderen etwas zahlen, d. h. das Element -2 in der rechten unteren Ecke der Auszahlungsmatrix wird durch 0 ersetzt.

Wir wollen für das Spiel ohne und mit Zusatzregel mit Hilfe des mathematischen Anwendersystems MATLAB, genauer der in der Optimization Toolbox enthaltenen Funktion `linprog`, jeweils optimale gemischte Strategien für die Spieler A und B berechnen und damit entscheiden, welches der beiden Spiele fair ist. So wird durch den Befehl

```
[x,feval]=linprog(f,A,b,A_0,b_0,l,u)
```

eine Lösung x und den zugehörigen Zielfunktionswert $f^T x$ der linearen Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f^T x \quad \text{unter den Nebenbedingungen} \\ Ax \leq b, \quad A_0 x \leq b_0, \quad l \leq x \leq u \end{cases}$$

ausgegeben. Die Aufgabe (A) mit der obigen Auszahlungsmatrix lautet dann

$$\text{Minimiere} \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}^T \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \alpha \end{pmatrix} \quad \text{unter den Nebenbedingungen}$$

$$\begin{pmatrix} -1 & 1 & -2 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & -1 & 2 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \alpha \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \alpha \end{pmatrix} = 1$$

und

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ -\infty \end{pmatrix} \leq \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \alpha \end{pmatrix} \leq \begin{pmatrix} \infty \\ \infty \\ \infty \\ \infty \end{pmatrix}.$$

Mit den MATLAB-Befehlen

```
>> f=[0;0;0;-1]; A=[-1 1 -2 1;1 -1 1 1;2 -1 2 1];  b=[0;0;0];
>> A_0=[1 1 1 0]; b_0=[1]; l=[0;0;0;-inf]; u=[];
>> [x,feval]=linprog(f,A,b,A_0,b_0,l,u);
Optimization terminated.
>> x
x =
    0.0000
    0.6667
    0.3333
    0.0000
>> feval
feval =
  -9.9476e-13
```

raten wir die Lösung

$$p^* = \begin{pmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}, \qquad \alpha^* = 0.$$

Das Spiel ist also fair. Mit Zusatzregel erhält man als Lösung für den Spieler A

$$p^* = \begin{pmatrix} 0 \\ \frac{3}{5} \\ \frac{2}{5} \end{pmatrix}, \qquad \alpha^* = \frac{1}{5},$$

dieses Spiel ist also unfair. Als Lösung für den Spieler B erhalten wir

$$q^* = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}, \qquad \beta^* = 0 \qquad \text{bzw.} \qquad q^* = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \\ 0 \end{pmatrix}, \qquad \beta^* = \frac{1}{5}.$$

Hierbei haben wir entsprechende MATLAB-Befehle benutzt. □

40 Das Public goods game

Viele Probleme unter den Menschen bestehen darin, dass es einen Konflikt zwischen dem Eigeninteresse von Personen, Gruppen, Organisationen oder Staaten und dem Wohl einer größeren Gesamtheit gibt. Man spricht von einem *sozialen Dilemma*, wenn einerseits jeder Beteiligte durch nicht-kooperative Handlung einen höheren Gewinn als durch eine kooperative Handlung erhält, und andererseits alle Beteiligten insgesamt besser gestellt sind, wenn sie kooperieren und nicht jeder egoistisch handelt. Man denke z. B. an die Überfischung der Meere oder die Klimaproblematik[51]. Wenn es in U-Bahnen nicht (gelegentlich) Kontrollen geben würde, würde zwar ein Schwarzfahrer davon profitieren, für die Gesamtheit wäre der öffentliche Nahverkehr bald aber nicht mehr finanzierbar. Unter dem Stichwort *tragedy of the commons* findet man weitere Beispiele, siehe z. B. `http://en.wikipedia.org/wiki/Tragedy_of_the_commons`.

Dieser Konflikt wird durch das Public goods game modelliert. Wir zitieren zunächst einmal wieder Wikipedia: Das Public goods game (auch Öffentliche-Güter-Spiel) ist Bestandteil der Spieltheorie und wird in der Experimentellen Ökonomik untersucht; in der Standardvariante des Spieles entscheiden die Teilnehmer im Geheimen wie viele ihrer eigenen Token (Jetons oder Spielmarken) sie in den öffentlichen Topf geben wollen. Jeder Spieler behält also seine nicht eingezahlten Token und erhält zusätzlich einen einheitlichen Anteil an Token aus dem Topf. Um einen Anreiz für Beitragsleistungen zu geben multipliziert der Spielleiter vor der Auszahlung die Anzahl der Token im öffentlichen Topf.

Etwas konkreter: Zu Beginn erhält jeder Mitspieler einen Betrag von 100 Euro, von dem er einen beliebigen Teil (er kann auch Null sein) in ein Projekt investieren kann. Dann werden alle eingegangenen Beträge verdoppelt und an alle Gruppenmitglieder zu gleichen Teilen aufgeteilt. Während des Spiels findet keine Kommunikation zwischen den Mitgliedern statt. Insbesondere ist jedem Mitglied unbekannt, wie viel die anderen Teilnehmer eingezahlt haben. Für die gesamte Gruppe ist es offensichtlich am besten, wenn alle Gruppenmitglieder den vollen Betrag von 100 Euro ins Projekt investieren. Allerdings ist es für den einzelnen Spieler eine *dominante Strategie*[52], Null zu setzen.

[51] In einem Gespräch mit dem Psychologen Michael Tomasello, erschienen in der Süddeutschen Zeitung vom 2.12.2011, kann man lesen: „Es würde uns leicht fallen, etwa beim Klimaschutz zu einer Lösung zu kommen, wenn die Erde von Invasoren aus dem Weltall bedroht werden würde." Hier wird also die Kooperation durch eine äußere Bedrohung erzwungen.

[52] Wikipedia: Die dominante Strategie in spieltheoretischen Modellen ist eine Strategie, die unter

Denn wir betrachten den j-ten Spieler und nehmen an, die Spieler $k \neq j$ hätten a_k^* Euro investiert. Bei einem Einsatz von a_j Euro hat der j-te Spieler nach dem Spiel

$$f(a_j) := 100 - a_j + \frac{2}{n}\Big(\sum_{k\neq j} a_k^* + a_j\Big) = 100 + \frac{2}{n}\sum_{k\neq j} a_k^* + \Big(\frac{2}{n} - 1\Big)a_j$$

Euro zur Verfügung. Für $n \geq 2$ ist $2/n - 1 \leq 0$, so dass f auf $[0, 100]$ für $a_j^* := 0$ maximal ist. Andererseits erhält man ein *Pareto-Optimum*[53], wenn alle Mitglieder den Maximalbetrag von 100 Euro investieren. Hierzu definieren wir

$$A := \{a = (a_j) \in \mathbb{R}^n : 0 \leq a_j \leq 100\}$$

als Menge der zulässigen Strategien für die n Spieler sowie $f\colon A \longrightarrow \mathbb{R}^n$ durch $f(a) := (f_j(a))$ mit

$$f_j(a) := 100 - a_j + \frac{2}{n}\sum_{k=1}^{n} a_k, \qquad j = 1, \ldots, n.$$

Man definiere $a^* = (a_j^*) \in A$ durch $a_j^* := 100$, $j = 1, \ldots, n$. Dann ist $f_j(a^*) = 200$, $j = 1, \ldots, n$. Ist nun $a \in A$ und $f_j(a) \geq 200$, $j = 1, \ldots, n$, bzw.

$$100 - a_j + \frac{2}{n}\sum_{k=1}^{n} a_k \geq 200, \qquad j = 1, \ldots, n,$$

so ist dies gleichbedeutend mit

$$100 - a_j \geq \frac{2}{n}\sum_{k=1}^{n}(100 - a_k), \qquad j = 1, \ldots, n.$$

Hieraus folgt aber $a = a^*$ und $f(a) = f(a^*)$, also ist a^* ein Pareto-Optimum.

Um die Kooperationsbereitschaft der Teilnehmer zu steigern, gibt es verschiedene Möglichkeiten. Müssen die einzelnen Beiträge offen gelegt werden, so werden sie i. Allg. höher sein, als wenn sie unbekannt sind. Weiter kann man an Belohnen und Strafen denken. Wir verweisen hierzu lediglich auf `http://postheroisch.wordpress.com/2009/09/11/`.

---

allen möglichen Strategien den höchsten Nutzen bietet, unabhängig davon, was die anderen Akteure (Spieler, Agenten) tun.

[53] Gegeben sei die Vektoroptimierungsaufgabe

(P) $$\text{Maximiere} \quad f(a), \quad a \in A.$$

Hierbei sei $A \subset \mathbb{R}^n$ und $f\colon A \longrightarrow \mathbb{R}^m$. Dann heißt $a^* \in A$ ein *Pareto-Optimum* von (P), wenn aus $a \in A$, $f(a) \geq f(a^*)$ (komponentenweise zu verstehen) folgt, dass $f(a) = f(a^*)$. Ist also $f_j(a)$ der Gewinn des j-ten Spielers, wenn von den n Spielern die Strategien $a_1, \ldots, a_n$ gewählt werden, so ist es in einem Pareto-Optimum nicht möglich, einen Spieler besser zu stellen ohne einen anderen schlechter zu stellen.

41 Die Berechnung der Quadratwurzel

Wie funktioniert die $\sqrt{\ }$-Taste bei einem Taschenrechner oder die `sqrt`-Funktion in MATLAB? Die Aufgabe besteht darin, bei gegebener positiver Zahl a die positive Quadratwurzel $\sqrt{a}$ von a bzw. die positive Nullstelle der durch $f(x) := x^2 - a$ definierten Funktion $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ zu bestimmen, wobei $\mathbb{R}$ die Menge der reellen Zahlen bzw. die reelle Zahlengerade bedeutet. Man kann sich relativ leicht überlegen, dass das sogenannte *Newton-Verfahren*[54]

$$x_{k+1} := \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big), \qquad k = 0, 1, \ldots$$

für jeden Startwert $x_0 > 0$ gegen $\sqrt{a}$ konvergiert. Denn für $k = 0, 1, \ldots$ ist

$$x_{k+1} - \sqrt{a} = \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big) - \sqrt{a} = \frac{1}{2x_k}(x_k - \sqrt{a})^2$$

und daher $\sqrt{a} \le x_k$ für $k = 1, 2, \ldots$. Hieraus wiederum folgt

$$x_k - x_{k+1} = \frac{1}{2}\Big(x_k - \frac{a}{x_k}\Big) \ge 0, \qquad k = 1, 2, \ldots .$$

Also ist $\{x_k\}_{k\in\mathbb{N}}$ eine nach unten (durch $\sqrt{a}$) beschränkte, monoton nicht wachsende Folge, wegen eines bekannten Satzes der Analysis existiert daher $x := \lim_{k\to\infty} x_k$. Es ist $0 < \sqrt{a} \le x$ und $x = \frac{1}{2}(x + a/x)$, woraus $x = \sqrt{a}$ folgt. Das Verfahren liefert also für *jeden* positiven Startwert x_0 eine gegen $\sqrt{a}$ konvergente Folge $\{x_k\}$. Aus

$$0 \le x_{k+1} - \sqrt{a} = \frac{1}{2x_k}(x_k - \sqrt{a})^2 \le \frac{1}{2\sqrt{a}}(x_k - \sqrt{a})^2, \qquad k = 1, 2, \ldots,$$

liest man ab, dass der absolute Fehler der $(k+1)$-ten Näherung bis auf eine (multiplikative) Konstante durch das Quadrat des Fehlers der k-ten Näherung abgeschätzt werden kann, eine für ein Iterationsverfahren sehr wünschenswerte Eigenschaft, da sich dann, grob gesagt, bei jedem Iterationsschritt die Anzahl der gültigen Dezimalstellen verdoppelt. In Abbildung 51 veranschaulichen wir das Verfahren. Mit $f(x) := x^2 - a$ und einer positiven Näherung x_k für $\sqrt{a}$ ist x_{k+1} die Nullstelle der Linearisierung $f_k(x) := f(x_k) + f'(x_k)(x - x_k)$ von $f(\cdot)$ in x_k.

[54] Statt vom Newton-Verfahren spricht man hier, also für den speziellen Fall der Quadratwurzelbestimmung, auch von dem *Heron-Verfahren* (benannt nach Heron von Alexandria) oder dem *babylonischen Wurzelziehen*. Eine geometrische Motivation ist die folgende: Ein Quadrat mit dem Flächeninhalt $a > 0$ hat die Seitenlänge $\sqrt{a}$. Man gehe von einem beliebigen Rechteck aus und verändere von Schritt zu Schritt das Seitenverhältnis so, dass sich die Rechtecke immer mehr einem Quadrat annähern, während der Flächeninhalt konstant gleich a bleibt. Am Anfang wähle man eine beliebige Seitenlänge $x_0 > 0$. Die zweite Seitenlänge ist $y_0 := a/x_0$. Um eine bessere Annäherung an ein Quadrat zu erhalten, wird die längere Seite gekürzt und die kürzere Seite verlängert. Als neue Länge der langen Seite wähle man $x_1 := (x_0 + y_0)/2$, die andere Seite ist $y_1 := a/x_1$.

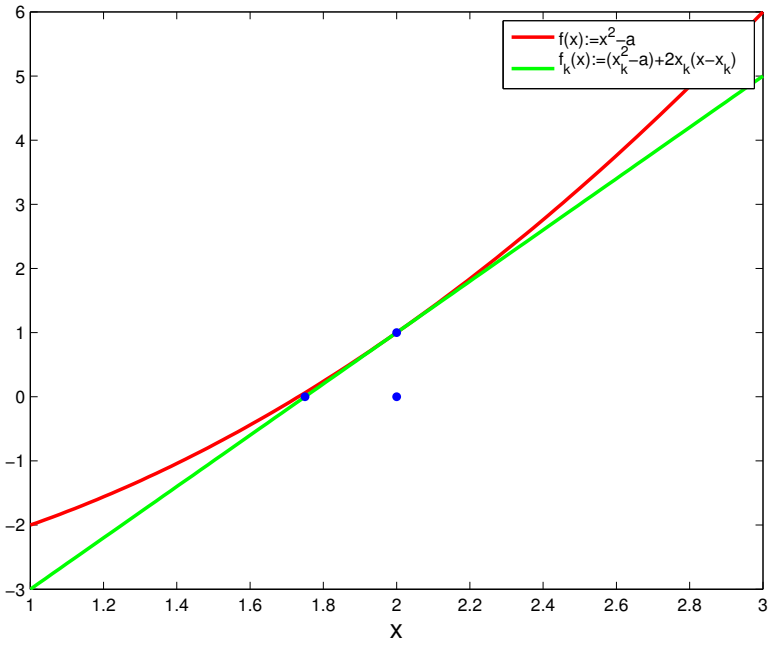


Abbildung 51: Veranschaulichung des obigen Verfahrens

Beispiel: Es sei $\sqrt{13}$ zu berechnen. Mit dem Startwert $x_0 := 4$ erhalten wir die Werte (wir verwenden das Computer-Algebra-System MuPAD mit `DIGITS:=40`).

k	x_k
0	4.0000000000000000000000000000000000000000
1	3.6250000000000000000000000000000000000000
2	3.6056034482758620689655172413793103448280
3	3.6055512758414576334068470845270729847270
4	3.6055512754639892931389800140558211796600
5	3.6055512754639892931192212674704959462510
6	3.6055512754639892931192212674704959462510

Man erkennt die außerordentlich gute Konvergenz. □

Jetzt stellt sich die Frage, welchen Startwert x_0 man bei gegebenem $a > 0$ wählen sollte, um mit möglichst wenigen Iterationen eine vorgegebene (absolute oder relative) Genauigkeit sichern zu können. Die Vorgehensweise hierbei ist die folgende:

1. In einem ersten Schritt wird die Berechnung der Quadratwurzel auf $(0, \infty)$ auf deren Berechnung auf einem Teilintervall, genauer auf $[\frac{1}{2}, 1]$ oder $[\frac{1}{4}, \frac{1}{2}]$, zurückgeführt.

2. Im zweiten Schritt wird die Wurzelfunktion auf diesem Teilintervall in einem geeigneten Sinne durch ein Polynom vom Grad kleiner oder gleich 1 approximiert und hierdurch eine erste Näherung für die Quadratwurzel erhalten.

3. Diese Näherung wird in einem dritten Schritt als Startwert für das oben angegebene Iterationsverfahren genommen, von dem je nach gewünschter Genauigkeit eine gewisse Anzahl von Schritten, etwa zwei oder drei, durchgeführt werden.

Es sei $a = 2^k t$ mit $k \in \mathbb{Z}$ und $t \in [\frac{1}{2}, 1)$. Z. B. ist $13 = 2^4 \frac{13}{16}$, also $k = 4$ und $t = \frac{13}{16}$. Nun wird eine Fallunterscheidung gemacht. Ist $k = 2p$ gerade, so ist $\sqrt{a} = 2^p \sqrt{t}$ mit $t \in [\frac{1}{2}, 1)$. Es kommt in diesem Falle also darauf an, die Quadratwurzel auf dem Intervall $[\frac{1}{2}, 1)$ zu approximieren. Ist dagegen $k = 2p + 1$ ungerade, so ist $a = 2^{2p+1} t = 2^{2(p+1)} s$ mit $s := \frac{1}{2} t$ und daher $\sqrt{a} = 2^{p+1} \sqrt{s}$ mit $s \in [\frac{1}{4}, \frac{1}{2})$ Hier muss also die Quadratwurzel auf dem Intervall $[\frac{1}{4}, \frac{1}{2})$ approximiert werden.

Bei der Beschreibung einer linearen Approximation an die Quadratwurzel beschränken wir uns im Folgenden auf den Fall, dass k gerade ist, das reduzierte Intervall, auf dem die Quadratwurzel zu approximieren ist, also durch $I := [\frac{1}{2}, 1]$ gegeben ist. Auf I soll $\sqrt{t}$ durch ein Polynom $p^*(t) = \alpha^* t + \beta^*$ vom Grade kleiner oder gleich 1 „möglichst gut“approximiert werden. Was soll das aber heißen? Naheliegend ist es, unter „möglichst gut“ zu verstehen, dass die auf I maximale betragsmäßige Abweichung $\max_{t \in I} |\alpha t + \beta - \sqrt{t}|$ minimal ist. Gesucht ist also eine Lösung (α^*, β^*) der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(\alpha, \beta) := \max_{t \in I} |\alpha t + \beta - \sqrt{t}|, \quad (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}.$$

Anschaulich ist ziemlich klar, dass die gesuchte Gerade parallel zu derjenigen ist, die durch $(\frac{1}{2}, \sqrt{\frac{1}{2}})$ und $(1, \sqrt{1}) = (1, 1)$ geht. Diese ist gegeben durch

$$p_L(t) := \sqrt{1} \cdot \frac{t - \frac{1}{2}}{1 - \frac{1}{2}} + \sqrt{\frac{1}{2}} \cdot \frac{1 - t}{1 - \frac{1}{2}} = (2 - \sqrt{2})t + \sqrt{2} - 1.$$

Da die Quadratwurzel-Funktion auf I konkav ist, ist $p_L(t) \leq \sqrt{t}$ für alle $t \in I$. „Verschiebt“man die durch p_L definierte Gerade bis sie zu einer Tangenten an $(t, \sqrt{t})$ wird, so erhält man eine Gerade, die durch

$$p_T(t) := (2 - \sqrt{2})t + \frac{2 + \sqrt{2}}{8}$$

gegeben ist. *Anschaulich* ist dann klar, dass der „goldene Mittelweg“zwischen p_L und p_T zum gesuchten Polynom p^* führt, d. h. dass

$$p^*(t) = \frac{p_L(t) + p_T(t)}{2} = (2 - \sqrt{2})t + \frac{3}{8}\Big(\frac{3}{2}\sqrt{2} - 1\Big)$$

und die gesuchten Koeffizienten α^* und β^* durch

$$\alpha^* := 2 - \sqrt{2}, \qquad \beta^* := \frac{3}{8}\Big(\frac{3}{2}\sqrt{2} - 1\Big)$$

gefunden sind. In Abbildung 52 geben wir die Defekte $p_L(t) - \sqrt{t}$, $p_T(t) - \sqrt{t}$ sowie $p^*(t) - \sqrt{t}$ an. Der Defekt $p^*(t) - \sqrt{t}$ nimmt an den Intervallenden $\frac{1}{2}$ und 1 von I sowie

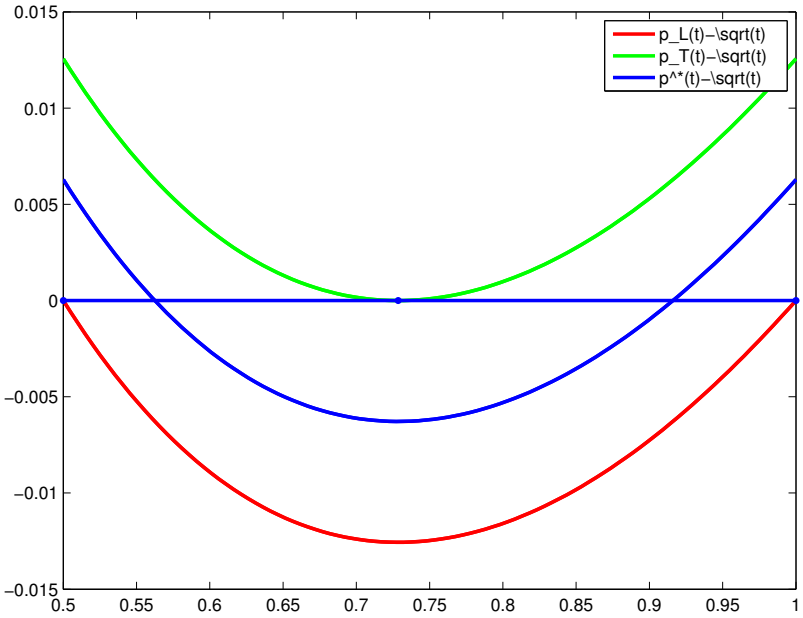


Abbildung 52: Approximation der Wurzelfunktion auf $[\frac{1}{2}, 1]$ durch ein lineares Polynom

dem Punkt $t^* := (3 + 2\sqrt{2})/8$, in dem p_T den Graphen der Wurzelfunktion tangential berührt, seinen betragsmäßig größten Wert an. Genauer ist

$$(*) \quad p^*(\tfrac{1}{2}) - \sqrt{\tfrac{1}{2}} = -(p^*(t^*) - \sqrt{t^*}) = p^*(1) - \sqrt{1} = \frac{5}{8} - \frac{7}{16}\sqrt{2} = \max_{t\in I} |p^*(t) - \sqrt{t}|.$$

Hiermit wollen wir nicht nur anschaulich klarmachen, sondern *beweisen*, dass p^* das gesuchtze Polynom vom Grade kleiner oder gleich 1 ist. Genauer formulieren wir die folgende Aussage:

Satz *Sei*

$$p^*(t) := (2 - \sqrt{2})t + \frac{3}{8}\Big(\frac{3}{2}\sqrt{2} - 1\Big)$$

und $I := [\frac{1}{2}, 1]$. *Dann ist*

$$\max_{t\in I} |p^*(t) - \sqrt{t}| \le \max_{t\in I} |p(t) - \sqrt{t}|$$

für alle Polynome vom Grad kleiner oder gleich 1.

Beweis: Sei p ein beliebiges Polynom vom Grad kleiner oder gleich 1. Im Widerspruch zur Behauptung nehmen wir an, es sei

$$|p(t) - \sqrt{t}| < \max_{t\in I} |p^*(t) - \sqrt{t}| =: \mu^*, \qquad t \in I,$$

insbesondere also $p \neq p^*$. Wir zeigen, dass das Vorzeichen von $p(\cdot) - p^*(\cdot)$ in den Punkten $\frac{1}{2}$, $t^* := (3 + 2\sqrt{2})/8$ und 1 alterniert, wobei wir obige Beziehung $(*)$ benutzen werden.

Hieraus folgt dann, dass $p(\cdot) - p^*(\cdot)$ in $(\frac{1}{2}, t^*)$ und $(t^*, 1)$ mindestens eine Nullstelle hat. Da aber $p(\cdot) - p^*(\cdot) \neq 0$ höchstens eine Nullstelle in I hat, ist der gewünschte Widerspruch erreicht. Nun ist

$$\begin{aligned} p(\tfrac{1}{2}) - p^*(\tfrac{1}{2}) &= [p(\tfrac{1}{2}) - \sqrt{\tfrac{1}{2}}] - [p^*(\tfrac{1}{2}) - \sqrt{\tfrac{1}{2}}] \\ &= [p(\tfrac{1}{2}) - \sqrt{\tfrac{1}{2}}] - \mu^* \\ &\leq |p(\tfrac{1}{2}) - \sqrt{\tfrac{1}{2}}| - \mu^* \\ &< 0. \end{aligned}$$

Entsprechend ist $p(t^*) - p^*(t^*) > 0$ und $p(1) - p^*(1) < 0$. Insgesamt ist obige Aussage bewiesen. □

Für $a = 2^{2p}t$ mit $p \in \mathbb{Z}$ und $t \in [\frac{1}{2}, 1)$ erhält man durch obige Überlegungen die Näherung $x_0(a) := 2^p p^*(t)$ für die Quadratwurzel $\sqrt{a}$. Die hierdurch erzielte Fehlerabschätzung

$$|x_0(a) - \sqrt{a}| = 2^p\,|p^*(t) - \sqrt{t}| \leq 2^p\Big(\frac{5}{8} - \frac{7}{16}\sqrt{2}\Big) \leq 2^p \cdot 0.0063$$

ist i. Allg. noch nicht gut genug. Bildet man dagegen

$$x_1(a) := \frac{1}{2}\Big(x_0(a) + \frac{a}{x_0(a)}\Big), \qquad x_2(a) := \frac{1}{2}\Big(x_1(a) + \frac{a}{x_1(a)}\Big),$$

so erhält man unter Berücksichtigung von $p^*(t) \geq p^*(\frac{1}{2}) = \frac{5}{8} + \frac{1}{16}\sqrt{2} \geq 0.7428$, dass

$$0 \leq x_1(a) - \sqrt{a} = \frac{1}{x_0(a)}[x_0(a) - \sqrt{a}]^2 = \frac{2^p[p^*(t) - \sqrt{t}]^2}{2p^*(t)} \leq \frac{2^p \cdot (0.0063)^2}{2 \cdot 0.7428} \leq 2^p \cdot 2.7 \cdot 10^{-5}$$

und anschließend

$$0 \leq x_2(a) - \sqrt{a} \leq \frac{1}{2\sqrt{a}}[x_1(a) - \sqrt{a}]^2 \leq 2^p \cdot 5.2 \cdot 10^{-10}.$$

Das dürfte für viele praktische Zwecke eine ausreichende Genauigkeit sein.

Beispiel: Wir berechnen $\sqrt{13} = 4\sqrt{13/16}$. Mit MuPAD und `DIGITS:=40` erhalten wir

k	$x_k(13)$
0	3.585786437626904951198311275790301921430
1	3.605605747310196728521372016476152591615
2	3.605551275875457190220842818343262545961
3	3.605551275463989293142699761796460259062
4	3.605551275463989293119221267470495946251
5	3.605551275463989293119221267470495946251

Der Gewinn gegenüber dem oben benutzten schlechteren Startwert $x_0 = 4$ ist also nicht übermäßig groß. □

42 Das Fagnano-Problem

Giovanni Francesco Fagnano dei Toschi veröffentlichte und löste 1775 das später nach ihm benannte Fagnano-Problem. Es lautet:

> Man bestimme auf jeder Seite eines gegebenen spitzwinkligen Dreiecks $\triangle ABC$ (hierbei denken wir uns die Ecken A, B und C jeweils als Punkte im $\mathbb{R}^2$, also der Ebene, gegeben) einen Punkt derart, dass das hierdurch bestimmte Dreieck minimalen Umfang besitzt.

Bei R. Courant, H. Robbins (1967, S. 264) heißt das entsprechende Problem das *Schwarzsche Dreiecksproblem*. Unser Ziel besteht darin, einen analytischen (einen eleganten elementargeometrischen Beweis von L. Fejer findet man bei H. S. M. Coxeter (1969, S. 20)) Beweis des folgenden Satzes anzugeben.

Satz Das durch das Fagnano-Problem gesuchte Dreieck zum gegebenen spitzwinkligen Dreieck $\triangle ABC$ ist das sogenannte *Höhenfußpunktdreieck* $\triangle U^*V^*W^*$, welches dadurch entsteht, dass das Lot jeder Ecke von $\triangle ABC$ auf die gegenüberliegende Seite gefällt wird.

Beweis: In Abbildung 53 haben wir ein Dreieck $\triangle ABC$ und das zugehörige Höhenfuß-

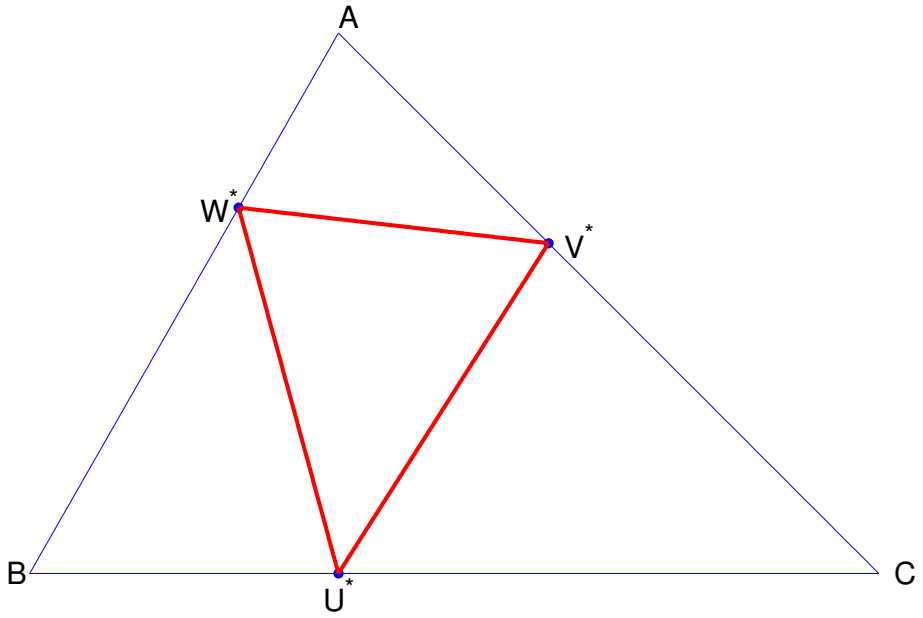


Abbildung 53: Das Fagnano-Problem

punktdreieck $\triangle U^*V^*W^*$ gezeichnet. Allgemein machen wir für die Punkte $U = U(\lambda)$,

$V = V(\mu)$ bzw. $W = W(\nu)$ auf den Seiten $\overline{BC}$, $\overline{CA}$ bzw. $\overline{AB}$ den Ansatz

$$U(\lambda) := (1-\lambda)B + \lambda C, \qquad V(\mu) := (1-\mu)C + \mu A, \qquad W(\nu) := (1-\nu)A + \nu B$$

mit $(\lambda, \mu, \nu) \in M := [0,1] \times [0,1] \times [0,1]$. Der Umfang des Dreiecks $\triangle UVW$ ist gegeben durch

$$\begin{aligned} f(\lambda,\mu,\nu) &:= \|U(\lambda) - V(\mu)\|_2 + \|V(\mu) - W(\nu)\|_2 + \|W(\nu) - U(\lambda)\|_2 \\ &= \|(1-\lambda)(B-C) + \mu(C-A)\|_2 + \|(1-\mu)(C-A) + \nu(A-B)\|_2 \\ &\quad + \|(1-\nu)(A-B) + \lambda(B-C)\|_2. \end{aligned}$$

Hierbei ist $\|\cdot\|_2$ die euklidische Norm auf dem $\mathbb{R}^2$, also z. B. $\|U\|_2 = \sqrt{U^T U}$. Das Fagnano-Problem kann dann auch als

$$\text{(P)} \qquad \text{Minimiere} \quad f(\lambda,\mu,\nu) \quad \text{auf} \quad M := [0,1] \times [0,1] \times [0,1]$$

formuliert werden. Da M kompakt und f stetig ist, besitzt diese Aufgabe (P) eine Lösung $(\lambda^*, \mu^*, \nu^*) \in M$. Wir nehmen zunächst an, es sei $(\lambda^*, \mu^*, \nu^*) \in (0,1) \times (0,1) \times (0,1)$, d. h. $(\lambda^*, \mu^*, \nu^*)$ liege im Inneren von M. Dann ist f in $(\lambda^*, \mu^*, \nu^*)$ stetig differenzierbar. Notwendigerweise verschwindet der Gradient von f in $(\lambda^*, \mu^*, \nu^*)$, also die partiellen Ableitungen von f nach λ, μ und ν in $(\lambda^*, \mu^*, \nu^*)$. Zur Abkürzung schreiben wir $U^* := U(\lambda^*)$, $V^* := V(\mu^*)$ bzw. $W^* := W(\nu^*)$ und erhalten

$$\begin{aligned} \frac{\partial f}{\partial \lambda}(\lambda^*,\mu^*,\nu^*) &= \Big(\frac{W^*-U^*}{\|W^*-U^*\|_2} - \frac{U^*-V^*}{\|U^*-V^*\|}\Big)^T (B-C) = 0, \\ \frac{\partial f}{\partial \mu}(\lambda^*,\mu^*,\nu^*) &= \Big(\frac{U^*-V^*}{\|U^*-V^*\|_2} - \frac{V^*-W^*}{\|V^*-W^*\|_2}\Big)^T (C-A) = 0, \\ \frac{\partial f}{\partial \nu}(\lambda^*,\mu^*,\nu^*) &= \Big(\frac{V^*-W^*}{\|V^*-W^*\|_2} - \frac{W^*-U^*}{\|W^*-U^*\|_2}\Big)^T (A-B) = 0. \end{aligned}$$

Da $B-C$ ein positives Vielfaches von $B-U^*$ und U^*-C ist, und entsprechendes für $C-A$ und $A-B$ gilt, ist

$$\begin{aligned} \frac{(W^*-U^*)^T(B-U^*)}{\|W^*-U^*\|_2\|B-U^*\|_2} &= \frac{(V^*-U^*)^T(C-U^*)}{\|V^*-U^*\|_2\|C-U^*\|_2}, \\ \frac{(U^*-V^*)^T(C-V^*)}{\|U^*-V^*\|_2\|C-V^*\|_2} &= \frac{(W^*-V^*)^T(C-V^*)}{\|W^*-V^*\|_2\|C-V^*\|_2}, \\ \frac{(V^*-W^*)^T(A-W^*)}{\|V^*-W^*\|_2\|A-W^*\|_2} &= \frac{(U^*-W^*)^T(B-W^*)}{\|U^*-W^*\|_2\|B-W^*\|_2}. \end{aligned}$$

Dies bedeutet, dass

$$\alpha := \sphericalangle W^*U^*B = \sphericalangle V^*U^*C, \ \beta := \sphericalangle U^*V^*C = \sphericalangle W^*V^*C, \ \gamma := \sphericalangle V^*W^*A = \sphericalangle U^*W^*B.$$

In Abbildung 54 haben wir neben dem Höhenfußpunktdreieck die entsprechenden Winkel eingetragen. Betrachtet man nun der Reihe nach die Dreiecke $\triangle W^*BU^*$, $\triangle U^*CV^*$ und $\triangle V^*AW^*$, so erhält man

$$\sphericalangle ABC = \pi - (\alpha+\gamma), \quad \sphericalangle BCA = \pi - (\beta+\alpha), \quad \sphericalangle CAB = \pi - (\gamma+\beta).$$

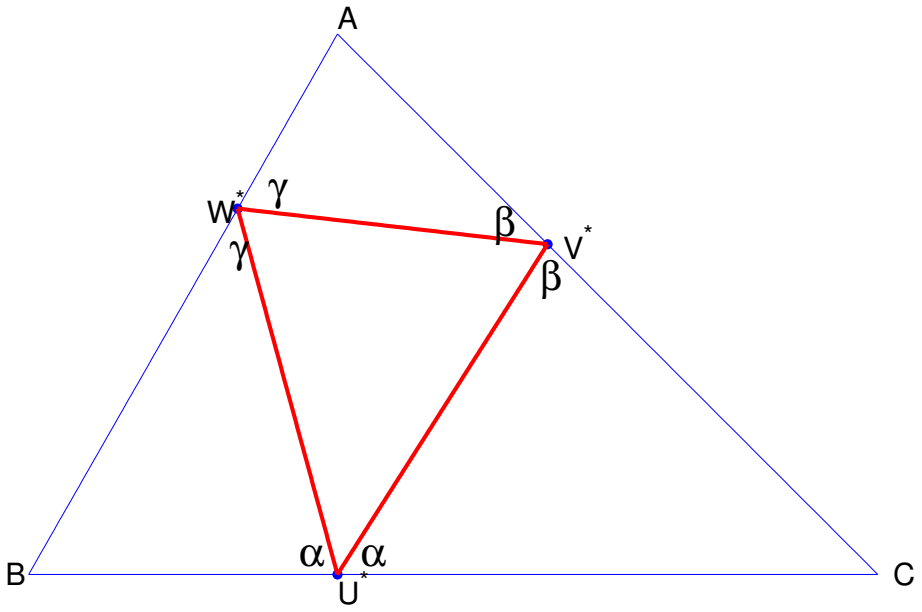


Abbildung 54: Winkel im Höhenfußpunktdreieck

Aufsummieren dieser drei Gleichungen liefert $\pi = 3\pi - 2(\alpha+\beta+\gamma)$ bzw. $\alpha+\beta+\gamma = \pi$, woraus dann

$$\begin{aligned}\alpha &:= \sphericalangle W^*U^*B = \sphericalangle V^*U^*C = \sphericalangle CAB,\\ \beta &:= \sphericalangle U^*V^*C = \sphericalangle W^*V^*C = \sphericalangle ABC,\\ \gamma &:= \sphericalangle V^*W^*A = \sphericalangle U^*W^*B = \sphericalangle BCA\end{aligned}$$

folgt. Zur Abkürzung schreiben wir jetzt z. B. $|BC|$ statt $\|B-C\|_2$ für die Länge der Strecke $\overline{BC}$. Der Sinussatz, angewandt auf das Dreieck $\triangle ABC$, sagt aus, dass

$$\frac{|BC|}{\sin\alpha} = \frac{|CA|}{\sin\beta} = \frac{|AB|}{\sin\gamma} = 2r,$$

wobei r der Radius des Umkreises zu $\triangle ABC$ ist. Eine entsprechende Anwendung auf die Dreiecke $\triangle U^*CV^*$, $\triangle V^*AW^*$ und $\triangle W^*BU^*$ liefert

$$\begin{aligned}\frac{\sin\alpha}{\sin\beta} &= \frac{|BC|}{|CA|} = \frac{|CV^*|}{|CU^*|} = \frac{|CA|-|AV^*|}{|CU^*|},\\ \frac{\sin\beta}{\sin\gamma} &= \frac{|CA|}{|AB|} = \frac{|AW^*|}{|AV^*|} = \frac{|AB|-|BW^*|}{|AV^*|},\\ \frac{\sin\gamma}{\sin\alpha} &= \frac{|AB|}{|BC|} = \frac{|BU^*|}{|BW^*|} = \frac{|BC|-|CU^*|}{|BW^*|}.\end{aligned}$$

Für $|CU^*|$, $|AV^*|$ und $|BW^*|$ erhalten wir ein System von drei linearen linearen Gleichungen, nämlich

$$\begin{pmatrix} |BC| & |CA| & 0 \\ 0 & |CA| & |AB| \\ |BC| & 0 & |AB| \end{pmatrix} \begin{pmatrix} |CU^*| \\ |AV^*| \\ |BW^*| \end{pmatrix} = \begin{pmatrix} |CA|^2 \\ |AB|^2 \\ |BC|^2 \end{pmatrix}$$

mit der Lösung

$$\begin{pmatrix} |CU^*| \\ |AV^*| \\ |BW^*| \end{pmatrix} = \begin{pmatrix} \frac{|BC|^2+|CA|^2-|AB|^2}{2|BC|} \\ \frac{|CA|^2+|AB|^2-|BC|^2}{2|CA|} \\ \frac{|AB|^2+|BC|^2-|CA|^2}{2|AB|} \end{pmatrix} = \begin{pmatrix} |CA|\cos\gamma \\ |AB|\cos\alpha \\ |BC|\cos\beta \end{pmatrix},$$

wobei wir für die zweite Gleichung den Kosinussatz angewandt haben. Hieraus kann man schließen, dass $|AU^*|$, $|BV^*|$ und $|CW^*|$ die Höhen im Dreieck $\triangle ABC$ sind bzw. $\triangle UVW$ das Höhenfußpunktdreieck ist. Denn wendet man den Kosinussatz auf $\triangle AU^*C$ an, so erhält man

$$\begin{aligned} |AU^*|^2 &= |CU^*|^2 + |CA|^2 - 2|CU^*|\,|CA|\cos\gamma \\ &= |CA|^2\cos^2\gamma + |CA|^2 - 2|CA|^2\cos^2\gamma \\ &= |CA|^2\sin^2\gamma \end{aligned}$$

und hieraus $|AU^*| = |CA|\sin\gamma$, das ist die Höhe im Dreieck $\triangle ABC$, welche A mit der gegenüberliegenden Seite verbindet. Ähnlich erhält man die entsprechenden Aussagen. Als Umfang $\mathrm{per}(\triangle U^*V^*W^*)$ von $\triangle U^*V^*W^*$ berechnet man

$$\begin{aligned} \mathrm{per}(\triangle U^*V^*W^*) &= |U^*V^*| + |V^*W^*| + |W^*U^*| \\ &= |CU^*|\,\frac{|AB|}{|CA|} + |AV^*|\,\frac{|BC|}{|AB|} + |BW^*|\,\frac{|CA|}{|BC|} \\ &\qquad \text{(wegen des Sinussatzes)} \\ &= |AB|\cos\gamma + |BC|\cos\alpha + |CA|\cos\beta \\ &= 2r(\sin\gamma\cos\gamma + \sin\alpha\cos\alpha + \sin\beta\cos\beta) \\ &= r[\sin(2\alpha) + \sin(2\beta) + \sin(2\gamma)] \\ &= 4r\sin\alpha\sin\beta\sin\gamma. \end{aligned}$$

Hierbei haben wir am Schluss ausgenutzt, dass (Produktformel für den Sinus)

$$\begin{aligned} \sin\alpha\sin\beta\sin\gamma &= \frac{1}{4}[\sin(\underbrace{\alpha+\beta-\gamma}_{=\pi-2\gamma}) + \sin(\underbrace{\beta+\gamma-\alpha}_{=\pi-2\alpha}) + \sin(\underbrace{\gamma+\alpha-\beta}_{\pi-2\beta}) \\ &\qquad + \sin(\underbrace{\alpha+\beta+\gamma}_{=\pi})] \\ &= \frac{1}{4}[\sin(2\alpha) + \sin(2\beta) + \sin(2\gamma)]. \end{aligned}$$

Bisher haben wir bewiesen:

- Das Fagnano-Problem besitzt eine Lösung $\triangle U^*V^*W^*$.
- Ist eine Lösung $\triangle U^*V^*W^*$ des Fagnano-Problems nichtentartet, stimmt also keine der Ecken von $\triangle U^*V^*W^*$ mit einer der Ecken von $\triangle ABC$ überein, so ist $\triangle U^*V^*W^*$ das Höhenfußpunktdreieck zu $\triangle ABC$ und sein Umfang gegeben durch
$$\mathrm{per}(\triangle U^*V^*W^*) = 4r\sin\alpha\sin\beta\sin\gamma,$$
wobei $\alpha := \sphericalangle CAB$, $\beta := \sphericalangle ABC$, $\gamma := \sphericalangle BCA$ und r der Radius des Umkreises zu $\triangle ABC$ ist.

Es bleibt zu zeigen, dass eine Lösung des Fagnano-Problems notwendigerweise nichtentartet ist. Der kleinste Umfang eines dem Dreieck $\triangle ABC$ einbeschriebenen entarteten „Dreiecks“ (eigentlich ist es eine Strecke) ist das Doppelte der kleinsten Höhe im Dreieck. Angenommen, diese kleinste Höhe sei die, die A mit der gegenüberliegenden Seite $\overline{BC}$ verbindet. Wir bezeichnen sie mit h_A. Das Doppelte dieser Höhe ist
$$2h_A = 2|CA|\sin\gamma = 4r\sin\beta\sin\gamma > 4r\sin\alpha\sin\beta\sin\gamma = \mathrm{per}(U^*V^*W^*),$$
wobei wir am Schluss ausgenutzt haben, dass $\triangle ABC$ spitzwinklig, also jeder Winkel kleiner als $\pi/2$ ist. Insgesamt ist der Satz bewiesen. □

43 Das Apfelmännchen

Die Menge
$$M := \{c \in \mathbb{C} : \text{ Die Folge } \{z_k\} \text{ mit } z_0 := 0,\ z_{k+1} := z_k^2 + c, \text{ ist beschränkt}\}$$
wird *Mandelbrot-Menge* (benannt nach dem Mathematiker Benoît Mandelbrot) genannt, ihre Visualisierung in der komplexen Zahlenebene wird häufig als *Apfelmännchen* bezeichnet. Wir wollen uns zunächst überlegen, dass
$$\left\{c \in \mathbb{C} : |c| \le \frac{3}{16}\right\} \subset M \subset \{c \in \mathbb{C} : |c| \le 2\}.$$
Denn: Sei $c \in \mathbb{C}$ mit $|c| \le \frac{3}{16}$ gegeben und die Folge $\{z_k\}$ durch $z_0 := 0$ und $z_{k+1} := z_k^2 + c$ definiert. Wir zeigen durch Induktion nach k, dass $|z_k| \le \frac{3}{4}$ für alle $k \in \mathbb{N}$ und insbesondere die Folge $\{z_k\}$ beschränkt ist bzw. $c \in M$ gilt. Es ist $|z_1| = |c| \le \frac{3}{16} < \frac{3}{4}$. Angenommen, es sei $|z_k| \le \frac{3}{4}$. Dann ist unter Benutzung dieser Induktionsvoraussetzung $|z_{k+1}| \le |z_k|^2 + |c| \le \frac{9}{16} + \frac{3}{16} = \frac{3}{4}$, womit die erste der oben behaupteten Inklusionen bewiesen ist. Nun nehmen wir an, es sei $|c| > 2$, etwa $|c| = 2 + \epsilon$ mit $\epsilon > 0$, und zeigen, dass $c \notin M$, also die durch $z_0 := 0$, $z_{k+1} := z_k^2 + c$ definierte Folge $\{z_k\}$ unbeschränkt ist. Hierzu zeigen wir durch Induktion nach k, dass $|z_k| \ge 2 + k\epsilon$ für alle $k \in \mathbb{N}$, woraus die Unbeschränktheit von $\{z_k\}$ folgt. Es ist $|z_1| = |c| = 2 + \epsilon$. Angenommen, es ist $|z_k| \ge 2 + k\epsilon$. Dann ist
$$|z_{k+1}| \ge |z_k|^2 - (2+\epsilon) \ge (2+k\epsilon)^2 - 2 - \epsilon \ge 2 + (k+1)\epsilon,$$
womit die Induktionsbehauptung bewiesen ist.

In Abbildung 55 wird die Mandelbrotmenge bzw. das Apfelmännchen dargestellt. In einem file `Apfel.m` haben wir die folgende MATLAB-Funktion abgelegt.

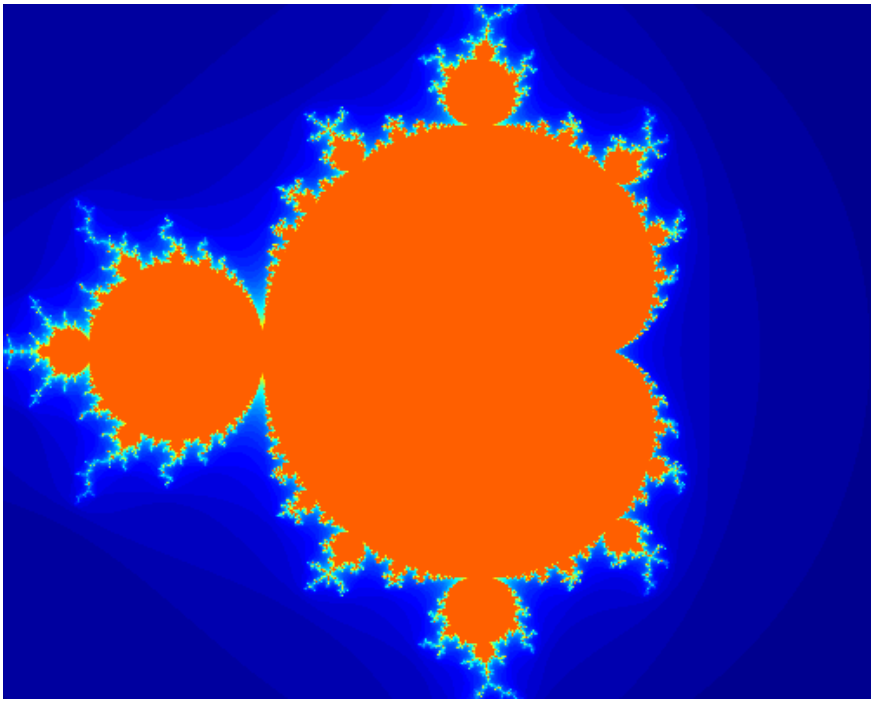


Abbildung 55: Das Apfelmännchen

```
function Apfel(max_iter,pixel,x_min,x_max,y_min,y_max);
r=(y_max-y_min)/(x_max-x_min);
x=linspace(x_min,x_max,pixel); y=linspace(y_min,y_max,round(r*pixel));
[Re,Im]=meshgrid(x,y);  C=Re+i*Im; B=zeros(round(r*pixel),pixel); Ck=B;
for k=1:max_iter
   Ck=Ck.*Ck+C;
   B=B+(abs(Ck)<2);
end;
image(B);  axis equal axis off
```

Die Abbildung 55 erhalten wir durch den Aufruf `Apfel(50,500,-1.5,1,-1,1);` In Abbildung 56 geben wir einen Ausschnitt des Apfelmännchens an, welchen wir mit `Apfel(50,500,-0.2,0,0.8,1);` erhalten haben.

44 Die erste Optimierungsaufgabe in der Geschichte der Mathematik

Bei M. Cantor (1880, S. 228) steht, dass die erste Extremwert- oder Optimierungsaufgabe in Euklids Elementen, Buch VI, Theorem 27 vorkommt. Im Prinzip handelt

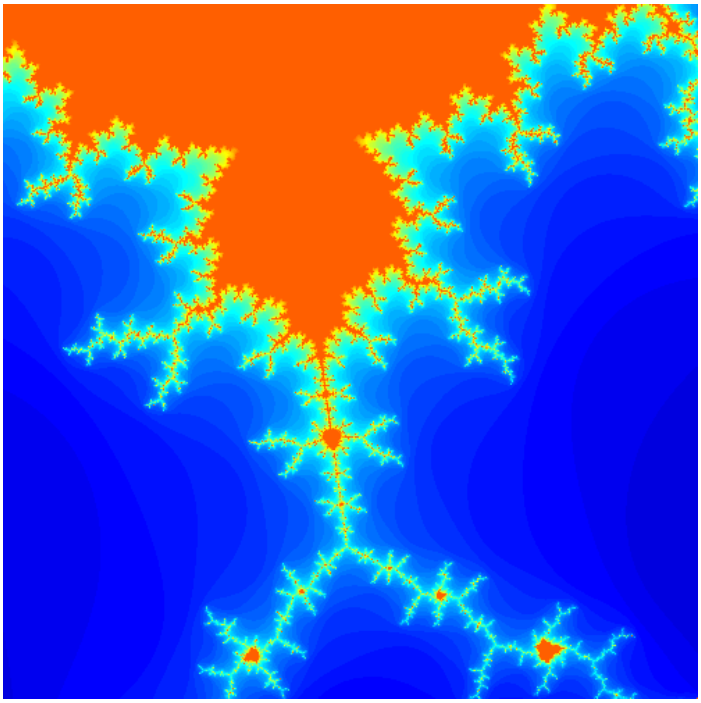


Abbildung 56: Ausschnitt des Apfelmännchens

es sich dabei um die folgende Aufgabe, siehe auch V. M. Tikhomirov (1990, S. 27):

- Finde einen Punkt E auf der Seite BC eines Dreiecks $\triangle ABC$ derart, dass das Parallelogramm $ADEF$ mit Eckpunkten D bzw. F auf den Seiten AB bzw. AC maximalen Flächeninhalt besitzt.

Die Lösung ist offensichtlich der Mittelpunkt der Strecke BC. In Abbildung 57 wird dies verdeutlicht. Denn ist E beliebig auf BC und

$$x := \frac{|BE|}{|BC|},$$

so ist

$$g(x) := \text{Flächeninhalt}(ADEF) = 2x(1-x) \cdot \text{Flächeninhalt}(\triangle ABC),$$

und diese Funktion g wird auf $M := [0,1]$ maximal für $x^* := \frac{1}{2}$. Bei M. Cantor kann man nachlesen: Satz 27 enthält das erste Maximum, welches in der Geschichte der Mathematik nachgewiesen worden ist, und welches als Function geschrieben besagen würde: $x(a-x)$ erhalte seinen grössten Werth durch $x = a/2$.

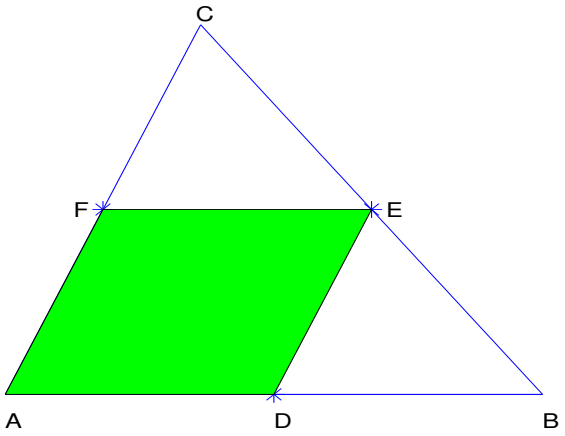


Abbildung 57: Die Lösung des ältesten Optimierungsproblems

45 Das Farkas Lemma

Das Lemma von Farkas wurde 1902 von Julius Farkas aufgestellt und gibt Bedingungen für die Lösbarkeit linearer Gleichungs-Ungleichungssysteme an. Dadurch werden entsprechende Aussagen[55] für lineare Gleichungssysteme verallgemeinert. Ich bin der Meinung, dass das Farkas Lemma ein bemerkenswertes Ergebnis ist. Es ist einfach formulierbar, nicht ganz einfach zu beweisen, es hat eine geometrische Interpretation und ist ein wichtiges Hilfsmittel zum Beweis des starken Dualitätssatzes der linearen Optimierung und des Satzes von Kuhn-Tucker.

Farkas Lemma *Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ gegeben. Dann ist genau eine der beiden folgenden Aussagen richtig:*

(a) Es existiert $x \in \mathbb{R}^n$ mit $Ax = b$ und $x \geq 0$.

(b) Es existiert $y \in \mathbb{R}^m$ mit $A^T y \geq 0$ und $b^T y < 0$.

Hierbei sind Ungleichungen für Vektoren, also $x \geq 0$ und $A^T y \geq 0$, komponentenweise zu verstehen.

Bemerkung: Wie wir sehen werden, ist im wesentlichen zu zeigen: Gilt (a) nicht, so gilt (b). Dass (a) nicht gilt, bedeutet, dass $b \notin K := \{Ax : x \geq 0\}$, sich also b nicht als eine nichtnegative Linearkombination der Spalten $a_1, \ldots, a_n$ von A darstellen lässt. Existiert ein $y \in \mathbb{R}^m$ mit $A^T y \geq 0$ und $b^T y < 0$, gilt also (b), so bedeutet dies, dass mit der Hyperebene durch den Nullpunkt $H := \{z \in \mathbb{R}^m : y^T z = 0\}$ der Vektor b in einem hierdurch erzeugten offenen Halbraum und K im gegenüberliegenden abgeschlossenen Halbraum liegt. In Abbildung 58 wird dies verdeutlicht. Diese Veranschaulichung kann auch als Beweisidee dienen. Hierzu zeigt man, dass $K \subset \mathbb{R}^m$ konvex und abgeschlossen ist und wendet auf $\{b\}$ und K einen sogenannten strikten Trennungssatz an. Wir werden einen geometrisch nicht so intuitiven, dafür aber insgesamt etwas kürzeren Beweis angeben. □

[55] Sind $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$, so ist das lineare Gleichungssystem $Ax = b$ genau dann lösbar, wenn $b^T y = 0$ für alle $y \in \mathbb{R}^m$ mit $A^T y = 0$.

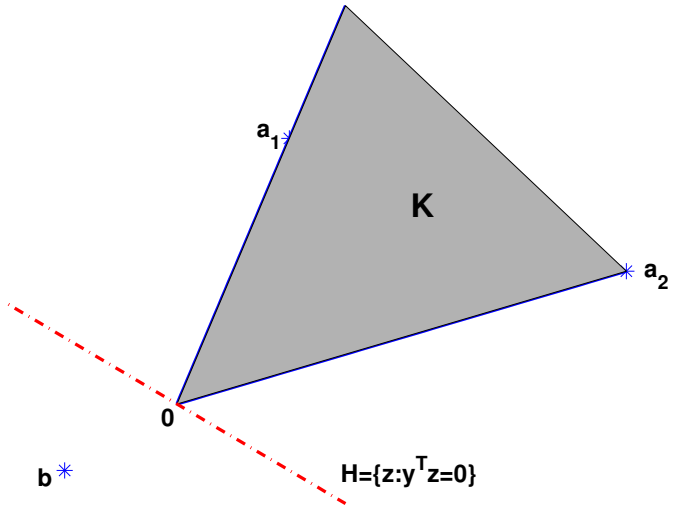


Abbildung 58: Veranschaulichung des Farkas Lemmas

Beweis des Farkas Lemmas: Die Aussagen (a) und (b) können nicht beide richtig sein. Denn wäre dies der Fall, so gäbe es $x \in \mathbb{R}^n$ und $y \in \mathbb{R}^m$ mit

$$0 > b^T y = (Ax)^T y = x^T A^T y \geq 0,$$

ein Widerspruch. Nun nehmen wir an, die Aussage (a) sei falsch, das Gleichungssystem $Ax = b$ habe also keine nichtnegative Lösung, und zeigen die Richtigkeit von (b). Wir unterscheiden zwei Fälle.

Im ersten Fall hat $Ax = b$ überhaupt keine Lösung, geschweige denn eine nichtnegative. Wegen der oben in einer Fußnote angegebenen Lösbarkeitsbedingung für lineare Gleichungsysteme gibt es dann ein $y \in \mathbb{R}^m$ mit $A^T y = 0$ und $b^T y \neq 0$. Indem man notfalls y durch $-y$ ersetzt, kann man $b^T y < 0$ erreichen und erkennt, dass (b) richtig ist.

Im zweiten Fall nehmen wir an, $Ax = b$ habe eine Lösung, aber keine nichtnegative, und zeigen auch in diesem Fall die Richtigkeit von (b). Dies geschieht durch Induktion nach n, der Anzahl der Spalten von A. Wir nehmen an, es sei $A = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}$, es sei also a_j die j-te Spalte von A. Der Induktionsanfang ist bei $n = 1$. Die Gleichung $Ax = a_1 x_1 = b$ hat nach Annahme nur eine Lösung $x_1 < 0$. Wir setzen $y := -b$. Dann ist

$$A^T y = a_1^T y = -\frac{b^T b}{x_1} > 0, \qquad b^T y = -b^T b < 0.$$

Für $n = 1$ ist (b) also richtig. Jetzt nehmen wir an, die Aussage (b) sei richtig für Matrizen mit höchstens $n-1$ Spalten. Wir haben zu zeigen, dass (b) auch für die Matrix $A = \begin{pmatrix} a_1 & \cdots & a_{n-1} & a_n \end{pmatrix}$ richtig ist. Da $Ax = \sum_{j=1}^{n} x_j a_j = b$ keine nichtnegative Lösung besitzt, hat auch $\sum_{j=1}^{n-1} x_j a_j = b$ keine nichtnegative Lösung. Wegen der Induktionsvoraussetzung gibt es ein $y^* \in \mathbb{R}^m$ mit $a_j^T y^* \geq 0$, $j = 1, \ldots, n-1$, und $b^T y^* < 0$.

Wir können annehmen, dass $a_n^T y^* < 0$, denn andernfalls hätten wir mit y^* eine Lösung von $A^T y \geq 0$, $b^T y < 0$ gefunden. Wir definieren

$$\hat{a}_j := (a_j^T y^*) a_n - (a_n^T y^*) a_j, \quad j = 1, \ldots, n-1, \qquad \hat{b} := (b^T y^*) a_n - (a_n^T y^*) b.$$

Dann hat $\sum_{j=1}^{n-1} \hat{x}_j \hat{a}_j = \hat{b}$ keine nichtnegative Lösung $\hat{x} \in \mathbb{R}^{n-1}$, denn andernfalls hätte $Ax = b$ wegen

$$\sum_{j=1}^{n-1} \underbrace{\hat{x}_j}_{\geq 0} a_j \underbrace{- \frac{1}{a_n^T y^*}}_{>0} \Big(\sum_{j=1}^{n-1} \underbrace{\hat{x}_j (a_j^T y^*)}_{\geq 0} \underbrace{- b^T y^*}_{>0} \Big) a_n = b$$

eine nichtnegative Lösung, im Widerspruch zur Annahme. Wegen der Induktionsannahme existiert $\hat{y} \in \mathbb{R}^m$ mit

$$\hat{a}_j^T \hat{y} \geq 0, \quad j = 1, \ldots, n-1, \qquad \hat{b}^T \hat{y} < 0.$$

Nun setze man

$$y := (a_n^T \hat{y}) y^* - (a_n^T y^*) \hat{y}.$$

Wir wollen zeigen, dass mit diesem y die Aussage (a) richtig ist, d. h. $A^T y \geq 0$ und $b^T y < 0$ gilt. Denn es ist

$$a_j^T y = (a_n^T \hat{y})(a_j^T y^*) - (a_n^T y^*)(a_j^T \hat{y}) = \hat{a}_j^T \hat{y} \geq 0, \quad j = 1, \ldots, n-1,$$

sowie

$$a_n^T y = 0, \qquad b^T y = \hat{b}^T \hat{y} < 0.$$

Damit ist das Farkas Lemma bewiesen. □

Bemerkung: Aus dem Farkas Lemma kann man weitere sogenannte *Alternativsätze* ableiten. So sagt der Alternativsatz von Gordan[56] aus: Ist $A \in \mathbb{R}^{m \times n}$, so ist genau eine der beiden folgenden Aussagen richtig:

(a) Es existiert $x \in \mathbb{R}^n$ mit $Ax = 0$, $x \geq 0$ und $x \neq 0$.

(b) Es existiert $y \in \mathbb{R}^m$ mit $A^T y > 0$ (d. h. alle Komponenten von $A^T y$ sind positiv.

Denn: Der Nachweis dafür, dass die Aussagen (a) und (b) nicht beide richtig sein können, ist einfach und soll übergangen werden. Angenommen, (a) gilt nicht. Mit $e := (1, \ldots, 1)^T \in \mathbb{R}^n$ besitzt dann auch

$$\begin{pmatrix} A \\ e^T \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad x \geq 0$$

keine Lösung. Aus dem Farkas Lemma folgt, dass

$$\begin{pmatrix} A^T & e \end{pmatrix} \begin{pmatrix} y \\ \delta \end{pmatrix} \geq 0, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} y \\ \delta \end{pmatrix} < 0$$

eine Lösung besitzt. Es ist $\delta < 0$ und folglich $A^T y \geq -\delta e > 0$, d. h. (b) ist richtig. □

[56] Paul Gordan (1837-1912) war Doktorvater (1907) von Emmy Noether (1882–1935).

46 Der Dualitätssatz der linearen Optimierung

Als Anwendung des Farkas Lemmas wollen wir jetzt den *Dualitätssatz* der linearen Optimierung beweisen. Hierbei wird einer sogenannten *primalen* linearen Optimierungsaufgabe eine hierzu *duale* lineare Optimierungsaufgabe zugeordnet und der enge Zusammenhang zwischen beiden Problemen geklärt. Im folgenden Dualitätssatz gehen wir bei dem primalen Problem von der sogenannten *Normalform* einer linearen Optimierungsaufgabe aus und zeigen in einer Bemerkung im Anschluss, dass sich die erhaltenen Ergebnisse auf *scheinbar* allgemeinere Formulierungen einer linearen Optimierungsaufgabe übertragen lassen.

Dualitätssatz der linearen Optimierung *Seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Hiermit betrachte man die beiden linearen Optimierungsaufgaben*

$$\text{(P)} \qquad \textit{Minimiere} \quad c^T x \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : Ax = b,\ x \geq 0\}$$

und

$$\text{(D)} \qquad \textit{Maximiere} \quad b^T y \quad \textit{auf} \quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Dann gilt[57]*:*

1. *Ist $x \in M$ und $y \in N$, so ist $b^T y \leq c^T x$.*
2. *Ist $x^* \in M$, $y^* \in N$ und $b^T y^* = c^T x^*$, so ist x^* eine Lösung von (P) und y^* eine Lösung von (D).*
3. *Ist $M \neq \emptyset$ und $N \neq \emptyset$, so besitzen (P) und (D) Lösungen $x^* \in M$ bzw. $y^* \in N$ mit $b^T y^* = c^T x^*$.*
4. *Ist $N \neq \emptyset$, $M = \emptyset$, so ist $\sup_{y\in N} b^T y = +\infty$.*
5. *Ist $M \neq \emptyset$, $N = \emptyset$, so ist $\inf_{x\in M} c^T x = -\infty$.*

Beweis: Sei $x \in M$ und $y \in N$. Dann ist

$$b^T y = (Ax)^T y = x^T A^T y \leq x^T c = c^T x,$$

womit der erste Teil des Satzes bewiesen ist.

Sei $x^* \in M$, $y^* \in N$ und $b^T y^* = c^T x^*$. Sind $x \in M$ und $y \in N$ beliebig, so ist nach der gerade eben bewiesenen Aussage

$$b^T y \leq c^T x^* = b^T y^* \leq c^T x,$$

woraus folgt, dass x^* eine Lösung von (P) und y^* eine Lösung von (D) ist. Die Aussagen des sogenannten schwachen Dualitätssatzes sind damit bewiesen.

[57] Die ersten beiden, leicht zu beweisenden Aussagen, heißen auch die Aussagen des *schwachen Dualitätssatzes*, während die restlichen den sogenannten *starken Dualitätssatz* bilden.

Wir setzen für den dritten Teil des Satzes $M \neq \emptyset$ sowie $N \neq \emptyset$ voraus und zeigen, dass es $x \in M$, $y \in N$ mit $c^T x = b^T y$ gibt bzw. das Gleichungs-Ungleichungssystem

$$(*) \qquad Ax = b, \quad x \geq 0, \quad A^T y \leq c, \quad c^T x = b^T y$$

lösbar ist. Wir nehmen das Gegenteil an. Durch die Einführung von sogenannten Schlupfvariablen z und die Darstellung der Variablen y, die nicht vorzeichenbeschränkt ist, als Differenz $y_+ - y_-$ von zwei vorzeichenbeschränkten Variablen y_+, y_- können wir das Gleichungs-Ungleichungssystem $(*)$ auf eine Form bringen, auf die das Farkas Lemma unmittelbar anwendbar ist. Wenn $(*)$ nicht lösbar ist, hat auch

$$\begin{pmatrix} A & 0 & 0 & 0 \\ 0 & A^T & -A^T & I \\ c^T & -b^T & b^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y_+ \\ y_- \\ z \end{pmatrix} = \begin{pmatrix} b \\ c \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} x \\ y_+ \\ y_- \\ z \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

keine Lösung. Wegen des Farkas Lemma existiert $(u, v, \delta) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ mit

$$\begin{pmatrix} A^T & 0 & c \\ 0 & A & -b \\ 0 & -A & b \\ 0 & I & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ \delta \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} b \\ c \\ 0 \end{pmatrix}^T \begin{pmatrix} u \\ v \\ \delta \end{pmatrix} < 0$$

bzw.

$$(**) \qquad A^T u + \delta c \geq 0, \quad Av - \delta b = 0, \quad v \geq 0, \quad b^T u + c^T v < 0.$$

Mit $x \in M$ und $y \in N$ (solche existieren nach Voraussetzung) ist

$$\begin{aligned} \delta(\underbrace{b^T y - c^T x}_{\leq 0}) &= (Av)^T y - \delta c^T x \\ &= v^T A^T y - \delta c^T x \\ &\leq c^T v + (A^T u)^T x \\ &= c^T v + (Ax)^T u \\ &= b^T u + c^T v \\ &< 0. \end{aligned}$$

Hieraus folgt $\delta > 0$. Nun definiere man $u^* = -u/\delta$ und $v^* := v/\delta$. Wegen $\delta > 0$ folgt $v^* \in M$, $u^* \in N$ und $c^T v^* < b^T u^*$, was ein Widerspruch zum ersten Teil dieses Satzes ist.

Nun sei $N \neq \emptyset$ und $M = \emptyset$. Wegen des Farkas Lemmas existiert ein $y^* \in \mathbb{R}^m$ mit $A^T y^* \geq 0$ und $b^T y^* < 0$. Sei $y \in N$ beliebig. Dann ist $y - ty^* \in N$ für alle $t \geq 0$ und $b^T(y - ty^*) \to +\infty$ mit $t \to +\infty$, also $\sup_{y \in N} b^T y = +\infty$.

Jetzt sei $M \neq \emptyset$ und $N = \emptyset$. Dann hat

$$\begin{pmatrix} A^T & -A^T & I \end{pmatrix} \begin{pmatrix} y_+ \\ y_- \\ z \end{pmatrix} = c, \qquad \begin{pmatrix} y_+ \\ y_- \\ z \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

keine Lösung. Wegen des Farkas Lemmas existiert ein $x^* \in \mathbb{R}^n$ mit $Ax^* = 0$, $x^* \geq 0$ und $c^T x^* < 0$. Mit einem beliebigen $x \in M$ ist $x + tx^* \in M$ für alle $t \geq 0$ und $c^T(x + tx^*) \to -\infty$ mit $t \to +\infty$, also $\inf_{x \in M} c^T x = -\infty$. Insgesamt ist der (starke) Dualitätssatz vollständig bewiesen. □

Bemerkung: Der Dualitätssatz war für *scheinbar* spezielle lineare Optimierungsaufgaben (P) und (D) formuliert, bei welchen die primale Aufgabe (P) in sogenannter Normalform vorliegt. Das bedeutet, dass *zulässige Lösungen* als nichtnegativ vorausgesetzt werden und die Restriktionen in Gleichungsform vorliegen. Nun gehen wir von einer *scheinbar* allgemeinen linearen Optimierungsaufgabe aus, bei welcher nur ein Teil der Variablen als nichtnegativ vorausgesetzt sind und sowohl Gleichungen als auch Ungleichungen unter den Restriktionen vorkommen. Genauer betrachten wir die lineare Optimierungsaufgabe

$$(\mathrm{P}_A) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad c_1^T x_1 + c_2^T x_2 \quad \text{auf} \\ M_A := \left\{ (x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : x_1 \geq 0, \quad \begin{array}{l} A_{11}x_1 + A_{12}x_2 \geq b_1, \\ A_{21}x_1 + A_{22}x_2 = b_2 \end{array} \right\}. \end{array} \right.$$

Hierbei seien $b_i \in \mathbb{R}^{m_i}$, $i = 1, 2$. Die Dimensionen der anderen Vektoren und Matrizen gehen aus dem Zusammenhang hervor. Z. B. sind $c_i \in \mathbb{R}^{n_i}$, $i = 1, 2$, ferner $A_{11} \in \mathbb{R}^{m_1 \times n_1}$, $A_{21} \in \mathbb{R}^{m_2 \times n_1}$ usw. Das Problem (P_A) wird in Normalform überführt, indem man die Schlupfvariable $u_1 \in \mathbb{R}^{m_1}$ einführt und die nicht vorzeichenbeschränkte Variable x_2 mittels $x_2 = x_2^+ - x_2-$ als Differenz nichtnegativer Variabler $x_2^+, x_2^- \in \mathbb{R}^{n_2}$ darstellt. Aus (P_A) erhalten wir damit die lineare Optimierungsaufgabe in Normalform

$$\text{Minimiere} \quad \begin{pmatrix} c_1 \\ c_2 \\ -c_2 \\ 0 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2^+ \\ x_2^- \\ u_1 \end{pmatrix} \quad \text{auf}$$

$$M := \left\{ \begin{pmatrix} x_1 \\ x_2^+ \\ x_2^- \\ u_1 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2^+ \\ x_2^- \\ u_1 \end{pmatrix} \geq 0, \begin{pmatrix} A_{11} & A_{12} & -A_{12} & -I \\ A_{21} & A_{22} & -A_{22} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2^+ \\ x_2^- \\ u_1 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\}.$$

Diese Aufgabe nennen wir (P). Die Aufgaben (P_A) und (P) sind im folgenden Sinne äquivalent:

1. Ist (x_1, x_2) zulässig für (P_A), $x_2 = x_2^+ - x_2^-$ mit $x_2^+, x_2^- \geq 0$ und $u_1 := A_{11}x_1 + A_{12}x_2 - b_1$, so ist (x_1, x_2^+, x_2^-, u_1) zulässig für (P) und die Zielfunktionswerte stimmen überein.

2. Ist (x_1, x_2^+, x_2^-, u_1) zulässig für (P) und $x_2 := x_2^+ - x_2^-$, so ist (x_1, x_2) zulässig für (P_A) (und die Zielfunktionswerte stimmen überein).

3. Entsprechend 1. und 2. erhält man aus einer Lösung von (P_A) eine von (P) und umgekehrt.

Nun bilden wir die zu (P) duale lineare Optimierungsaufgabe (D). Diese ist gegeben durch

$$\text{(D)} \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{auf} \\ N := \left\{ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : \begin{pmatrix} A_{11}^T & A_{21}^T \\ A_{12}^T & A_{22}^T \\ -A_{12}^T & -A_{22}^T \\ -I & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \leq \begin{pmatrix} c_1 \\ c_2 \\ -c_2 \\ 0 \end{pmatrix} \right\} \end{array} \right.$$

bzw.

$$\text{(D)} \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad b_1^T y_1 + b_2^T y_2 \quad \text{auf} \\ N := \left\{ (y_1, y_2) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} : y_1 \geq 0, \quad \begin{array}{l} A_{11}^T y_1 + A_{21}^T y_2 \leq c_1, \\ A_{12}^T y_1 + A_{22}^T y_2 = c_2 \end{array} \right\}. \end{array} \right.$$

Klar ist jedenfalls, dass die Aussagen des Dualitätssatzes für das Paar linearer Optimierungsaufgaben (P_A) und (D) ganz entsprechend gelten. Weiter ist es nicht schwer nachzuweisen, dass das zu (D) duale Problem gerade wieder (P) bzw. (P_A) ist. Denn hierzu schreibe man (D) nach Überführung in eine Minimierungsaufgabe durch Einführung einer Schlupfvariablen v_1 und Darstellung von y_2 als Differenz nichtnegativer y_2^+, y_2^- in Normalform:

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \begin{pmatrix} -b_1 \\ -b_2 \\ b_2 \\ 0 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2^+ \\ y_2^- \\ v_1 \end{pmatrix} \quad \text{auf} \\ N_A := \left\{ \begin{pmatrix} y_1 \\ y_2^+ \\ y_2^- \\ v_1 \end{pmatrix} : \begin{pmatrix} y_1 \\ y_2^+ \\ y_2^- \\ v_1 \end{pmatrix} \geq 0, \begin{pmatrix} A_{11}^T & A_{21}^T & -A_{21}^T & I \\ A_{12}^T & A_{22}^T & -A_{22}^T & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2^+ \\ y_2^- \\ v_1 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\}. \end{array} \right.$$

Die duale Aufgabe hierzu ist

$$\left\{ \begin{array}{c} \text{Maximiere} \quad \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}^T \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \text{auf} \\ \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ -A_{21} & -A_{22} \\ I & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \leq \begin{pmatrix} -b_1 \\ -b_2 \\ b_2 \\ 0 \end{pmatrix} \right\} \end{array} \right.$$

bzw.

$$\left\{ \begin{array}{c} \text{Maximiere} \quad c_1^T z_1 + c_2^T z_2 \quad \text{auf} \\ \left\{ (z_1, z_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : z_1 \leq 0, \quad \begin{array}{l} A_{11} z_1 + A_{12} z_2 \leq -b_1, \\ A_{21} z_1 + A_{22} z_2 = -b_2 \end{array} \right\}. \end{array} \right.$$

Ersetzt man (z_1, z_2) durch $(-x_1, -x_2)$, so erkennt man, dass diese Aufgabe mit dem Ausgangsproblem (P_A) übereinstimmt. □

Beispiel: Wie man bei G. B. Dantzig (1966, S. 625 ff.) nachlesen kann, war eine der ersten Anwendungen der linearen Optimierung die Bestimmung eines ausreichenden Speiseplans bei möglichst niedrigen Kosten. Dieses Problem wurde zuerst von G. J. Stigler untersucht. Gegeben seien n Lebensmittel (bei Stigler ist $n = 77$, die Lebensmittel sind u. a. Weizenmehl, Maismehl, Kondensmilch, Margarine, Cheddar-Käse, Erdnussbutter, Schmalz, Rindsleber, Schweinelendenbraten, Lachs, grüne Bohnen, Kohl, Zwiebeln, Kartoffeln, Spinat, Süßkartoffeln, getrocknete Pfirsiche, getrocknete Pflaumen, Lima-Bohnen, weiße Bohnen), welche m Nährwerte (bei Stigler ist $m = 9$, die Nährwerte sind Kalorien, Eiweiß, Kalzium, Eisen, Vitamin A, Vitamin B_1, Vitamin B_2, Nicotinsäureamid, Vitamin C) enthalten. Sei a_{ij} die Anzahl der Mengeneinheiten des i-ten Nährwertes in einer Mengeneinheit des j-ten Lebensmittels. Eine zulässige Diät muss mindestens b_i Einheiten des i-ten Nährwertes enthalten. So wird z. B. bei Dantzig angegeben, dass ein mäßig aktiver Mann, der 70 kg wiegt, als täglichen Bedarf 3000 Kalorien, 70 g Eiweiß, 0.8 g Kalzium, 12 g Eisen usw. benötigt. Bekannt sei ferner der Preis c_j (etwa in Euro) einer Einheit des j-ten Lebensmittels. Ein Diätplan besteht in der Angabe eines Vektors $x = (x_j) \in \mathbb{R}^n$, wobei x_j die Anzahl der Einheiten des j-ten Hilfsmittels angibt. Dieser ist zulässig, wenn durch ihn die Mindestanforderungen an die enthaltenen Nährmittel sämtlich erfüllt sind, wenn also

$$\sum_{j=1}^{n} a_{ij} x_j \geq b_i, \qquad i = 1, \ldots, m.$$

Die zugehörigen Kosten sind $\sum_{j=1}^{n} c_j x_j$. Ferner berücksichtige man, dass die Komponenten x_j eines Diätplans x sämtlich nichtnegativ sein müssen, schließlich soll einem auch bei einer Diät etwas zugeführt und nicht weggenommen werden. Geht man noch zur Vektor- bzw. Matrixschreibweise über, setzt also $A := (a_{ij}) \in \mathbb{R}^{m \times n}$, $b := (b_i) \in \mathbb{R}^m$ und $c := (c_j) \in \mathbb{R}^n$, so kommt man zur linearen Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax \geq b\}.$$

Die hierzu duale lineare Optimierungsaufgabe ist (siehe obige Bemerkung)

$$\text{(D)} \qquad \text{Maximiere} \quad b^T y \quad \text{auf} \quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Dieses duale Problem wollen wir nun interpretieren. Hierzu stellen wir uns vor, ein Pharmazieunternehmen D mache den Vorschlag, die benötigten Nährwerte in Pillenform herzustellen und damit in Konkurrenz zu dem Unternehmen P zu treten, welches sozusagen mit konventionellen Speiseplänen arbeitet. Hierzu verlangt das Pharmazieunternehmen D für eine Mengeneinheit des i-ten Nährwertes y_i Geldeinheiten (etwa Euro). Der von den Pillenherstellern geforderte Preis für eine "Mahlzeit", die die verlangten b_i Einheiten des i-ten Nährwerts enthält, ist $b^T y = \sum_{i=1}^{m} b_i y_i$. Diesen Preis wird D versuchen zu maximieren. Dann ist $\sum_{i=1}^{m} a : ijy_i = (A^T y)_j$ der Preis der Pillen, die einer Einheit des j-ten Lebensmittels entsprechen. Das Unternehmen P wird auf den Vorschlag von D daher nur eingehen, wenn $A^T y \leq c$. Daher hat das Pharmazieunternehmen genau die duale lineare Optimierungsaufgabe (D) zu lösen. □

47 Das Minimalkosten-Fluss-Problem

Ein Produkt (Öl oder Orangen oder ...) wird in gewissen Orten in einer bestimmten Menge angeboten und an anderen Orten verlangt. Weiter gibt es Orte, in denen das Produkt weder angeboten noch verlangt wird, in denen es aber umgeladen werden kann. Gewisse Orte sind miteinander durch Verkehrswege miteinander verbunden. Die Kosten für den Transport einer Mengeneinheit des Gutes längs eines Verkehrsweges sind bekannt, ferner sind Kapazitätschranken für jeden Transportweg vorgegeben. Diese geben obere und eventuell auch untere Schranken für die zu transportierende Menge auf dem Weg an. Gesucht ist ein zulässiger Transportplan mit minimalen Kosten. *Zulässig* bedeutet hierbei:

1. Die Kapazitätsbeschränkungen auf den Verkehrswegen werden respektiert.
2. Der Bedarf in allen Orten, die das Produkt verlangen, wird gedeckt.
3. In jedem Ort, in dem das Produkt angeboten wird, wird nicht mehr wegtransportiert als dort vorhanden ist.

Es wird in diesem Abschnitt darauf ankommen, dieser Aufgabe ein mathematisches Modell zuzuordnen. Hierzu wird den Orten *Knoten*, den Verkehrswegen *Pfeile* eines *gerichteten Graphen* (oder auch *Digraphen*) $(\mathcal{N}, \mathcal{A})$ zugeordnet. Hier steht $\mathcal{N}$ für die (endliche) Menge der *Knoten* (**N**odes) und $\mathcal{A}$ für die Menge der *Pfeile* (**A**rcs), also *geordneten* Paaren von Knoten. Mit jedem Knoten $k \in \mathcal{N}$ ist eine Mengenangabe b_k des im gerichteten Graphen $(\mathcal{N}, \mathcal{A})$ zu transportierenden Gutes verbunden. Ist $b_k > 0$, so sind b_k Mengeneinheiten dieses Gutes im Knoten k vorhanden und Knoten k wird ein *Angebotsknoten* genannt. Ist dagegen $b_k < 0$, so werden dort $|b_k|$ Mengeneinheiten benötigt, man spricht von einem *Bedarfsknoten*. Im Fall $b_k = 0$ handelt es sich um einen *Umladeknoten*.

Zu jedem Pfeil $(i, j) \in \mathcal{A}$ des gerichteten Graphen gehören die Kosten c_{ij} für den Fluss einer Mengeneinheit auf ihm. Mit x_{ij} wird der Fluss auf diesem Pfeil bezeichnet. Untere Schranken für die *Kapazität* des Pfeils $(i, j) \in \mathcal{A}$ sind durch l_{ij} (oft ist $l_{ij} = 0$), obere Schranken durch $u_{ij} \geq l_{ij}$ gegeben. Gesucht wird ein Fluss im Digraphen, der unter Berücksichtigung der Kapazitätsbeschränkungen die Angebote und Bedürfnisse mengenmäßig ausgleicht und die dafür erforderlichen Kosten minimiert. Dabei ist in jedem Knoten der Fluss zu erhalten. Dies bedeutet für den Knoten $k \in \mathcal{N}$, dass die Summe der Flüsse auf seinen eingehenden Pfeilen plus der in ihm verfügbaren (wenn k ein Angebotsknoten) beziehungsweise minus der von ihm benötigten (wenn k ein Bedarfsknoten) Menge $|b_k|$ gleich der Summe der Flüsse auf seinen ausgehenden Pfeilen ist. Die Flusserhaltungsbedingung für den Knoten k lautet daher

$$\sum_{i:(i,k)\in\mathcal{A}} x_{ik} + b_k = \sum_{j:(k,j)\in\mathcal{A}} x_{kj}.$$

Das kapazitierte lineare Minimalkosten-Fluss-Problem lässt sich daher wie folgt formu-

lieren:

$$\left\{\begin{array}{c} \text{Minimiere} \quad \displaystyle\sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \\ \\ \text{unter den Nebenbedingungen} \\ \\ \displaystyle\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = b_k \qquad (k\in\mathcal{N}), \qquad l_{ij}\le x_{ij}\le u_{ij} \qquad ((i,j)\in\mathcal{A}). \end{array}\right.$$

Diese Aufgabe wollen wir nun in Matrix-Vektorschreibweise formulieren. Der Fluss $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ hat so viele Komponenten wie es Pfeile gibt, ihre Anzahl ist also $|\mathcal{A}|$. Entsprechend sind die Kosten $c = (c_{ij})_{(i,j)\in\mathcal{A}}$ und die Kapazitäten $l = (l_{ij})_{(i,j)\in\mathcal{A}}$ und $u = (u_{ij})_{(i,j)\in\mathcal{A}}$ Vektoren des $\mathbb{R}^{|\mathcal{A}|}$. Der Vektor $b = (b_k)_{k\in\mathcal{N}}$ hat $|\mathcal{N}|$ Komponenten. Definiert man die *Knoten-Pfeil-Inzidenzmatrix* $A = (a_{k,(i,j)})_{k\in\mathcal{N},(i,j)\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{A}|}$ durch

$$a_{k,(i,j)} := \left\{\begin{array}{rl} 1, & \text{falls } k = i, \\ -1, & \text{falls } k = j, \\ 0, & \text{sonst,} \end{array}\right.$$

so erkennt man, dass obiges Netzwerkflussproblem, das sogenannte (kapazitierte) Minimale-Kosten-Fluss-Problem, in der Form

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x\in\mathbb{R}^n : l\le x\le u,\ Ax = b\}$$

geschrieben werden kann.

Beispiel: Gegeben sei der *gerichtete Graph* in Abbildung 59. In den blauen Orten bzw.

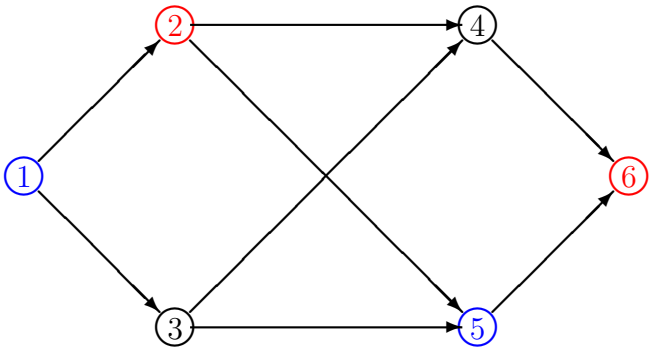


Abbildung 59: Ein gerichteter Graph

Knoten 1 und 5 wird ein gewisses Produkt hergestellt, das in den roten Knoten 2 und 6 benötigt wird. Die schwarzen Knoten 3 und 4 sind Umladeknoten. Das Angebot bzw. Bedarf in den einzelnen Knoten ist gegeben durch

Knoten k	1	2	3	4	5	6
b_k (in Tonnen)	30	-10	0	0	5	-25

Die Kosten (pro Mengeneinheit) c_{ij} und die unteren sowie oberen Kapazitätsschranken l_{ij} bzw. u_{ij} auf den Pfeilen (i,j) werden jetzt angegeben:

Pfeil	$(1,2)$	$(1,3)$	$(2,4)$	$(2,5)$	$(3,4)$	$(3,5)$	$(4,6)$	$(5,6)$
l_{ij} (in Tonnen)	0	0	0	0	0	0	0	0
u_{ij} (in Tonnen)	25	18	10	12	7	10	10	15
c_{ij} (in Euro)	120	130	130	50	70	70	11	110

Die zu dem in Abbildung 59 angegebenen gerichteten Graph gehörende Knoten-Pfeil-Inzidenzmatrix (die Nummerierung der Pfeile erfolgt wie in obiger Tabelle) ist

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

Mit

$$b := \begin{pmatrix} 30 \\ -10 \\ 0 \\ 0 \\ 5 \\ -25 \end{pmatrix}, \qquad c := \begin{pmatrix} 120 \\ 130 \\ 130 \\ 50 \\ 70 \\ 70 \\ 11 \\ 110 \end{pmatrix}, \qquad u := \begin{pmatrix} 25 \\ 18 \\ 10 \\ 12 \\ 7 \\ 10 \\ 10 \\ 15 \end{pmatrix}, \qquad x = \begin{pmatrix} x_{12} \\ x_{13} \\ x_{24} \\ x_{25} \\ x_{34} \\ x_{35} \\ x_{46} \\ x_{56} \end{pmatrix}$$

ist die zum Minimalkosten-Fluss-Problem gehörende lineare Optimierungsaufgabe gegeben durch

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x : Ax = b,\ 0 \le x \le u\}.$$

Als Lösung (z. B. mit Hilfe der MATLAB-Funktion `linprog`) erhält man

$$x^* = (23, 7, 3, 10, 7, 0, 10, 15)^T.$$

In Abbildung 60 haben wir den kostenminimalen Fluss in dem gerichteten Graphen eingetragen. □

Auf die numerische Behandlung des Minimalkosten-Fluss-Problems wollen wir nicht eingehen, das würde den Rahmen sprengen. Es ist dabei einleuchtend, dass die spezielle Struktur des Problems ausgenutzt werden sollte. Uns kam es lediglich darauf an, den Weg von einer noch nicht genau formulierten Aufgabenstellung zu einem klar formulierten zugehörigen mathematischen Modell aufzuzeigen.

Bemerkung: Addiert man die $|\mathcal{N}|$ Zeilen der Knoten-Pfeil-Inzidenzmatrix A eines gerichteten Graphen $(\mathcal{N}, \mathcal{A})$, so erhält man einen Nullvektor, da in jeder Spalte von A genau eine 1 und genau eine -1 steht. Der Rang von A ist daher $\le |\mathcal{N}| - 1$. Wir wollen uns überlegen:

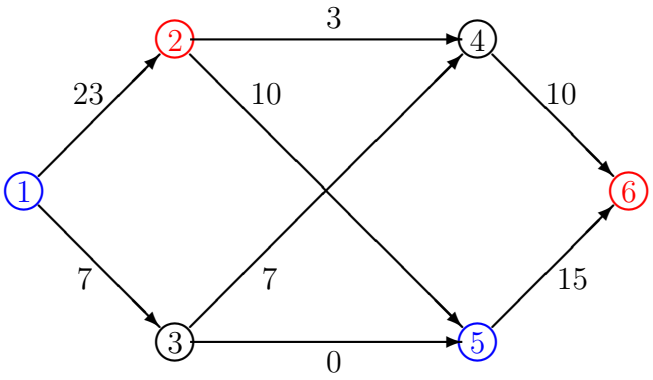


Abbildung 60: Ein kostenminimaler Fluss

- *Ist $(\mathcal{N}, \mathcal{A})$ schwach zusammenhängend, d. h. als ungerichteter Graph zusammenhängend, so hat die zugehörige Knoten-Pfeil-Inzidenzmatrix $A \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{A}|}$ den Rang $|\mathcal{N}| - 1$.*

Denn: Sei $G = (V, E)$ der zu $(\mathcal{N}, \mathcal{A})$ gehörende *ungerichtete Graph*, der als zusammenhängend vorausgesetzt ist. Man bestimme einen G *aufspannenden Baum* (V, E') mit $E' \subset E$, also einen zusammenhängenden Teilgraphen mit derselben Knotenmenge und der Eigenschaft, dass in (V, E') keine Kreise existieren. Die *Existenz* eines aufspannenden Baumes zu einem beliebigen zusammenhängenden Graphen ist einfach einzusehen. Enthält nämlich ein zusammenhängender Graph einen Kreis, so entferne man aus diesem Kreis eine Kante. Der hierdurch entstehende Graph hat eine Kante weniger und ist immer noch zusammenhängend. Nach endlich vielen Schritten erhält man einen zusammenhängenden, kreisfreien Graphen (V, E') mit derselben Knotenmenge wie der Ausgangsgraph, also einen aufspannenden Baum. Nun zeigen wir, dass es in $G' = (V, E')$ eine Ecke vom Grad 1 gibt und $|E'| = |V| - 1$ gilt. Ist nämlich $P = x, x_1, \ldots, y$ ein längster Weg in G, also ein Weg mit maximal vielen Kanten, so hat x (und y) nur einen Nachbarn aus P. Es kann keinen weiteren aus P geben, da G' kreisfrei ist, und keinen weiteren außerhalb von P, da P maximale Länge hat. Man setze $G'_1 := (V_1, E'_1)$ mit $V_1 := (V \setminus \{x\}$, $E'_1 := E' \setminus \{(x, x_1)\}$. Auch G_1 ist ein Baum, d. h. zusammenhängend und kreisfrei. Nach $m - 2$ Schritten mit $m := |V|$ erhalten wir einen Baum $G'_{m-2} = (V_{m-2}, E'_{m-2})$ auf 2 Knoten (und eine Kante). Also ist $|V| - |E'| = |V_{m-2}| - |E'_{m-2}| = 2 - 1 = 1$, also $|E'| = |V| - 1$. Nun zeigen wir, dass die $|V| - 1 = |\mathcal{N}| - 1$ Spalten der Knoten-Pfeil-Inzidenzmatrix A, die zu den Kanten E' eines G aufspannenden Baumes gehören, linear unabhängig sind. Wie wir gerade eben bewiesen haben, gibt es in $G' = (V, E')$ einen Knoten mit dem Grad 1, der also in G' genau einen Nachbarn besitzt. In der dem Knoten entsprechenden Zeile haben die Spalten genau einen von Null verschiedenen Eintrag, nämlich 1 oder -1. Bei einer zu Null sich addierenden Linearkombination der Spalten muss also der Koeffizient verschwinden, der zu der Spalte gehört, in der 1 bzw. -1 steht. Streichen des Knotens vom Grad 1 und der einzigen inzidierenden Kante liefert wieder einen Baum, auf den dasselbe Argument sinngemäß angewandt werden kann. Damit haben wir gezeigt, dass

es $|\mathcal{N}| - 1$ linear unabhängige Spalten in A gibt, womit schließlich Rang $(A) = |\mathcal{N}| - 1$ bewiesen ist. □

Beispiel: Fassen wir den gerichteten Graphen in Abbildung 59 als ungerichtet auf, so ist der in Abbildung 61 angegebene Graph ein diesen Graphen aufspannender Baum.Die

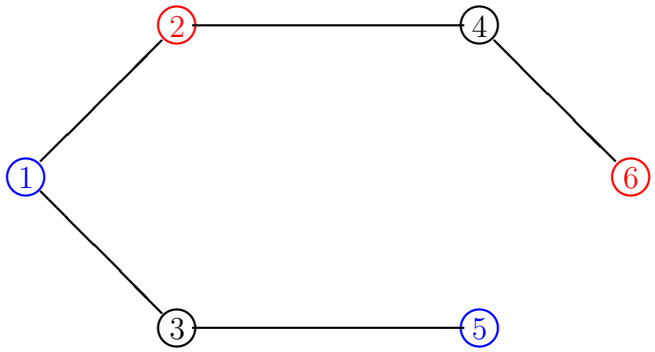


Abbildung 61: Ein aufspannender Baum

den Kanten $(1,2)$, $(1,3)$, $(2,4)$, $(3,5)$ und $(4,6)$ entsprechenden Spalten von A sind zusammengefasst in

$$A_B := \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

Offensichtlich ist Rang $(A_B) = 5$, die fünf Spalten von A_B sind linear unabhängig. □

48 Das Max Flow-Min Cut Theorem von Ford-Fulkerson

Beim *Maximaler-Fluss-Problem* (*maximum flow problem*) ist ein gerichteter Graph $(\mathcal{N}, \mathcal{A})$ gegeben, in dem zwei Knoten s (Quelle, **s**ource, kein Pfeil endet in s) und t (Senke, **t**erminal, kein Pfeil startet in t) ausgezeichnet sind. Längs der Pfeile sind wieder Kapazitäten festgelegt, d. h. es ist ein Vektor $u = (u_{ij})_{(i,j)\in\mathcal{N}}$ gegeben. Es wird angenommen, dass es eine die Quelle s und die Senke t verbindenden (Vorwärts-) Pfad gibt. Es wird nach dem maximalen Fluss von s nach t gefragt, also nach der maximalen Anzahl der Mengeneinheiten, die bei s losgeschickt werden können und in t ankommen, wobei natürlich die Kapazitätsbeschränkungen zu berücksichtigen sind und der Fluss in jedem Knoten erhalten bleibt. Unser Ziel in diesem Abschnitt ist es, das Maximaler-Fluss-Problem als eine lineare Optimierungsaufgabe zu formulieren und anschließend

das Max Flow-Min Cut Theorem von Ford-Fulkerson mit Hilfe des Dualitätssatzes der linearen Optimierung zu beweisen.

Beispiel: In Abbildung 62 geben wir einen gerichteten Graphen mit 8 Knoten und 14 Pfeilen an, eingetragen sind ferner die Kapazitäten längs der Pfeile. Was ist der

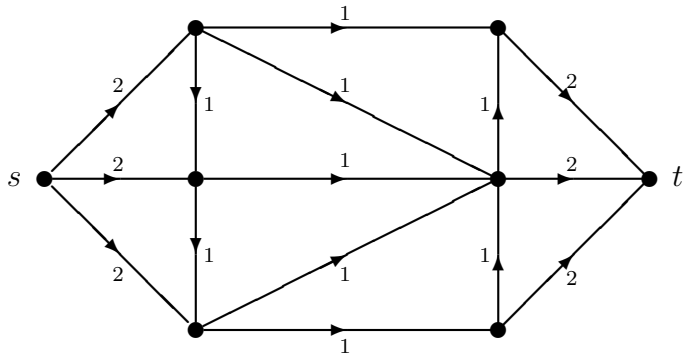


Abbildung 62: Ein gerichteter Graph mit 8 Knoten und 14 Pfeilen

maximale Fluss? Klar ist, dass dieser nicht größer als 6 sein kann, da die drei Pfeile weg von der Quelle nur eine Gesamtkapazität von 6 besitzen.

In der Abbildung 63 geben wir einen Fluss mit dem Wert 5 an. Gibt es auch einen mit einem größeren Wert? □

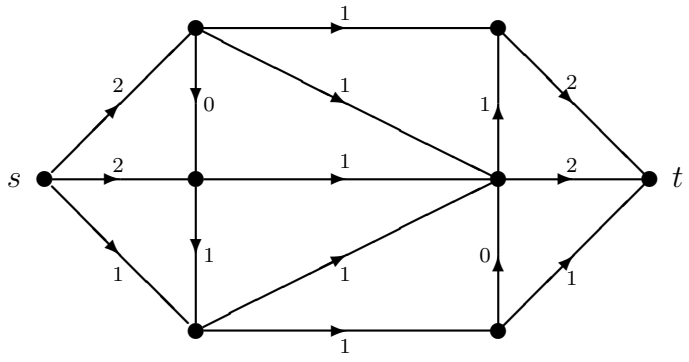


Abbildung 63: Ein Fluss mit dem Wert 5

Mit $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ bezeichnen wir einen *Fluss* auf den Pfeilen $\mathcal{A}$. Das Maximaler-Fluss-Problem lässt sich dann formulieren als

$$\text{Maximiere} \sum_{j:(s,j)\in\mathcal{A}} x_{sj} \quad \text{unter den Nebenbedingungen}$$

$$\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = 0 \qquad (k \in \mathcal{N} \setminus \{s,t\}), \qquad 0 \le x_{ij} \le u_{ij} \qquad ((i,j) \in \mathcal{A}).$$

Mit $A \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{A}|}$ sei wie im letzten Abschnitt 47 die Knoten-Pfeil-Inzidenzmatrix zu $(\mathcal{N}, \mathcal{A})$ bezeichnet. Wir denken uns dabei die Knoten und die Pfeile in einer bestimmten Weise durchnummeriert. Für $k \in \mathcal{N}$ und $(i,j) \in \mathcal{A}$ ist dann

$$a_{k,(i,j)} := \begin{cases} 1, & \text{falls } k = i, \\ -1, & \text{falls } k = j, \\ 0, & \text{sonst.} \end{cases}$$

Dann ist

$$(Ax)_s = \sum_{j:(s,j)\in\mathcal{A}} x_{sj}$$

der an der Quelle austretende Fluss, da es keinen in s endenden Pfeil gibt. In jedem Knoten $k \in \mathcal{N} \setminus \{s,t\}$ gilt die Flusserhaltungsgleichung

$$(Ax)_k = \sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = 0,$$

ferner ist

$$\sum_{k\in\mathcal{N}} (Ax)_k = \sum_{k\in\mathcal{N}} \sum_{(i,j)\in\mathcal{A}} a_{k,(i,j)} x_{ij} = \sum_{(i,j)\in\mathcal{A}} \Big(\underbrace{\sum_{k\in\mathcal{N}} a_{k,(i,j)}}_{=0} \Big) x_{ij} = 0,$$

da jeder Pfeil in genau einem Knoten startet und in genau einem Knoten endet bzw. in jeder Spalte von A genau eine 1 und genau eine -1 steht.. Daher ist

$$(Ax)_t = -(Ax)_s,$$

womit lediglich ausgedrückt wird, dass alles in der Senke ankommt, was an der Quelle startet, da zwischendurch nichts versickern kann. Definiert man den Vektor $d \in \mathbb{R}^{|\mathcal{N}|}$ durch

$$d_k := \begin{cases} -1 & \text{für} \quad k = s, \\ 0 & \text{für} \quad k \in \mathcal{N} \setminus \{s,t\}, \\ 1 & \text{für} \quad k = t, \end{cases}$$

so erkennt man, dass das Maximaler-Fluss-Problem als lineare Optimierungsaufgabe

$$\text{(D)} \qquad \text{Maximiere} \quad v \quad \text{auf} \quad \{(x,v) \in \mathbb{R}^{|\mathcal{A}|} \times \mathbb{R} : Ax + dv = 0,\ 0 \le x \le u\}$$

formuliert werden kann. Wir haben dieses Problem mit (D) bezeichnet, da es das **d**uale Problem zu

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad u^T w \quad \text{auf} \\ M := \{(y,w) \in \mathbb{R}^{|\mathcal{N}|} \times \mathbb{R}^{|\mathcal{A}|} : w \ge 0,\ w + A^T y \ge 0,\ d^T y = 1\} \end{cases}$$

ist, siehe die Bemerkung im Anschluss an den Dualitätssatz in Abschnitt 46. Äquivalente komponentenweise Formulierungen sind

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \sum_{(i,j)\in\mathcal{A}} u_{ij} w_{ij} \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad\qquad\qquad w_{ij} \ge 0, \qquad (i,j) \in \mathcal{A}, \\ y_i - y_j + w_{ij} \ge 0, \qquad (i,j) \in \mathcal{A}, \\ -y_s + y_t = 1 \end{cases}$$

und

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \sum_{j:(s,j)\in\mathcal{A}} x_{sj} \quad \text{unter den Nebenbedingungen} \\ \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = \sum_{j:(k,j)\in\mathcal{A}} x_{kj}, \qquad k \in \mathcal{N} \setminus \{s,t\}, \\ 0 \le x_{ij} \le u_{ij}, \qquad (i,j) \in \mathcal{A}. \end{cases}$$

Ist in dieser Notation $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ ein zulässiger Fluss, so ist sein *Wert* durch $v := \sum_{j:(s,j)\in\mathcal{A}} x_{sj}$ gegeben.

Bisher haben wir das Maximaler-Fluss-Problem (max-flow problem) geschildert und gezeigt, dass es sich in natürlicher Weise als eine lineare Optimierungsaufgabe formulieren lässt. Diese besitzt eine Lösung, denn die zulässige Menge in (D) ist nichtleer (der triviale Fluss $x = 0$ genügt allen Nebenbedingungen) und kompakt (jedenfalls dann, wenn alle Kapazitätsschranken endlich sind).

Nun kommen wir zum Minimaler-Schnitt-Problem (min-cut problem). Wieder ist der gerichtete Graph $(\mathcal{N}, \mathcal{A})$ mit den zwei ausgezeichneten Knoten s (Quelle) und t (Senke) sowie (i. allg. positive) Kapazitäten u_{ij} längs der Pfeile $(i,j) \in \mathcal{A}$ gegeben. Ein *Schnitt* ist eine Partition der Knotenmenge $\mathcal{N}$ in zwei (disjunkte) Mengen $\mathcal{N}_1$ und $\mathcal{N}_2$ mit $s \in \mathcal{N}_1$ und $t \in \mathcal{N}_2$. Zu einem Schnitt $(\mathcal{N}_1, \mathcal{N}_2)$ definieren wir die zugehörige *Kapazität* $C(\mathcal{N}_1, \mathcal{N}_2)$ als die Summe aller Kapazitätsschranken über Pfeilen, die in $\mathcal{N}_1$ starten und in $\mathcal{N}_2$ enden, also in der oben eingeführten Notation durch

$$C(\mathcal{N}_1, \mathcal{N}_2) := \sum_{\substack{(i,j)\in\mathcal{A} \\ i\in\mathcal{N}_1,\, j\in\mathcal{N}_2}} u_{ij}.$$

Unter dem *Minimaler-Schnitt-Problem* (min-cut problem) versteht man die Aufgabe, einen Schnitt mit minimaler Kapazität zu bestimmen.

In Abbildung 64 geben wir einen Schnitt in dem gerichteten Graphen aus Abbildung

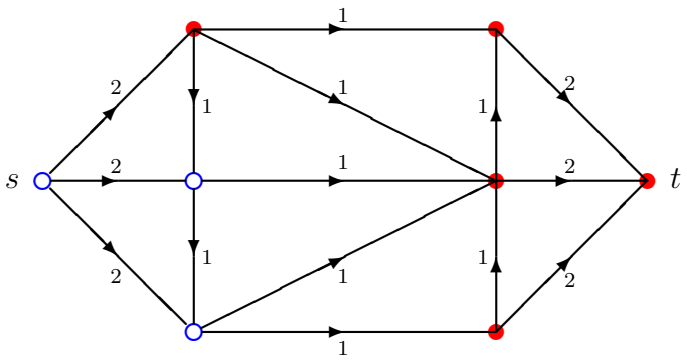


Abbildung 64: Ein Schnitt mit der Kapazität 5

62 an. Die zu $\mathcal{N}_1$ gehörenden Knoten sind durch ∘, solche zu $\mathcal{N}_2$ durch • gekennzeichnet. Hier gibt es vier Pfeile, die Knoten aus $\mathcal{N}_1$ mit Knoten aus $\mathcal{N}_2$ verbinden, die zugehörige Kapazität ist 5.

Unser Ziel ist es, das Max-Flow Min-Cut Theorem von Ford-Fulkerson zu beweisen. Dieses sagt unter Benutzung obiger Bezeichnungen aus:

Max Flow-Min Cut Theorem *Sei $(\mathcal{N}, \mathcal{A})$ ein gerichteter Graph mit einer Quelle $s \in \mathcal{N}$ und einer Senke $t \in \mathcal{N}$. Gegeben seien nichtnegative Kapazitäten $u = (u_{ij})_{(i,j)\in\mathcal{A}}$ auf den Pfeilen. Dann gilt:*

(a) Ist $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ ein zulässiger Fluss mit dem Wert $v = \sum_{j:(s,j)\in\mathcal{A}} x_{sj}$ und ist $(\mathcal{N}_1, \mathcal{N}_2)$ ein Schnitt mit Kapazität $C(\mathcal{N}_1, \mathcal{N}_2)$, so ist $v \leq C(\mathcal{N}_1, \mathcal{N}_2)$.

(b) Ist x^* eine Lösung des Maximaler-Fluss-Problems mit dem Wert

$$v^* = \sum_{j:(s,j)\in\mathcal{A}} x^*_{sj},$$

so existiert ein Schnitt $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ mit der Kapazität $C(\mathcal{N}_1^*, \mathcal{N}_2^*) = v^*$. Dieser Schnitt ist eine Lösung des Minimaler-Schnitt-Problems.

(c) Ist $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ eine Lösung des Minimaler-Schnitt-Problems (da es nur endlich viele Schnitte gibt, und mindestens einen, wenn es eine die Quelle und die Senke verbindende Pfeilfolge gibt, hat das Minimaler-Schnitt-Problem trivialerweise eine Lösung), so gibt es einen zulässigen Fluss x^* mit dem Wert $v^* = C(\mathcal{N}_1^*, \mathcal{N}_2^*)$. Dieser Fluss ist eine Lösung des Maximaler-Fluss-Problems.

(d) Ein zulässiger Fluss x^* und ein Schnitt $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ sind genau dann optimal für das Maximaler-Fluss- bzw. das Minimaler-Schnitt-Problem, wenn

$$x^*_{ij} = \begin{cases} 0 & \text{falls } (i,j) \in \mathcal{A} \text{ mit } i \in \mathcal{N}_2^* \text{ und } j \in \mathcal{N}_1^*, \\ u_{ij} & \text{falls } (i,j) \in \mathcal{A} \text{ mit } i \in \mathcal{N}_1^* \text{ und } j \in \mathcal{N}_2^*. \end{cases}$$

Beweis: Zum Beweis von (a) definieren wir zu dem Schnitt $(\mathcal{N}_1, \mathcal{N}_2)$ ein Paar $(y, w) \in \mathbb{R}^{|\mathcal{N}|} \times \mathbb{R}^{|\mathcal{A}|}$ durch

$$y_k := \begin{cases} 0 & \text{für alle } k \in \mathcal{N}_1, \\ 1 & \text{für alle } k \in \mathcal{N}_2. \end{cases}$$

$$w_{ij} := \begin{cases} 1 & \text{für alle } (i,j) \in \mathcal{A} \text{ mit } i \in \mathcal{N}_1 \text{ und } j \in \mathcal{N}_2, \\ 0 & \text{für alle anderen } (i,j) \in \mathcal{A}, \end{cases}$$

Offenbar ist (y, w) für die obige lineare Optimierungsaufgabe (P) zulässig und

$$C(\mathcal{N}_1, \mathcal{N}_2) = \sum_{(i,j)\in\mathcal{A}} u_{ij} w_{ij}.$$

Der schwache Dualitätssatz (siehe die erste Aussage im Dualitätssatz in Abschnitt 46) der linearen Optimierung liefert[58] (a).

[58] Natürlich kann man den Beweis auch direkt, ohne den Umweg über den auf (P) und (D) angewandten schwachen Dualitätssatz führen. Sei hierzu x ein zulässiger Fluss und $(\mathcal{N}_1, \mathcal{N}_2)$ ein Schnitt.

Zum Beweis von (b) nehmen wir an, x^* sei Lösung des Maximaler-Fluss-Problems mit zugehörigem Wert v^*. Dann ist (x^*, v^*) auch eine Lösung der linearen Optimierungsaufgabe (D). Wegen des (starken) Dualitätssatzes der linearen Optimierung besitzt die lineare Optimierungsaufgabe (P) eine Lösung (y^*, w^*) mit

$$\sum_{(i,j)\in\mathcal{A}} u_{ij} w^*_{ij} = v^*.$$

Wir konstruieren einen Schnitt $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ mit

$$v^* = C(\mathcal{N}_1^*, \mathcal{N}_2^*),$$

womit wir dann auch (b) bewiesen haben, da ein solcher Schnitt wegen (a) notwendigerweise eine Lösung des Minimaler-Schnitt-Problems ist. O. B. d. A. können wir annehmen, dass $y_s^* = 0$ (und dann $y_t^* = 1$), da man notfalls y^* durch $y^* - y_s^* e$ ersetzen kann, ohne etwas an der Optimalität von (y^*, w^*) zu ändern. Wir definieren

$$\mathcal{N}_1^* := \{k \in \mathcal{N} : y_k^* < 1\}, \qquad \mathcal{N}_2^* := \{k \in \mathcal{N} : y_k^* \geq 1\},$$

wodurch offenbar ein Schnitt im gerichteten Graphen $(\mathcal{N}, \mathcal{A})$, gegeben ist. Zunächst beachten wir, dass die Gleichgewichtsbedingungen zwischen den beiden zueinander dualen linearen Optimierungsaufgaben die Beziehungen

$$w^*_{ij}[u_{ij} - x^*_{ij}] = 0, \qquad x^*_{ij}[y^*_i - y^*_j + w^*_{ij}] = 0, \qquad ((i,j) \in \mathcal{A})$$

Dann gilt:

$$\begin{aligned}
v &= \sum_{j:(s,j)\in\mathcal{A}} x_{sj} - \underbrace{\sum_{i:(i,s)\in\mathcal{A}} x_{is}}_{=0} + \sum_{k\in\mathcal{N}_1\setminus\{s\}} \Bigg(\underbrace{\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik}}_{=0} \Bigg) \\
&\qquad \text{(Definition von } v \text{ und Aufsummieren der Flussgleichung für alle } k \in \mathcal{N}_1 \setminus \{s\}.) \\
&= \sum_{k\in\mathcal{N}_1} \Bigg(\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} \Bigg) \\
&= \sum_{k\in\mathcal{N}_1} \Bigg(\sum_{\substack{j\in\mathcal{N}_2 \\ (k,j)\in\mathcal{A}}} x_{kj} - \sum_{\substack{i\in\mathcal{N}_2 \\ (i,k)\in\mathcal{A}}} x_{ik} \Bigg) + \underbrace{\sum_{k\in\mathcal{N}_1} \Bigg(\sum_{\substack{j\in\mathcal{N}_1 \\ (k,j)\in\mathcal{A}}} x_{kj} - \sum_{\substack{i\in\mathcal{N}_1 \\ (i,k)\in\mathcal{A}}} x_{ik} \Bigg)}_{=0} \\
&\leq \sum_{k\in\mathcal{N}_1} \sum_{\substack{j\in\mathcal{N}_2 \\ (k,j)\in\mathcal{A}}} x_{kj} \\
&= \sum_{i\in\mathcal{N}_1} \sum_{\substack{j\in\mathcal{N}_2 \\ (i,j)\in\mathcal{A}}} x_{ij} \\
&\leq \sum_{\substack{(i,j)\in\mathcal{A} \\ i\in\mathcal{N}_1,\, j\in\mathcal{N}_2}} u_{ij} \\
&= C(\mathcal{N}_1, \mathcal{N}_2).
\end{aligned}$$

Diese Argumentation wird auch beim Beweis von (b) und in (d) eine Rolle spielen.

liefern. Um diese Beziehungen, die ja für alle Pfeile aus $\mathcal{A}$ gelten, auszunutzen, unterscheiden wir zwischen "Vorwärts-Pfeilen" und "Rückwärts-Pfeilen" bezüglich des Schnittes $(\mathcal{N}_1^*, \mathcal{N}_2^*)$. Hierbei nennen wir ein $(i,j) \in \mathcal{A}$ einen *Vorwärts-Pfeil* (bezüglich $(\mathcal{N}_1^*, \mathcal{N}_2^*)$), wenn $i \in \mathcal{N}_1^*$ und $j \in \mathcal{N}_2^*$. Für einen solchen ist

$$w_{ij}^* \geq \underbrace{y_j^*}_{\geq 1} - \underbrace{y_i^*}_{<1} > 0$$

und daher $x_{ij}^* = u_{ij}$. Ist dagegen $(i,j) \in \mathcal{A}$ ein "Rückwärts-Pfeil", also $i \in \mathcal{N}_2^*$ und $j \in \mathcal{N}_1^*$, so ist

$$y_i^* - y_j^* + \underbrace{w_{ij}^*}_{\geq 0} \geq \underbrace{y_i^*}_{\geq 1} - \underbrace{y_j^*}_{<1} > 0,$$

und folglich $x_{ij}^* = 0$. Der Trick besteht jetzt darin, den maximalen Fluss als Summe der Flüsse auf den Vorwärts-Pfeilen bzw. (wegen $x_{ij}^* = u_{ij}$ auf den Vorwärts-Pfeilen) als Kapazität des Schnittes $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ zu "entlarven". Es ist nämlich

$$\begin{aligned}
v^* &= \sum_{j:(s,j)\in\mathcal{A}} x_{sj}^* - \underbrace{\sum_{i:(i,s)\in\mathcal{A}} x_{is}^*}_{=0} \\
&\qquad \text{(In der Quelle } s \text{ endet kein Pfeil)} \\
&= \sum_{j:(s,j)\in\mathcal{A}} x_{sj}^* - \sum_{i:(i,s)\in\mathcal{A}} x_{is}^* + \sum_{k\in\mathcal{N}_1^*\setminus\{s\}} \Big(\underbrace{\sum_{j:(k,j)\in\mathcal{A}} x_{kj}^* - \sum_{i:(i,k)\in\mathcal{A}} x_{ik}^*}_{=0} \Big) \\
&\qquad \text{(Flussbedingung)} \\
&= \sum_{k\in\mathcal{N}_1^*} \Big(\sum_{j:(k,j)\in\mathcal{A}} x_{kj}^* - \sum_{i:(i,k)\in\mathcal{A}} x_{ik}^* \Big) \\
&= \sum_{k\in\mathcal{N}_1^*} \Big(\sum_{\substack{j\in\mathcal{N}_2^*: \\ (k,j)\in\mathcal{A}}} \underbrace{x_{kj}^*}_{=u_{kj}} - \sum_{\substack{i\in\mathcal{N}_2^*: \\ (i,k)\in\mathcal{A}}} \underbrace{x_{ik}^*}_{=0} \Big) + \underbrace{\sum_{k\in\mathcal{N}_1^*} \Big(\sum_{\substack{j\in\mathcal{N}_1^*: \\ (k,j)\in\mathcal{A}}} x_{kj}^* - \sum_{\substack{i\in\mathcal{N}_1^*: \\ (i,k)\in\mathcal{A}}} x_{ik}^* \Big)}_{=0} \\
&= \sum_{i\in\mathcal{N}_1^*} \sum_{\substack{j\in\mathcal{N}_2^*: \\ (i,j)\in\mathcal{A}}} u_{ij} \\
&= C(\mathcal{N}_1^*, \mathcal{N}_2^*),
\end{aligned}$$

womit schließlich auch (b) bewiesen ist.

In (c) wird angenommen, $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ sei eine Lösung des Minimaler-Schnitt-Problems. Sei x^* eine Lösung des Maximaler-Fluss-Problems mit dem Fluss v^*. Nach (b) existiert zu x^* ein Schnitt $(\mathcal{N}_1^{**}, \mathcal{N}_2^{**})$ mit $v^* = C(\mathcal{N}_1^{**}, \mathcal{N}_2^{**})$. Wegen (a) ist

$$C(\mathcal{N}_1^{**}, \mathcal{N}_2^{**}) = v^* \leq C(\mathcal{N}_1^*, \mathcal{N}_2^*).$$

Da weiter $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ eine Lösung des Minimalschnittproblems ist, ist $v^* = C(\mathcal{N}_1^*, \mathcal{N}_2^*)$, also mit x^* der gesuchte Fluss gefunden.

In (d) haben wir nur noch einmal das zusammengefasst, was wir wegen der Gleichgewichtsbedingung schon wissen. Ist nämlich x^* ein Fluss, v^* sein Wert und $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ ein Schnitt mit

$$x_{ij}^* = \begin{cases} 0 & \text{falls } (i,j) \in \mathcal{A} \text{ mit } i \in \mathcal{N}_2^* \text{ und } j \in \mathcal{N}_1^*, \\ u_{ij} & \text{falls } (i,j) \in \mathcal{A} \text{ mit } i \in \mathcal{N}_1^* \text{ und } j \in \mathcal{N}_2^*, \end{cases}$$

so ist (siehe obige Rechnung)

$$v^* = \sum_{k \in \mathcal{N}_1^*} \Big(\sum_{\substack{j \in \mathcal{N}_2^*: \\ (k,j) \in \mathcal{A}}} \underbrace{x_{kj}^*}_{=u_{kj}} - \sum_{\substack{i \in \mathcal{N}_2^*: \\ (i,k) \in \mathcal{A}}} \underbrace{x_{ik}^*}_{=0} \Big) = C(\mathcal{N}_1^*, \mathcal{N}_2^*),$$

also x^* und $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ optimal für das Maximaler-Fluss- bzw. das Minimaler-Schnitt-Problem. Sind umgekehrt x^* und $(\mathcal{N}_1^*, \mathcal{N}_2^*)$ optimal, so ist $v^* = C(\mathcal{N}_1^*, \mathcal{N}_2^*)$ und die Behauptung folgt (siehe einmal wieder obige Rechnung) aus

$$v^* = \sum_{i \in \mathcal{N}_1^*} \Big(\sum_{\substack{j \in \mathcal{N}_2^*: \\ (i,j) \in \mathcal{A}}} x_{ij}^* - \sum_{\substack{j \in \mathcal{N}_2^*: \\ (j,i) \in \mathcal{A}}} x_{ji}^* \Big) = \sum_{i \in \mathcal{N}_1^*} \sum_{\substack{j \in \mathcal{N}_2^*: \\ (i,j) \in \mathcal{A}}} u_{ij} = C(\mathcal{N}_1^*, \mathcal{N}_2^*),$$

dass

$$0 = \sum_{i \in \mathcal{N}_1^*} \sum_{\substack{j \in \mathcal{N}_2^*: \\ (i,j) \in \mathcal{A}}} \underbrace{(u_{ij} - x_{ij}^*)}_{\geq 0} + \sum_{\substack{j \in \mathcal{N}_2^*: \\ (j,i) \in \mathcal{A}}} \underbrace{x_{ji}^*}_{\geq 0}$$

und daher $x_{ij}^* = u_{ij}$ für alle $(i,j) \in \mathcal{A}$ mit $i \in \mathcal{N}_1^*$, $j \in \mathcal{N}_2^*$ und $x_{ij}^* = 0$ für alle $(i,j) \in \mathcal{A}$ mit $i \in \mathcal{N}_2^*$, $j \in \mathcal{N}_1^*$. Damit ist der Satz schließlich bewiesen. □

49 Trennungssätze für konvexe Mengen im $\mathbb{R}^n$

Eine *Hyperebene im* $\mathbb{R}^n$ ist ein affin linearer (also verschobener linearer) $(n-1)$-dimensionaler Teilraum des $\mathbb{R}^n$ und lässt sich daher mit $(y, \gamma) \in (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$ in der Form

$$H := \{x \in \mathbb{R}^n : y^T x = \gamma\}$$

darstellen. Für $n = 2$, also die Ebene, stimmen Hyperebenen und Geraden überein. In Abbildung 65 wird dies mit einem $\gamma > 0$ veranschaulicht. Außerdem haben wir noch eine parallele Hyperebene durch den Nullpunkt eingezeichnet. Wir nennen zwei nichtleere Mengen $A, B \subset \mathbb{R}^n$ *trennbar* bzw. *stark trennbar*, wenn $y \in \mathbb{R}^n \setminus \{0\}$ mit

$$\sup_{a \in A} y^T a \leq \inf_{b \in B} y^T b \qquad \text{bzw.} \qquad \sup_{a \in A} y^T a < \inf_{b \in B} y^T b$$

existiert. Eine Hyperebene $H = \{x \in \mathbb{R}^n : y^T x = \gamma\}$ induziert zwei (abgeschlossene) Halbräume, nämlich

$$H^- := \{x \in \mathbb{R}^n : y^T x \leq \gamma\}, \qquad H^+ := \{x \in \mathbb{R}^n : y^T x \geq \gamma\}.$$

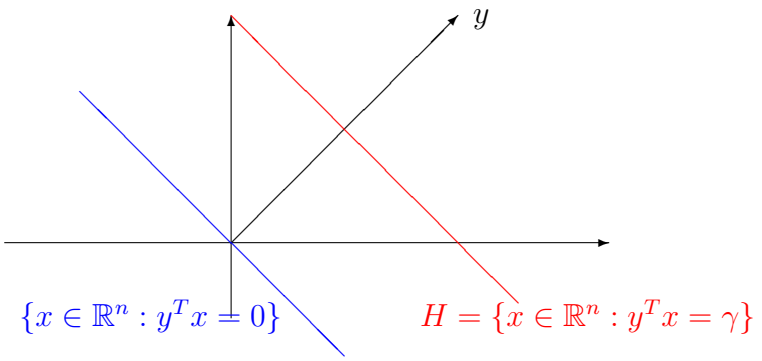


Abbildung 65: Hyperebene

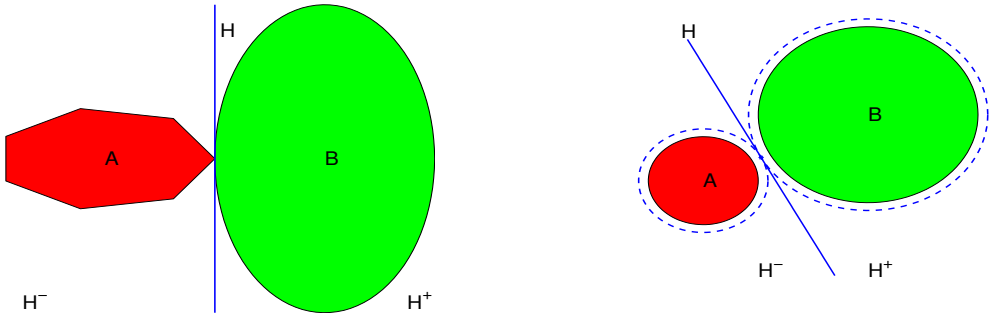


Abbildung 66: Trennbare und stark trennbare Mengen

Dann sind A und B genau dann trennbar, wenn eine Hyperebene H mit $A \subset H^-$ und $B \subset H^+$ existiert. Dies veranschaulichen wir uns in Abbildung 66 links. Entsprechend sind A und B genau dann stark trennbar, wenn ein $\epsilon > 0$ mit $A + B[0;\epsilon] \subset H^-$ und $B + B[0;\epsilon] \subset H^+$ existiert. Hierbei bedeutet $B[0;\epsilon] := \{x \in \mathbb{R}^n : \|x\|_2 \leq \epsilon\}$ die abgeschlossene (euklidische) ϵ-Kugel um den Nullpunkt. In Abbildung 66 rechts wird dies veranschaulicht.

Zwei disjunkte Mengen sind natürlich nicht notwendig trennbar (Beispiel?). Sind aber beide konvex[59], so zeigt der folgende *Trennungssatz*, dass dies möglich ist. Den Beweis haben wir O. L. Mangasarian (1969, S. 47 ff.) entnommen.

Trennungssatz für konvexe Mengen im $\mathbb{R}^n$ *Seien* $A, B \subset \mathbb{R}^n$ *nichtleer, konvex und disjunkt. Dann sind* A *und* B *trennbar.*

Beweis: Die Menge $B-A$ ist naheliegenderweise durch $B-A := \{b-a : b \in B,\ a \in A\}$ definiert. Es ist $0 \notin C := B - A$, da A und B disjunkt sind, ferner ist C konvex[60]. Wir

[59] Eine Teilmenge A eines linearen Raumes heißt bekanntlich *konvex*, wenn mit zwei Punkten aus A die gesamte Verbindungsstrecke zu A gehört, wenn also aus $x, y \in A$ und $t \in [0,1]$ folgt, dass $(1-t)x + ty \in A$.

[60] Denn sind $b_1 - a_1,\ b_2 - a_2 \in C$ und $t \in [0,1]$, so ist

$$(1-t)(b_1-a_1) + t(b_2-a_2) = \underbrace{(1-t)b_1 + tb_2}_{\in B} - \underbrace{(1-t)a_1 + ta_2}_{\in A} \in B - A = C.$$

zeigen die Existenz eines $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T x \geq 0$ für alle $x \in C$, woraus offenbar die Behauptung folgt.

Für $x \in C$ definieren wir

$$\Lambda_x := \{y \in \mathbb{R}^n : \|y\| = 1,\ y^T x \geq 0\},$$

eine nichtleere, abgeschlossene Teilmenge der kompakten Einheitssphäre. Wir wollen zeigen, dass $\bigcap_{x \in C} \Lambda_x \neq \emptyset$, denn ein Element aus diesem Durchschnitt ist der gesuchte Vektor y. Wegen der Kompaktheit der Einheitssphäre (sogenannte finite intersection property kompakter Mengen) genügt es zu zeigen: Sind $x_1, \ldots, x_m \in C$, so ist $\bigcap_{i=1}^m \Lambda_{x_i} \neq \emptyset$. Dies sieht man wiederum folgendermaßen ein. Angenommen, es wäre $\bigcap_{i=1}^m \Lambda_{x_i} = \emptyset$. Dann hätte das Ungleichungssystem $y^T x_i \geq 0$, $i = 1, \ldots, m$, keine nichttriviale Lösung. Mit $X := (\ x_1 \ \cdots \ x_m\ ) \in \mathbb{R}^{n \times m}$ und $e := (1, \ldots, 1)^T \in \mathbb{R}^m$ bedeutet dies, dass das Ungleichungssystem

$$X^T y \geq 0, \qquad (-Xe)^T y < 0$$

nicht lösbar ist. Das Farkas-Lemma (siehe Abschnitt 45) liefert die Existenz eines nichtnegativen Vektors $\lambda \in \mathbb{R}^m$ mit $X\lambda = -Xe$ bzw. $X(\lambda + e) = 0$. Also ist der Nullpunkt eine positive Linearkombination und dann auch ein Konvexkombination der Punkte $x_1, \ldots, x_m \in C$. Aus der Konvexität von C folgt $0 \in C$, was ein Widerspruch ist. □

Es folgt

Starker Trennungssatz für konvexe Mengen im $\mathbb{R}^n$ *Seien $A, B \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Sind A und B disjunkt und eine der beiden Mengen kompakt, so sind A und B stark trennbar.*

Beweis: Sei $M := B - A$, wobei die Differenz der beiden Mengen B und A natürlich durch

$$B - A := \{b - a : a \in A,\, b \in B\}$$

definiert ist. Dann ist M nichtleer, konvex und abgeschlossen (Beweis?), ferner $0 \notin M$, da $A \cap B = \emptyset$. Sei $x^* = P_M(0) \in M$ die wegen des Projektionssatzes für konvexe Mengen (siehe Abschnitt 37) existierende Projektion von $z := 0$ auf M. Insbesondere ist $(x^*)^T x \geq \|x^*\|^2$ für alle $x \in M$ und $x^* \neq 0$. Definiert man daher $y := x^*$, so ist

$$y^T(b - a) \geq \|x^*\|^2 \qquad \text{für alle } a \in A,\, b \in B$$

und daher

$$\sup_{a \in A} y^T a < \sup_{a \in A} y^T a + \|x^*\|^2 \leq \inf_{b \in B} y^T b.$$

Also sind A und B stark trennbar.

Das folgende Korollar ist eine unmittelbare Folgerung aus dem starken Trennungssatz.

Korollar *Sei $K \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Dann kann jedes $z \notin K$ von K stark getrennt werden, d. h. zu jedem $z \notin K$ existiert ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T z < \inf_{x \in K} y^T x$.*

50 Die Lagrangesche Multiplikatorenregel

Dieser Abschnitt wird mathematisch etwas anspruchsvoller sein, da wir insbesondere einige Begriffe und Grundlagen der (linearen) Funktionalanalysis als bekannt voraussetzen werden. Unser Ziel besteht darin, die *Lagrangesche Multiplikatorenregel*, eine notwendige Optimalitätsbedingung bei restringierten Optimierungsaufgaben, wenigstens für einen Spezialfall, nämlich einer Optimierungsaufgabe mit endlich vielen Gleichungen als Nebenbedingungen, zu formulieren und zu beweisen. Hierbei werden wir z. B. ohne eine explizite Anwendung des Satzes über implizite Funktionen auskommen. Im darauffolgenden Abschnitt über das Problem der Dido werden diese Aussagen angewandt.

Lagrangesche Multiplikatorenregel *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in X : x \in U,\ g(x) = 0\}.$$

Hierbei sei $(X, \|\cdot\|)$ ein (reeller) linearer, normierter Raum, $U \subset X$ eine offene Teilmenge, $f\colon X \longrightarrow \mathbb{R}$ (Zielfunktion) und $g\colon X \longrightarrow \mathbb{R}^m$ (Restriktionsabbildung). Sei $x^* \in M$ eine lokale Lösung[61] *von (P). Die Zielfunktion f sei in x^* Fréchet-differenzierbar, besitze also ein Fréchet-Differential $f'(x^*)$, während die Restriktionsabbildung g als in x^* stetig Fréchet-differenzierbar*[62] *vorausgesetzt wird. Die Abbildung $g'(x^*)\colon X \longrightarrow \mathbb{R}^m$ sei surjektiv, d. h. es sei $g'(x^*)(X) = \mathbb{R}^m$. Dann existiert ein Vektor $y^* \in \mathbb{R}^m$, der sogenannte Lagrange-Multiplikator, mit*

$$f'(x^*)(h) + (y^*)^T g'(x^*)(h) = 0 \quad \textit{für alle } h \in X.$$

Beweis: Man sollte sich nicht wundern, dass die offene Menge U im Beweis nicht auftaucht. Da alle Argumente *lokaler* Art sind bzw. sich alles in einer Umgebung von x^* abspielt, spielt die Nebenbedingung $x \in U$ keine Rolle.

[61] D. h. es existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^* \cap M$.

[62] Die Abbildung $g\colon X \longrightarrow \mathbb{R}^m$ (entsprechendes gilt für die Abbildung $f\colon X \longrightarrow \mathbb{R}$, also den Spezialfall $m = 1$) heißt in $x \in X$ *Fréchet-differenzierbar*, wenn eine lineare, stetige Abbildung $g'(x)\colon X \longrightarrow \mathbb{R}^m$ existiert, das sogenannte *Fréchet-Differential*, mit

$$\lim_{h \to 0} \frac{\|g(x+h) - g(x) - g'(x)(h)\|}{\|h\|} = 0$$

bzw. es zu jedem $\epsilon > 0$ ein $\delta > 0$ mit

$$\|h\| \leq \delta \Longrightarrow \|g(x+h) - g(x) - g'(x)(h)\| \leq \epsilon \|h\|$$

gibt. Das Fréchet-Differential ist, wenn es existiert, eindeutig bestimmt, wie man sich leicht überlegt. Die Abbildung $g\colon X \longrightarrow \mathbb{R}^m$ heißt in x^* *stetig Fréchet-differenzierbar*, wenn es eine Umgebung U^* von x^* gibt derart, dass g für jedes $x \in U^*$ Fréchet-differenzierbar ist und die Abbildung $g'\colon X \longrightarrow L(X, \mathbb{R}^m)$ in x^* stetig ist. Hierbei bezeichne $L(X, \mathbb{R}^m)$ die Menge der linearen stetigen Abbildungen von X nach $\mathbb{R}^m$. Diese Menge ist in kanonischer Weise selbst ein linearer normierter Raum, wobei die Norm eines $T \in L(X, \mathbb{R}^m)$ durch $\|T\| := \sup_{h \neq 0} \|T(h)\|/\|h\|$ definiert ist. Man beachte, dass wir es mit drei normierten Räumen zu tun haben: Mit dem Ausgangsraum X, dem Raum $\mathbb{R}^m$, und mit $L(X, \mathbb{R}^m)$. Die jeweilige Norm wird stets mit $\|\cdot\|$ bezeichnet.

Der sogenannte *Tangentialkegel* (bzw. der Kegel[63] der tangentialen Richtungen) $T(M;x^*)$ an M in x^* wird definiert durch

$$T(M;x^*) := \left\{ h \in X : \begin{array}{c} \text{Es existieren Folgen } \{t_k\} \subset \mathbb{R}_+,\ \{r_k\} \subset X \text{ mit} \\ x^* + t_k h + r_k \in M \text{ für alle } k,\ t_k \to 0,\ r_k/t_k \to 0 \end{array} \right\}.$$

Zunächst zeigen wir:

(1) Es ist $\{h \in X : g'(x^*)(h) = 0\} \subset T(M;x^*)$.

Denn: Der $\mathbb{R}^m$ wird z. B. von den m Einheitsvektoren $e_1, \ldots, e_m \in \mathbb{R}^m$ aufgespannt. Da $g'(x^*) : X \longrightarrow \mathbb{R}^m$ nach Voraussetzung surjektiv ist, existieren $\{x_1, \ldots, x_m\} \subset X$ mit $g'(x^*)(x_i) = e_i$, $i = 1, \ldots, m$. Man definiere den linearen Teilraum

$$X_m := \operatorname{span}\{x_1, \ldots, x_m\} = \Big\{ \sum_{i=1}^m \xi_i x_i : (\xi_1, \ldots, \xi_m)^T \in \mathbb{R}^m \Big\}$$

von X. Offenbar sind $\{x_1, \ldots, x_m\}$ linear unabhängig bzw. X_m ein m-dimensionaler linearer Teilraum von X. Denn ist $\sum_{i=1}^m \xi_i x_i = 0$, so ist

$$0 = g'(x^*)\Big(\sum_{i=1}^m \xi_i x_i \Big) = \sum_{i=1}^m \xi_i \underbrace{g'(x^*)(x_i)}_{=e_i} = \sum_{i=1}^m \xi_i e_i,$$

woraus natürlich $\xi_1 = \cdots = \xi_m = 0$ und damit die lineare Unabhängigkeit von $\{x_1, \ldots, x_m\}$ folgt. Weiter zeigt diese Argumentation, dass $g'(x^*)$ eine umkehrbar eindeutige Abbildung von X_m auf den $\mathbb{R}^m$ ist. Nun definieren wir die positive Zahl

$$\rho := \min_{x \in X_m \cap B[0;1]} \|g'(x^*)(x)\|$$

und überlegen uns, dass

$$(*) \qquad B[0;\rho] \subset g'(x^*)(X_m \cap B[0;1]).$$

Hierbei ist

$$B[0;\rho] := \{y \in \mathbb{R}^m : \|y\| \le \rho\}$$

die abgeschlossene Kugel im $\mathbb{R}^m$ um den Nullpunkt mit dem Radius ρ, während

$$B[0;1] := \{x \in X : \|x\| \le 1\}$$

die abgeschlossene Einheitskugel in X ist. Der Einfachheit halber gehen wir davon aus, dass die Norm $\|\cdot\|$ auf dem $\mathbb{R}^m$ die *euklidische Norm* ist. Wegen der sogenannten Äquivalenz der Normen im $\mathbb{R}^m$ ist das keine Einschränkung. Zum Nachweis von $(*)$

[63] Unter einem *Kegel* versteht man eine Menge, die mit einem Punkt auch jedes nichtnegative Vielfache enthält. Durch Wikipedia erfährt man: Die Redewendung *Kind und Kegel* bedeutet eigentlich alle ehelichen und unehelichen Kinder. Heute steht der Begriff für die gesamte Verwandtschaft oder auch teilweise für Kinder, Haustiere und Gepäck. Wenn jemand mit Kind und Kegel reist, so ist der Ausdruck scherzhaft zu verstehen und derjenige hat die gesamte Familie dabei.

geben wir uns ein $y \in B[0;\rho]$ vor. Es existiert genau ein $x \in X_m$ mit $y = g'(x^*)(x)$, da $g'(x^*)$ eine umkehrbar eindeutige Abbildung von X_m auf den $\mathbb{R}^m$ ist. Angenommen, es wäre $\|x\| > 1$. Dann wäre

$$\|g'(x^*)(x/\|x\|)| \leq \frac{\rho}{\|x\|} < \rho,$$

ein Widerspruch zur Definition von ρ. Nun zeigen wir:

- Zu $h \in X$ existieren $t^* > 0$, $c_0 > 0$ und eine Abbildung $r\colon [0,t^*] \longrightarrow X_m$ mit

  (a) $\|r(t)\| \leq c_0 \|g(x^* + th) - g(x^*) - g'(x^*)(h)\|$,

  (b) $g(x^*) + g'(x^*)(h) = g(x^* + th + r(t))$

  für alle $t \in [0,t^*]$.

Ist dies bewiesen, so ist mit einer beliebigen Nullfolge $\{t_k\} \subset \mathbb{R}_+$ und $r_k := r(t_k)$ offensichtlich, dass ein h mit $g'(x^*)(h) = 0$ zum Tangentialkegel $T(M;x^*)$ gehört, der erste Beweisteil (1) also abgeschlossen ist.

Jetzt kommen wir zum Nachweis von • und geben uns ein beliebiges $h \in X$ vor, wobei wir $h \neq 0$ annehmen können (andernfalls wähle man t^* und c_0 beliebig positiv und setze $r(t) := 0$ für alle $t \in [0,t^*]$). Da g in x^* stetig differenzierbar ist, gibt es ein $\delta > 0$ mit

$$\|g'(x) - g'(x^*)\| \leq \frac{\rho}{2} \quad \text{für alle } x \in B[x^*;\delta] := \{x \in X : \|x - x^*\| \leq \delta\}.$$

Hieraus folgt, dass

$$(**) \qquad \|g(x) - g(x') - g'(x^*)(x - x')\| \leq \frac{\rho}{2}\|x - x'\| \quad \text{für alle } x, x' \in B[x^*;\delta].$$

Denn: Seien $x, x' \in B[x^*;\delta]$ vorgegeben. Es gibt [64] ein $z^* \in \mathbb{R}^m$ mit $\|z^*\| = 1$ und

$$(z^*)^T[g(x) - g(x') - g'(x^*)(x - x')] = \|g(x) - g(x') - g'(x^*)(x - x')\|.$$

Nun definiere man die Abbildung $\phi : [0,1] \longrightarrow \mathbb{R}$ durch

$$\phi(t) := (z^*)^T[g(x' + t(x - x')) - tg'(x^*)(x - x')].$$

Dann ist ϕ auf $[0,1]$ differenzierbar und

$$\phi'(t) = (z^*)^T[g'(x' + t(x - x'))(x - x') - g'(x^*)(x - x')].$$

[64] Zu jedem $y \in \mathbb{R}^m$ gibt es ein $z^* \in \mathbb{R}^m$ mit $\|z^*\| = 1$ und $(z^*)^T y = \|y\|$. Um dies einzusehen (o. B. d. A. ist $y \neq 0$) setze man einfach $z^* := y/\|y\|$. Hier nutzen wir aus, dass vereinbarungsgemäß $\|\cdot\|$ im $\mathbb{R}^m$ die euklidische Norm ist.

Der Mittelwertsatz der Differentialrechnung liefert die Existenz eines $t_0 \in (0,1)$ mit

$$\begin{aligned}
\|g(x) - g(x') - g'(x^*)(x-x')\| &= (z^*)^T[g(x) - g(x') - g'(x^*)(x-x^*)] \\
&= \phi(1) - \phi(0) \\
&= \phi'(t_0) \\
&= (z^*)^T[g'(x' + t_0(x-x'))(x-x') - g'(x^*)(x-x')] \\
&\leq \underbrace{\|z^*\|}_{=1} \|[g'(x' + t_0(x-x')) - g'(x^*)](x-x')\| \\
&\leq \|g'(\underbrace{x' + t_0(x-x')}_{\in B[x^*;\delta]}\| \, \|x-x'\| \\
&\leq \frac{\rho}{2}\|x-x'\|.
\end{aligned}$$

Damit ist $(**)$ bewiesen.

Nach diesen Vorbereitungen kommen wir schließlich zum Beweis von $\bullet$ und damit von (1). Wir definieren $t^* := \delta/(2\|h\|)$, wählen ein $t \in [0, t^*]$ fest und konstruieren eine Folge $\{r_k\} \subset X_m$ (nicht zu verwechseln mit $\{r_k\}$ in der Definition des Tangentialkegels), von der wir nachweisen werden, dass sie eine Cauchyfolge und damit konvergent gegen ein $r(t) \in X_m$ ist. Die so konstruierte Abbildung $t \mapsto r(t)$ wird sich als die gesuchte herausstellen. Und nun kommt die Vorschrift, nach der $\{r_k\}$ gebildet wird.

> Input: $h \in X \setminus \{0\}$, $t \in [0, t^*]$.
>
> Setze $r_0 := 0$.
>
> Für $k = 0, 1, \ldots$:
>
> > Berechne $y_k := g(x^*) + tg'(x^*)(h) - g(x^* + th + r_k)$.
> > Bestimme $u_k \in X_m$ mit $y_k = g'(x^*)(u_k)$ und setze $r_{k+1} := r_k + u_k$.
>
> Output: Folgen $\{r_k\} \subset X_m$, $\{u_k\} \subset X_m$ und $\{y_k\} \subset \mathbb{R}^m$.

Wir zeigen, dass die Folge $\{r_k\} \subset \mathbb{R}^m$ eine Cauchyfolge ist und ihr daher existierender Limes $r(t)$ den Bedingungen (a) und (b) in $\bullet$ genügt. Hierzu setzen wir zur Abkürzung

$$d(t) := \|g(x^* + th) - g(x^*) - tg'(x^*)(h)\|$$

und beachten, dass $d(t)$ wegen $(**)$ und $\|th\| \leq t^*\|h\| = \delta/2$ durch

$$d(t) = ||g(x^* + th) - g(x^*) - tg'(x^*)(h)\| \leq \frac{\rho}{2} t^* \|h\| \leq \frac{\rho}{4}\delta$$

abgeschätzt werden kann. Durch vollständige Induktion nach k zeigen wir nun, dass die folgenden Aussagen richtig sind:

(i) $\|r_k\| \leq (2/\rho)[1 - (1/2)^k]d(t)$.

(ii) $x^* + th + r_k \in B[x^*; \delta]$.

(iii) $\|y_k\| \leq (1/2)^k d(t)$.

(iv) $\|u_k\| \le (1/\rho)(1/2)^k d(t)$.

Ist uns dieser Induktionsbeweis gelungen, so ist offenbar auch $\bullet$ bewiesen. Denn $\{\|y_k\|\}$ und $\{\|u_k\|\}$ gehen so schnell wie $\{(1/2)^k\}$ gegen Null. Wegen $u_k = r_{k+1} - r_k$ und (iv) ist $\{r_k\}$ eine Cauchyfolge, deren Limes $r(t)$ der Beziehung $\|r(t)\| \le (1/\rho)d(t)$ genügt. Hieraus folgt offenbar $r(t)/t \to 0$ für $t \to 0+$. Nun zum Induktionsbeweis. Für $k = 0$ sind die Aussagen (i)–(iv) richtig. Und zwar ist (i) wegen $r_0 = 0$ richtig, (ii) wegen $\|th\| \le \delta/2$, (iii) wegen $\|y_0\| = d(t)$ und (iv) wegen $B[0;\rho] \subset g'(x^*)(X_m \cap B[0;1])$. Der Induktionsschluss von k nach $k+1$ ist nicht schwieriger. Wegen $r_{k+1} = r_k + u_k$, ferner (i) und (iv) für k ist

$$\begin{aligned} \|r_{k+1}\| &\le |r_k\| + \|u_k\| \\ &\le (2/\rho)[1 - (1/2)^k]d(t) + (1/\rho)(1/2)^k d(t) \\ &= (2/\rho)[1 - (1/2)^{k+1}]d(t) \end{aligned}$$

und das ist (i) für $k+1$. Wie haben oben $d(t) \le (\rho/4)\delta$ nachgewiesen. Mit Hilfe von (i) für $k+1$ erhalten wir

$$\|r_{k+1}\| \le \frac{2}{\rho}[1 - (1/2)^{k+1}]d(t) \le \frac{\delta}{2}.$$

Hieraus und aus $\|th\| \le \delta/2$ folgt $x^* + th + r_{k+1} \in B[x^*;\delta]$ und das ist (ii) für $k+1$. Weiter ist

$$\begin{aligned} \|y_{k+1}\| &= \|g(x^*) + tg'(x^*) - g(x^* + th + r_{k+1})\| \\ &= \|y_k + g(x^* + th + r_k) - g(x^* + th + r_{k+1}\| \\ &= \|g(x^* + th + r_{k+1}) - g(x^* + th + r_k) - g'(x^*)(u_k)\| \\ &\le \frac{\rho}{2}\|u_k\| \\ &\qquad \text{(wegen } (**) \text{ und } x^* + th + r_k,\ x^* + th + r_{k+1} \in B[x^*;\delta]) \\ &\le (1/2)^{k+1} d(t) \qquad \text{(wegen der Induktionsannahme (iv))} \end{aligned}$$

und das ist (iii) für $k+1$. Schließlich ist

$$\underbrace{\rho(y_{k+1}/\|y_{k+1}\|)}_{\in B[0;\rho]} = g'(x^*)(\rho u_{k+1}/\|y_{k+1}),$$

wegen $(*)$ ist daher

$$\|u_{k+1}\| \le (1/\rho)\|y_{k+1}\| \le (1/\rho)(1/2)^{k+1} d(t),$$

wobei wir die Induktionsannahme (iii) benutzt haben. Damit ist auch (iv) für $k+1$ nachgewiesen, der Induktionsbeweis und der Beweisteil (1) abgeschlossen.

(2) Es ist $f'(x^*)(h) = 0$ für alle $h \in X$ mit $g'(x^*)(h) = 0$.

Denn: Sei $h \in X$ mit $g'(x^*)(h) = 0$ gegeben. Da $x^* \in M$ eine lokale Lösung von (P) ist, existiert eine Umgebung U^* von x^* derart, dass $f(x^*) \leq f(x)$ für alle $x \in U^* \cap M$. Wegen (1) und der Definition des Tangentialkegels $T(M; x^*)$ existieren Folgen $\{t_k\} \subset \mathbb{R}_+$ und $\{r_k\} \subset X$ mit $t_k \to 0$, $r_k/t_k \to 0$ und $x^* + t_k h + r_k \in M$ für alle k. Da $h_k := t_k h + r_k = t_k(h + r_k/t_k) \to 0$, ist $x^* + h_k \in U^*$ und folglich $f(x^*) \leq f(x^* + h_k)$ für alle hinreichend großen k. Es ist

$$\lim_{k\to\infty} \frac{|f(x^* + h_k) - f(x^*) - f'(x^*)(h_k)|}{\|h_k\|} = 0,$$

da f in x^* Fréchet-differenzierbar ist. Für alle hinreichend großen k ist daher

$$\begin{aligned}
0 &\leq \frac{f(x^* + h_k) - f(x^*)}{t_k} \\
&= \frac{f(x^* + h_k) - f(x^*) - f'(x^*)(h_k)}{t_k} + f'(x^*)(h_k/t_k) \\
&\leq \underbrace{\frac{|f(x^* + h_k) - f(x^*) - f'(x^*)(h_k)|}{\|h_k\|}}_{\to 0} \|\underbrace{h_k/t_k}_{\to h}\| + f'(x^*)(\underbrace{h_k/t_k}_{\to h}) \\
&\to f'(x^*)(h).
\end{aligned}$$

Also ist $f'(x^*)(h) \geq 0$ für jedes $h \in X$ mit $g'(x^*)(h) = 0$. Wegen der Linearität von $f'(x^*)$ und $g'(x^*)$ ist $f'(x^*)(h) = 0$ für alle $h \in X$ mit $g'(x^*)(h) = 0$. Damit ist (2) bewiesen.

(3) Es existiert ein $y^* \in \mathbb{R}^m$ mit $f'(x^*)(h) + (y^*)^T g'(x^*)(h) = 0$ für alle $h \in X$.

Denn: Wir definieren die Menge

$$\Lambda_+ := \{(f'(x^*)(h) + r, g'(x^*)(h)) \in \mathbb{R} \times \mathbb{R}^m : h \in X,\, r > 0\}.$$

Dann ist Λ_+ eine nichtleere, konvexe (nachrechnen!) Teilmenge von $\mathbb{R} \times \mathbb{R}^m$. Weiter ist $(0,0) \notin \Lambda_+$, denn andernfalls gäbe es ein $h \in X$ mit $f'(x^*)(h) < 0$ und $g'(x^*)(h) = 0$, ein Widerspruch zum Beweisteil (2). Der Trennungssatz für konvexe Mengen im $\mathbb{R}^n$ aus Abschnitt 49 (angewandt auf $A := \{(0,0)\}$ und $B := \Lambda_+$) liefert die Existenz von $(y_0^*, y^*) \in \mathbb{R} \times \mathbb{R}^m \setminus \{(0,0)\}$ mit

$$0 \leq y_0^*(f'(x^*)(h) + r) + (y^*)^T g'(x^*)(h) \qquad \text{für alle } h \in X,\, r > 0.$$

Dann ist insbesondere $0 \leq y_0^*$, indem man $h := 0$ und $r := 1$ setzt. Wäre $y_0^* = 0$, so wäre $0 \leq (y^*)^T g'(x^*)(h)$ für alle $h \in X$. Nach Voraussetzung ist $g'(x^*)\colon X \longrightarrow \mathbb{R}^m$ surjektiv. Folglich existiert ein $h^* \in X$ mit $g'(x^*)(h^*) = -y^*$. Wir erhalten $0 \leq -\|y^*\|^2$ und damit $y^* = 0$, ein Widerspruch zu $(y_0^*, y^*) \neq (0,0)$. Also ist $y_0^* > 0$ und daher o. B. d. A. $y_0^* = 1$ (notfalls ersetze man y^* durch y^*/y_0^*). Mit $r \to 0+$ erhalten wir

$$0 \leq f'(x^*)(h) + (y^*)^T g'(x^*)(h) \qquad \text{für alle } h \in X.$$

Da aber $f'(x^*)\colon X \longrightarrow \mathbb{R}$ und $g'(x^*)\colon X \longrightarrow \mathbb{R}^m$ lineare Abbildungen sind, ist

$$0 = f'(x^*)(h) + (y^*)^T g'(x^*)(h) \qquad \text{für alle } h \in X.$$

Damit ist die Lagrangesche Multiplikatorenregel bewiesen. □

Zum Schluss wollen wir auf den Spezialfall $X := \mathbb{R}^n$ gesondert eingehen. Hierzu beweisen wir zunächst einen wohlbekannten Satz, in dem der Zusammenhang zwischen partieller Differenzierbarkeit und Fréchet-Differenzierbarkeit geklärt wird.

Satz *Sei $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ eine Abbildung, die in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar ist. D. h. es existiere eine Kugel um x^*, auf der die partiellen Ableitungen $\partial g_i/\partial x_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, existieren und in x^* stetig sind. Hierbei sei*

$$g(x) = \begin{pmatrix} g_1(x_1, \ldots, x_n) \\ \vdots \\ g_m(x_1, \ldots, x_n) \end{pmatrix}.$$

Dann ist g in x^* Fréchet-differenzierbar und besitzt das Fréchet-Differential $g'(x^*) \in L(R^n, \mathbb{R}^m) = \mathbb{R}^{m\times n}$ mit

$$g'(x^*) = \Big(\frac{\partial g_i}{\partial x_j}(x^*)\Big)_{\substack{1\le i\le m\\ 1\le j\le n}}.$$

Ferner gilt: Ist g auf einer Kugel um x^* stetig partiell differenzierbar, so ist g in x^* stetig Fréchet-differenzierbar.

Beweis: Nach Voraussetzung existiert eine Kugel B um x^*, auf der die *Funktionalmatrix*

$$g'(x) := \Big(\frac{\partial g_i}{\partial x_j}(x)\Big)_{\substack{1\le i\le m\\ 1\le j\le n}}$$

existiert und in x^* stetig ist. Wir haben zu zeigen, dass durch $g'(x^*)$, als lineare (und natürlich stetige) Abbildung vom $\mathbb{R}^n$ in den $\mathbb{R}^n$ aufgefasst, das Fréchet-Differential von g in x^* gegeben ist. Im Folgenden sei $\|\cdot\|$ die euklidische Norm auf dem $\mathbb{R}^n$ bzw. $\mathbb{R}^m$ oder die zugeordnete Matrixnorm. Zu vorgegebenem $\epsilon > 0$ existiert dann ein $\delta > 0$ mit

$$\|x - x^*\| \le \delta \Longrightarrow x \in B,\ \|g'(x) - g'(x^*)\| \le \epsilon.$$

Sei nun $h \in \mathbb{R}^n$ mit $\|h\| \le \delta$ gegeben. Dann ist

$$\begin{aligned} \|g(x^* + h) - g(x^*) - g'(x^*)h\| &= \Big\| \int_0^1 [g'(x^* + th) - g'(x^*)]h\, dt \Big\| \\ &\le \int_0^1 \|g'(x^* + th) - g'(x^*)\|\, dt\ \|h\| \\ &\le \epsilon\, \|h\|. \end{aligned}$$

Dies zeigt, dass $g'(x^*)$ auch Fréchet-Differential ist. Der Rest der Behauptungen ist trivial. □

Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar, so ist $f'(x^*)h = \nabla f(x^*)^T h$, wobei

$$\nabla f(x^*) := \Big(\frac{\partial f}{\partial x_1}(x^*), \ldots, \frac{\partial f}{\partial x_n}(x^*)\Big)^T$$

der sogenannte *Gradient* von f in x^* ist. Die Funktionalmatrix $g'(x^*) \in \mathbb{R}^{m\times n}$ einer in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbaren Abbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ lässt sich dann schreiben als

$$g'(x^*) = \begin{pmatrix} \nabla g_1(x^*)^T \\ \vdots \\ \nabla g_m(x^*)^T \end{pmatrix}.$$

Als Spezialisierung der obigen Lagrangeschen Multiplikatorenregel auf den Spezialfall $X := \mathbb{R}^n$ erhalten wir:

- *Gegeben sei die Optimierungsaufgabe*

  (P) $\qquad$ *Minimiere* $\quad f(x) \quad$ *auf* $\quad M := \{x \in \mathbb{R}^n : x \in U,\ g(x) = 0\}.$

  Hierbei seien $U \subset \mathbb{R}^n$ offen, $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ bzw. $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in der lokalen Lösung $x^* \in M$ von (P) stetig partiell differenzierbar bzw. auf einer Kugel um x^* stetig partiell differenzierbar. Sei $g'(x^*)(\mathbb{R}^n) = \mathbb{R}^m$ bzw. $\{\nabla g_1(x^*), \ldots, \nabla g_m(x^*)\}$ linear unabhängig. Dann existiert ein $y^* \in \mathbb{R}^m$ mit

$$\nabla f(x^*) + g'(x^*)^T y^* = \nabla f(x^*) + \sum_{i=1}^{m} y_i^* \nabla g_i(x^*) = 0.$$

Beispiel: Wir stellen uns die Aufgabe, das maximale Volumen eines Quaders mit der Oberfläche $L\,\mathrm{m}^2$ zu berechnen, wobei $L > 0$ gegeben ist. Die Kantenlängen des Quaders, in Metern gemessen, werden mit x_1, x_2, x_3 bezeichnet. Dann haben wir offenbar die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -x_1x_2x_3 \quad \text{unter der Nebenbedingung} \\ \qquad x > 0, \quad g(x) := 2(x_1x_2 + x_2x_3 + x_3x_1) - L = 0 \end{cases}$$

zu lösen. Hierbei ist $x > 0$ komponentenweise zu verstehen. Wir nehmen an, x^* sei eine (lokale) Lösung von (P). Offenbar ist $\nabla g(x^*) \neq 0$, die Voraussetzungen für die Lagrangesche Multiplikatorenregel sind erfüllt. Daher existiert ein $y^* \in \mathbb{R}$ mit

$$0 = \nabla f(x^*) + y^* \nabla g(x^*) = \begin{pmatrix} -x_2^*x_3^* \\ -x_1^*x_3^* \\ -x_1^*x_2^* \end{pmatrix} + y^* \begin{pmatrix} 2(x_2^* + x_3^*) \\ 2(x_1^* + x_3^*) \\ 2(x_1^* + x_2^*) \end{pmatrix},$$

außerdem muss natürlich noch die Gleichung

$$x_1^*x_2^* + x_2^*x_3^* + x_1^*x_3^* = \frac{L}{2}$$

gelten. Damit haben wir vier Gleichungen für die vier Unbekannten x_1^*, x_2^*, x_3^*, y^*. Aus

$$2y^* = \frac{x_2^*x_3^*}{x_2^* + x_3^*} = \frac{x_1^*x_3^*}{x_1^* + x_3^*} = \frac{x_1^*x_2^+}{x_1^* + x_2^*}$$

erhalten wir $x_1^* = x_2^* = x_3^*$. Aus $3(x_1^*)^2 = L/2$ ergibt sich $x_1^* = \sqrt{L/6}$. Der gesuchte Quader ist also ein Würfel mit der Kantenlänge $\sqrt{L/6}\,\mathrm{m}$. □

51 Das Problem der Dido

In den Sagen des klassischen Altertums von Gustav Schwab (oder auch der Aeneis von Vergil) wird von der phönizischen Prinzessin Dido berichtet, die nach dem Mord ihres reichen Mannes Sychaeus durch ihren Bruder Pygmalion (nicht zu verwechseln mit Pygmalion aus Ovids Metamorphosen) aus Tyros fliehen musste und an den Golf von Tunis kam.

- *Hier erkaufte sie anfangs nur ein Stück Landes, welches Byrsa oder Stierhaut genannt wurde; mit diesem Namen aber verhielt es sich so: Dido, in Afrika angekommen, verlangte nur so viel Feldes, als sie mit einer Stierhaut zu umspannen vermochte. Diese Haut aber schnitt sie in so dünne Riemen, dass dieselbe den ganzen Raum einschloss, den jetzt Byrsa, die Burg Karthagos, einnimmt.*

Die Aufgabe der Dido kann mathematisch folgendermaßen formuliert werden:

- Welche Gestalt muss eine ebene, einfach geschlossene, glatte Kurve gegebener Länge haben, damit die von ihr berandete Fläche maximal wird?

Die entsprechende Aufgabe heißt auch das *isoperimetrische Problem*. Wir betrachten ein einfacheres Problem, bei dem berücksichtigt wird, dass ein geradliniger Küstenabschnitt einer vorgegebenen Länge zu der gesuchten Kurve gehört.

- Welche Kurve fester Länge zwischen zwei Punkten umschließt zusammen mit der Strecke zwischen diesen Punkten den größten Flächeninhalt?

Eine exakte mathematische Formulierung ist dann:

- Seien positive Zahlen a und L mit $2a < L < \pi a$ gegeben. Gesucht ist eine Lösung der Optimierungsaufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := -\displaystyle\int_{-a}^{a} x(t)\,dt \quad \text{auf} \\ M := \{x \in C^1[-a,a] : x(-a) = x(a) = 0,\ \displaystyle\int_{-a}^{a} \sqrt{1 + x'(t)^2}\,dt = L\}. \end{cases}$$

Wie man richtig rät, ist die Lösung ein Teil eines Kreises. Wir nehmen an, dies sei schon gesichert und überlegen uns, welcher Kreisabschnitt die gesuchte Lösung ist. In Abbildung 67 veranschaulichen wir uns die Aufgabe. Die Strecke $\overline{PR}$ mit Mittelpunkt Q ist die Küstenlinie der Länge $2a$, eingetragen ist ein Kreisbogen der Länge L. Der Kreis habe den Radius r und O als Mittelpunkt. Ferner sei $\alpha := \sphericalangle POR$ und daher $\alpha/2 = \sphericalangle POQ = \sphericalangle QOR$. Aus $\alpha r = L$ und $\sin(\alpha/2) = a/r$ erhalten wir durch Eliminieren von r, dass $\alpha \in (0,\pi)$ als Lösung von

$$\frac{a}{L}\alpha - \sin\Big(\frac{\alpha}{2}\Big) = 0$$

zu bestimmen ist. Wir wollen uns überlegen, dass diese Gleichung genau eine Lösung in $(0,\pi)$ besitzt. Hierzu definieren wir $\psi(\alpha) := (a/L)\alpha - \sin(\alpha/2)$. Es ist $\psi(0) = 0$, $\psi'(0) =$

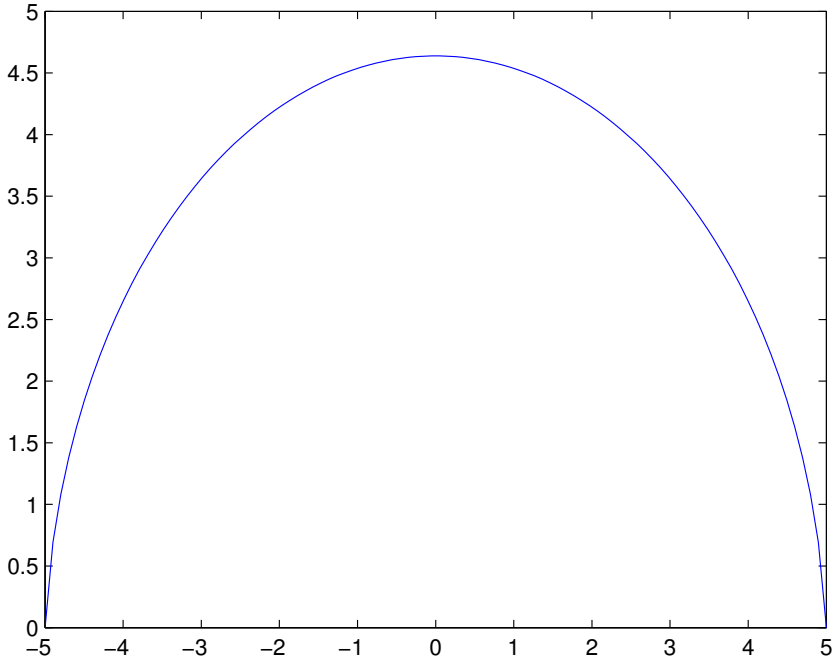


Abbildung 67: Das Problem der Dido

$a/L - 1/2 < 0$ und $\psi(\pi) = (a/L)\pi - 1 > 0$. Daher besitzt ψ mindestens eine Nullstelle $\alpha^* \in (0, \pi)$. Diese ist auch eindeutig, wie man nach leichter Argumentation feststellt[65]. Für $a := 1$ und $L := 2.5$ haben wir in Abbildung 68 die Funktion ψ geplottet. Man erkennt noch einmal das, was wir uns eben schon überlegt haben, insbesondere besitzt ψ genau eine positive Nullstelle $\alpha^* \in (0, \pi)$. Wir erhalten $\alpha^* = 2.2622$ und anschließend $r^* = 1.1051$ als Radius des gesuchten Kreises.

Wir wollen mit Hilfe der Lagrangeschen Multiplikatorenregel (siehe Abschnitt 50) nachweisen, dass eine (lokale) Lösung von (P) *notwendigerweise* durch einen Kreisabschnitt gegeben ist. Man beachte, dass hierdurch *nicht* nachgewiesen ist, dass dieser Kreisabschnitt wirklich eine Lösung ist. Hierzu müssen wir die Daten in der Lagrangeschen Multiplikatorenregel festlegen. Sei

$$X := \{x \in C^1[-a, a] : x(-a) = x(a) = 0\}$$

die Menge der auf $[-a, a]$ stetig differenzierbaren, reellwertigen Funktionen, die an den Intervallenden verschwinden. Als Norm auf X definieren wir

$$\|x\| := \max\Big(\max_{t\in[-a,a]} |x(t)|, \max_{t\in[-a,a]} |x'(t)|\Big).$$

[65] Denn ψ' besitzt genau eine Nullstelle $\hat{\alpha} \in (0, \pi)$, dort hat ψ ein Minimum. Auf $(0, \hat{\alpha}]$ ist ψ negativ und monoton fallend, rechts von $\hat{\alpha}$ ist ψ monoton wachsend. Da $\psi(\pi)$ positiv ist, besitzt ψ genau eine Nullstelle in $(\hat{\alpha}, \pi)$.

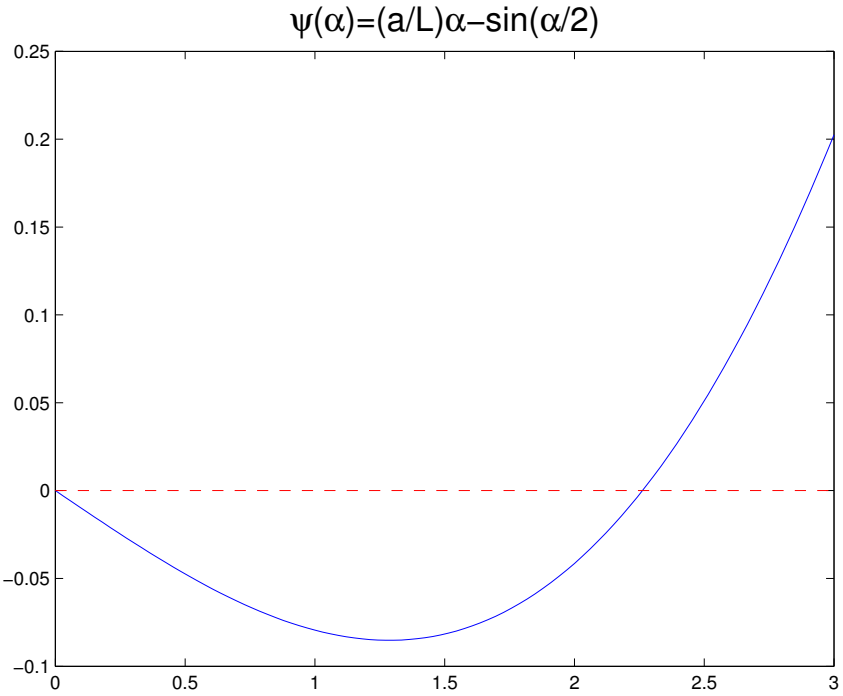


Abbildung 68: Die Funktion ψ beim Problem der Dido

Versehen mit dieser Norm ist X offenbar ein linearer normierter Raum. Sei ferner $U := X$, die Abbildungen $f\colon X \longrightarrow \mathbb{R}$ und $g\colon X \longrightarrow \mathbb{R}$ sind durch

$$f(x) := -\int_{-a}^{a} x(t)\,dt, \qquad g(x) := \int_{-a}^{a} \sqrt{1 + x'(t)^2}\,dt - L$$

gegeben. Wir haben nachzuweisen, dass f bzw. g in der lokalen Lösung x^* Fréchet-differenzierbar bzw. stetig Fréchet-differenzierbar sind. Für $h \in X$ ist

$$f(x^* + h) - f(x^*) = -\int_{-a}^{a} h(t)\,dt,$$

das Fréchet-Differential $f'(x^*)$ ist daher durch

$$f'(x^*)(h) = -\int_{-a}^{a} h(t)\,dt$$

gegeben[66]. Die Abbildung $g\colon X \longrightarrow \mathbb{R}$ ist in x^* stetig Fréchet-differenzierbar mit dem Fréchet-Differential

$$g'(x^*)(h) = \int_{-a}^{a} \frac{(x^*)'(t)h'(t)}{\sqrt{1 + (x^*)'(t)^2}}\,dt.$$

[66]Man beachte, dass die Abbildung $f'(x^*)\colon X \longrightarrow \mathbb{R}$ nicht nur linear, sondern auch stetig ist, wie man aus $|f'(x^*)(h)| \le 2a\,\|h\|$ für alle $h \in X$ erkennt.

Um dies einzusehen beachten wir: Ist bei gegebenem $y \in \mathbb{R}$ (bei der Anwendung gleich ist $y = (x^*)'(t)$) die Abbildung $\phi\colon \mathbb{R} \longrightarrow \mathbb{R}$ durch $\phi(k) := \sqrt{1+(y+k)^2}$ definiert, so ist

$$\phi'(k) = \frac{y+k}{\sqrt{1+(y+k)^2}}$$

und

$$\phi''(k) = \frac{\sqrt{1+(y+k)^2} - (y+k)^2/\sqrt{1+(y+k)^2}}{1+(y+k)^2} = \frac{1}{(1+(y+k)^2)^{3/2}} > 0.$$

Folglich ist

$$|\phi(k) - \phi(0) - \phi'(0)k| = \frac{\phi''(k_0)}{2!}k^2 \leq \frac{1}{2}k^2.$$

Daher ist

$$\begin{aligned}
& \Big| g(x^*+h) - g(x^*) - \int_{-a}^{a} \frac{(x^*)'(t)h'(t)}{\sqrt{1+(x^*)'(t)^2}}\,dt \Big| \\
= \; & \Big| \int_{-a}^{a} \Big[\sqrt{1+((x^*)'(t)+h'(t))^2} - \sqrt{1+(x^*)'(t)^2} - \frac{(x^*)'(t)h'(t)}{\sqrt{1+(x^*)'(t)^2}} \Big]\,dt \Big| \\
\leq \; & \int_{-a}^{a} \Big| \sqrt{1+((x^*)'(t)+h'(t))^2} - \sqrt{1+(x^*)'(t)^2} - \frac{(x^*)'(t)h'(t)}{\sqrt{1+(x^*)'(t)^2}} \Big|\,dt \\
\leq \; & \frac{1}{2}\int_{-a}^{a} h'(t)^2\,dt \\
\leq \; & a\,\|h\|^2.
\end{aligned}$$

Hieran erkennen wir, dass die angegebene Abbildung $g'(x^*)$ in der Tat das Fréchet-Differential von g in x^* ist. Dass g in x^* sogar stetig Fréchet-differenzierbar ist, erkennt man an

$$\begin{aligned}
|g'(x)(h) - g'(x^*)(h)| \; &\leq \; \int_{-a}^{a} \Big| \frac{x'(t)}{\sqrt{1+x'(t)^2}} - \frac{(x^*)'(t)}{\sqrt{1+(x^*)'(t)^2}} \Big|\,|h'(t)|\,dt \\
&\leq \; \frac{1}{2}\int_{-a}^{a} |x'(t) - (x^*)'(t)|\,|h'(t)|\,dt \\
&\leq \; a\,\|x - x^*\|\,\|h\|,
\end{aligned}$$

woraus $\|g'(x) - g'(x^*)\| \leq a\,\|x - x^*\|$ und damit die Behauptung folgt. Schließlich bleibt nachzuweisen, dass $g'(x^*)(X) = \mathbb{R}$ bzw. $g'(x^*)\colon X \longrightarrow \mathbb{R}$ surjektiv ist. Dies ist aber einfach. Denn definiert man $h \in X$ durch $h(t) := \alpha\,x^*(t)$, so ist

$$g'(x^*)(h) = \alpha \int_{-a}^{a} \frac{(x^*)'(t)^2}{\sqrt{1+(x^*)'(t)^2}}\,dt,$$

woraus man die Gültigkeit der Behauptung ablesen kann. Damit sind alle Voraussetzungen für die Anwendung der Lagrangeschen Multiplikatorenregel erfüllt und es

existiert daher ein $y^* \in \mathbb{R}$ mit

$$\begin{aligned} 0 &= f'(x^*)(h) + y^* g'(x^*)(h) \\ &= \int_{-a}^{a} \Big[-h(t) + y^* \frac{(x^*)'(t) h'(t)}{\sqrt{1 + (x^*)'(t)^2}} \Big] \, dt \\ &= \int_{-a}^{a} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} \Big] h'(t) \, dt \end{aligned}$$

für alle $h \in X$. Nun definiere man

$$c := \frac{1}{2a} \int_{-a}^{a} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} \Big] \, dt$$

und anschließend

$$h(t) := \int_{-a}^{t} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} - c \Big] \, dt.$$

Dann ist $h \in X$ und

$$\begin{aligned} & \int_{-a}^{a} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} - c \Big]^2 \, dt \\ = & \int_{-a}^{a} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} - c \Big] h'(t) \, dt \\ = & \underbrace{\int_{-a}^{a} \Big[(t+a) + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} \Big] h'(t) \, dt}_{=0} - \underbrace{c(h(a) - h(-a))}_{=0} \\ = & \, 0 \end{aligned}$$

und folglich

$$t + a - c + y^* \frac{(x^*)'(t)}{\sqrt{1 + (x^*)'(t)^2}} = 0.$$

Daher ist

$$(y^*)^2 \frac{(x^*)'(t)^2}{1 + (x^*)'(t)^2} = (t + a - c)^2$$

Auflösen nach $(x^*)'(t)$ ergibt

$$(x^*)'(t)^2 = \frac{(t+a-c)^2}{(y^*)^2 - (t+a-c)^2} = \Big(\frac{d}{dt} \sqrt{(y^*)^2 - (t+a-c)^2} \Big)^2.$$

Damit folgt die Existenz einer Konstanten d mit

$$(x^*(t) - d)^2 + (t + a - c)^2 = (y^*)^2 \qquad \text{für alle } t \in [-a, a].$$

Zur Bestimmung der noch unbekannten Konstanten c, y^* und d berücksichtigen wir die Nebenbedingungen

$$x^*(-a) = x^*(a) = 0, \qquad g(x^*) := \int_{-a}^{a} \sqrt{1 + (x^*)'(t)^2} \, dt - L = 0.$$

Wegen $x^*(-a) = x^*(a) = 0$ ist $c = a$ und $(y^*)^2 = d^2 + a^2$. Also ist

$$(x^*(t) - d)^2 + t^2 = d^2 + a^2 \quad \text{bzw.} \quad x^*(t) = d + \sqrt{d^2 + a^2 - t^2}.$$

Die Konstante d (offenbar ist $(0, d)$ der Mittelpunkt des gesuchten Kreises und daher $d < 0$) ist zu bestimmen aus $g(x^*) = 0$, was auf

$$g(x^*) = \int_{-a}^{a} \sqrt{\frac{d^2 + a^2}{d^2 + a^2 - t^2}}\, dt - L = 2\sqrt{d^2 + a^2} \arcsin \frac{a}{\sqrt{d^2 + a^2}} - L = 0$$

bzw.

$$\frac{a}{\sqrt{d^2 + a^2}} = \sin \frac{L}{2\sqrt{d^2 + a^2}}$$

führt. Mit $r := \sqrt{d^2 + a^2}$ (Radius des gesuchten Kreisabschnitts, siehe Abbildung 67) und $\alpha = L/r$ lässt sich diese Gleichung auch als

$$\frac{a}{L}\alpha - \sin\Big(\frac{\alpha}{2}\Big) = 0$$

schreiben. Genau diese Gleichung hatten wir zu Beginn dieses Abschnitts schon erhalten. Insbesondere hatten wir uns überlegt, dass sie wegen $2a < L < \pi a$ genau eine Nullstelle in $(0, \pi)$ besitzt.

52 Das diskrete Problem der Dido

In diesem Abschnitt betrachten zu dem kontinuierlichen Problem der Dido

$$\text{(P)} \quad \begin{cases} \text{Minimiere} f(x) := -\displaystyle\int_{-a}^{a} x(t)\, dt \quad \text{auf} \\ M := \{x \in C^1[-a, a] : x(-a) = x(a) = 0,\ \displaystyle\int_{-a}^{a} \sqrt{1 + x'(t)^2}\, dt - L = 0\} \end{cases}$$

eine *diskrete Version*, bei welchem in (P) die Integrale

$$f(x) := -\int_{-a}^{a} x(t)\, dt, \qquad g(x) := \int_{-a}^{a} \sqrt{1 + x'(t)^2}\, dt$$

durch Summen ersetzt werden. Nach wie vor setzen wir voraus, dass $2a < L < \pi a$. Mit $n \in \mathbb{N}$ sei

$$h := \frac{2a}{n+1}$$

und

$$t_i := -a + ih, \qquad i = 0, \ldots, n+1.$$

Wir betrachten die *endlich dimensionale* Optimierungsaufgabe

$$\text{(}\Pi\text{)} \quad \begin{cases} \text{Minimiere} \quad \phi(\xi) := -h \displaystyle\sum_{i=1}^{n} \xi_i \quad \text{auf} \\ M := \Big\{\xi \in \mathbb{R}^n : \gamma(\xi) := \displaystyle\sum_{i=0}^{n} \sqrt{h^2 + (\xi_{i+1} - \xi_i)^2} - L = 0\Big\}. \end{cases}$$

Hierbei ist $\xi_0 = 0$ und $\xi_{n+1} = 0$, was den Bedingungen $x(-a) = 0$ bzw. $x(a) = 0$ entspricht. Wie kommt man von der kontinuierlichen Aufgabe (P) zum diskreten Problem (Π)? Hierzu gehen wir von einer stückweise linearen Funktion $x(\cdot)$ aus, d. h. es sei

$$x(t) := \frac{1}{h}[(t - t_i)\xi_{i+1} + (t_{i+1} - t)\xi_i] \quad \text{für } t \in [t_i, t_{i+1}], \quad i = 0, \ldots, n.$$

Dann ist

$$\begin{aligned} f(x) &= -\int_{-a}^{a} x(t)\,dt \\ &= -\sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} \frac{1}{h}[(t - t_i)\xi_{i+1} + (t_{i+1} - t)\xi_i]\,dt \\ &= -\frac{h}{2}\sum_{i=0}^{n}(\xi_{i+1} + \xi_i) \\ &= -h\sum_{i=1}^{n}\xi_i \\ &= \phi(\xi). \end{aligned}$$

Entsprechend ist $g(x) = \gamma(\xi)$. Man beachte, dass

$$\gamma(\xi) = \sum_{i=0}^{n} \|(t_{i+1}, \xi_{i+1}) - (t_i, \xi_i)\|_2 - L,$$

so dass $\gamma(\xi)$ die um L verminderte Länge des Streckenzuges ist, der von (t_0, ξ_0) über $(t_1, \xi_1), \ldots, (t_n, \xi_n)$ nach (t_{n+1}, ξ_{n+1}) führt.

Nun wollen wir die Lagrangesche Multiplikatorenregel anwenden, um einen Kandidaten ξ^* für eine lokale Lösung von (Π) zu berechnen. Der Gradient von ϕ ist konstant und gegeben durch $\nabla\phi(\xi^*) = -he$, wobei $e \in \mathbb{R}^n$ der Vektor im $\mathbb{R}^n$ ist, dessen Komponenten alle gleich 1 sind. Die k-te Komponente des Gradienten von γ in ξ^* ist

$$\frac{\partial\gamma}{\partial\xi_k}(\xi^*) = \frac{\xi_k^* - \xi_{k-1}^*}{\sqrt{h^2 + (\xi_k^* - \xi_{k-1}^*)^2}} - \frac{\xi_{k+1}^* - \xi_k^*}{\sqrt{h^2 + (\xi_{k+1}^* - \xi_k^*)^2}}, \qquad k = 1, \ldots, n.$$

Hierbei hat man für die erste und die letzte Komponente $\xi_0^* = 0$ bzw. $\xi_n^* = 0$ zu berücksichtigen. Es ist $\nabla\gamma(\xi^*) \neq 0$. [67] Daher ist die Lagrangesche Multiplikatorenregel

[67] Wäre $\nabla\gamma(\xi^*) = 0$, so wäre

$$\frac{\xi_{k+1}^* - \xi_k^*}{\sqrt{h^2 + (\xi_{k+1}^* - \xi_k^*)^2}} = c, \qquad k = 0, \ldots, n$$

mit einer von k unabhängigen Konstanten c mit $|c| < 1$. Hieraus schließen wir auf $(\xi_{k+1}^* - \xi_k^*)^2 = c^2h^2/(1 - c^2)$ und hieraus auf

$$\xi_{k+1}^* - \xi_k^* = \frac{ch}{\sqrt{1 - c^2}}, \qquad k = 0, \ldots, n.$$

anwendbar und liefert die Existenz eines $y^* \in \mathbb{R}$ mit $\nabla\phi(\xi^*) + y^*\nabla\gamma(\xi^*) = 0$ bzw.

$$(*) \qquad -h + y^*\Big[\frac{\xi_k^* - \xi_{k-1}^*}{\sqrt{h^2 + (\xi_k^* - \xi_{k-1}^*)^2}} - \frac{\xi_{k+1}^* - \xi_k^*}{\sqrt{h^2 + (\xi_{k+1}^* - \xi_k^*)^2}}\Big] = 0, \qquad k = 1, \dots, n.$$

Offenbar ist notwendigerweise $y^* \neq 0$. Durch $(*)$ und $\gamma(\xi^*) = 0$ hat man $n+1$ Gleichungen für die $n+1$ Unbekannten $\xi_1^*, \dots, \xi_n^*$ und y^*. Die Idee für das weitere Vorgehen besteht darin, in Abhängigkeit von $y \in \mathbb{R}$ einen Vektor $\xi(y) \in \mathbb{R}^n$ so zu bestimmen, dass die Gleichungen

$$-h + y\Big[\frac{\xi_k(y) - \xi_{k-1}(y)}{\sqrt{h^2 + (\xi_k(y) - \xi_{k-1}(y))^2}} - \frac{\xi_{k+1}(y) - \xi_k(y)}{\sqrt{h^2 + (\xi_{k+1}(y) - \xi_k(y))^2}}\Big] = 0, \qquad k = 1, \dots, n,$$

wobei $\xi_0(y) = \xi_{n+1}(y) = 0$. Anschließend bestimmen wir numerisch (bei vorgegebenen Daten a und n, damit h, sowie L) eine Lösung y^* von $\gamma(\xi(y)) = 0$ und setzen $\xi^* := \xi(y^*)$. Das Vorgehen ist also ähnlich dem im kontinuierlichen Fall.

Für $y \in \mathbb{R}$ mit $y > hn/2$ definieren wir

$$\alpha_k(y) := \frac{h}{y}\Big(\frac{n}{2} - k\Big), \qquad k = 0, \dots, n.$$

Für diese y ist $|\alpha_k(y)| < 1$, $k = 0, \dots, n$ und daher

$$\eta_k(y) := \frac{\alpha_k(y)h}{\sqrt{1 - \alpha_k(y)^2}}, \qquad k = 0, \dots, n,$$

reell und wohldefiniert. Für $k = 1, \dots, n$ ist dann

$$\frac{\eta_{k-1}(y)}{\sqrt{h^2 + \eta_{k-1}(y)^2}} - \frac{\eta_k(y)}{\sqrt{h^2 + \eta_k(y)^2}} = \alpha_{k-1}(y) - \alpha_k(y) = \frac{h}{y}$$

bzw.

$$-h + y\Big[\frac{\eta_{k-1}(y)}{\sqrt{h^2 + \eta_{k-1}(y)^2}} - \frac{\eta_k(y)}{\sqrt{h^2 + \eta_k(y)^2}}\Big] = 0.$$

Jetzt definieren wir sukzessive $\xi_0(y), \xi_1(y), \dots, \xi_n(y), \xi_{n+1}(y)$ durch $\xi_0(y) := 0$ und

$$\xi_{k+1}(y) := \xi_k(y) + \eta_k(y), \qquad k = 0, \dots, n.$$

Wegen der gerade eben bewiesenen Beziehung ist

$$-h + y\Big[\frac{\xi_k(y) - \xi_{k-1}(y)}{\sqrt{h^2 + (\xi_k(y) - \xi_{k-1}(y))^2}} - \frac{\xi_{k+1}(y) - \xi_k(y)}{\sqrt{h^2 + (\xi_{k+1}(y) - \xi_k(y))^2}}\Big] = 0$$

Aufsummieren liefert

$$0 = \xi_{n+1}^* - \xi_0^* = \sum_{k=0}^{n}(\xi_{k+1}^* - \xi_k^*) = \frac{ch}{\sqrt{1 - c^2}}(n+1).$$

Daher ist $c = 0$ und folglich $\xi^* = 0$. Aus $0 = \gamma(\xi^*) = 2a - L$ erhalten wir einen Widerspruch zu $2a < L$.

für $k = 1, \ldots, n$. Weiter ist $\xi_0(y) = 0$. Wir zeigen, dass auch $\xi_{n+1}(y) = 0$. Denn es ist

$$\xi_{n+1}(y) = \xi_n(y) + \eta_n(y) = \underbrace{\xi_0(y)}_{=0} + \sum_{k=0}^{n} \eta_k(y) = \sum_{k=0}^{n} \eta_k(y).$$

Nun ist $\alpha_0(y) = -\alpha_n(y)$ und daher $\eta_0(y) = -\eta_n(y)$, ferner $\alpha_1(y) = -\alpha_{n-1}(y)$ und daher $\eta_1(y) = -\eta_{n-1}(y)$. Da weiter für gerades n offenbar $\alpha_{n/2}(y) = 0$ und daher $\eta_{n/2}(y) = 0$ ist insgesamt $\sum_{k=0}^{n} \eta_k(y) = 0$ und daher $\xi_{n+1}(y) = 0$.

Für $y > hn/2$ ist

$$\begin{aligned}
\gamma(\xi(y)) &= \sum_{i=0}^{n} \sqrt{h^2 + (\xi_{i+1}(y) - \xi_i(y))^2} - L \\
&= \sum_{i=0}^{n} \sqrt{h^2 + \eta_i(y)^2} - L \\
&= h \sum_{i=0}^{n} \frac{1}{\sqrt{1 - \alpha_i(y)^2}} - L \\
&= yh \sum_{i=0}^{n} \frac{1}{\sqrt{y^2 - (n/2 - i)^2 h^2}} - L.
\end{aligned}$$

Mit

$$\chi(y) := y \sum_{i=0}^{n} \frac{1}{\sqrt{y^2 - (n/2 - i)^2 h^2}}$$

haben wir also eine Lösung y^* von $\chi(y) = L/h$ mit $y^* > nh/2$ zu bestimmen. Offenbar ist $\chi(y) \to +\infty$ für $y \searrow hn/2$ und $\lim_{y \to +\infty} \chi(y) = n + 1$. Nun ist

$$\frac{L}{h} = \frac{L}{2a}(n + 1) > n + 1,$$

da wir $L > 2a$ vorausgesetzt haben. Folglich besitzt $\chi(y) = L/h$ eine Lösung y^* in $(hn/2, +\infty)$. Weiter ist

$$\begin{aligned}
\chi'(y) &= \sum_{i=0}^{n} \Big[\frac{1}{\sqrt{y^2 - (n/2 - i)^2 h^2}} - \frac{y^2}{(y^2 - (n/2 - i)^2 h^2)^{3/2}} \Big] \\
&= -h^2 \sum_{i=0}^{n} \frac{(n/2 - i)^2}{(y^2 - (n/2 - i)^2 h^2)^{3/2}} \\
&< 0
\end{aligned}$$

für alle $y \in (nh/2, +\infty)$. Daher ist $\chi(\cdot)$ auf $(nh/2, +\infty)$ monoton fallend, womit insgesamt gezeigt ist, dass die Gleichung $\chi(y) = L/h$ bzw. $\gamma(\xi(y)) = 0$ genau eine Lösung $y^* \in (nh/2, +\infty)$ besitzt.

Beispiel: Sei $a := 5$ und $L := 15$. Als Lösung (genauer: Lösungskandidaten) des (kontinuierlichen) Problems der Dido erhalten wir

$$x^*(t) = d + \sqrt{r^2 - t^2},$$

wobei

$$d := -0.375778920234108, \qquad r := 5.014101095599520.$$

Nun kommen wir zur Lösung des diskreten Problems. Der Lagrange-Multiplikator y^* ist positive Lösung der Gleichung

$$\chi(y) := y \sum_{i=0}^{n} \frac{1}{\sqrt{y^2 - (n/2 - i)^2 h^2}} = \frac{L}{h},$$

wobei $h = 2a/(n+1)$. Für $n = 20$ haben wir in Abbildung 69 die Funktion $\chi(\cdot)$ und die konstante Funktion mit dem Wert L/h gestrichelt aufgezeichnet. Wenden

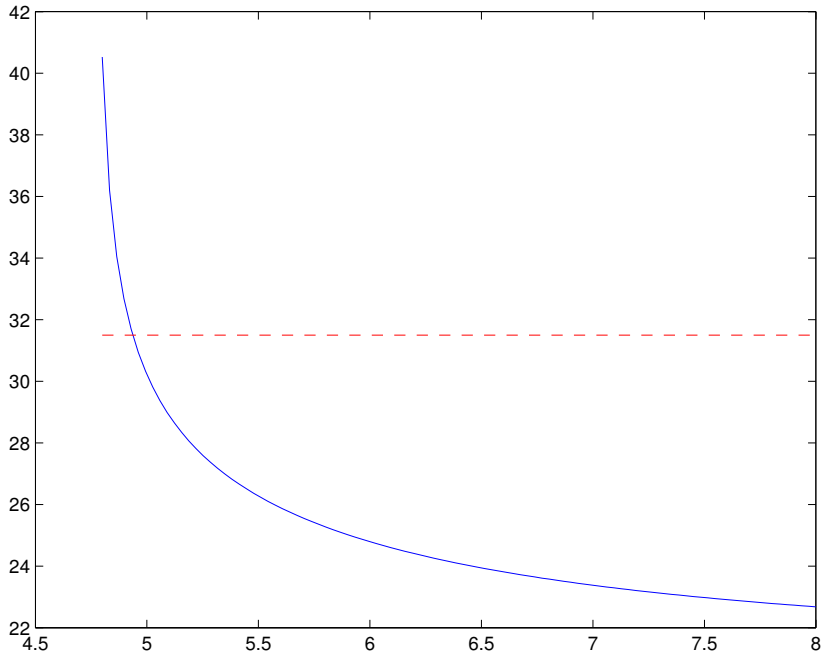


Abbildung 69: Die Funktion $\chi(\cdot)$ für $n = 20$

wir auf die Gleichung $\chi(y) = L/h$ das Newton-Verfahren an, so erhalten wir $y^* = 4.937066810618908$ als Lösung. Die Lösung des diskreten Problems der Dido für $n = 20$ haben wir in Abbildung 70 zusammen mit der Lösung der kontinuierlichen Lösung aufgetragen. Die Vermutung, dass die Lösung des diskreten Problems auf dem Kreissegment liegt, welche die Lösung des kontinuierlichen Problems ist, ist falsch. In Abbildung 71 haben wir für $n = 3$ die Lösung des diskreten zusammen mit der des kontinuierlichen Problems eingetragen. □

53 Der Satz von Kuhn-Tucker

Im Satz von Kuhn-Tucker, siehe H. W. Kuhn, A. W. Tucker (1951), werden notwendige Optimalitätsbedingungen erster Ordnung (d. h. in den Bedingungen, den

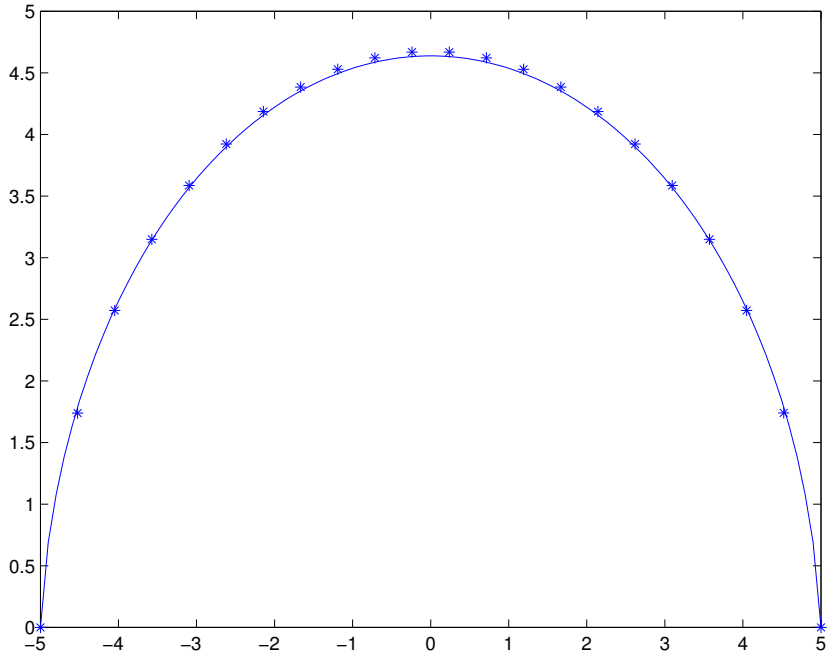


Abbildung 70: Die Lösung des diskreten Problems der Dido für $n = 20$

sogenannten *Kuhn-Tucker-Bedingungen*, treten nur Ableitungen erster Ordnung auf) dafür angegeben, dass ein für eine Optimierungsaufgabe zulässiger Punkt x^* eine lokale Lösung ist. Ist z. B. eine Funktion $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ in $x^* \in \mathbb{R}$ stetig differenzierbar und hat $f(\cdot)$ in x^* ein lokales Minimum oder Maximum, so ist notwendigerweise $f'(x^*) = 0$, d. h. $f'(x^*) = 0$ ist die Kuhn-Tucker-Bedingung für die (unrestringierte) Optimierungsaufgabe, $f(\cdot)$ zu minimieren (oder zu maximieren). Komplizierter wird die Situation bei restringierten Optimierungsaufgaben.

Bei der Herleitung der Kuhn-Tucker-Bedingungen für die Optimierungsaufgabe, die Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer (durch Ungleichungen und Gleichungen gegebenen) Menge $M \subset \mathbb{R}^n$ zu minimieren, spielen zwei Mengen eine ausgezeichnete Rolle. Dies ist der *Kegel der zulässigen Richtungen* und der *Tangentialkegel*, den wir schon im Beweis der Lagrangeschen Multiplikatorenregel in Abschnitt 50 eingeführt haben. Wir geben jetzt die Definitionen an:

Definition Sei $M \subset \mathbb{R}^n$ und $x^* \in M$. Dann heißt

$$F(M; x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{c} \text{Es existiert eine Folge } \{t_k\} \subset \mathbb{R}_+ \text{ mit} \\ t_k \to 0 \text{ und } x^* + t_k p \in M \text{ für alle } k \end{array} \right\}$$

der *Kegel der zulässigen Richtungen* an M in x^*. Dagegen heißt

$$T(M; x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{c} \text{Es existieren Folgen } \{t_k\} \subset \mathbb{R}_+,\ \{r_k\} \subset \mathbb{R}^n \text{ mit} \\ x^* + t_k p + r_k \in M \text{ für alle } k,\ t_k \to 0,\ r_k/t_k \to 0. \end{array} \right\}$$

der *Tangentialkegel* bzw. der *Kegel der tangentialen Richtungen* an M in x^*.

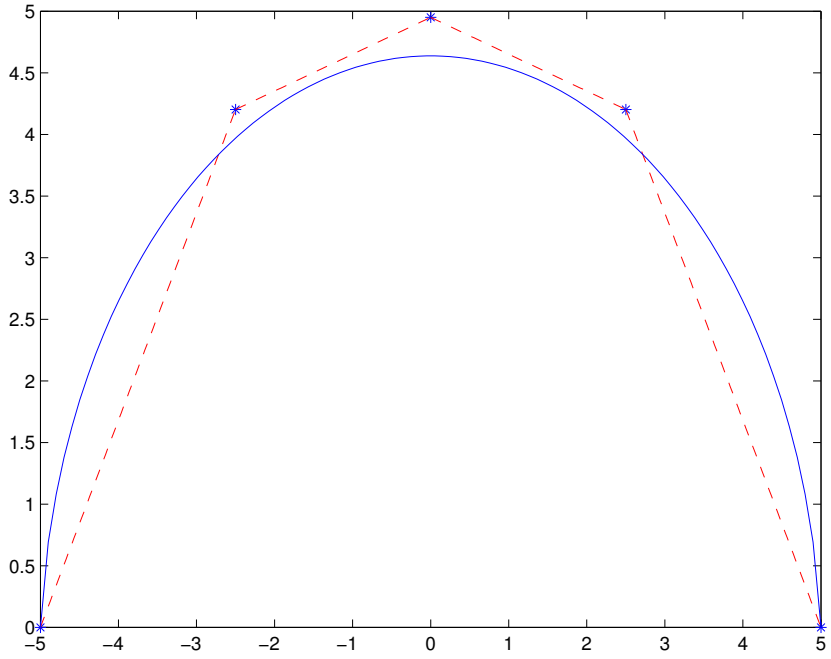


Abbildung 71: Die Lösung des diskreten Problems der Dido für $n = 3$

Ist z. B. $M := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ die Oberfläche der euklidischen Einheitskugel, so ist $F(M; x^*)$ offenbar trivial, also $F(M; x^*) = \{0\}$, aber (Beweis?)

$$T(M; x^*) = \{p \in \mathbb{R}^n : (x^*)^T p = 0\}.$$

Wir notieren:

Lemma *Sei $M \subset \mathbb{R}^n$ und $x^* \in M$. Dann gilt:*

- *Es ist $F(M; x^*) \subset T(M; x^*)$.*
- *Der Tangentialkegel $T(M; x^*)$ ist abgeschlossen und enthält daher den Abschluss* $\operatorname{cl} F(M; x^*)$ *des Kegels der zulässigen Richtungen.*

Beweis: Die erste Aussage, dass der Kegel der zulässigen Richtungen $F(M; x^*)$ im Tangentialkegel $T(M; x^*)$ enthalten ist, ist trivialerweise richtig. Zum Beweis der zweiten Aussage sei $\{p^{(j)}\} \subset T(M; x^*)$ eine gegen $p \in \mathbb{R}^n$ konvergente Folge. Nach Definition des Tangentialkegels existieren zu jedem $j \in \mathbb{N}$ Folgen $\{t_k^{(j)}\} \subset \mathbb{R}_+$ und $\{r_k^{(j)}\} \subset \mathbb{R}^n$ mit

$$x^* + t_k^{(j)} p^{(j)} + r_k^{(j)} \in M \quad \text{für alle } k$$

und

$$\lim_{k\to\infty} t_k^{(j)} = 0, \qquad \lim_{k\to\infty} \frac{r_k^{(j)}}{t_k^{(j)}} = 0.$$

Zu jedem $j \in \mathbb{N}$ existiert ein $k(j) \in \mathbb{N}$ mit

$$0 < t_k^{(j)} \leq \frac{1}{j}, \qquad \frac{\|r_k^{(j)}\|}{t_k^{(j)}} \leq \frac{1}{j} \qquad \text{für alle } k \geq k(j).$$

Nun definiere man die Folgen $\{t_j\} \subset \mathbb{R}_+$ und $\{r_j\} \subset \mathbb{R}^n$ durch

$$t_j := t_{k(j)}^{(j)}, \qquad r_j := r_{k(j)}^{(j)} + t_{k(j)}^{(j)}(p^{(j)} - p).$$

Dann ist

$$x^* + t_j p + r_j = x^* + t_{k(j)}^{(j)} p^{(j)} + r_{k(j)}^{(j)} \in M \qquad \text{für alle } j \in \mathbb{N}$$

und

$$t_j = t_{k(j)}^{(j)} \to 0, \qquad \frac{r_j}{t_j} = \underbrace{\frac{r_{k(j)}^{(j)}}{t_{k(j)}^{(j)}}}_{\to 0} + \underbrace{p^{(j)} - p}_{\to 0} \to 0.$$

Insgesamt ist damit $p \in T(M; x^*)$, die Abgeschlossenheit des Tangentialkegels $T(M; x^*)$ ist damit bewiesen. □

Der folgende Satz gibt notwendige Optimalitätsbedingungen erster Ordnung bei einer *linear restringierten* Optimierungsaufgabe an.

Satz 1 *Gegeben sei die linear restringierte Optimierungsaufgabe*

(P) $\qquad$ *Minimiere* $\quad f(x) \quad$ *auf* $\quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\}.$

Hierbei seien $A \in \mathbb{R}^{l \times n}$, $A_0 \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^l$ und $b_0 \in \mathbb{R}^m$ gegeben. Ist $x^* \in M$ eine lokale Lösung von (P) und ist die Zielfunktion $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar, so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, sogenannte Lagrange-Multiplikatoren, mit

$$u^* \geq 0, \qquad \nabla f(x^*) + A^T u^* + A_0^T v^* = 0, \qquad (b - Ax^*)^T u^* = 0.$$

Beweis: Wir definieren die Menge der in x^* *aktiven* Ungleichungsrestriktionen durch

$$I^* := \{i \in \{1, \ldots, l\} : (Ax^*)_i = (b)_i\}.$$

Der Kegel $F(M; x^*)$ der zulässigen Richtungen an M in x^* kann leicht angegeben werden, es ist nämlich

$$F(M; x^*) = \{p \in \mathbb{R}^n : (Ap)_i \leq 0,\ i \in I^*,\ A_0 p = 0\}.$$

Es bezeichne $A_{I^*} \in \mathbb{R}^{|I^*| \times n}$ die Untermatrix von A, die nur zu I^* gehörende Zeilen enthält. Da $x^* \in M$ eine lokale Lösung von (P) ist, ist $\nabla f(x^*)^T p \geq 0$ für alle $p \in F(M; x^*)$ bzw. das System

$$\nabla f(x^*)^T p < 0, \qquad A_{I^*} p \leq 0, \qquad A_0 p = 0$$

nicht lösbar. Dies ist äquivalent dazu, dass

$$\begin{pmatrix} -A_{I^*} \\ -A_0 \\ A_0 \end{pmatrix} p \geq 0, \qquad \nabla f(x^*)^T p < 0$$

nicht lösbar ist. Das Farkas-Lemma (siehe Abschnitt 45) liefert die Existenz einer Lösung $(u_{I^*}, v_+, v_-) \in \mathbb{R}^{|I^*|} \times \mathbb{R}^m \times \mathbb{R}^m$ von

$$-A_{I^*}^T u_{I^*} - A_0^T(v_+ - v_-) = \begin{pmatrix} -A_{I^*} \\ -A_0 \\ A_0 \end{pmatrix}^T \begin{pmatrix} u_{I^*} \\ v_+ \\ v_- \end{pmatrix} = \nabla f(x^*), \qquad \begin{pmatrix} u_{I^*} \\ v_+ \\ v_- \end{pmatrix} \geq 0.$$

Definiert man $u^* = (u_i^*) \in \mathbb{R}^l$ durch

$$u_i^* := \begin{cases} u_i, & \text{falls } i \in I^*, \\ 0, & \text{sonst,} \end{cases} \qquad i = 1, \ldots, l,$$

und $v^* \in \mathbb{R}^m$ durch $v^* := v_+ - v_-$, so hat man in $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ ein gesuchtes Paar gefunden. □

Ist $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$, so wird mit $g_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ die i-te Komponentenabbildung bezeichnet, d. h. es ist $g(x) = (g_1(x), \ldots, g_l(x))^T$. Ist $g(\cdot)$ in x^* differenzierbar, so wird mit

$$g'(x^*) = \Big(\frac{\partial g_i}{\partial x_j}(x^*)\Big)_{\substack{i=1,\ldots,l \\ j=1,\ldots,n}} = \begin{pmatrix} \nabla g_1(x^*)^T \\ \vdots \\ \nabla g_l(x^*)^T \end{pmatrix} \in \mathbb{R}^{l \times n}$$

die *Funktionalmatrix* oder auch *Jacobi-Matrix* (engl.: Jacobian) von g in x^* bezeichnet. Entsprechende Bezeichnungen werden natürlich auch für eine Abbildung $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ benutzt.

Satz 2 *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ A_0 x = b_0\}$.

Hierbei seien $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *sowie* $A_0 \in \mathbb{R}^{m \times n}$, $b_0 \in \mathbb{R}^m$, *gegeben. Sei* $x^* \in M$ *eine lokale Lösung von (P) und* $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *in* x^* *stetig differenzierbar. Ferner sei*

(CQ) $$L_+(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p < 0,\ i \in I^*,\ A_0 p = 0\} \neq \emptyset,$$

wobei

$$I^* := \{i \in \{1, \ldots, l\} : g_i(x^*) = 0\}$$

die Menge der in x^* *aktiven Ungleichungsrestriktionen bezeichnet. Dann existiert ein Paar* $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, *sogenannte Lagrange-Multiplikatoren, mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + A_0^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Wir wollen uns überlegen, dass

$$L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0,\ i \in I^*,\ A_0 p = 0\} \subset T(M; x^*).$$

Da das System

$$\nabla f(x^*)^T p < 0, \qquad p \in L_0(M; x^*)$$

keine Lösung besitzt, folgt dann die Behauptung genau wie im letzten Satz.

Offenbar ist $L_+(M;x^*) \subset F(M;x^*)$. Sei $p \in L_0(M;x^*)$ und $\hat{p} \in L_+(M;x^*)$. Für alle $t \in (0,1]$ ist

$$(1-t)p + t\hat{p} \in L_+(M;x^*) \subset F(M;x^*)$$

und daher

$$p = \lim_{t \to 0+} ((1-t)p + t\hat{p}) \in \text{cl } F(M;x^*) \subset T(M;x^*),$$

wobei wir am Schluss die Abgeschlossenheit des Tangentialkegels benutzt haben. □

Bemerkung: Die Bedingung (CQ) in Satz 2 nennt man eine *Constraint Qualification* oder auch *Regularitätsbedingung*. Ohne eine solche Zusatzbedingung ist die Aussage bei nichtlinear restringierten Problemen i. Allg. nicht richtig. Hierzu betrachte man die Aufgabe

$$\text{Minimiere} \quad f(x) := -x_1 \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^2 : g(x) := \begin{pmatrix} -x_1 \\ -x_2 \\ x_2 - (1-x_1)^3 \end{pmatrix} \leq 0 \right\},$$

die offenbar die eindeutige Lösung $x^* = (1,0)^T$ besitzt. In Abbildung 72 stellen wir die Menge M der zulässigen Lösungen dar. Offenbar ist $I^* = \{2,3\}$ die Menge der in x^*

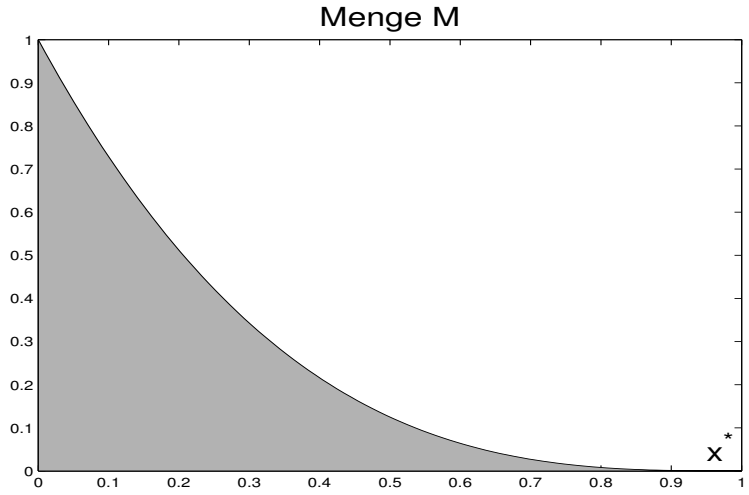


Abbildung 72: Eine Constraint Qualification ist nötig!

aktiven Restriktionen. Es ist

$$\nabla f(x^*) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \qquad \nabla g_2(x^*) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \qquad \nabla g_3(x^*) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Daher ist $-\nabla f(x^*)$ *keine* nichtnegative Linearkombination von $\nabla g_2(x^*)$, $\nabla g_3(x^*)$. Dies liegt daran, dass die Constraint Qualification (CQ) in Satz 2 nicht erfüllt ist. □

Beispiel: Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -x_1 \quad \text{unter den Nebenbedingungen} \\ g(x) := \begin{pmatrix} (x_1-1)^2 + (x_2 - \frac{1}{2})^2 - 2 \\ (x_1-1)^2 + (x_2 + \frac{1}{2})^2 - 2 \end{pmatrix} \leq 0. \end{cases}$$

In Abbildung 73 links stellen wir den zulässigen Bereich dar. Die Lösung von (P) ist

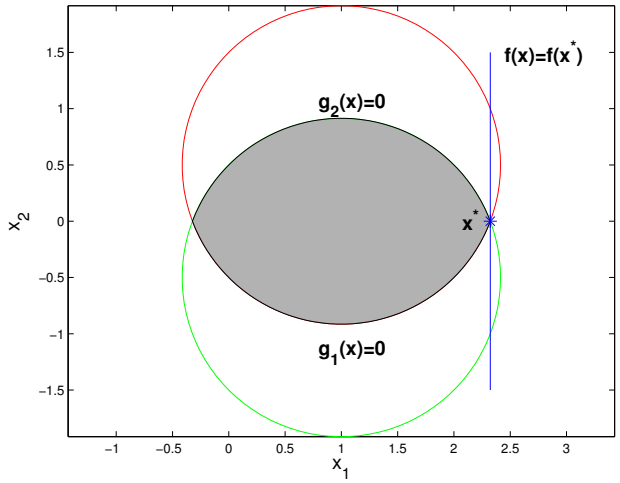


Abbildung 73: Ein nichtlinear restringiertes Optimierungsproblem

offenbar $x^* = (1 + \frac{1}{2}\sqrt{7}, 0)$. Das liest man aus Abbildung 73 ab. Beide Ungleichungen sind in x^* aktiv. Nach leichter Rechnung erhält man

$$-\nabla f(x^*) = \frac{1}{2\sqrt{7}} \nabla g_1(x^*) + \frac{1}{2\sqrt{7}} \nabla g_2(x^*),$$

d. h. $-\nabla f(x^*)$ ist eine nichtnegative (sogar positive) Linearkombination von $\nabla g_1(x^*)$ und $\nabla g_2(x^*)$. □

Zum Schluss geben wir Kuhn-Tucker-Bedingungen für eine Optimierungsaufgabe mit nichtlinearen Gleichungsrestriktionen an. Beim Beweis werden wir Aussagen benutzen, die wir zum Beweis der Lagrangeschen Multiplikatorenregel in Abschnitt 50 gewonnen haben.

Satz 3 *Sei x^* eine lokale Lösung von*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ seien auf einer Umgebung von x^* stetig differenzierbar. Mit $I^* := \{i \in \{1, \ldots, l\} : g_i(x^*) = 0\}$ werde die Indexmenge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Es sei

$$L_+(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p < 0,\ i \in I^*,\ h'(x^*)p = 0\} \neq \emptyset,$$

ferner sei Rang $h'(x^*) = m$. *Dann existiert ein Paar* $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, *sogenannte Lagrange-Multiplikatoren, mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Beim Beweis der Lagrangeschen Multiplikatorenregel in Abschnitt 50 haben wir (mit geringfügig anderen Bezeichnungen) unter der auch hier gemachten Voraussetzung, dass $h'(x^*)\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ surjektiv bzw. Rang $(h'(x^*)) = m$ ist, nachgewiesen, dass es zu jedem $p \in \mathbb{R}^n$ mit $h'(x^*)p = 0$ und einer beliebigen Nullfolge $\{t_k\} \subset \mathbb{R}_+$ eine Folge $\{r_k\} \subset \mathbb{R}^n$ mit $r_k/t_k \to 0$ und $h(x^* + t_k p + r_k) = 0$, $k = 0, 1, \ldots$, existiert. Hieraus folgt sofort, dass $L_+(M; x^*) \subset T(M; x^*)$. Wegen der Abgeschlossenheit des Tangentialkegels ist

$$\begin{aligned} L_0(M; x^*) &:= \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0,\ i \in I^*,\ h'(x^*)p = 0\} \\ &\subset \operatorname{cl} L_+(M; x^*) \\ &\subset T(M; x^*). \end{aligned}$$

Man erhält die Behauptung mit Hilfe des Farkas Lemmas wie im Beweis von Satz 1. □

Nur erwähnt sei, dass die Kuhn-Tucker-Bedingungen bei einer *konvexen Optimierungsaufgabe* (d. h. die Zielfunktion und die Ungleichungsrestriktionsabbildungen sind konvex, die Gleichungsrestriktionen sind affin linear) hinreichend für (globale) Optimalität sind. Dies ist ziemlich einfach zu beweisen.

54 Die kreisrunde Wiese von Bauer Lindemann

Gibt man bei Google das Stichwort *kreisrunde Wiese* ein, so stößt man auf die folgende Aufgabe[68]:

- Bauer Lindemann hat eine kreisrunde Wiese mit Radius R. Er kauft sich eine Ziege und will diese so am Rand der Wiese anpflocken, dass die Ziege genau die Hälfte der Wiese abgrasen kann. Wie lang muss das Seil sein?

In Abbildung 74 verdeutlichen wir uns die Situation. Die Wiese sei der Kreis mit dem Mittelpunkt M und dem Radius R. Die Ziege ist im Punkt B angepflockt, ihr Seil habe die Länge r. Die beiden Kreise mögen sich in C und D schneiden. In H sei das Lot von C auf die Seite MH, sei $h := |CH|$, ferner $qR := |HB|$ und daher $(1-q)R = |MH|$. Schließlich sei $\beta := \sphericalangle(CBM)$ und $\delta := \sphericalangle(BMC)$ (jeweils im Bogenmaß gemessen).

Zu berechnen ist der Flächeninhalt F des Durchschnitts der beiden Kreise. Dieser setzt sich zusammen aus den Flächeninhalt des Kreissektors BCD (in der Abbildung gestrichelt gezeichnet) plus der Summe der Flächeninhalte der beiden Kreisabschnitte BC und BD im Kreis um M mit dem Radius R. Die Summe der beiden letzteren ist

[68] Diese Aufgabe heißt manchmal auch das *Ziegenproblem*, was nicht mit dem in Abschnitt 34 geschilderten Problem verwechselt werden sollte.

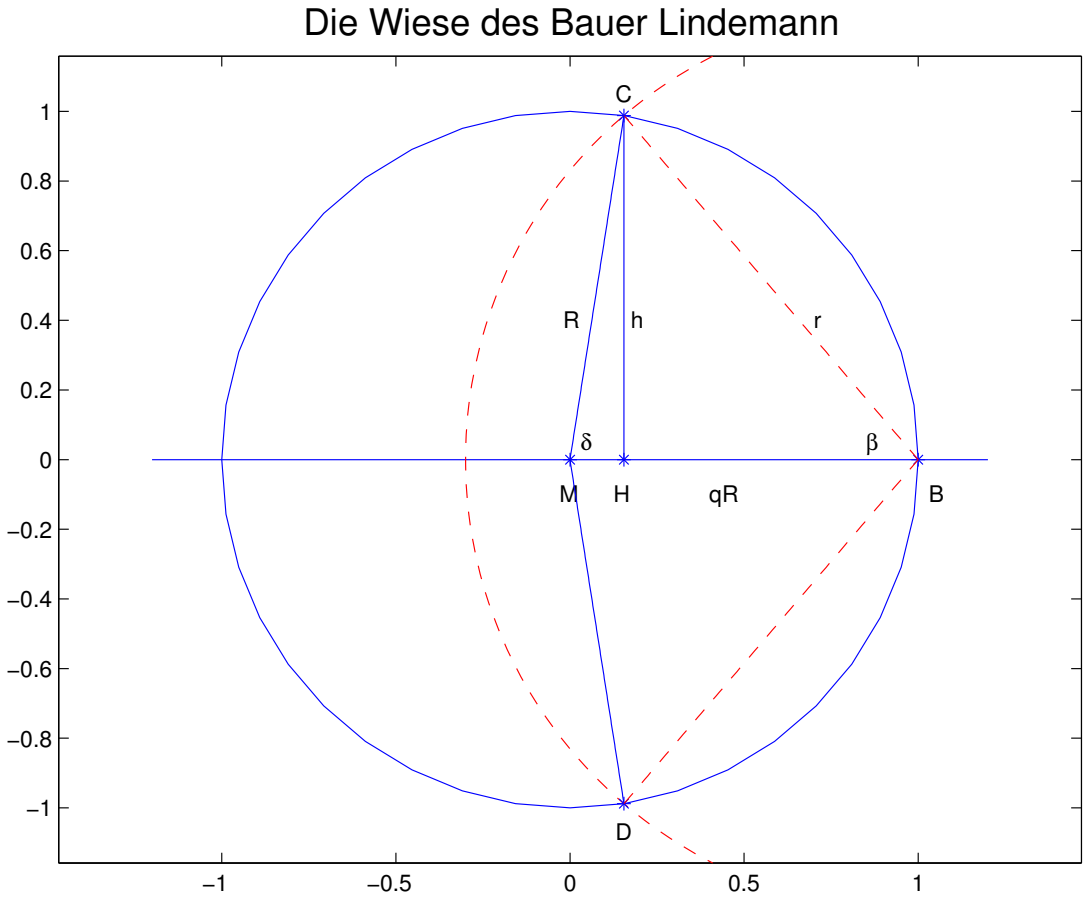


Abbildung 74: Eine kreisrunde Wiese

aber der Flächeninhalt des Kreissektors MCD minus der Summe der Flächeninhalte der beiden Dreiecke $\triangle MCB$ und $\triangle MDB$. Daher ist

$$F = \beta r^2 + \delta R^2 - hR.$$

Der Satz von Pythagoras, angewandt auf die beiden Dreiecke $\triangle MHC$ und $\triangle MCB$, liefert die beiden Beziehungen

$$h^2 + (1-q)^2 R^2 = R^2, \qquad h^2 + q^2 R^2 = r^2.$$

Hieraus erhält man

$$q = \frac{1}{2}\Big(\frac{r}{R}\Big)^2, \qquad h = r\Big[1 - \frac{1}{4}\Big(\frac{r}{R}\Big)^2\Big]^{1/2}.$$

Wir führen die Bestimmung des gesuchten Radius r auf die Bestimmung des Winkels β zurück und beachten hierzu, dass die Winkelsumme im gleichschenkligen Dreieck $\triangle BMC$ gerade (im Bogenmaß) π ist, so dass $\delta + 2\beta = \pi$ bzw. $\delta = \pi - 2\beta$. Weiter beachte man, dass

$$\cos\beta = \frac{qR}{r} = \frac{1}{2}\Big(\frac{r}{R}\Big),$$

so dass $r = 2R\cos\beta$. Dann kann der Flächeninhalt F alleine in Abhängigkeit des Radius R des gegebenen Kreises und des Winkels β angegeben werden:

$$\begin{aligned} F &= \beta r^2 + \delta R^2 - rR\Big[1 - \frac{1}{4}\Big(\frac{r}{R}\Big)^2\Big]^{1/2} \\ &= 4R^2\beta\cos 2\beta + (\pi - 2\beta)R^2 - 2R^2\cos\beta\sin\beta \\ &= R^2[2\beta(1+\cos(2\beta)) + \pi - 2\beta - \sin(2\beta)] \\ &= R^2[2\beta\cos(2\beta) + \pi - \sin(2\beta)]. \end{aligned}$$

Die Forderung, dass der Flächeninhalt F die Hälfte des Flächeninhaltes des Kreises mit dem Radius R ist, führt auf die Gleichung

$$R^2[2\beta\cos(2\beta) + \pi - \sin(2\beta)] = \frac{1}{2}\pi R^2.$$

Der Winkel $\beta \in (0, \pi/2)$ ist also als Lösung von

$$2\beta\cos(2\beta) + \frac{\pi}{2} - \sin(2\beta) = 0$$

zu bestimmen. Macht man die Variablentransformation $\gamma = 2\beta$, so ist also eine Nullstelle γ^* von

$$\phi(\gamma) := \gamma\cos\gamma + \frac{\pi}{2} - \sin\gamma$$

in $(0, \pi)$ zu bestimmen, anschließend ist der gesuchte Radius r^* durch

$$r^* = 2R\cos\frac{\gamma^*}{2}$$

gegeben. Die mathematische Modellierung ist damit abgeschlossen. Das gegebene Problem ist auf ein Standardproblem der numerischen Mathematik, nämlich die Bestimmung einer Nullstelle einer bestimmten Funktion, zurückgeführt worden.

In Abbildung 75 geben wir einen Plot von $\phi(\cdot)$ über dem Intervall $[0, \pi]$ an. An

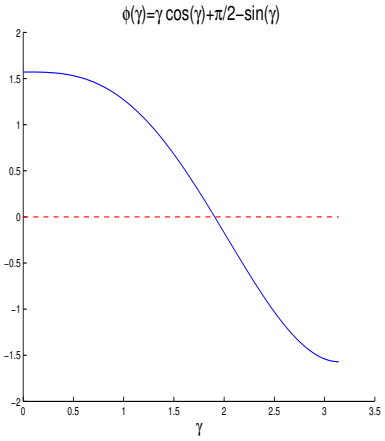


Abbildung 75: Die Funktion $\phi(\gamma) = \gamma\cos\gamma + \pi/2 - \sin\gamma$

diesem Plot erkennt man, dass ϕ genau eine Nullstelle $\gamma^* \approx 2$ im Intervall $(0, \pi)$ besitzt. Ein exakter Beweis für ersteres ist einfach, denn es ist $\phi(0) = \pi/2 > 0$, $\phi(\pi) = -\pi/2$, so dass ϕ wegen des Zwischenwertsatzes in $(0, \pi)$ eine Nullstelle γ^* besitzt. Ferner ist $\phi'(\gamma) = -\gamma \sin(\gamma) < 0$ auf $(0, \pi)$, also ist ϕ auf $(0, \pi)$ monoton fallend und die Nullstelle γ^* ist eindeutig. Damit ist die Existenz und Eindeutigkeit einer Lösung bewiesen.

Nun kommt die numerische Mathematik ins Spiel. Wie kann die Nullstelle einer nichtlinearen Gleichung bestimmt werden? Natürlich kann man es sich einfach machen und ein mathematisches Anwendersystem bemühen. Benutzt man z. B. MATLAB, so erhält man etwa

```
>> format long
>> gamma_stern=fzero(@(gamma) gamma*cos(gamma)+pi/2-sin(gamma),2)

gamma_stern =

   1.905695729309884

>> faktor=2*cos(gamma_stern/2)

faktor =

   1.158728473018122
```

Benutzt man MuPAD und wendet man auf die Nullstellenaufgabe $\phi(\gamma) = 0$ das Newton-Verfahren an, iteriert man also mit einer Startnäherung γ_0, etwa $\gamma_0 := 2$, gemäß $\gamma_{k+1} := \gamma_k - \phi(\gamma_k)/\phi'(\gamma_k)$, so erhält man mit `DIGITS:=20` die folgenden Werte:

k	γ_k
0	2.0
1	1.9060842095851370242
2	1.9056957426183945333
3	1.9056957293098839105
4	1.9056957293098838949
5	1.9056957293098838949

55 Wie rechnete Adam Ries(e)?

Jeder hat den Ausspruch „Macht nach Adam Ries(e)“ schon einmal gemacht oder wenigstens gehört. Er soll die Richtigkeit eines einfachen Ergebnisses besonders deutlich hervorheben. Wer war aber Adam Ries (die Schreibweisen seines Nachnamens sind nicht einheitlich, wir werden Ries benutzen)? Adam Ries, geboren 1492 in Staffelstein in Oberfranken, gestorben 1559 in Annaberg im Erzgebirge, war ein deutscher Rechenmeister, siehe `http://de.wikipedia.org/wiki/Adam_Ries`. Im Internet findet man sehr viel über ihn, das wollen wir hier nicht wiederholen. Uns interessieren aber die Rechenmethoden, die er in seinen einflussreichen Büchern *Rechenung auff der linihen (1518)* und *Rechenung auff der linihen und federn... (1522)* dargestellt hat. Motiviert, diesen

Abschnitt zu schreiben, bin ich vor allem dadurch, dass ich Anfang September 2010 mit meinem Bruder Bodo (auch Mathematiker) das Adam-Ries-Museum in Annaberg besuchte. Wer keine Gelegenheit hat, nach Annaberg zu fahren, um sich dieses kleine, sehr empfehlenswerte Museum anzusehen, sollte sich wenigstens den Internet-Auftritt des Museums nicht entgehen lassen, siehe `http://www.adam-ries-museum.de/`.

Zunächst muss man wissen, wie man Zahlen *auf Linien* darstellt. Auf dem Rechenbrett oder Rechenfeld befinden sich mindestens vier Linien, die von unten nach oben die Wertigkeit 1, 10, 100 und 1000 haben. Es kann aber natürlich auch noch eine weitere Linie der Wertigkeit 10000 hinzukommen. Die Zwischenräume (Bezeichnung: spacio) haben die Wertigkeit 5, 50, 500 (und 5000). Eine Zahl wird auf dem Rechenbrett durch Rechenpfennige auf den Linien und den Zwischenräumen repräsentiert. Auf jeder Linie dürfen höchstens vier, auf jedem Zwischenraum darf höchstens ein Rechenpfennig liegen. Es gelten die folgenden Ersetzungsregeln für das Höher- oder Tiefersetzen (Elevieren bzw. Resolvieren) von Rechenpfennigen:

(H1) Zwei Rechenpfennige, die in einem Zwischenraum liegen, können weggenommen werden. Dafür wird dann ein Rechenpfennig auf die nächsthöhere Linie gelegt.

(H2) Fünf Rechenpfennige, die auf einer Linie liegen, können weggenommen werden. Dafür wird dann ein Rechenpfennig in den nächsthöheren Zwischenraum gelegt.

(T1) Ein Rechenpfennig, der auf einer Linie liegt, kann weggenommen werden. Dafür werden dann zwei Rechenpfennige in den nächsttieferen Zwischenraum gelegt.

(T2) Ein Rechenpfennig, der in einem Zwischenraum liegt, kann weggenommen werden. Dafür werden dann fünf Rechenpfennige auf die nächsttiefere Linie gelegt.

(T3) Ein Rechenpfennig, der auf einer Linie liegt, kann weggenommen werden. Dafür werden dann ein Rechenpfennig in den nächsttieferen Zwischenraum und fünf Rechenpfennige auf die nächsttiefere Linie gelegt.

In Abbildung 76 demonstrieren wir die Addition zweier Zahlen. Wir addieren die

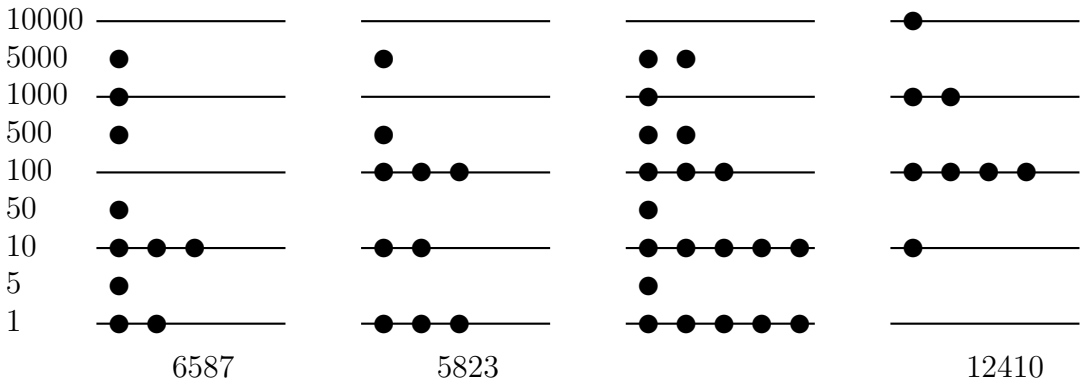


Abbildung 76: Addition der Zahlen 6587 und 5823

Rechenpfennige auf den Linien und den Zwischenräumen, wobei aus fünf Pfennigen

auf einer Linie ein Pfennig auf dem darüberliegenden Zwischenraum (Ersetzungsregel (H2)) und aus zwei Pfennigen auf einem Zwischenraum ein Pfennig auf der darüberliegenden Linie wird (Ersetzungsregel (H1)). In welcher Reihenfolge das Höherlegen von Pfennigen geschieht ist unerheblich. Man sehe sich zur Addition zweier Zahlen nach Adam Ries das Applet `http://www.adam-ries-bund.de/eradd.html` an. Nun kommen wir zur Subtraktion beim Rechnen auf Linien, siehe das Applet `http://www.adam-ries-bund.de/ersub.html`. Wir nehmen an, der Minuend, also die Zahl, von der etwas abgezogen wird, sei größer als der Subtrahend. Beide seien natürliche Zahlen. Wir geben wieder nur ein Beispiel an, siehe Abbildung 77. Durch Tieferlegen

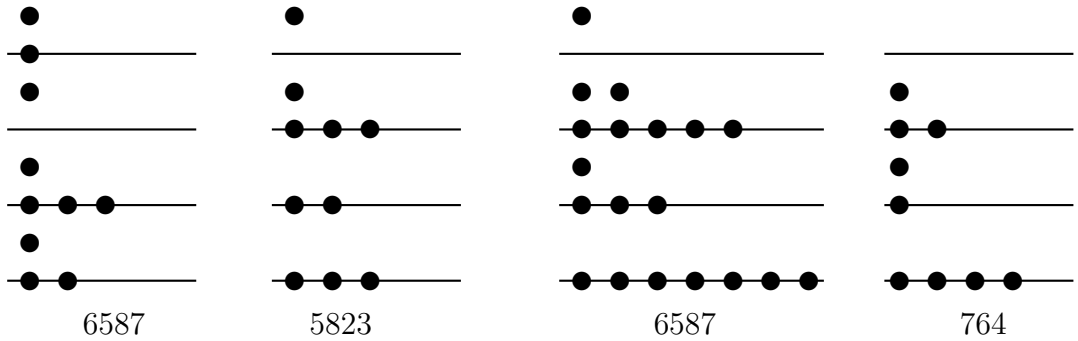


Abbildung 77: Berechne $6587 - 5823 = 764$

von Pfennigen sorgen wir dafür, dass auf jeder Linie und in jedem Zwischenraum der Minuend nicht weniger Rechenpfennige enthält als der Subtrahend. Man kann aber auch schon in einem ersten Schritt auf allen Linien und Zwischenräumen, bei denen links nicht weniger Pfennige als rechts beim Subtrahenden liegen, dieselbe Anzahl von Pfennigen entfernen. Anschließend ist das Ergebnis der Subtraktion einfach abzulesen.

Etwas komplizierter wird das Rechnen auf den Linien bei der Multiplikation. Auch hier schildern wir die Methode lediglich anhand eines Beispiels. Es sei $368 \cdot 123$ zu berechnen. Diesmal legen wir nur Pfennige für den ersten Multiplikator auf die Linien und merken uns den zweiten Faktor. Im Prinzip berechnen wir

$$368 \cdot 123 = 368 \cdot 100 + 368 \cdot 20 + 368 \cdot 3.$$

Das erste Produkt $368 \cdot 100$ erhalten wir, indem wir alle Pfennige jeweils zwei Linien höher legen. Das zweite Produkt $368 \cdot 20$ wird erhalten, in dem wir alle Pfennige jeweils eine Linie nach oben schieben und alle Pfennige auf den Zwischenräumen und Linien verdoppeln und dann die Ersetzungsregeln (H1) und (H2) anwenden. Entsprechend wird $368 \cdot 3$ berechnet. Wir illustrieren dies in Abbildung 78. Die Division beim Rechnen auf den Linien wird auf die mehrfache Subtraktion des Divisors vom Dividenden zurückgeführt. Hierauf wollen wir aber nicht mehr eingehen.

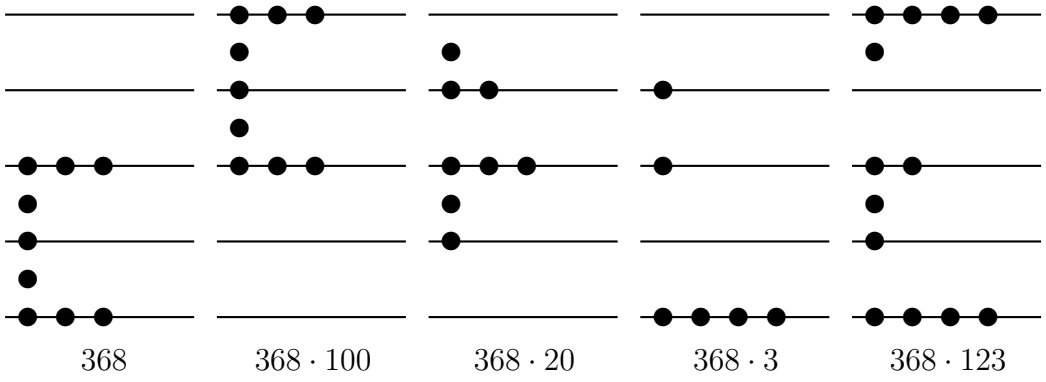


Abbildung 78: Berechne $368 \cdot 123 = 45264$

56 Die Neunerprobe

Die *Neunerprobe* war schon Leonardo da Pisa (Fibonacci) und Adam Ries bekannt. Mit ihr kann getestet werden, ob eine Addition, Division oder Multiplikation innerhalb ganzer Zahlen korrekt durchgeführt wurde. Wird der Test nicht bestanden, so war die Rechnung falsch. Wird der Test bestanden, so *kann* die Rechnung richtig sein, *muss* es aber nicht. Wichtig für das weitere ist der sogenannte *Neunerrest*[69], d. h. der Rest $r \in \{0, 1, \ldots, 8\}$, der sich bei der Division einer ganzen Zahl durch 9 ergibt. Allgemein gilt die Aussage[70]:

Satz (Division mit Rest) *Sei $a \in \mathbb{Z}$ und $b \in \mathbb{N}$. Dann existieren eindeutig bestimmte Zahlen q, $r \in \mathbb{Z}$ mit*

$$a = b \cdot q + r, \qquad 0 \leq r < b.$$

Wir schreiben auch $r = a \bmod b$ für den Rest bei der Division von a durch b..

Beweis: Wir beweisen zunächst die *Existenz* von q und r mit den angegebenen Eigenschaften für $a \geq 0$ durch vollständige Induktion nach a. Der Induktionsanfang liegt bei $a = 0$. Hier können wir $q := 0$ und $r := 0$ setzen. Nehmen wir an, die Aussage sei für nichtnegative ganze Zahlen $\leq a$ richtig. Ist $a + 1 < b$, so haben wir durch $a + 1 = b \cdot 0 + (a + 1)$ eine gewünschte Darstellung erhalten.Ist dagegen $a + 1 \geq b$, so wenden wir auf $a + 1 - b \leq a$ die Induktionsannahme an, gehen also von einer Darstellung

$$a + 1 - b = b \cdot q + r, \qquad 0 \leq r < b$$

aus. Also ist $a + 1 = b \cdot (q + 1) + r$ die gewünschte Darstellung von $a + 1$. Damit ist die Existenzaussage für nichtnegative ganze Zahlen a bewiesen. Nun nehmen wir an, a sei eine negative ganze Zahl. Eine Anwendung der gerade eben bewiesenen Aussage auf $-a$ liefert die Existenz ganzer Zahlen q', r' mit

$$-a = b \cdot q' + r', \qquad 0 \leq r' < b.$$

[69] Der Neunerrest wird von Adam Ries *prob* genannt.

[70] Mit $\mathbb{Z}$ bezeichnen wir die Menge der ganzen Zahlen und mit $\mathbb{N}$ die Menge der natürlichen Zahlen.

Für a erhalten wir dann die gewünschte Darstellung

$$a = \begin{cases} b \cdot (-q' - 1) + (b - r'), & \text{falls } r' > 0, \\ b \cdot (-q'), & \text{falls } r' = 0. \end{cases}$$

Und nun kommen wir zum *Eindeutigkeitsbeweis*. Hierzu nehmen wir an, dass es ein zweites Paar (q', r') ganzer Zahlen mit $a = b \cdot q' + r'$, $0 \leq r' < b$ gibt. O. B. d. A. ist $r \geq r'$ (notfalls vertausche man die Rollen von (q, r) und (q', r')). Dann ist $0 \leq r - r' = -b \cdot (q - q')$ und daher $q - q' \leq 0$. Wegen $r - r' < b$ ist andererseits $b(1 + q - q') > 0$ und daher $q - q' \geq 0$. Insgesamt ist $q = q'$ und dann auch $r = r'$. Insgesamt ist die obige Aussage bewiesen. □

Die *Quersumme* einer Zahl

$$a := \pm \sum_{i=0}^{n} a_i \cdot 10^i, \qquad a_i \in \{0, 1, \ldots, 9\},$$

ist bekanntlich gegeben durch die Summe ihrer Ziffern (in der Dezimaldarstellung), also durch

$$Q(a) := \pm \sum_{i=0}^{n} a_i.$$

Wir wollen uns überlegen, dass die folgende Aussage richtig ist:

Satz *Der Neunerrest einer Zahl $a \in \mathbb{Z}$ stimmt mit dem Neunerrest ihrer Quersumme $Q(a)$ überein. Insbesondere ist eine ganze Zahl genau dann durch 9 teilbar, wenn ihre Quersumme es ist.*

Beweis: Wir beschränken uns zunächst auf *nichtnegative* ganze Zahlen und nehmen daher an, es sei

$$a = \sum_{i=0}^{n} a_i \cdot 10^i, \qquad a_i \in \{0, \ldots, 9\}.$$

Es ist

$$10^i = 9 \cdot \frac{10^i - 1}{10 - 1} + 1 = 9 \cdot \sum_{k=0}^{i-1} 10^k + 1$$

und daher

$$a = \sum_{i=0}^{n} a_i \cdot 10^i = 9 \cdot \sum_{i=0}^{n} a_i \cdot \sum_{k=0}^{i-1} 10^k + Q(a).$$

Wegen des vorigen Satzes über die Division mit Rest hat $Q(a)$ die eindeutige Darstellung $Q(a) = 9 \cdot q + r$ mit ganzen Zahlen q, r mit $0 \leq r < 9$. Also ist

$$a = 9 \cdot \Big(\sum_{i=0}^{n} a_i \cdot \sum_{k=0}^{i-1} 10^k + q \Big) + r, \qquad 0 \leq r < 9.$$

Hieraus liest man ab, dass r nicht nur Neunerrest von $Q(a)$, sondern auch von a ist. Da der Neunerrest eindeutig bestimmt ist, stimmen die Neunerreste von a und $Q(a)$ überein.

Sei nun a eine negative ganze Zahl. Wegen des gerade eben bewiesenen ersten Teils des Beweises ist der Neunerrest r' von $-a$ derselbe wie der von $-Q(a)$. Dann ist aber $9 - r'$ (für $r' > 0$) bzw. 0 (für $r' = 0$) der Neunerrest von a bzw. $Q(a)$ und die Aussage ist bewiesen. □

Der Neunerrest einer ganzen Zahl stimmt mit dem Neunerrest der Quersumme $Q(a)$ von a überein[71]. Dasselbe gilt für die $Q^2(a) = Q(Q(a))$, also die Quersumme der Quersumme. So kann man fortfahren bis man zu einer Zahl mit nur einer Ziffer kommt. Ist diese kleiner als 9, so hat man den Neunerrest erhalten, andernfalls ist dieser gleich 0.

Beispiel: Sei $a := 1\,447\,827\,112$. Dann ist $a = 9 \cdot 160\,869\,679 + 1$, der Neunerrest von a ist also 1. Weiter ist $Q(a) = 37$, $Q^2(a) = 10$ und $Q^3(a) = 1$, was (natürlich) mit obigem Ergebnis übereinstimmt. Der Neunerrest einer ganzen Zahl kann also ohne eine Division berechnet werden. □

Die folgende Aussage ist Grundlage für die Neunerprobe. Mit $r_9(a)$ bezeichnen wir hierbei den Neunerrest von $a \in \mathbb{Z}$.

Satz *Seien $a, b \in \mathbb{Z}$. Dann ist*

1. $r_9(a + b) = r_9(r_9(a) + r_9(b))$,
2. $r_9(a \cdot b) = r_9(r_9(a) \cdot r_9(b))$.

Beweis: Sei $a = 9 \cdot q_a + r_9(a)$, $b = 9 \cdot q_b + r_9(b)$ mit ganzen Zahlen q_a, q_b. Dann ist

$$a + b = 9 \cdot (q_a + q_b) + (r_9(a) + r_9(b)).$$

Daher stimmt der Neunerrest von $a + b$ mit dem von $r_9(a) + r_9(b)$ überein, d. h. es ist $r_9(a+b) = r_9(r_9(a)+r_9(b))$. Ganz ähnlich funktioniert der Beweis für die entsprechende Aussage bei der Multiplikation ganzer Zahlen. Denn es ist

$$a \cdot b = 9 \cdot (9q_aq_b + q_ar_9(b) + q_br_9(a)) + r_9(a) \cdot r_9(b).$$

Daher stimmen die Neunerreste von $a \cdot b$ und $r_9(a) \cdot r_9(b)$ überein, d. h. es ist $r_9(a \cdot b) = r_9(r_9(a) \cdot r_9(b))$. Damit ist der einfache Satz bewiesen. □

Bemerkung: Bezeichnet man mit $r_n(a)$ den Rest bezüglich $n \in \mathbb{N}$ bei Division von $a \in \mathbb{Z}$ durch n, so gelten die obigen Aussagen natürlich ganz entsprechend. Für $n = 9$ hat man lediglich den Vorteil, dass sich der Rest über Quersummen einfach berechnen lässt. Ähnlich ist dies für $n = 11$. Grundlage ist hierbei die leicht durch vollständige Induktion beweisbare Beziehung

$$\frac{10^i - (-1)^i}{11} = (-1)^{i+1} \sum_{k=0}^{i-1} (-10)^k.$$

[71] Aus der Gleichung

$$a = 3 \cdot 3 \sum_{i=0}^{n} a_i \cdot \sum_{k=0}^{i-1} 10^k + Q(a)$$

liest man natürlich auch ab, dass der Dreierrest von a mit dem von $Q(a)$ übereinstimmt. Daher ist eine ganze Zahl genau dann durch 3 teilbar, wenn es die Quersumme ist.

Für eine nichtnegative ganze Zahl a mit

$$a = \sum_{i=0}^{n} a_i \cdot 10^i, \qquad a_i \in \{0, \ldots, 9\}$$

ist dann

$$\begin{aligned} a &= \sum_{i=0}^{n} a_i \cdot \left(11 \cdot \frac{10^i - (-1)^i}{11} + (-1)^i\right) \\ &= \underbrace{11 \cdot \sum_{i=0}^{n} (-1)^{i+1} a_i \cdot \sum_{k=0}^{i-1} (-10)^k}_{\in \mathbb{Z}} + \underbrace{\sum_{i=0}^{n} (-1)^i a_i}_{=:Q_\pm(a)}. \end{aligned}$$

Definiert man daher die *alternierende Quersumme* von a durch

$$Q_\pm(a) := \sum_{i=0}^{n} (-1)^i a_i,$$

so besitzen a und $Q_\pm(a)$ denselben Elferrest. Insbesondere ist a genau dann durch 11 teilbar, wenn es $Q_\pm(a)$ ist. Z. B. ist $a := 918\,291$ durch 11 teilbar, weil $Q_\pm(a) = (1 + 2 + 1) - (9 + 8 + 9) = -22$ durch 11 teilbar ist. □

Beispiel: Wir zeigen, dass $42\,454 \cdot 34\,103 \neq 1\,447\,827\,112$ ohne die Multiplikation durchzuführen. Sei also $a := 42\,454$, $b := 34\,103$ und $c := 1\,447\,827\,112$. Dann ist $r_9(a) = 1$, $r_9(b) = 2$, $r_9(c) = 1$ und daher $a \cdot b \neq c$, da

$$r_9(a \cdot b) = r_9(r_9(a) \cdot r_9(b)) = 2 \neq 1 = r_9(c).$$

□

57 Wofür steht ISBN? Was ist eine IBAN?

Unter `http://de.wikipedia.org/wiki/Internationale_Standardbuchnummer` können wir nachlesen: Die *Internationale Standardbuchnummer* (International Standard Book Number), abgekürzt ISBN, ist eine Nummer zur eindeutigen Kennzeichnung von Büchern und anderen selbstständigen Veröffentlichungen mit redaktionellem Anteil, wie beispielsweise Multimedia-Produkte und Software. Seit dem 1. Januar 2007 ist die Angabe der ISBN-13 verbindlich, ferner wurden die bisherigen 10-stelligen ISBN (ISBN-10) zu 13-stelligen ISBN (ISBN-13) erweitert, damit der ISBN-Zahlenraum nicht zur Neige geht.

Beispiele: *Krieg und Frieden* von Lew Tolstoi in der Neuübersetzung von Barbara Conrad ist in zwei Bänden, die nicht getrennt gekauft werden können, bei Hanser erschienen. Beide haben die ISBN 978-3-446-23575-5. Der erste Band von *Auf der Suche nach der verlorenen Zeit* von Marcel Proust in der bei Suhrkamp herausgekommenen Frankfurter Ausgabe hat die ISBN 3-518-2777-8, der zweite Band hat eine andere ISBN.

Die beiden Bände können aber auch getrennt erworben werden. *Maigret in New York* von Georges Simenon, Band 27 sämtlicher Maigret-Romane, hat die ISBN 978-3-257-23827-3, erschienen bei Diogenes in Zürich. Der Band 46 aus demselben Verlag hat die ISBN 978-3-257-23846-4. *The TEXbook* von Donald E. Knuth ist bei Addison Wesley herausgekommen und hat die ISBN 0-201-13447-0. □

Wie wir schon wissen, ist die ISBN eine 10- oder 13-stellige Zahl. Sie ist auf die folgende Weise formatiert, wobei die Trennstriche auch wegfallen können:

ISBN (Präfix–)Gruppennummer–Verlagsnummer–Titelnummer–Prüfziffer

Das *Präfix* ist je nach Buch 978 oder 979. Bei der ISBN-10 gab es kein Präfix. Die *Gruppennummer* (auch *Ländernummer* genannt) ist eine Kennzahl für eine nationale, geographische, Sprach- oder sonstige geeignete Gruppe. Sie ist z. B. 0 oder 1 für den englischsprachigen Raum, 3 für den deutschsprachigen Raum (Deutschland, Österreich und die deutschsprachige Schweiz). Die *Verlagsnummer* ist eine Kennzahl für den Verlag. Z. B. steht 446 für den Hanser Verlag, 518 für den Suhrkamp Verlag und 8270 für den Berlin Verlag. Dann folgt die vom Verlag vergebene Titelnummer (auch Bandnummer genannt). Der Verlag ist frei in der Verwendung, allerdings müssen verschiedene Produkte differenziert werden, also separat verkäufliche Bände, unterschiedliche Einbände usw. Zum Schluss wird eine *Prüfziffer* angegeben. Die Prüfziffer ermöglicht das Erkennen von Eingabe- und Lesefehlern: Erkannt werden ein Einzelfehler (genau eine Ziffer falsch) und die meisten Vertauschungen von zwei Nachbarziffern[72]. Nur bei der Berechnung der Prüfziffer aus den 12 (bei ISBN-13) oder 9 (bei ISBN-10) anderen Ziffern und dem Nachweis, dass durch die Prüfziffer gewisse Fehler bei der Datenerfassung erkannt werden können, kommt ein wenig Mathematik vor.

Zunächst betrachten wir die Berechnung der Prüfziffer bei ISBN-13. Mit $z_1, \ldots, z_{12}$ bezeichnen wir die ersten 12 Ziffern (von links nach rechts gelesen) der ISBN, zu berechnen ist die Prüfziffer z_{13}. Dies geschieht durch

$$z_{13} := \{10 - [(z_1 + z_3 + z_5 + z_7 + z_9 + z_{11}) + 3 \cdot (z_2 + z_4 + z_6 + z_8 + z_{10} + z_{12})] \bmod 10\} \bmod 10.$$

Hierbei ist $a \bmod 10 \in \{0, \ldots, 9\}$ der Rest von a bei Division durch 10. Die zweite modulo-Operation sichert, dass $z_{13} \in \{0, \ldots, 9\}$ (andernfalls wäre auch 10 als Prüfziffer möglich).

Beispiel: Wir berechnen die Prüfziffer zu 978-3-446-23575-? und erhalten

$$\begin{aligned} z_{13} &= \{10 - [(9 + 8 + 4 + 6 + 3 + 7) + 3 \cdot (7 + 3 + 4 + 2 + 5 + 5)] \bmod 10\} \bmod 10 \\ &= \{10 - [37 + 3 \cdot 26] \bmod 10\} \bmod 10 \\ &= \{10 - 115 \bmod 10\} \bmod 10 \\ &= \{10 - 5\} \bmod 10 \\ &= 5. \end{aligned}$$

Die Prüfziffer ist also 5, die korrekte ISBN ist ISBN 978-3-446-23575-5. □

[72] Wir folgen nach wie vor wörtlich der obigen Wikipedia-Quelle.

Nun kommen wir zur Berechnung der Prüfziffer bei ISBN-10. Bezeichnen wir die ersten neun Ziffern mit $z_1, \ldots, z_9$ (wieder von links nach rechts gelesen), so ist die Prüfziffer an der zehnten Stelle gegeben durch

$$z_{10} := \Big(\sum_{i=1}^{9} i \cdot z_i\Big) \bmod 11.$$

Bei einem Ergebnis von 0 bis 9 wird daraus unmittelbar die Prüfziffer; ergibt die Formel den Wert 10, wird ein X als letztes Zeichen verwendet, welches als römische Zahl 10 interpretiert werden kann. Ist also $z_1 \cdots z_9 z_{10}$ eine *korrekte* ISBN-10-Zahl, so ist $\sum_{i=1}^{10} i z_i$ eine durch 11 teilbare Zahl, also $\sum_{i=1}^{10} i z_i \equiv 0 \bmod 11$. Denn es ist $\sum_{i=1}^{9} i z_i = z_{10} + p \cdot 11$ mit $z_{10} \in \{0, \ldots, 10\}$ und $p \in \mathbb{Z}$. Daher ist

$$\begin{aligned} \sum_{i=1}^{10} i z_i &= \sum_{i=1}^{9} i z_i + 10 \cdot 10 \\ &= 11 \cdot z_{10} + p \cdot 11 \\ &= (p + z_{10}) \cdot 11, \end{aligned}$$

also

$$\sum_{i=1}^{10} i z_i \equiv 0 \bmod 11.$$

Beispiel: Wir berechnen die Prüfziffer zu 3-936384-07-? und erhalten

$$z_{10} = 208 \bmod 11 = 10.$$

Die Prüfziffer ist also 10 bzw. X. Die korrekte ISBN ist 3-936384-07-X, das entsprechende Buch ist *Venedig unter vier Augen* von Louis Begley und Anka Muhlstein im Marebuchverlag. □

Nun wollen wir etwas genauer darauf eingehen, welche Darstellungsfehler durch die Prüfziffer erkannt werden. Was soll das überhaupt heißen? Bei der Datenerfassung treten die folgenden Fehlertypen am häufigsten auf:

1. Einzelfehler: Eine Ziffer a wird durch eine andere Ziffer b (aber an derselben Stelle) ersetzt.
2. Zifferndreher: Aufeinanderfolgende Ziffern ab werden durch ba ersetzt.
3. Sprungtransposition: Aufeinanderfolgende Ziffern acb werden durch bca ersetzt.
4. Zwillingsfehler: Zwei gleiche, aufeinanderfolgende Ziffern aa werden durch bb ersetzt.

Wir sagen nun, der ISBN-10-Code *erkenne einen Fehler*, z. B. einen Einzelfehler, wenn durch den Fehler aus einer korrekten eine inkorrekte ISBN-Zahl entsteht. Damit in diesem Abschnitt ein mathematischer Satz enthalten ist, formulieren wir:

Satz *Der ISBN-10-Code entdeckt die Fehler vom Typ 1, 2 und 3. Ein Fehler vom Typ 4 wird erkannt, es sei denn, der Fehler tritt an den Stellen 5 und 6 auf.*

Beweis: Die Zahl $z_1' \cdots z_9' z_{10}'$ entstehe aus der korrekten ISBN-Zahl $z_1 \cdots z_9 z_{10}$ durch einen Einzelfehler, indem die j-te Ziffer $z_j = a$ durch b ersetzt wird. Dann ist

$$\sum_{i=1}^{10} i z_i - \sum_{i=1}^{10} i z_i' = j(a-b).$$

Nun ist $j(a-b)$ für $j \in \{1, \ldots, 10\}$ und $a, b \in \{0, \ldots, 9\}$ mit $a \neq b$ *nicht* durch 11 teilbar, womit bewiesen sein wird, dass $z_1' \cdots z_9' z_{10}'$ keine korrekte ISBN-Zahl ist und daher Fehler vom Typ 1 erkannt werden.

Nun entstehe die Zahl $z_1' \cdots z_9' z_{10}'$ aus der korrekten ISBN-Zahl $z_1 \cdots z_9 z_{10}$ durch einen Zifferndreher, indem an den Positionen $(j, j+1)$ die Ziffern ab durch ba mit $a \neq b$ ersetzt werden. Dann ist

$$\sum_{i=1}^{10} i z_i - \sum_{i=1}^{10} i z_i' = j(a-b) + (j+1)(b-a) = b - a \neq 0.$$

Da $b - a$ kein ganzzahliges Vielfaches von 11 ist, ist $z_1' \cdots z_9' z_{10}'$ keine korrekte ISBN-Zahl.

Jetzt betrachten wir Fehler vom Typ 3. Die Zahl $z_1' \cdots z_9' z_{10}'$ entstehe also aus der korrekten ISBN-Zahl $z_1 \cdots z_9' z_{10}'$, indem an den Positionen $(j, j+1, j+2)$ die Ziffern acb durch bca mit $a \neq b$ ersetzt werden. Dann ist

$$\sum_{i=1}^{10} i z_i - \sum_{i=1}^{10} i z_i' = j(a-b) + (j+2)(b-a) = 2(b-a).$$

Da $2(b-a)$ nicht durch 11 teilbar ist, ist $z_1' \cdots z_9' z_{10}'$ keine korrekte ISBN-Zahl.

Schließlich betrachten wir einen Zwillingsfehler, also einen Fehler vom Typ 4. Die Ziffernfolge $z_1' \cdots z_9' z_{10}'$ gehe aus der korrekten ISBN-Zahl $z_1 \cdots z_9 z_{10}$ hervor, indem an den Positionen $(j, j+1)$ die Ziffern aa durch bb mit $a \neq b$ ersetzt werden. Wir erhalten

$$\sum_{i=1}^{10} i z_i - \sum_{i=1}^{10} i z_i' = j(a-b) + (j+1)(a-b) = (2j+1)(a-b).$$

Nur für $j = 5$ ist diese Zahl durch 11 teilbar, in allen anderen Fällen ist $z_1' \cdots z_9' z_{10}'$ keine korrekte ISBN-Zahl. □

Prüfziffern kommen aber auch bei Kreditkarten oder anderen Karten, wie z. B. der Bahncard vor. Die Kartennummer der Bahncard hat z. B. 16 Ziffern, von denen die ersten beiden den festen Präfix 70 bilden und die letzte Ziffer eine Prüfziffer ist. Diese Prüfziffer wird folgendermaßen berechnet: Die erste, dritte bis zur 15. Ziffer werden mit zwei multipliziert. Von diesen wird jeweils die Quersumme genommen. Zur Summe dieser Quersummen werden die zweite, vierte bis zur 14. Ziffer addiert. Der Rest dieser Summe bei Division durch 10 wird von 10 subtrahiert. Dies ergibt die Prüfziffer, falls sich eine Zahl < 10 ergibt, andernfalls ist die Prüfziffer gleich 0.

Beispiel: Wir berechnen die Prüfziffer zu der Bahncard mit der Kartennummer 7081-4101-1240-151?. Wir berechnen (5+7+8+0+2+8+2+2)+(0+1+1+1+2+0+5) = 44. Der Rest modulo 10 ist 4, die Prüfziffer ist daher $10 - 4 = 6$. Das stimmt in diesem Fall, denn dies ist die Kartennummer meiner Bahncard (ohne Kreditkartenfunktion!). Dasselbe Verfahren zur Berechnung der Prüfziffer wird offenbar auch bei Kreditkarten angewandt. Jedenfalls erhalte ich nach diesem Verfahren bei meiner Mastercard die korrekte Prüfziffer. □

Die internationale Kontonummer IBAN (International Bank Account Number, Internationale Bankkontonummer) soll den grenzüberschreitenden Zahlungsverkehr vereinheitlichen und damit vereinfachen. Durch die weltweit einheitliche Form ist eine vollautomatische Abwicklung der Transfers möglich, was zu Kosteneinsparungen führt. Die IBAN für Deutschland besteht aus genau 22 Stellen und ist folgendermaßen aufgebaut:

DE pp Bankleitzahl Kontonummer

Hierbei ist pp eine zweistellige Prüfsumme, die Bankleitzahl besteht aus 8 Ziffern (z. B. 26050001 für die Göttinger Sparkasse), die Kontonummer besteht aus 10 Ziffern, wobei diese notfalls mit führenden Nullen aufgefüllt wird. Die Berechnung der IBAN-Prüfsumme wird z. B. bei `http://www.iban.de/iban-pruefsumme.html` geschildert. Sie verläuft folgendermaßen, wobei wir die einzelnen Schritte anhand der Kontonummer des Finanzamts Göttingen (BBK Göttingen, BLZ 26000000, Kto 26001500) demonstrieren:

1. Zunächst bilde man aus der Bankleitzahl und der (durch Nullen ergänzten) 10-stelligen Kontonummer die 18-stellige BBAN (Basic Bank Account Number). Im Beispiel wäre die BBAN also 260000000026001500.

2. Die Buchstaben der Länderkennung, im Falle von Deutschland also DE, werden in numerische Zeichen verwandelt[73] Die Grundlage für die Zahlen, die aus den Buchstaben gebildet werden sollen, bildet ihre Position der jeweiligen Alpha-Zeichen im lateinischen Alphabet. Zu diesem Zahlenwert wird 9 addiert. Die Summe ergibt die Zahl, die den jeweiligen Buchstaben ersetzen soll. Dementsprechend steht für A (Position 1+9) die Zahl 10, für D (Position 4+9) die 13 und für E (Position 5+9) die 14. Der Länderkennung DE entspricht also die Ziffernfolge 1314. Dieser numerischen Länderkennung werden zwei Nullen angehängt. Diese Ziffernfolge wird der 18-stelligen BBAN angehängt. In unserem Beispiel erhalten wir 260000000026001500131400.

3. Die so erhaltene Zahl (BBAN+numerische Länderkennung+00) modulo 97, also der Rest bei Division durch 97, wird von 98 subtrahiert und ergibt die Prüfziffer. Ist dieses Resultat einstellig bzw. kleiner als 10, so wird der Zahl eine Null vorangestellt, so dass sich wieder ein zweistelliger Wert ergibt. In unserem Beispiel ist
$$260000000026001500131400 = 2680412371402077320942 * 97 + 26,$$

[73] Wir gehen davon aus, dass in der Kontonummer keine alphanumerischen Zeichen vorkommen, was nicht überall der Fall zu sein scheint.

es ist also

$$26 = 260000000026001500131400 \bmod 97.$$

Die gesuchte Prüfziffer ist daher $98 - 26 = 72$. Die IBAN des Göttinger Finanzamts ist folglich DE72260000000026001500.

58 Der Satz von Perron

Der Satz von Perron ist für mich einer der interessantesten Sätze der linearen Algebra. Mit ein Grund hierfür ist sicherlich, dass er einfach zu formulieren, aber nicht ganz einfach zu beweisen ist. Außerdem gibt es nichttriviale Verallgemeinerungen (Satz von Perron-Frobenius) sowie interessante Anwendungen, worauf wir in den beiden folgenden Abschnitten eingehen wollen. In diesem Abschnitt konzentrieren wir uns auf den Satz von O. Perron (1907) . Beim Beweis orientieren wir uns an P. Deuflhard, A. Hohmann (2008, S. 153 ff.). Erwähnt sei aber für diesen und die nächsten beiden Abschnitte auch die vorzügliche Darstellung in Chapter 8 von C. Meyer (2000).

Satz (Perron) *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine positive Matrix, d. h. es ist $a_{ij} > 0$, $i, j = 1, \ldots, n$. Mit*

$$\rho(A) := \max_{\lambda \text{ ist Eigenwert von } A} |\lambda|$$

bezeichnen wir den Spektralradius von A. Dann gilt:

1. *Es ist $\rho(A)$ ein positiver Eigenwert von A, zu dem es einen positiven Eigenvektor gibt. D. h. es gibt einen Vektor $p \in \mathbb{R}^n$, den sogenannten Perron-Vektor zu A, dessen Komponenten alle positiv sind, mit $Ap = \rho(A)p$.*
2. *Für jeden Eigenwert λ von A mit $\lambda \neq \rho(A)$ ist $|\lambda| < \rho(A)$.*
3. *$\rho(A)$ ist ein einfacher Eigenwert von A.*
4. *Ein nichtnegativer Eigenvektor zu A ist ein positives Vielfaches des Perron-Vektors von A.*
5. *Es gilt die Collatz-Wielandt-Formel: Es ist $\rho(A) = \max_{x\in\mathcal{N}} f(x)$, wobei*

$$f(x) := \min_{\substack{1\le i\le n\\ x_i\neq 0}} \frac{(Ax)_i}{x_i}, \qquad \mathcal{N} := \{x \in \mathbb{R}^n : x \geq 0,\ x \neq 0\}.$$

Beweis: Zunächst zeigen wir, dass $\rho(A) > 0$. Angenommen, im Widerspruch zur Behauptung wäre $\rho(A) = 0$ bzw. jeder Eigenwert von A würde verschwinden. Da A auf Jordansche Normalform gebracht werden kann, existiert eine nichtsinguläre Matrix $P \in \mathbb{C}^{n\times n}$ mit $P^{-1}AP = J$, wobei $J = \operatorname{diag}(J_1, \ldots, J_p)$ eine Blockdiagonalmatrix ist mit

$$J_i = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_i\times n_i}, \qquad i = 1, \ldots, p.$$

Offenbar ist J und damit auch A nilpotent, d. h. eine geeignete Potenz von J und damit auch von A verschwindet. Eine positive Matrix kann aber nicht nilpotent sein. Damit ist $\rho(A) > 0$ bewiesen.

Sei λ ein Eigenwert von A mit $|\lambda| = \rho(A)$ (einen solchen muss es nach Definition des Spektralradius geben) und x ein zugehöriger Eigenvektor. Ist $x = (x_j)$, so bezeichnen wir mit $|x| := (|x_j|)$ den Absolutbetrag von x, entsprechende Vereinbarungen benutzen wir für Matrizen. Dann ist

$$\rho(A)\,|x| = |\lambda|\,|x| = |\lambda x| = |Ax| \le |A|\,|x| = A|x|.$$

Wir zeigen, dass hier Gleichheit gilt. Definiere hierzu den nichtnegativen Vektor y durch

$$y := A|x| - \rho(A)\,|x|.$$

Angenommen, es sei $y \neq 0$. Da A eine positive Matrix ist, ist $Ay > 0$ und aus demselben Grund $z := A|x| > 0$. Dann gibt es ein $\epsilon > 0$ mit $Ay > \epsilon\rho(A)\,z$. Da $Ay = Az - \rho(A)\,z$ ist

$$\Big(\underbrace{\frac{A}{\rho(A)(1+\epsilon)}}_{=:B}\Big)z > z$$

und daher $B^k z > z$ für $k = 1, \ldots$. Andererseits ist $\rho(B) = 1/(1+\epsilon) < 1$ und folglich[74] $\lim_{k\to\infty} B^k = 0$ und $\lim_{k\to\infty} B^k z = 0$, womit wir den gewünschten Widerspruch erhalten haben. Also ist $\rho(A)\,|x| = A|x| > 0$, also $\rho(A)$ ein Eigenwert von A mit dem positiven Eigenvektor, dem Perron-Vektor, $p := |x|$. Damit ist der erste Teil des Satzes bewiesen.

Sei λ ein Eigenwert von A mit $|\lambda| = \rho(A)$ und x ein zugehöriger Eigenvektor. Wie wir gerade eben bewiesen haben, ist $|x|$ ein positiver Eigenvektor zum Eigenwert $\rho(A)$, ferner ist $|Ax| = A|x|$ bzw.

$$(*) \qquad \Big|\sum_{j=1}^{n} a_{kj}x_j\Big| = \sum_{j=1}^{n} a_{kj}|x_j| = \sum_{j=1}^{n} |a_{kj}x_j|, \qquad k = 1, \ldots, n.$$

Nun gilt:

- Sind $z_1, \ldots, z_n \in \mathbb{C} \setminus \{0\}$ und $|\sum_{j=1}^n z_j| = \sum_{j=1}^n |z_j|$, so existieren $\alpha_j > 0$ mit $z_j = \alpha_j z_1$, $j = 1, \ldots, n$.

Denn: Für $n = 1$ ist die Aussage trivial. Wir beweisen die Aussage für $n = 2$ und erhalten das allgemeine Ergebnis durch vollständige Induktion. Sei also $|z_1 + z_2| = |z_1| + |z_2|$ mit $z_1, z_2 \neq 0$ und $z_j = x_j + iy_j$ mit reellen x_j, y_j, $j = 1, 2$. Dann ist

$$\begin{aligned}
(x_1+x_2)^2 + (y_1+y_2)^2 &= |z_1+z_2|^2 \\
&= (|z_1| + |z_2|)^2 \\
&= |z_1|^2 + 2|z_1|\,|z_2| + |z_2|^2 \\
&= (x_1^2+y_1^2) + 2\sqrt{(x_1^2+y_1^2)(x_2^2+y_2^2)} + (x_2^2+y_2^2).
\end{aligned}$$

[74] Ist $\rho(B) < 1$, so existiert eine natürliche Matrixnorm $\|\cdot\|$ mit $\|B\| < 1$, siehe z. B. J. Werner (1992, S. 23). Dann ist aber $\|B^k\| \le \|B\|^k \to 0$, also $\lim_{k\to\infty} B^k = 0$.

Hieraus folgt

$$(x_1x_2 + y_1y_2) = \sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2)},$$

nach erneutem Quadrieren folgt

$$2x_1x_2y_1y_2 = x_1^2y_2^2 + y_1^2x_2^2$$

und damit

$$(x_1y_2 - y_1x_2)^2 = 0$$

bzw. $x_1y_2 = y_1x_2$. Hieraus folgt: Ist $x_1 \neq 0$, so ist $z_2 = (x_2/x_1)z_1$. Ist dagegen $x_1 = 0$, so ist notwendig $y_1 \neq 0$ (da $z_1 \neq 0$) und $z_2 = (y_2/y_1)z_1$. In jedem Fall ist $z_2 = \alpha_2 z_1$ mit reellem, von Null verschiedenem α_2 (wäre $\alpha_2 = 0$, so wäre $z_2 = 0$, was wir ausgeschlossen hatten). Folglich ist

$$|1 + \alpha_2|\,|z_1| = |z_1 + z_2| = |z_1| + |z_2| = (1 + |\alpha_2|)\,|z_1|$$

und daher $|1 + \alpha_2| = 1 + |\alpha_2|$. Hieraus folgt aber $\alpha_2 \geq 0$, wegen $\alpha_2 \neq 0$ ist $\alpha_2 > 0$. Damit ist die Behauptung für $n = 2$ bewiesen. Nun nehmen wir an, die Behauptung sei für n richtig. Ist $|\sum_{j=1}^{n+1} z_j| = \sum_{j=1}^{n+1} |z_j|$, so ist notwendig $\sum_{j=1}^{n} z_j| = \sum_{j=1}^{n} |z_j|$. Aus der Induktionsannahme folgt die Existenz von positiven α_j mit $z_j = \alpha_j z_1$, $j = 1, \ldots, n$. Folglich ist

$$\Big|\Big(\sum_{j=1}^{n} \alpha_j\Big) z_1 + z_n\Big| = \Big(\sum_{j=1}^{n} \alpha_j\Big)|z_1| + |z_n|.$$

Aus der schon bewiesenen Behauptung für $n = 2$ folgt die Existenz von $\alpha_n > 0$ mit $z_n = \alpha_n z_1$. Die (eigentlich wohlbekannte und wesentlich allgemeiner gültige) Aussage • ist damit bewiesen.

Nun fahren wir mit dem Beweis des zweiten Teiles des Satzes fort. Wenden wir • auf $(*)$ an (man beachte, dass $a_{kj}x_j \neq 0$, $j, k = 1, \ldots, n$), so erhält man die Existenz von $\alpha_{jk} > 0$ mit

$$a_{kj}x_j = \alpha_{jk}a_{k1}x_1, \qquad j, k = 1, \ldots, n.$$

Da A eine positive Matrix ist, ist

$$x_j = \frac{\alpha_{jk}a_{k1}}{a_{kj}}x_1, \qquad j, k = 1, \ldots, n.$$

Die linke Seite hängt nicht von k ab und daher ist auch die rechte Seite von k unabhängig. Folglich existieren positive α_j mit $x_j = \alpha_j x_1$, $j = 1, \ldots, n$. Definiert man den positiven Vektor $a \in \mathbb{R}^n$ durch $a := (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$, so ist $a = (1/x_1)x$ ebenfalls ein Eigenvektor von A zum Eigenwert λ und daher

$$\lambda a = Aa = |Aa| = |\lambda a| = |\lambda|\,a = \rho(A)a,$$

womit $\lambda = \rho(A)$ und damit der zweite Teil des Satzes bewiesen ist.

Nun kommen wir zum Beweis des dritten Teiles des Satzes, dass nämlich $\rho(A)$ ein *einfacher* Eigenwert von A ist. Indem wir notfalls A durch $(1/\rho(A))A$ ersetzen, können wir o. B. A. annehmen, dass $\rho(A) = 1$. Die Matrix A kann auf Jordansche Normalform

transformiert werden, d. h. es existiert eine nichtsinguläre Matrix $P \in \mathbb{C}^{n\times n}$ derart, dass

$$P^{-1}AP = J = \operatorname{diag}(J_1, \ldots, J_p)$$

eine Blockdiagonalmatrix ist mit

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix} \in \mathbb{R}^{n_i\times n_i}, \qquad i = 1, \ldots, p.$$

Wegen $\rho(A^k) = \rho(A)^k = 1$ ist die Folge $\{\|A^k\|\}$ beschränkt[75]. Wegen $J^k = P^{-1}A^kP$ und $\|J^k\| \le \|P\|\,\|P^{-1}\|\,\|A^k\|$ ist dann auch $\{\|J^k\|\}$ beschränkt. Dies ist genau dann der Fall, wenn $\{\|J_i^k\|\}$ beschränkt, $i = 1, \ldots, p$. Für $\lambda_i \neq \rho(A) = 1$ ist $|\lambda_i| < 1$ und folglich $\lim_{k\to\infty} J_i^k = 0$. Daher betrachten wir jetzt ein i mit $\lambda_i = \rho(A) = 1$. Angenommen, für den entsprechenden Jordan-Block J_i, also

$$J_i = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{n_i\times n_i},$$

gelte $n_i \ge 2$. Wegen

$$J_i^k = \begin{pmatrix} 1 & \binom{k}{1} & \cdots & \binom{k}{n_i-1} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \binom{k}{1} \\ 0 & \cdots & 0 & 1 \end{pmatrix},$$

wobei

$$\binom{k}{m} = \frac{k(k-1)\cdots(k-(m-1))}{1\cdot 2\cdots m},$$

ist $\{\|J_i^k\|\}$ unbeschränkt, ein Widerspruch. Also ist $n_i = 1$ für jedes λ_i mit $\lambda_i = \rho(A) = 1$. Insbesondere ist die algebraische Vielfachheit des Eigenwertes $\rho(A)$, also die Vielfachheit als Nullstelle des charakteristischen Polynoms, gleich der geometrischen Vielfachheit des Eigenwertes $\rho(A)$, also der Dimension des Eigenraumes von $\rho(A)$ bzw. des Nullraumes von $A - \rho(A)I$. Nun nehmen wir an, diese (geometrische oder algebraische) Vielfachheit sei größer als 1. Angenommen, die geometrische Vielfachheit des Eigenwertes $\rho(A)$ sei größer als 1. Dann gibt es zwei linear unabhängige Eigenvektoren x und y zu $\rho(A) = 1$. Sei $y_i \neq 0$ (da y ein Eigenvektor von A ist, ist wenigstens eine Komponente von y von Null verschieden). Wir definieren

$$z := x - \frac{x_i}{y_i}y \neq 0.$$

[75] Wieder z. B. nach J. Werner (1992, S. 23) gibt es zu A^k und jedem $\epsilon > 0$ eine Matrixnorm $\|\cdot\|$ mit $\|A^k\| \le \rho(A^k) + \epsilon = 1 + \epsilon$. Wegen der Äquivalenz der Normen ist $\{A^k\}$ beschränkt.

Dann ist z ebenfalls ein Eigenvektor zum Eigenwert $\rho(A) = 1$. Wie wir am Anfang bewiesen haben, ist dann $|z|$ ein Eigenvektor von A mit Eigenwert $\rho(A) = 1$ und folglich $|z| > 0$. Dies ist aber ein Widerspruch zu $z_i = x_i - (x_i/y_i)y_i = 0$. Daher ist $\rho(A)$ ein einfacher Eigenwert von A und der dritte Teil des Satzes ist bewiesen.

Im vierten Teil des Satzes nehmen wir an, λ sei ein Eigenwert von A mit einem nichtnegativen Eigenvektor $y \neq 0$. Mit A ist natürlich auch A^T eine positive Matrix. Sei $x > 0$ ein Perron-Vektor zu der positiven Matrix A^T, also $A^T x = \rho(A)x$. Hieraus folgt

$$\rho(A)x^T y = (A^T x)^T y = x^T Ay = \lambda x^T y.$$

Da $x^T y > 0$ folgt $\lambda = \rho(A)$. Also ist $\rho(A)$ der einzige Eigenwert von A, zu dem es einen nichtnegativen Eigenvektor gibt.

Jetzt kommen wir zum Beweis der Collatz-Wielandt-Formel. Zunächst zeigen wir, dass $f(x) \leq \rho(A)$ für alle $x \in \mathcal{N}$. Sei also $x \in \mathcal{N}$. Dann ist $0 \leq f(x)x \leq Ax$. Sei p bzw. q ein Perronvektor zu A bzw. A^T, also p, $q > 0$ sowie $Ap = \rho(A)p$, $A^T q = \rho(A)q$. Aus $f(x)x \leq Ax$ folgt nach Multiplikation mit q^T von links, dass

$$f(x)q^T x \leq q^T Ax = (A^T q)^T x = \rho(A)q^T x.$$

Da $q^T x > 0$ folgt $f(x) \leq \rho(A)$. Zum anderen ist $p \in \mathcal{N}$ und $f(p) = \rho(A)$, insgesamt also $\rho(A) = \max_{x \in \mathcal{N}} f(x)$. □

59 Der Satz von Perron-Frobenius

Der Satz von Perron-Frobenius ist eine Verallgemeinerung des Satzes von Perron (siehe Abschnitt 58) auf *nichtnegative* Matrizen. Ohne weitere Voraussetzungen sind die Aussagen des Satzes von Perron für nichtnegative Matrizen nicht richtig. Z. B. hat die Matrix

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

die Null als Eigenwert der algebraischen Vielfachheit 2, während die Matrix

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

mit 1 und -1 zwei Eigenwerte besitzt, die maximalen Betrag haben. Die entscheidende zusätzliche Voraussetzung ist, dass A *irreduzibel* ist.

Definition Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt *reduzibel*, wenn es nichtleere Teilmengen I, J von $N := \{1, \ldots, n\}$ mit $I \cap J = \emptyset$, $I \cup J = N$ sowie $a_{ij} = 0$ für alle $(i, j) \in I \times J$ gibt. Die Matrix A heißt *irreduzibel*, wenn sie nicht reduzibel ist.

Um zu entscheiden, ob eine gegebene nichtnegative Matrix irreduzibel ist, ist es zur Veranschaulichung vorteilhaft, Begriffe aus der *Graphentheorie* einzusetzen. Hierzu wird einer nichtnegativen Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ ein *gerichteter Graph* $G = (V, E)$ bestehend aus den Knoten (oder Ecken) $V := \{1, \ldots, n\}$ und den Pfeilen bzw. gerichteten Kanten $E := \{(i, j) : a_{ij} \neq 0\}$ zugeordnet. Dieser gerichtete Graph heißt

stark zusammenhängend, wenn es von jedem Knoten einen gerichteten Weg[76] zu jedem anderen Knoten gibt. Wir werden im nächsten Satz zeigen, dass die nichtnegative Matrix A genau dann irreduzibel ist, wenn der zugeordnete gerichtete Graph stark zusammenhängend ist.

Beispiel: Sei

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

Wir wollen zeigen, dass A reduzibel ist. Hierzu definieren wir $I := \{1, 2, 3.4, 5, 7, 8\}$ und $J := \{6, 9\}$. Offenbar ist $I \cap J = \emptyset$ sowie $I \cup J = \{1, \ldots, 9\}$, ferner $a_{ij} = 0$ für alle $(i, j) \in I \times J$. Betrachtet man den A zugeordneten gerichteten Graphen, so erkennt man, dass in I die Knoten zusammengefasst sind, die man vom Knoten 1 durch einen gerichteten Weg erreichen kann. Die Indexmenge J ist das Komplement von I, also die Menge der Knoten, die vom Knoten 1 nicht erreicht werden können. □

Der Beweis des Satzes von Perron-Frobenius wird die Aussagen des folgenden Lemmas benutzen.

Lemma *Sei $A \in \mathbb{R}^{n \times n}$ eine nichtnegative Matrix. Dann gilt:*

1. *Die Matrix A ist genau dann irreduzibel, wenn der zugeordnete gerichtete Graph stark zusammenhängend ist.*

2. *Ist A irreduzibel, so ist $(I + A)^{n-1} > 0$.*

Beweis: Wir nehmen an, A sei reduzibel und zeigen, dass der A zugeordnete Graph G nicht stark zusammenhängend ist. Da A reduzibel ist, existieren nichtleere Indexmengen $I, J \subset \{1, \ldots, n\}$ mit $I \cap J = \emptyset$, $I \cup J = \{1, \ldots, n\}$ und $a_{ij} = 0$ für alle $(i, j) \in I \times J$. Dann ist klar, dass es keinen gerichteten Weg von einem Knoten $i \in I$ zu einem Knoten $j \in J$ geben kann. Denn ein von $i \in I$ ausgehender gerichteter Weg trifft nur Knoten aus I. Daher ist der Graph G nicht stark zusammenhängend. Damit ist gezeigt: Ist G stark zusammenhängend, so ist A irreduzibel. Nun nehmen wir an, G sei nicht stark zusammenhängend. Dann gibt es Knoten k und l, die nicht durch einen gerichteten Weg miteinander verbunden werden können. Es sei $I \subset \{1, \ldots, n\}$ die Menge aller Knoten, die durch einen von k ausgehenden gerichteten Weg erreicht werden können, mit J wird das Komplement bezüglich $\{1, \ldots, n\}$ bezeichnet. Dann sind I und J nichtleere Teilmengen von $\{1, \ldots, n\}$, es gilt $I \cap J = \emptyset$, $I \cup J = \{1, \ldots, n\}$. Weiter ist $a_{ij} = 0$ für alle $(i, j) \in I \times J$. Denn der Knoten $i \in I$ kann von k ausgehend durch einen gerichteten Weg in G erreicht werden. Wäre $a_{ij} \neq 0$, so wäre $(i, j) \in E$ und auch

[76] Unter einem *gerichteten Weg* vom Knoten i zum Knoten j versteht man eine Folge von Pfeilen bzw. gerichteter Kanten $(i, i_1), (i_1, i_2), \ldots, (i_k, j) \in E$.

j könnte vom Knoten k durch einen gerichteten Weg erreicht werden. Dies ist aber wegen der Definition von J als Komplement von I ausgeschlossen. Also ist A reduzibel. Damit ist der erste Teil des Satzes bewiesen.

Nun setzen wir voraus, die nichtnegative Matrix $A \in \mathbb{R}^{n\times n}$ sei irreduzibel bzw. der zugeordnete gerichtete Graph G stark zusammenhängend. Sei $A^k = (a_{ij}^{(k)})$ die k-te Potenz der nichtnegativen Matrix A. Es ist $a_{ij}^{(k)} > 0$ genau dann, wenn es vom Knoten i zum Knoten j einen gerichteten Weg der Länge k gibt. Entsprechend ist $[(I+A)^k]_{ij} > 0$, wenn es vom Knoten i zum Knoten j einen Weg der Länge $\leq k$ gibt, denn zu jedem Knoten gehört eine "Schleifenkante", welche von einem Knoten zum selben Knoten führt. Bei vorzeitigem Eintreffen im Knoten j kann man daher in der Schleife verweilen bis man einen Weg der Länge k erreicht hat. Folglich ist $(I+A)^{n-1} > 0$, da es zwischen je zwei Knoten einen gerichteten Weg der Länge $\leq n-1$ gibt. □

Nach diesen Vorbereitungen kommen wir zum Satz von Perron-Frobenius, der eine Verallgemeinerung des Satzes von Perron auf nichtnegative, irreduzible Matrizen darstellt. Man beachte, dass sich die beiden Sätze nur in der zweiten Aussage unterscheiden.

Satz (Perron-Frobenius) *Sei $A \in \mathbb{R}^{n\times n}$ nichtnegativ und irreduzibel. Mit $\rho(A)$ bezeichnen wir den Spektralradius von A. Dann gilt:*

1. *Es ist $\rho(A)$ ein positiver Eigenwert von A, zu dem es einen positiven Eigenvektor gibt. D. h. es gibt einen Vektor $p \in \mathbb{R}^n$, den sogenannten Perron-Vektor zu A, dessen Komponenten alle positiv sind, mit $Ap = \rho(A)p$.*

2. *Ist λ mit $|\lambda| = \rho(A)$ ein Eigenwert von A mit Eigenvektor x, so ist $|x|$ ein positiver Eigenvektor zum Eigenwert $\rho(A)$ von A.*

3. *$\rho(A)$ ist ein einfacher Eigenwert von A.*

4. *Ein nichtnegativer Eigenvektor zu A ist ein positives Vielfaches des Perron-Vektors von A.*

5. *Es gilt die Collatz-Wielandt-Formel: Es ist $\rho(A) = \max_{x\in\mathcal{N}} f(x)$, wobei*

$$f(x) := \min_{\substack{1\leq i\leq n\\ x_i\neq 0}} \frac{(Ax)_i}{x_i}, \qquad \mathcal{N} := \{x \in \mathbb{R}^n : x \geq 0,\ x \neq 0\}.$$

Beweis: Für $k \in \mathbb{N}$ definieren wir die positive Matrix $A_k := A + (1/k)I$. Mit $p_k > 0$ bezeichnen wir den durch $\|p_k\|_1 = 1$ normierten Perron-Vektor zum positiven Eigenwert $\rho_k := \rho(A_k)$. Aus Kompaktheitsgründen besitzt $\{p_k\}$ eine konvergente Teilfolge $\{p_{k_i}\}$. Der Limes p dieser Teilfolge ist nichtnegativ und vom Nullvektor verschieden (da $p_{k_i} > 0$ und $\|p_{k_i}\|_1 = 1$). Da $A_1 > A_2 > \cdots > A$ ist $\rho_1 \geq \rho_2 \geq \cdots \geq \rho(A)$. Dies ist eine direkte Folge der Collatz-Wielandt-Formel für positive Matrizen. Also existiert der Limes $\rho^* = \lim_{k\to\infty} \rho_k$ und es ist $\rho^* \geq \rho(A)$. Weiter ist

$$Ap = \Big(\lim_{i\to\infty} A_{k_i}\Big)\Big(\lim_{i\to\infty} p_{k_i}\Big) = \lim_{i\to\infty} A_{k_i}p_{k_i} = \lim_{i\to\infty} \rho_{k_i}p_{k_i} = \rho^* p.$$

Folglich ist ρ^* ein (nichtnegativer) Eigenwert von A und damit $\rho^* \leq \rho(A)$. Inssgesamt haben wir gezeigt, dass $\rho^* = \rho(A)$ ein Eigenwert von A ist, zu dem es einen nichtnegativen Eigenvektor p gibt. Bisher haben wir *nicht* ausgenutzt, dass A irreduzibel ist. Für den Beweis des ersten Teiles des Satzes bleibt zu zeigen, dass $\rho(A) > 0$ und ein zugehöriger nichtnegativer Eigenvektor sogar positiv ist. Die Matrix $B := (I + A)^{n-1}$ ist wegen des zweiten Teils des vorangegangenen Lemmas positiv. Es existiert eine nichtsinguläre Matrix $P \in \mathbb{C}^{n\times n}$, welche A auf ähnliche Jordansche Normalform transformiert, d. h. es ist $P^{-1}AP = J$, wobei $J = \operatorname{diag}(J_1, \ldots, J_p)$ eine Blockdiagonalmatrix ist mit

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}, \qquad i = 1, \ldots, p.$$

Dann ist

$$(I + A) = P(I + P^{-1}AP)P^{-1} = P(I + J)P^{-1}$$

und folglich nach vollständiger Induktion

$$(I + A)^{n-1} = P(I + J)^{n-1}P^{-1}.$$

Daher haben $(I + A)^{n-1}$ und $(I + J)^{n-1}$ dieselben Eigenwerte. Die Matrix $(I + J)^{n-1}$ ist aber eine obere Dreiecksmatrix mit $(1 + \lambda_i)^{n-1}$, $i = 1, \ldots, p$, in der Diagonalen. Dies zeigt, dass λ genau dann ein Eigenwert von A ist, wenn $(1 + \lambda)^{n-1}$ Eigenwert von B ist. Darüberhinaus folgt $\rho(B) = (1 + \rho(A))^{n-1}$. Der nichtnegative Eigenvektor p zum Eigenwert $\rho(A)$ von A ist auch ein Eigenvektor zum Eigenwert $\rho(B)$ der positiven Matrix B. Aus dem vierten Teil des Satzes von Perron (angewandt auf die positive Matrix B) folgt, dass p ein positives Vielfaches des Perron-Vektors von B ist und daher insbesondere selbst positiv ist. Für den ersten Teil des Satzes bleibt zu zeigen, dass $\rho(A) > 0$. Andernfalls wäre $Ap = 0$ mit $p > 0$. Dies ergibt nur für die Nullmatrix keinen Widerspruch und diese ist reduzibel. Damit ist der erste Teil des Satzes von Perron-Frobenius bewiesen.

Sei nun λ mit $|\lambda| = \rho(A)$ ein Eigenwert von A mit Eigenvektor x. Dann ist

$$\rho(A)|x| = |\lambda x| = |Ax| \leq |A|\,|x| = A|x|$$

und folglich $y := A|x| - \rho(A)|x| \geq 0$. Angenommen $y \neq 0$. Mit A ist auch A^T nichtnegativ und irreduzibel, wegen des ersten Teils dieses Satzes existiert daher $q > 0$ mit $A^T q = \rho(A)q$. Folglich ist

$$0 < y^T q = (A|x| - \rho(A)|x|)^T q = |x|^T (A^T q - \rho(A)q) = 0,$$

ein Widerspruch. Also ist $y = 0$ bzw. $|x|$ ein (nichtnegativer) Eigenvektor von A zum Eigenwert $\rho(A)$. Weiter ist sogar $|x| > 0$, wie wir fast zum Schluss des ersten Teiles des Satzes uns überlegten.

Weiter ist $\rho(A)$ ein einfacher Eigenwert von A, denn andernfalls wäre $\rho(B)$ ein Eigenwert der positiven Matrix B mit einer algebraischen Vielfachheit > 1, ein Widerspruch. Damit ist auch der dritte Teil des Satzes von Perron-Frobenius bewiesen.

Der vierte Teil des Satzes kann genauso wie der entsprechende Teil des Satzes von Perron bewiesen werden. Auch die Collatz-Wielandt-Formel ist einfach nachzuweisen, wobei die Irreduzibilität von A nicht benutzt zu werden braucht. □

Bemerkung: Nur in der zweiten Aussage unterscheiden sich (bis auf die unterschiedlichen Voraussetzungen) die Sätze von Perron und Perron-Frobenius. Bei einer nichtnegativen irreduziblen Matrix kann tatsächlich mehr als ein Eigenwert auf dem Rand des Spektralkreises liegen, wie die (nichtnegative, irreduzible) Matrix

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

mit den Eigenwerten ± 1 zeigt. □

Bemerkung: Mit Hilfe des Brouwerschen Fixpunktsatzes (siehe Abschnitt 37) kann man leicht einen Teil des Satzes von Perron-Frobenius beweisen. Hierauf haben G. Debreu, I. N. Herstein (1953) hingewiesen.

- *Sei $A \in \mathbb{R}^{n\times n}$ nichtnegativ und irreduzibel. Dann besitzt A einen positiven Eigenwert λ^* mit einem zugehörigen positiven Eigenvektor x^*.*

Denn: Sei

$$\Sigma_n := \|x \in \mathbb{R}^n : x \geq 0,\, e^T x = 1\},$$

wobei e einmal wieder der Vektor im $\mathbb{R}^n$ ist, dessen Komponenten alle gleich 1 sind. Die Menge Σ_n ist nichtleer, konvex und kompakt. Wir definieren die Abbildung $F\colon \Sigma_n \longrightarrow \mathbb{R}^n$ durch

$$F(x) := \frac{1}{e^T Ax} Ax.$$

Ist $x \in \Sigma_n$, so ist $Ax \geq 0$, da A nichtnegativ ist, und $Ax \neq 0$. Letzteres zeigen wir durch Widerspruch. Angenommen, es sei $Ax = 0$ bzw.

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j = 0, \qquad i = 1, \ldots, n.$$

Dann ist $a_{ij} x_j = 0$, $i, j = 1, \ldots, n$. Wir definieren die beiden Indexmengen

$$I := \{i \in \{1, \ldots, n\} : x_i = 0\}, \qquad J := \{j \in \{1, \ldots, n\} : x_j > 0\}.$$

Dann ist $a_{ij} = 0$, $(i, j) \in I \times J$. Da A nach Voraussetzung irreduzibel ist und $J \neq \emptyset$, ist $I = \emptyset$ bzw. $A = 0$, ein Widerspruch. Für $x \in \Sigma_n$ ist daher $e^T Ax > 0$. Folglich ist die Abbildung F wohldefiniert, stetig und es ist $F(\Sigma_n) \subset \Sigma_n$. Der Brouwersche Fixpunktsatz liefert die Existenz eines $x^* \in \Sigma_n$ mit $F(x^*) = x^*$ bzw. $Ax^* = (e^T Ax^*)x^*$. Daher ist $\lambda^* := e^T Ax^*$ ein positiver Eigenwert von A mit dem zugehörigen Eigenvektor $x^* \in \Sigma_n$. Zu zeigen bleibt, dass $x^* > 0$. Seien die Indexmengen I^* und J^* durch

$$I^* := \{i \in \{1, \ldots, n\} : x_i^* = 0\}, \qquad J^* := \{j \in \{1, \ldots, n\} : x_j^* > 0\}$$

definiert. Für $i \in I^*$ ist dann

$$0 = \lambda^* x_i^* = (Ax^*)_i = \sum_{j \in J^*} a_{ij} x_j^*$$

und daher $a_{ij} = 0$, $(i,j) \in I^* \times J^*$. Da A irreduzibel ist, folgt aus $J^* \neq \emptyset$, dass $I^* = \emptyset$ und daher $x^* > 0$ gilt. Wir haben also obige Behauptung mit Hilfe des Brouwerschen Fixpunktsatzes bewiesen. □

Zur Abrundung wollen wir noch einige wenige Bemerkungen zu nichtnegativen *primitiven* Matrizen machen.

Definition Eine nichtnegative, irreduzible Matrix $A \in \mathbb{R}^{n\times n}$, welche außer $\rho(A)$ keinen weiteren Eigenwert λ mit $|\lambda| = \rho(A)$ besitzt, heißt *primitiv*.

Z. B. sind positive Matrizen primitiv. Aber nicht jede nichtnegative, irreduzible Matrix ist primitiv, wie das Beispiel

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

mit den Eigenwerten 1 und -1 zeigt.

Bevor wir hinreichende Bedingungen dafür angeben, dass eine Matrix primitiv ist, formulieren und beweisen wir das folgende Ergebnis.

Satz *Eine nichtnegative irreduzible Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann primitiv, wenn $\lim_{k\to\infty}(A/\rho(A))^k$ existiert. Ist dies der Fall, so ist*

$$\lim_{k\to\infty}\Big(\frac{A}{\rho(A)}\Big)^k = \frac{pq^T}{p^Tq} > 0,$$

wobei die positiven Vektoren p bzw. q Perron-Vektoren zu A bzw. A^T sind.

Beweis: Sei A primitiv. Die Matrix $A/\rho(A)$ hat 1 als einzigen Eigenwert auf dem Einheitskreis und dieser ist wegen des Satzes von Perron-Frobenius einfach. Die Jordansche Normalform von A hat also die Form

$$J = P^{-1}(A/\rho(A))P = \begin{pmatrix} 1 & 0^T \\ 0 & K \end{pmatrix},$$

wobei $\rho(K) < 1$. Hieran erkennt man, dass $\lim_{k\to\infty}(A/\rho(A))^k$ existiert, denn es ist

$$\Big(\frac{A}{\rho(A)}\Big)^k = P\begin{pmatrix} 1 & 0^T \\ 0 & K^k \end{pmatrix}P^{-1} \to P\begin{pmatrix} 1 & 0^T \\ 0 & 0 \end{pmatrix}P^{-1}.$$

Aus

$$PJ = \frac{1}{\rho(A)}AP$$

lesen wir ab (multipliziere diese Gleichung auf beiden Seiten mit e_1, dem ersten Einheitsvektor), dass die erste Spalte p_1 von P Eigenvektor zum Eigenwert 1 von $A/\rho(A)$ ist. Entsprechend folgt aus

$$(P^{-1})^T J^T = \frac{1}{\rho(A)}A^T(P^{-1})^T,$$

dass die erste Spalte q_1 von $(P^{-1})^T$ Eigenvektor zum Eigenwert 1 von $A^T/\rho(A)$ ist. Denken wir uns P und P^{-1} in der Form

$$P = \begin{pmatrix} p_1 & P_2 \end{pmatrix}, \qquad P^{-1} = \begin{pmatrix} q_1^T \\ Q_2 \end{pmatrix}$$

partitioniert, so ist also

$$\lim_{k\to\infty}\Big(\frac{A}{\rho(A)}\Big)^k = \begin{pmatrix} p_1 & P_2 \end{pmatrix}\begin{pmatrix} 1 & 0^T \\ 0 & 0 \end{pmatrix}\begin{pmatrix} q_1^T \\ Q_2 \end{pmatrix} = p_1 q_1^T.$$

Seien p bzw. q (positive) Perron-Vektoren zu A bzw. A^T. Wegen

$$p = \Big(\frac{A}{\rho(A)}\Big)^k p = (p_1 q_1^T)p = (q_1^T p)p_1, \qquad q = \Big(\frac{A^T}{\rho(A)}\Big)^k q = (q_1 p_1^T)q = (p_1^T q)q_1$$

ist

$$p^T q = (q_1^T p)(p_1^T q)$$

und daher

$$p_1 q_1^T = \frac{pq^T}{(q_1^T p)(p_1^T q)} = \frac{pq^T}{p^T q}.$$

Damit ist eine Richtung im obigen Satrz bewiesen, dass nämlich aus der Primitivität der nichtnegativen, irreduziblen Matrix A die Existenz von $\lim_{k\to\infty}(A/\rho(A))^k$ folgt und dass dieser Limes sich in einfacher Weise durch die Perron-Vektoren zu A und A^T ausdrücken lässt.

Nun sei $A \in \mathbb{R}^{n\times n}$ eine nichtnegative, irreduzible Matrix, für die $\lim_{k\to\infty}(A/\rho(A))^k$ existiert. Sei

$$J = P^{-1}(A/\rho(A))P$$

Jordansche Normalform von $A/\rho(A)$. Auch $\lim_{k\to\infty} J^k$ existiert. Dann kann aber J und damit auch $A/\rho(A)$ außer 1 keinen Eigenwert vom Betrag 1 haben, da λ^k mit $\lambda = e^{i\theta}$ und $\theta \in (0, 2\pi)$ nicht konvergiert. Also ist A primitiv, der Satz ist bewiesen. □

Im folgenden Satz werden primitive Matrizen charakterisiert.

Satz *Sei $A \in \mathbb{R}^{n\times n}$ nichtnegativ. Dann ist A genau dann primitiv, wenn ein $m \in \mathbb{N}$ existiert derart, dass A^m positiv ist.*

Beweis: Sei A primitiv. Wegen des vorigen Satzes ist $\lim_{k\to\infty}(A/\rho(A))^k$ eine positive Matrix, also existiert ein $m \in \mathbb{N}$ mit $A^m > 0$. Umgekehrt existiere ein $m \in \mathbb{N}$ derart, dass $A^m = (a_{ij}^{(m)})$ eine positive Matrix ist. Dann ist A irreduzibel, denn in dem A zugeordneten Graphen G gibt es von jedem Knoten i zu jedem Knoten j einen gerichteten Weg der Länge m. Also ist G stark zusammenhängend und damit A irreduzibel. Die Eigenwerte von A^m sind genau die m-ten Potenzen der Eigenwerte von A und die algebraischen Vielfachheiten stimmen überein. Als positive Matrix besitzt A^m den einfachen Eigenwert $\rho(A)^m$ und alle anderen Eigenwerte von A^m sind betragsmäßig kleiner als $\rho(A)^m$. Daher besitzt A außer $\rho(A)$ keinen weiteren Eigenwert λ mit $|\lambda| = \rho(A)$. Also ist A eine primitive Matrix. □

Beispiel: Wir betrachten (siehe C. Meyer (2000, S. 679)) die Matrix

$$A := \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 3 & 4 & 0 \end{pmatrix}.$$

Dann ist

$$A^5 = \begin{pmatrix} 48 & 64 & 12 \\ 36 & 96 & 128 \\ 192 & 274 & 96 \end{pmatrix}$$

eine positive Matrix, also A primitiv. □

60 Anwendungen des Satzes von Perron-Frobenius

60.1 Konvergenz der Potenzmethode

Der Satz von Perron liefert insbesondere, dass der Spektralradius $\rho(A)$ einer positiven Matrix $A \in \mathbb{R}^{n\times n}$ ein einfacher Eigenwert ist, welcher betragsmäßig größer als alle anderen Eigenwerte von A ist. Sind also $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A, so ist $\lambda_1 = \rho(A) > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Daher kann der Spektralradius einer positiven Matrix mit Hilfe der *Potenzmethode* bzw. der *Vektoriteration* berechnet werden. Hierbei nehmen wir an, dass es ein System linear unabhängiger Eigenvektoren $u_1, \ldots, u_n$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$ gibt. Das Verfahren (Iterationsindizes schreiben wir ausnahmsweise in Klammern nach oben) der Vektoriteration bzw. die Potenzmethode sieht folgendermaßen aus:

- Sei $y^{(0)} = \sum_{j=1}^{n} \alpha_j u_j$ mit $\alpha_1 \neq 0$ gegeben.
- Berechne $x^{(0)} := y^{(0)}/\|y^{(0)}\|_2$.
- Für $k = 1, 2, \ldots$:

  Berechne $y^{(k)} := Ax^{(k-1)}, \quad x^{(k)} := y^{(k)}/\|y^{(k)}\|_2$.

Beispiel: Wir wenden das Verfahren der Vektoriteration auf die positive Matrix

$$A := \begin{pmatrix} 16 & 2 & 1 & 7 \\ 5 & 10 & 3 & 8 \\ 9 & 29 & 5 & 12 \\ 4 & 35 & 14 & 1 \end{pmatrix}$$

an und erhalten mit dem Startwert $y^{(0)} := (1, 1, 1, 1)^T$ die Werte:

k	$x^{(k)}$				$\|y^{(k)}\|_2$
0	0.50000000000	0.50000000000	0.50000000000	0.50000000000	2.00000000000
1	0.30445298833	0.30445298833	0.64403516762	0.63232543730	42.69953161336
2	0.30344897514	0.33241302827	0.64360820366	0.61902243764	34.76850437108
3	0.29308576665	0.32736232902	0.64267849168	0.62761373093	35.81468675857
4	0.29208258217	0.32888602585	0.64370383573	0.62623248005	35.53813943846
5	0.29107440105	0.32849888517	0.64348879322	0.62712546471	35.58686207049
6	0.29095840473	0.32864503475	0.64360827628	0.62698009103	35.56369249042
7	0.29085237270	0.32860561061	0.64358439103	0.62707446256	35.56819345386
8	0.29083990366	0.32862091531	0.64359714743	0.62705913300	35.56580332773
9	0.29082866503	0.32861676425	0.64359461387	0.62706912127	35.56627186843
10	0.29082733359	0.32861838147	0.64359596501	0.62706750452	35.56601954877

Mit Hilfe des Befehl `eig(A)` ermittelt MATLAB die Eigenwerte

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} = \begin{pmatrix} 35.566044910949216 \\ 11.592953677039640 \\ -11.564612550793525 \\ -3.594386037195332 \end{pmatrix}.$$

Als Eigenvektor zu λ_1 ermittelt man mit MATLAB (der Befehl `[V,D]=eig(A)` liefert neben den Eigenwerten von A in der Diagonalen der Diagonalmatrix D eine Matrix V, deren Spalten Eigenvektoren sind, mit $AV = VD$) den Vektor (genauer ist die erste Spalte von V genau das Negative)

$$u_1 = \begin{pmatrix} 0.290825840957337 \\ 0.328618082407270 \\ 0.643595825780208 \\ 0.627068496405870 \end{pmatrix}.$$

die nach 10 Iterationen durch die Vektoriteration ermittelten Werte sind also in etwa den ersten 7 Ziffern korrekt. □

In diesem Beispiel beobachten wir Konvergenz der Vektoriteration. Im folgenden Satz formulieren und beweisen wir hierzu eine Konvergenzaussage, wobei wir uns (zum Teil unnötigerweise) auf den Fall positiver Matrizen beschränken.

Satz *Sei $A \in \mathbb{R}^n$ eine positive Matrix, zu der es linear unabhängige Eigenvektoren $u_1, \ldots, u_n$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$ mit $\lambda_1 = \rho(A) > |\lambda_2| \geq \cdots \geq |\lambda_n|$ gibt. Startet man die Vektoriteration mit einem positiven Vektor $y^{(0)} = \sum_{j=1}^n \alpha_j u_j$ mit $\alpha_1 \neq 0$, so erzeugt dieses Folgen $\{x^{(k)}\}$ und $\{y^{(k)}\}$ derart, dass $\{x^{(k)}\}$ gegen einen positiven Eigenvektor x von A mit $\|x\|_2 = 1$ und $\{\|y^{(k)}\|_2\}$ gegen $\lambda_1 = \rho(A)$ konvergiert.*

Beweis: Durch vollständige Induktion erhält man leicht, dass $x^{(k)} = A^k y^{(0)} / \|A^k y^{(0)}\|_2$, $k = 0, 1, \ldots$. Denn für $k = 0$ ist dies nach Definition von $x^{(0)}$ richtig. Die Aussage sei für $k-1$ richtig. Dann ist

$$y^{(k)} = Ax^{(k-1)} = \frac{A^k y^{(0)}}{\|A^{k-1} y^{(0)}\|_2}$$

und daher

$$x^{(k)} = \frac{y^{(k)}}{\|y^{(k)}\|_2} = \frac{A^k y^{(0)}}{\|A^k y^{(0)}\|_2},$$

womit die Induktionsbehauptung bewiesen ist. Wegen $A^k u_j = \lambda_j^k u_j$ ist weiter

$$A^k y^{(0)} = \sum_{j=1}^n \alpha_j \lambda_j^k u_j = \lambda_1^k \Big[\alpha_1 u_1 + \alpha_2 \Big(\frac{\lambda_2}{\lambda_1}\Big)^k u_2 + \cdots \alpha_n \Big(\frac{\lambda_n}{\lambda_1}\Big)^k u_n \Big].$$

Wegen $\lambda_1 > 0$ ist

$$x^{(k)} = \frac{A^k y^{(0)}}{\|A^k y^{(0)}\|_2} = \frac{\alpha_1 u_1 + \alpha_2 (\lambda_2/\lambda_1)^k u_2 + \cdots + \alpha_n (\lambda_n/\lambda_1)^k u_n}{\|\alpha_1 u_1 + \alpha_2 (\lambda_2/\lambda_1)^k u_2 + \cdots + \alpha_n (\lambda_n/\lambda_1)^k u_n\|_2}.$$

Wegen $|\lambda_j/\lambda_1| < 1$, $j = 2, \ldots, n$, existiert $\lim_{k\to\infty} x^{(k)}$ und es ist

$$x := \lim_{k\to\infty} x^{(k)} = \frac{\alpha_1 u_1}{|\alpha_1|\,\|u_1\|_2} = \operatorname{sign}(\alpha_1)\,\frac{u_1}{\|u_1\|_2}.$$

Hier ist x ein (nichttriviales) Vielfaches von u_1, also selbst ein Eigenvektor zum Eigenwert λ_1, ferner ist $\|x\|_2 = 1$. Da offenbar $x^{(k)}$ wegen $y^{(0)} > 0$ und $A > 0$ ein positiver Vektor ist, ist der Limes x von $\{x^{(k)}\}$ nichtnegativ. Wegen Teil 4 des Satzes von Perron ist x positiv. Daher ist die Konvergenzaussage bezüglich der Folge $\{x^{(k)}\}$ bewiesen.

Es ist

$$\|y^{(k)}\|_2 = \frac{\|A^k y^{(0)}\|_2}{\|A^{k-1} y^{(0)}\|_2} = \lambda_1 \frac{\|\alpha_1 u_1 + \alpha_2(\lambda_2/\lambda_1)^k u_2 + \cdots + \alpha_n(\lambda_n/\lambda_1)^k u_n\|_2}{\|\alpha_1 u_1 + \alpha_2(\lambda_2/\lambda_1)^{k-1} u_2 + \cdots + \alpha_n(\lambda_n/\lambda_1)^{k-1} u_n\|_2}.$$

Wieder nutzen wir aus, dass $|\lambda_j/\lambda_1| < 1$, $j = 2, \ldots, n$, und erhalten

$$\lim_{k\to\infty} \|y^{(k)}\|_2 = \lambda_1.$$

Damit ist der Satz über die Konvergenz der Vektoriteration bei positiven Matrizen bewiesen. □

Bemerkung: Die Geschwindigkeit, mit der die Folgen $\{x^{(k)}\}$ bzw. $\{\|y^{(k)}\|_2\}$ gegen einen positiven Eigenvektor x bzw. $\lambda_1 = \rho(A)$ konvergieren, hängt von dem Quotienten $|\lambda_2|/\lambda_1$ ab. Je kleiner dieser ist, desto besser die Konvergenz. Dies schimmert schon im obigen Konvergenzbeweis durch, kann aber natürlich präzisiert werden. Mit

$$x := \frac{\alpha_1}{|\alpha_1|\,\|u_1\|_2} u_1$$

und

$$u^{(k)} := \alpha_1 u_1 + \sum_{j=2}^{n} \alpha_j(\lambda_j/\lambda_1)^k u_j, \qquad v := \alpha_1 u_1$$

ist für alle k nämlich

$$\begin{aligned}
\|x^{(k)} - x\|_2 &= \Big\| \frac{u^{(k)}}{\|u^{(k)}\|_2} - \frac{v}{\|v\|_2} \Big\|_2 \\
&= \frac{\|\|v\|_2 u^{(k)} - \|u^{(k)}\|_2 v\|_2}{\|u^{(k)}\|_2 \|v\|_2} \\
&= \frac{\|\|v\|_2 (u^{(k)} - v) + (\|v\|_2 - \|u^{(k)}\|_2) v\|_2}{\|u^{(k)}\|_2 \|v\|_2} \\
&\le \frac{\|v\|_2 \|u^{(k)} - v\|_2 + |\|v\|_2 - \|u^{(k)}\|_2|\,\|v\|_2}{\|u^{(k)}\|_2 \|v\|_2} \\
&\le \frac{\|v\|_2 \|u^{(k)} - v\|_2 + \|u^{(k)} - v\|_2 \|v\|_2}{\|u^{(k)}\|_2 \|v\|_2} \\
&= \frac{2\|u^{(k)} - v\|_2}{\|u^{(k)}\|_2} \\
&= \frac{2\|\sum_{j=2}^{n} \alpha_j(\lambda_j/\lambda_1)^k u_j\|_2}{\|\alpha_1 u_1 + \sum_{j=2}^{n} \alpha_j(\lambda_j/\lambda_1)^k u_j\|_2} \\
&\le C\Big(\frac{|\lambda_2|}{\lambda_1}\Big)^k
\end{aligned}$$

mit einer hinreichend großen Konstanten $C > 0$. Hierbei haben wir ausgenutzt, dass $|\lambda_j|/\lambda_1 \le |\lambda_2|/\lambda_1 < 1$, $j = 2, \dots, n$. Wieder für alle k ist entsprechend

$$\begin{aligned}
|\|y^{(k)}\|_2 - \lambda_1| &= \Big|\lambda_1 \frac{\|u^{(k)}\|_2}{\|u^{(k-1)}\|_2} - \lambda_1\Big| \\
&= \lambda_1 \frac{|\|u^{(k)}\|_2 - \|u^{(k-1)}\|_2|}{\|u^{(k-1)}\|_2} \\
&\le \lambda_1 \frac{\|u^{(k)} - u^{(k-1)}\|_2}{\|u^{(k-1)}\|_2} \\
&= \lambda_1 \frac{\sum_{j=2}^n \alpha_j (\lambda_j/\lambda_1)^{k-1}((\lambda_j/\lambda_1) - 1)u_j\|_2}{\|\alpha_1 u_1 + \sum_{j=2}^n \alpha_j (\lambda_j/\lambda_1)^{k-1} u_j\|_2} \\
&\le D\Big(\frac{|\lambda_2|}{\lambda_1}\Big)^k
\end{aligned}$$

mit einer hinreichend großen Konstanten $D > 0$. □

60.2 Das Leontief-Modell in den Wirtschaftswissenschaften

Bisher sind wir auf eine sozusagen innermathematische Anwendung des Satzes von Perron eingegangen, nämlich die Konvergenz der Potenzmethode bei positiven Matrizen. Dies soll nun anders werden, da wir in diesem und dem nächsten Unterabschnitt auf das *Leontief-Modell* in den Wirtschaftswissenschaften und das *Leslie-Modell* in der Populationsdynamik eingehen werden.

Durch das (offene) Leontief[77]-Modell werden Zusammenhänge zwischen der Produktion und der Nachfrage in einer Volkswirtschaft beschrieben. Es seien n Firmen F_j, $j = 1, \dots, n$, gegeben. Jede dieser Firmen stelle ein Produkt P_j her. Für die Herstellung einer Werteinheit von P_k mögen $a_{jk} \ge 0$ Werteinheiten von P_j benötigt. Vorgegeben sei ein *Bedarfsvektor* $b \ge 0$, wobei die j-te Komponente b_j angibt, wie viele Werteinheiten vom Produkt P_j am Markt verlangt werden. Für einen *Produktionsvektor* $x \ge 0$ gibt die j-te Komponente $(Ax)_j = \sum_{k=1}^n a_{jk}x_k$ von Ax den Verbrauch des Produktes P_j für die Herstellung von x_k Einheiten des Produktes P_k an. Soll durch den Produktionsvektor $x \ge 0$ der gesamte Bedarf b gedeckt werden, so hat man x aus

$$x - Ax = b$$

bzw.

$$x_j - \sum_{k=1}^n a_{jk}x_k = b_j, \quad j = 1, \dots, n,$$

zu bestimmen. Die Aufgabe $x - Ax = b$ bzw. $(I - A)x = b$ besitzt für jedes $b \in \mathbb{R}^n$ genau dann eine eindeutige Lösung, wenn 1 kein Eigenwert von A ist. Ist dies der Fall, so ist diese Lösung genau dann für jedes $b \ge 0$ ebenfalls nichtnegativ, wenn $(I - A)^{-1} \ge 0$. Für eine nichtnegative Matrix A ist hierfür $\rho(A) < 1$ eine hinreichende Bedingung. Ist nämlich $\rho(A) < 1$, so existiert eine natürliche Matrixnorm $\|\cdot\|$ mit $\|A\| < 1$, wie wir

[77] Wassily Leontief (1906–1999) erhielt 1973 den Wirtschaftsnobelpreis.

schon auf Seite 210 feststellten. Dann existiert die geometrische Reihe $\sum_{k=0}^{\infty} A^k$ und man weist leicht nach, dass $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$. Die rechtsstehende Reihe hat nur nichtnegative Summanden, ist also selbst nichtnegativ. Ist A zusätzlich irreduzibel, so kann eine natürliche Matrixnorm $\|\cdot\|$ mit $\rho(A) = \|A\|$ angegeben werden. Denn ist $v > 0$ ein nach dem Satz von Perron-Frobenius existierender positiver Eigenvektor von A zum Eigenwert $\rho(A)$, so definiere man die (transformierte Maximum-) Vektornorm $\|\cdot\|_v$ durch

$$\|x\|_v := \max_{j=1,\ldots,n} \frac{|x_j|}{v_j} = \|V^{-1}x\|_\infty$$

mit $V := \operatorname{diag}(v_1, \ldots, v_n)$. Die zugeordnete (natürliche) Matrixnorm $\|\cdot\|_v$ ist durch

$$\|A\|_v := \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v}$$

definiert. Man berechnet leicht, siehe z.B. J. Werner (1992, S. 20) *Numerische Mathematik 1, Vieweg-Verlag, Braunschweig-Wiesbaden*, dass

$$\|A\|_v = \|V^{-1}AV\|_\infty = \max_{i=1,\ldots,n} \Big(\frac{1}{v_i} \underbrace{\sum_{j=1}^{n} a_{ij} v_j}_{=\rho(A)v_i} \Big) = \rho(A).$$

Während $\rho(A) \leq \|A\|$ für jede natürliche Matrixnorm gilt, kann hier, d. h. für eine nichtnegative, irreduzible Matrix $A \in \mathbb{R}^{n\times n}$, mit Hilfe des Perron-Vektors eine natürliche Matrixnorm $\|\cdot\|_v$ mit $\rho(A) = \|A\|_v$ angegeben werden. Weiter gilt:

- *Ist $A \in \mathbb{R}^{n\times n}$ nichtnegativ und irreduzibel sowie $\rho(A) < 1$, so ist $(I-A)^{-1} > 0$.*

Denn: Wegen der Irreduzibilität von A gibt es für jede Position $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\}$ ein $k \in \{0,\ldots,n-1\}$ mit $a_{ij}^{(k)} > 0$, wobei $A^k = (a_{ij}^{(k)})$ (siehe den Beweis des zweiten Teil des Lemmas in Abschnitt 59). Wegen $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$ folgt dann die Behauptung. Dies bedeutet insbesondere: Ist der Bedarfsvektor b nichtnegativ und vom Nullvektor verschieden, so ist der Produktionsvektor $x = (I-A)^{n-1}b$ positiv.

60.3 Das Leslie-Modell in der Populationsdynamik

Beim *Leslie-Modell* handelt es sich um ein mathematisches Modell zur Beschreibung der zeitlichen Änderung einer Population, wobei man sich diese in n Altersklassen unterteilt denkt. Ein Populationsvektor ist ein Vektor $x(t) = (x_1(t), \ldots, x_n(t))^T \in \mathbb{R}^n$. Dieser sagt aus, dass es zur Zeit t genau $x_1(t)$ Individuen in der Altersklasse 1, $x_2(t)$ Individuen in der Altersklasse 2 usw. gibt. Wir gehen davon aus, dass die Anfangspopulation $x(0) \geq 0$ zur Zeit 0 bekannt sei. Weiter wird angenommen, die Geburtenrate $b_j \geq 0$, $j = 1, \ldots, n$, und die Überlebenswahrscheinlichkeit $s_j \geq 0$, $j = 1, \ldots, n-1$, in der Altersklasse j, also die Wahrscheinlichkeit, von der Altersklasse j in die Altersklasse $j+1$ zu kommen, sei bekannt. Das bedeutet genauer: Es ist

$$x_1(t+1) = b_1x_1(t) + b_2x_2(t) + \cdots + b_nx_n(t)$$

und

$$x_j(t+1) = s_{j-1}x_{j-1}(t), \quad j = 2, \ldots, n.$$

In Matrixschreibweise ist also

$$x(t+1) = Lx(t)$$

mit

$$L := \begin{pmatrix} b_1 & b_2 & \cdots & b_{n-1} & b_n \\ s_1 & 0 & \cdots & \cdots & 0 \\ 0 & s_2 & 0 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{n-1} & 0 \end{pmatrix}.$$

Die Matrix L heißt die zugehörige *Leslie-Matrix*. Natürlich ist L eine nichtnegative Matrix. Weiter ist L irreduzibel, wenn $s_1, \ldots, s_{n-1}$ und $b_2, \ldots, b_n$ positiv sind, was wir jetzt voraussetzen. Denn dann ist der zugehörige Graph stark zusammenhängend. Weiter ist in diesem Fall sogar $L^{n+2} > 0$, denn zu je zwei Knoten i und j gibt es in dem zu L gehörenden Graphen einen gerichteten Weg von i nach j der Länge $n+2$. Insbesondere ist L dann primitiv. Durch $x(k) = L^k x(0)$ ist die Population nach k Zeitschritten gegeben. Wir definieren

$$X(t) := \frac{1}{e^T x(t)} x(t).$$

Die j-te Komponente $X_j(t) = x_j(t)/e^T x(t)$ gibt das Verhältnis zwischen der Population in der j-ten Altersklasse und der gesamten Population zur Zeit t an. Mit p bzw. q seien Perron-Vektoren zu L bzw. L^T bezeichnet. Wegen der Primitivität von L und (siehe Satz auf S. 218)

$$\lim_{k\to\infty} \Big(\frac{L}{\rho(L)}\Big)^k = \frac{pq^T}{p^T q}$$

ist

$$\begin{aligned} X(k) &= \frac{1}{e^T x(k)} x(k) \\ &= \frac{1}{e^T L^k x(0)} L^k x(0) \\ &= \frac{1}{e^T (L/\rho(L))^k x(0)} (L/\rho(L))^k x(0) \\ &\to \frac{1}{e^T (pq^T/(p^T q)) x(0)} (pq^T/(p^T q)) x(0) \\ &= \frac{p}{e^T p}. \end{aligned}$$

Die Folge $\{X(k)\}$, in welchem das Verhältnis zwischen der Population der jeweiligen Altersklassen und der gesamten Population dargestellt ist, konvergiert also gegen einen positiven Vektor, und zwar den normierten Perron-Vektor von L. Insbesondere ist dieser Limes von der Ausgangspopulation unabhängig.

Beispiel: Das folgende Beispiel findet man an verschiedenen Stellen im Internet. In einem Nationalpark Australiens wurde eine bestimmte Possum-Art (ein Mitglied der Beuteltierfamilie) beobachtet. Dabei wurden folgende Daten gewonnen:

	1	2	3	4	5
Alter in Jahren	$0-1$	$1-2$	$2-3$	$3-4$	$4-5$
Ausgangspopulation $x(0)$	403	157	102	52	11
Überlebenswahrscheinlichkeit s_j	0.6	0.8	0.8	0.4	0
Geburtenrate b_j	0	1.3	1.8	0.9	0.2

Die zugehörige Leslie-Matrix sowie die Ausgangspopulation sind durch

$$L := \begin{pmatrix} 0 & 1.3 & 1.8 & 0.9 & 0.2 \\ 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \end{pmatrix}, \qquad x(0) := \begin{pmatrix} 403 \\ 157 \\ 102 \\ 52 \\ 11 \end{pmatrix}$$

gegeben. Jetzt überprüfen wir obiges Ergebnis, dass nämlich $\lim_{k\to\infty} X(k) = p/e^T p$ mit einem Perron-Vektor p zu L, indem wir $X(0), \ldots, X(5)$ sowie $p/e^T p$ ausrechnen.

$X(0)$	$X(1)$	$X(2)$	$X(3)$	$X(4)$	$X(5)$	$p/e^T p$
0.4817	0.5122	0.5033	0.5022	0.5047	0.5037	0.5038
0.2667	0.2172	0.2375	0.2352	0.2327	0.2342	0.2339
0.1386	0.1603	0.1343	0.1480	0.1453	0.1439	0.1448
0.0900	0.0833	0.0991	0.0837	0.0914	0.0899	0.0897
0.0229	0.0271	0.0257	0.0309	0.0258	0.0283	0.0278

Die Übereinstimmung ist zufriedenstellend. Wie man leicht nachrechnet, wird die Gesamtpopulation der Possum-Art immer größer. Denn ist $x(k) = L^k x(0)$ die Population zur k-ten Zeitstufe, so ist

$$x(k) = \rho(L)^k \underbrace{\left(\frac{L}{\rho(L)}\right)^k x(0)}_{\to (q^T x(0)/p^T q)p > 0} .$$

Da aber $\rho(L) > 1$, ist $\lim_{k\to\infty} x_j(k) = +\infty$ für alle Altersklassen j. □

Beispiel: Auch das Populationsmodell zur Hawaiischen Green Sea Schildkröte findet man an verschiedenen Stellen im Netz, z. B. bei `http://isolatium.uhh.hawaii.edu/linear/ch6/green.htm`. Die Population wird in 5 Altersgruppen aufgeteilt.

1. Eier, Geschlüpfte (< 1).
2. Jungtiere ($1 - 16$).
3. Fast ausgewachsene Tiere ($17 - 24$).
4. Erstbrüter (25).

5. Brüter $(26 - 50)$.

Die Leslie-Matrix wird hier modifiziert, weil in einer Altersklasse ein Anteil auch in der Altersklasse verbleiben kann. Genauer hat die Leslie-Matrix jetzt die Form

$$L = \begin{pmatrix} b_1 & b_2 & b_3 & b_4 & b_5 \\ q_1 & p_2 & 0 & 0 & 0 \\ 0 & q_2 & p_3 & 0 & 0 \\ 0 & 0 & q_3 & p_4 & 0 \\ 0 & 0 & 0 & q_4 & p_5 \end{pmatrix}.$$

Hierbei gibt b_j die Anzahl der Eier, die eine Schildkröte in der Altersklasse j legt. Durch p_j wird der Anteil der Population in der j-ten Altersklasse angegeben, der ein weiteres Jahr in derselben Altersklasse verbleibt, während q_j der Anteil der Population in der j-ten Altersklasse gegeben ist, der überlebt und in die $(j + 1)$-te Altersklasse wechselt. In der oben angegebenen Internet-Quelle ist die Leslie-Matrix durch

$$L := \begin{pmatrix} 0 & 0 & 0 & 280 & 70 \\ 0.23 & 0.679 & 0 & 0 & 0 \\ 0 & 0.001 & 0.711 & 0 & 0 \\ 0 & 0 & 0.039 & 0 & 0 \\ 0 & 0 & 0 & 0.89 & 0.907 \end{pmatrix}$$

gegeben. Man stellt fest, dass L primitiv ist und $\rho(L) < 1$, daher wird die Population letztendlich aussterben. □

60.4 Google's PageRank

Ein weiterer Anwendungsbereich der Perron-Frobenius-Theorie findet sich beim sogenannten *Ranking* bzw. der Bestimmung einer Rangordnung vergleichbarer Objekte, wie z. B. Sportmannschaften, Universitäten, Seiten im Web oder Algorithmen zur Lösung bestimmter Aufgaben.

Wir wollen jetzt das Grundprinzip von Googles PageRank-Verfahren zur Bewertung von Webseiten schildern. Eine ausführliche Darstellung findet man bei A. N. Langville, C. D. Meyer (2006), siehe auch K. Bryan, T. Leise (2006). Zur Erläuterung des Titels dieses Aufsatzes erfährt man: $ 25,000,000,000 was the approximate market value of Google when the company went public in 2004. Hier ist eine schöne Motivation aus einer Webseite des Department of Mathematics der University of California at San Diego:

- *We quickly sketch the backbone of the Google ranking algorithm (officially called the PageRank algorithm). ..., but this is the basic idea that made Google into such a success. It's also a demonstration of how sometimes there are old mathematical results, even ones that can be understood and used by undergraduates, that can be turned into multi-million dollar ideas!*

Im Internet gibt es sehr viele Seiten zu Google's PageRank-Algorithmus, siehe z. B. `http://www.ams.org/samplings/feature-column/fcarc-pagerank`.

Wir gehen aus von n Webseiten (mit sehr großem n), die zum Teil miteinander verlinkt sind. Diese denken wir uns als Knoten in einem gerichteten Graphen, wobei es einen Pfeil von j nach i gibt, wenn in j ein Verweis bzw. Link auf die Seite i existiert. Bei der Bewertung einer Webseite i sollten die Bewertungen der Seiten j eine Rolle spielen, die einen Link auf i besitzen. Hierbei sollte berücksichtigt werden, wie viele Links in der Seite j enthalten sind. Um dies zu formalisieren definieren wir die sogenannte *Hyperlink-Matrix* $H = (h_{ij}) \in \mathbb{R}^{n\times n}$ (Vorsicht: manchmal wird auch die *transponierte* Matrix so bezeichnet) durch

$$h_{ij} := \begin{cases} 1, & \text{falls es einen Link von Seite } j \text{ zur Seite } i \text{ gibt,} \\ 0, & \text{sonst.} \end{cases}$$

Dann ist

$$n_j := \sum_{i=1}^{n} h_{ij}, \qquad j = 1, \ldots, n,$$

die Gesamtzahl der von j ausgehenden Links. Anschließend definieren wir $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ durch $a_{ij} := h_{ij}/n_j$ und bezeichnen diese Matrix als *normalisierte Hyperlink-Matrix*. Geht von j kein Link aus, ist also $n_j = 0$, so vereinbaren wir, dass $a_{ij} = 1/n$, $i = 1, \ldots, n$. Mit einem Dämpfungsfaktor $d \in (0, 1)$, z. B. $d := 0.85$, ist der PageRank-Vektor x zu bestimmen aus der Gleichung

$$(*) \qquad [(1-d)\frac{ee^T}{n} + dA]x = x.$$

Hierbei ist $e \in \mathbb{R}^n$ einmal wieder der Vektor, dessen Komponenten sämtlich gleich 1 sind. Daher ist $ee^T \in \mathbb{R}^{n\times n}$ die $n \times n$-Matrix, deren Einträge sämtlich gleich 1 sind. Komponentenweise geschrieben bedeutet $(*)$, dass

$$(1-d)\frac{1}{n}\sum_{j=1}^{n} x_j + d\sum_{j\in L_i} \frac{x_j}{n_j} = x_i, \qquad i = 1, \ldots, n,$$

wobei L_i die Menge der Webseiten j ist, die einen Link auf die Seite i besitzen. Durch Einführung des Dämpfungsfaktors $d \in (0, 1)$ ist die Matrix $A_d := (1-d)ee^T/n + dA$ eine positive Matrix. Weiter wird durch d gesteuert, wie sehr außer den verlinkten Seiten auch andere Seiten für die Bedeutung bzw. den PageRank relevant sind. Ist d nahe bei 1, so wird die wahre Link-Struktur stark bewertet. Durch obige Modifikation der normalisierten Hyperlink-Matrix für den Fall, dass von einer Seite kein Link ausgeht (ein Verweis auf eine solche Seite heißt ein *dangling link*), haben wir erreicht, dass A spaltenstochastisch, also die Spaltensummen sämtlich gleich 1 ist bzw. $e^T A = e^T$ gilt. Dann ist auch A_d spaltenstochastisch, also $A_d^T e = e$ bzw. 1 ein Eigenwert von A_d^T. Folglich ist in diesem Falle 1 auch ein Eigenwert von A_d. Die folgende Aussage formulieren wir als

Lemma *Ist $A \in \mathbb{R}^{n\times n}$ nichtnegativ und sind die Spaltensummen von A sämtlich gleich 1, so ist $\rho(A) = 1$. Ist A zusätzlich irreduzibel, so besitzt die Gleichung $Ax = x$ eine bis auf eine positive multiplikative Konstante eindeutige positive Lösung.*

Beweis: Da die Spaltensummen von A sämtlich gleich 1 sind, ist 1 ein Eigenwert von A^T (mit zugehörigem Eigenvektor e, dessen Komponenten sämtlich gleich 1 sind) und damit auch ein Eigenwert von A. Wir zeigen, dass $\rho(A) = 1$. Da 1 ein Eigenwert von A ist, genügt es nachzuweisen, dass $|\lambda| \leq 1$ für jeden Eigenwert λ von A. Sei also λ ein Eigenwert von A und x ein zugehöriger Eigenwert. O. B. d. A. ist $\|x\|_1 = 1$, wobei die Betragssummennorm $\|\cdot\|_1$ durch $\|x\|_1 := \sum_{i=1}^n |x_i|$ definiert ist. Dann ist $\sum_{j=1}^n a_{ij}x_j = \lambda x_i$, $i = 1, \ldots, n$, und daher

$$\begin{aligned}
|\lambda| &= |\lambda| \sum_{i=1}^n |x_i| \\
&= \sum_{i=1}^n |\lambda x_i| \\
&= \sum_{i=1}^n \Big|\sum_{j=1}^n a_{ij}x_j\Big| \\
&\leq \sum_{i=1}^n \sum_{j=1}^n a_{ij}|x_j| \\
&= \sum_{j=1}^n \Big(\underbrace{\sum_{i=1}^n a_{ij}}_{=1}\Big)|x_j| \\
&= \sum_{j=1}^n |x_j| \\
&= 1.
\end{aligned}$$

Ist zusätzlich A irreduzibel, so gibt es wegen des Satzes von Perron-Frobenius zu $\rho(A) = 1$ einen positiven Eigenvektor, der bis auf ein positives Vielfaches eindeutig bestimmt ist. Das Lemma ist damit bewiesen. □

Insbesondere wissen wir, dass die Gleichung $(*)$ bis auf ein positives Vielfaches genau eine positive Lösung x besitzt, den PageRank-Vektor. Der Dämpfungsfaktor ist u. a. deshalb eingeführt worden, um diese Eindeutigkeitsaussage machen zu können. Wenn z. B. von einer Webseite kein Link ausgeht, so ist die entsprechende Spalte in der Hyperlink-Matrix eine Nullspalte. Bei der normalisierten Hyperlink-Matrix haben wir diese Spalte aber durch $(1/n)e$ ersetzt. Ein Verweis auf eine Seite, die selbst keinen ausgehenden Link besitzt, nennt man einen *dangling link*. Weiter sind wir sicher, dass die Power-Methode, angewandt auf die Gleichung $(*)$ bzw. die positive Matrix A_d, konvergiert, wenn es zu A_d ein vollständiges System von Eigenvektoren gibt bzw. A_d diagonalisierbar ist. Dies haben wir als Satz zu Beginn dieses Abschnitts formuliert und bewiesen.

Beispiel: Wir betrachten sechs Webseiten 1, 2, 3, 4, 5, 6 über einen bestimmten Begriff

(z. B. *ranking perron*) und wollen bezüglich ihrer Wichtigkeit ein Ranking bzw. eine Rangfolge bilden. In Abbildung 79 geben wir die Linkstruktur der Webseiten an. Die

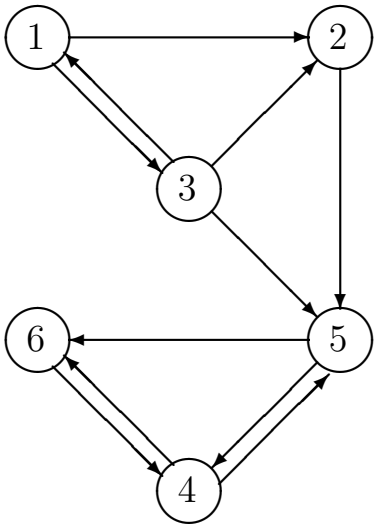


Abbildung 79: Beziehungen zwischen 6 Webseiten

normalisierte Hyperlink-Matrix ist

$$A := \begin{pmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 1 \\ 0 & 1 & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

In diesem Fall ist die Matrix A reduzibel, da der zugehörige gerichtete Graph nicht stark zusammenhängend ist. Normieren wir den PageRank-Vektor auf die euklidische Länge gleich 1, so erhalten wir für verschiedene Dämpfungsfaktoren die folgenden Ergebnisse:

$d = 0.85$		$d = 0.9$		$d = 1.0$	
x	Rang	x	Rang	x	Rang
0.0706	6	0.0464	6	0.0000	4
0.1006	4	0.0673	4	0.0000	4
0.0784	5	0.0518	5	0.0000	4
0.6970	1	0.7142	1	0.7428	1
0.4524	3	0.4284	3	0.3714	3
0.5369	2	0.5450	2	0.5571	2

Wäre die Matrix A irreduzibel, so müsste hätte A zu $\rho(A) = 1$ einen positiven Eigenvektor. Das ist hier nicht der Fall, da $x = (0, 0, 0, 1, \frac{1}{2}, \frac{3}{4})^T$ ein (noch nicht auf die euklidische Länge 1 normierter) Eigenvektor von A zum Eigenwert 1 ist. □

Beispiel: In Abbildung 80 geben wir einen gerichteten Graphen an, der die Link-Struktur zwischen sieben Webseiten angibt, siehe A. N. Langville, C. D. Meyer (2006, S. 60). Die zugehörige normalisierte Hyperlink-Matrix (man beachte, dass sich

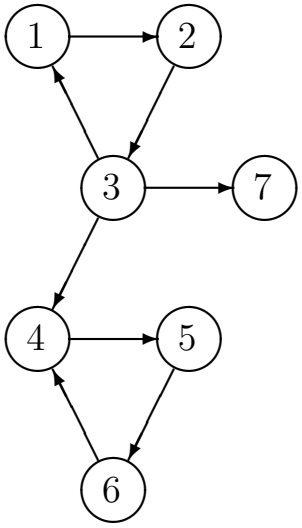


Abbildung 80: Gerichteter Graph für sieben Seiten des Web

in Webseite 7 kein Verweis auf eine andere Seite findet und daher die entsprechende ursprüngliche Nullspalte modifiziert wird) ist

$$A := \begin{pmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{7} \\ 1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{7} \\ 0 & 1 & 0 & 0 & 0 & 0 & \frac{1}{7} \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 1 & \frac{1}{7} \\ 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{7} \\ 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{7} \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{7} \end{pmatrix}$$

Für verschiedene Dämpfungsfaktoren geben wir den PageRank-Vektor und den Rang der Webseite an, wobei wir im Gegensatz zu Langville-Meyer eine Normierung auf euklidische Länge 1 (statt Normierung auf Betragssummennorm gleich 1) vornehmen.

$d = 0.8$		$d = 0.9$		$d = 0.99$	
x	Rang	x	Rang	x	Rang
0.1506	6	0.0859	6	0.0095	6
0.2048	5	0.1187	5	0.0133	5
0.2483	4	0.1482	4	0.0171	4
0.5576	1	0.5782	1	0.5791	1
0.5304	2	0.5618	2	0.5772	2
0.5087	3	0.5470	3	0.5753	3
0.1506	6	0.0859	6	0.0095	6

Die Matrix A ist reduzibel. Denn mit $I := \{1, 2, 3, 7\}$, $J := \{4, 5, 6\}$ ist $a_{ij} = 0$ für alle $(i, j) \in I \times J$. Daher kann man nicht erwarten, dass es zu $\rho(A) = 1$ einen *positiven* Eigenvektor von A gibt. In der Tat ist $x = (0, 0, 0, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0)^T$ der auf euklidische Länge 1 normierte Eigenvektor von A zu $\rho(A) = 1$. □

60.5 Weitere Ranking-Probleme

Wir betrachten jetzt die Aufgabe, eine Rangfolge unter Sportmannschaften anzugeben. Eine der ersten Arbeiten hierzu stammt von J. P. Keener (1993). Diese Arbeit bezieht sich vor allem auf das Ranking amerikanischer College Football Teams. Hier gibt es über 100 Mannschaften in der besten Division I-A, die in Conferences mit zehn bis zwölf Mannschaften unterteilt wird. Jede Mannschaft macht zwölf bis vierzehn Spiele, vorwiegend, aber nicht nur gegen Mannschaften der eigenen Conference. Wichtig ist also, dass nicht wie etwa in der Fußball-Bundesliga jede Mannschaft mit Hin- und Rückspielen gegen jede andere spielt. Ein Ranking der besten 25 Teams wird durch Umfragen bei Journalisten durch Associated Press und United Press International veröffentlicht. Hierbei wird ein Gewinn über eine starke Mannschaft stärker bewertet als einer gegen eine relativ schwache Mannschaft. Nicht notwendig wird also eine Mannschaft mit dem besten record an erster Stelle eines Rankings stehen. Ähnlich dem PageRank einer Webseite, die von dem PageRank der auf diese Seite verweisenden Seiten abhängt, sollte die Bewertung eines Teams von der Bewertung der gegnerischen Mannschaften und natürlich von dem gegen diese Gegner erzielten Spielergebnis abhängen. Die Idee bei der Berechnung eines Ranking ist im Prinzip fast immer dieselbe. Sind insgesamt n Mannschaften beteiligt, so stelle man eine nichtnegative Matrix $A \in \mathbb{R}^{n\times n}$ auf. Hier gibt der Eintrag a_{ij} an, wie erfolgreich Mannschaft i gegen Mannschaft j war. Je größer a_{ij}, desto besser hat Mannschaft i gegen Mannschaft j abgeschnitten. Außerdem sollte hier die Anzahl der von Mannschaft i bestrittenen Spiele eingehen, worauf man natürlich verzichten kann, wenn die Gesamtzahl der Spiele für alle Mannschaften gleich ist, wie das etwa in der Fußballbundesliga der Fall ist (nicht aber im amerikanischen College-Football). In der Wahl der Bewertungsmatrix hat man Spielraum. Der Rang x_i von Verein i sollte (unabhängig von i) proportional zu $\sum_{j=1}^{n} a_{ij}x_j$ sein, also einer gewichteten Summe der Ränge der Gegner. Dies führt auf die Eigenwertaufgabe $Ax = \lambda x$ und die Bestimmung eines positiven (bzw. nichtnegativen) Eigenvektors von A. Ordnet man die Komponenten des Vektor x der Größe nach, so erhält man das (natürlich von der Bewertungsmatrix A abhängige) Ranking der beteiligten Mannschaften.

Jetzt wollen wir obige Idee zur Bestimmung eines Ranking auf die deutsche Fußball-Bundesliga anwenden und untersuchen, ob wir hierdurch eine andere Abschlusstabelle der Saison 2010/2011 erhalten. In der Bewertungsmatrix $A = (a_{i,j}) \in \mathbb{R}^{18\times 18}$ gibt der nichtnegative Eintrag $a_{i,j}$ an, wie viele Punkte der Verein i in Hin- und Rückspielen gegen den Verein j erreicht hat. Natürlich sind andere Bewertungen möglich, in die z. B. auch erzielte Tore eingehen könnten. Hierauf gehen wir in einem späteren Beispiel noch näher ein. In Tabelle 2 geben wir die beteiligten Mannschaften und ihren Platz in der offiziellen Abschlusstabelle an, außerdem einen auf euklidische Länge gleich 1 normierten Perron-Vektor einer gleich zu definierenden nichtnegativen Matrix A sowie den daraus resultierenden "Perron-Platz". Wir glauben nicht,

dass diese auf dem Satz von Perron-Frobenius basierende Platzierungsmethode gegen die übliche durchgesetzt werden kann. Es ist aber trotzdem interessant, die (geringfügigen) Unterschiede zu untersuchen. Meisterschaft und Abstieg sind unabhängig von der Berechnungsmethode. Aber nach der hier durchgeführten Berechnung hätte Wolfsburg und nicht Mönchengladbach das Relegationsspiel bestreiten müssen. Das ist auch kein Wunder, denn Mönchengladbach hat gegen besser postierte Mannschaften relativ gut abgeschnitten, jedenfalls besser als Wolfsburg. Weiter wird nach der "Perron-Methode" der 1. FC Köln statt auf dem 10. Rang auf dem 7. platziert. Wir

Platz	Punkte	Verein	Perron-Vektor	Perron-Platz
1	75	Borussia Dortmund	0.3725	1
2	68	Bayer 04 Leverkusen	0.3208	2
3	65	FC Bayern München	0.3058	3
4	60	Hannover 96	0.2855	4
5	58	1. FSV Mainz 05	0.2738	5
6	47	1. FC Nürnberg	0.2251	6
7	46	1. FC Kaiserslautern	0.2132	8
8	45	Hamburger SV	0.2124	9
9	44	SC Freiburg	0.2074	11
10	44	1. FC Köln	0.2196	7
11	43	1899 Hoffenheim	0.2094	10
12	42	VfB Stuttgart	0.1940	13
13	41	SV Werder Bremen	0.2016	12
14	40	FC Schalke 04	0.1913	15
15	38	VfL Wolfsburg	0.1833	16
16	36	Borussia Mönchengladbach	0.1924	14
17	34	Eintracht Frankfurt	0.1641	17
18	29	FC St. Pauli	0.1401	18

Tabelle 2: Die Saison 2010/2011 in der Fußballbundesliga

bilden nun die nichtnegative Matrix $A = (a_{i,j}) \in \mathbb{R}^{18\times 18}$, wobei $a_{i,j}$ die Anzahl der Punkte ist, die der Verein i gegen den Verein j in den beiden Hin- und Rückspielen erkämpft hat. Z.B. ist $a_{11,13} = 3$ und $a_{13,11} = 3$, da Hoffenheim im Hinspiel (4:1) gegen Werder Bremen gewonnen und im Rückspiel (1:2) verloren hat. Wir erhalten die folgende Bewertungsmatrix A, wobei wir (das ist etwas mühsam) die Ergebnisse `http://www.bundesliga.de/de/liga/saisonrueckblick/2010/spieltag1.php` und ent-

sprechenden Seiten entnommen haben:

$$A = \begin{pmatrix}
0 & 3 & 6 & 6 & 4 & 6 & 4 & 4 & 6 & 6 & 1 & 4 & 3 & 4 & 6 & 3 & 3 & 6 \\
3 & 0 & 1 & 4 & 3 & 1 & 6 & 4 & 4 & 3 & 4 & 6 & 2 & 6 & 6 & 3 & 6 & 6 \\
0 & 4 & 0 & 3 & 3 & 4 & 3 & 4 & 6 & 1 & 6 & 6 & 4 & 3 & 4 & 4 & 4 & 6 \\
0 & 1 & 3 & 0 & 6 & 3 & 6 & 4 & 6 & 3 & 3 & 3 & 4 & 3 & 3 & 3 & 6 & 3 \\
1 & 3 & 3 & 0 & 0 & 4 & 6 & 3 & 1 & 3 & 6 & 3 & 4 & 3 & 3 & 6 & 3 & 6 \\
0 & 4 & 1 & 3 & 1 & 0 & 3 & 4 & 1 & 3 & 1 & 6 & 3 & 4 & 6 & 1 & 3 & 3 \\
1 & 0 & 3 & 0 & 0 & 3 & 0 & 1 & 3 & 4 & 1 & 4 & 6 & 6 & 4 & 6 & 1 & 3 \\
1 & 1 & 1 & 1 & 3 & 1 & 4 & 0 & 0 & 3 & 4 & 3 & 3 & 6 & 3 & 4 & 6 & 1 \\
0 & 1 & 0 & 0 & 4 & 4 & 3 & 6 & 0 & 3 & 6 & 6 & 0 & 0 & 3 & 3 & 4 & 1 \\
0 & 3 & 4 & 3 & 3 & 3 & 1 & 3 & 3 & 0 & 2 & 3 & 3 & 3 & 1 & 0 & 6 & 3 \\
4 & 1 & 0 & 3 & 0 & 4 & 4 & 1 & 0 & 2 & 0 & 1 & 3 & 6 & 1 & 3 & 6 & 4 \\
1 & 0 & 0 & 3 & 3 & 0 & 1 & 3 & 0 & 3 & 4 & 0 & 4 & 4 & 1 & 6 & 3 & 6 \\
3 & 2 & 0 & 1 & 1 & 3 & 0 & 3 & 6 & 3 & 3 & 1 & 0 & 1 & 1 & 4 & 2 & 6 \\
1 & 0 & 3 & 3 & 3 & 1 & 0 & 0 & 6 & 3 & 0 & 1 & 4 & 0 & 4 & 1 & 4 & 6 \\
0 & 0 & 1 & 3 & 3 & 0 & 1 & 3 & 3 & 4 & 4 & 4 & 4 & 1 & 0 & 4 & 1 & 2 \\
3 & 3 & 1 & 3 & 0 & 4 & 0 & 1 & 3 & 6 & 3 & 0 & 1 & 4 & 1 & 0 & 3 & 0 \\
3 & 0 & 1 & 0 & 3 & 3 & 4 & 0 & 1 & 0 & 0 & 3 & 2 & 1 & 4 & 3 & 0 & 6 \\
0 & 0 & 0 & 3 & 0 & 3 & 3 & 4 & 4 & 3 & 1 & 0 & 0 & 0 & 2 & 6 & 0 & 0
\end{pmatrix}$$

Den auf die euklidische Länge 1 normierten Eigenvektor zum Spektralradius $\rho(A) \approx 44.5265$ haben wir in Tabelle 2 als Perron-Vektor angegeben.

Nun kommen wir zu einem anderen Ranking-Problem, und zwar betrachten wir einen Wettkampf unter n Mannschaften, bei dem nicht jede Mannschaft gegen jede andere spielt. Wir geben ein Beispiel an, bei dem kein Unentschieden auftritt bzw. nicht auftreten kann (wie etwa beim Basketball). Außerdem gehen wir davon aus, dass die Höhe eines Gewinns bzw. einer Niederlage für das Ranking keine Rolle spielt.

Beispiel: In Abbildung 81 geben wir einen gerichteten Graphen mit sechs Knoten an, die für sechs Mannschaften stehen. Ein Pfeil von Knoten i zum Knoten j bedeutet, dass Mannschaft i die Mannschaft j geschlagen hat. Nun wollen wir ein Ranking der

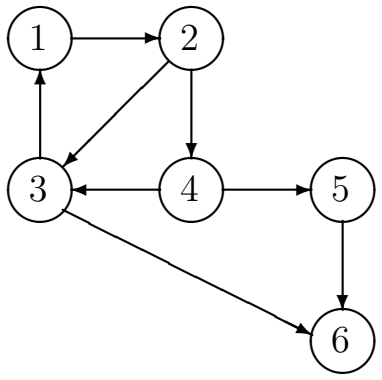


Abbildung 81: Was ist die beste Mannschaft?

sechs Mannschaften angeben. Klar ist, dass 6 die schlechteste Mannschaft ist, da sie

ihre beiden Spiele verloren hat. Was aber ist die beste Mannschaft? Sei a_{ij} die Anzahl der Spiele, die Mannschaft i gegen Mannschaft j gewonnen hat, dividiert durch die Gesamtzahl der Spiele von Mannschaft i. Z. B. ist $a_{12} = \frac{1}{2}$, da Mannschaft 1 gegen 2 einmal gewonnen und insgesamt zwei Spiele gemacht hat. Dann ist

$$A = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

die Bewertungsmatrix zu den angegebenen Spielausgängen. Diese ist nichtnegativ und offensichtlich reduzibel. Der auf euklidische Länge 1 normierte Eigenvektor x zum Eigenwert $\rho(A) \approx 0.4067$ ist

Team	x	Rang
1	0.6769	1
2	0.5506	2
3	0.4161	3
4	0.2558	4
5	0.0000	5
6	0.0000	5

Hiernach ist also 1 die beste Mannschaft. Ungerecht mag erscheinen, dass Team 5 genauso schlecht eingestuft wird wie Team 6. □

Wir geben ein weiteres Beispiel an, das wir A. Y. Govan, C. D. Meyer, R. Albright (2008) entnehmen.

Beispiel: Wir betrachten sechs Mannschaften der NFL (National Football League) und ihre Spiele in der Saison 2007. Die Mannschaften sind

1. Carolina Panthers (Car),
2. Dallas Cowboys (Dal),
3. Houston Texans (Hou),
4. New Orleans Saints (NO),
5. Philadelphia Eagles (Phi),
6. Washington Redskins (Was).

In der Saison 2007 gab es zwischen diesem Mannschaften in der regulären Saison insgesamt 12 Spiele, wobei es die folgenden Spielresultate gab (traditionell wird im amerikanischen Sport die Heimmannschaft als zweite Mannschaft genannt), die wir in Tabelle 3 auflisten. Die Ergebnisse haben wir `http://en.wikipedia.org/wiki/2007_NFL_season` entnommen. Zunächst gehen wir ähnlich vor wie im letzten Beispiel, berücksichtigen also nicht, wie hoch ein Sieg bzw. eine Niederlage ausgefallen ist. Wie oben

Spielpaarung	Ergebnis
Hou-Car	$34-21$
Car-NO	$16-13$
NO-Car	$31-\ 6$
Dal-Car	$20-13$
Dal-Phi	$38-17$
Was-Dal	$23-28$
Phi-Dal	$10-\ 6$
Dal-Was	$6-27$
NO-Hou	$10-23$
Phi-NO	$38-23$
Was-Phi	$20-12$
Phi-Was	$33-25$

Tabelle 3: Spielergebnisse in der NFL-Saison 2007

sei a_{ij}, der (i,j)-Eintrag der Matrix A, die Anzahl der Spiele, die Mannschaft i gegen Mannschaft j gewonnen hat, dividiert durch die Gesamtzahl der Spiele von Mannschaft i. Für die sechs Mannschaften der NFL in der oben angegebenen Reihenfolge erhalten wir

$$A := \begin{pmatrix} 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

Als auf euklidische Länge 1 normierten Eigenvektor zu $\rho(A) \approx 0.3449$ erhalten wir $x = (0, 0.3564, 0, 0, 0.6146, 0.7037)^T$. Die ersten drei Mannschaften wären also Washington, Philadelphia und Dallas, die anderen drei Mannschaften landen gleichauf an letzter Stelle. Da A reduzibel ist, können wir nicht erwarten, dass der Eigenvektor x positiv ist. Das Ergebnis scheint uns nicht angemessen zu sein, da Houston trotz zweier Siege in zwei Spielen eigentlich einen besseren Rang verdient hätte. Obwohl berücksichtigt werden muss, dass die zwei Siege gegen die ziemlich schlechten Mannschaften Carolina und New Orleans gelangen. Eine Möglichkeit, die Irreduzibilität der Bewertungsmatrix zu erzwingen, besteht darin, diese durch $A + \epsilon e e^T$ zu ersetzen. Ist z. B. $\epsilon = 0.01$, so ist

$$x = (0.0171, 0.3630, 0.0534, 0.0171, 0.6152, 0.6974)^T$$

der zugehörige Perron-Vektor. Nach wie vor wären also Washington, Philadelphia und Dallas in dieser Reihenfolge die ersten drei Mannschaften. An die vierte Stelle schiebt sich Houston, während Carolina und New Orleans gemeinsam an letzter Stelle bleiben.

Bei der gerade eben angewandte Methode spielen die erzielten Punkte oder Tore keine Rolle, wichtig ist nur Sieg oder Niederlage. Das wird nun anders. Hierzu bezeichnen wir mit S_{ij} die Anzahl der von Mannschaft i im Spiel gegen Mannschaft j erzielten Punkte oder Tore, entsprechend ist S_{ji} definiert. Wir gehen davon aus, dass die Mannschaften i und j nur ein Spiel gegeneinander austragen. Ist dies nicht der Fall, so werden die erzielten Tore bzw. Punkte für jede Mannschaft aufaddiert[78]. Mit n_i bezeichnen wir die Anzahl der von Team i gespielten Spiele. In der NFL machen 32 Teams jeweils 16 Spiele. Wenn wir also alle Spiele betrachten, können wir auf eine Normalisierung mittels der jeweiligen Spieleanzahl verzichten. Bei A. N. Langville, C. D. Meyer (2012, S. 42–44) werden *alle* Spiele der Saison 2009 zu Grunde gelegt. In unserem obigen kleinen NFL-Beispiel ist die Anzahl der Spiele unterschiedlich, deswegen führen wir mittels n_i, der Anzahl der von Team i gemachten Spiele, eine Normalisierung durch. Eine Möglichkeit besteht darin,

$$a_{ij} := \frac{1}{n_i} \cdot \frac{S_{ij}}{S_{ij} + S_{ji}}$$

zu setzen. Für ein torloses Unentschieden (was für ein Football-Spiel sehr ungewöhnlich ist) müsste man eine gesonderte Definition finden. Hiervon einmal abgesehen, hat diese Definition den Nachteil, dass ein knappes Ergebnis wie 3-0 im Football (oder 1-0 im Fußball) für Team i gegen Team j zu $a_{ij} = 1/n_i$ und $a_{ji} = 0$ führt. In diesem Fall erhielte Team j trotz der knappen Niederlage überhaupt keine Belohnung für seine Bemühungen, jedenfalls weniger als bei einer hohen Niederlage, bei welcher es wenigstens einmal gepunktet hat. Dies könnte Anlass zu einer Modifikation mittels

$$a_{ij} := \frac{1}{n_i} \cdot \frac{S_{ij} + 1}{S_{ij} + S_{ji} + 2}$$

sein. Aber auch diese Bewertung hat Nachteile. Es wäre wünschenswert, dass zwischen der Bewertung eines hohen und eines sehr hohen Sieges kein zu großer Unterschied besteht. Denn einem klar überlegenen Team sollte es nicht zum Nachteil gereichen, dass es sozusagen großmütig ist und nicht aus Bewertungsgründen gezwungen ist, die Niederlage demütigend zu machen. Genau diese Gefahr besteht aber bei obigem Modell nicht der Fall. Um dies deutlich zu machen, geben wir in Abbildung 82 die Funktion $a(S) := (S + 1)/(S + S_{ji} + 2)$ für $S_{ji} = 0, 3$ auf $[0, 10]$ an. Von J. P. Keener (1993) stammt der Vorschlag, die Funktion h durch

$$h(x) := \frac{1}{2} + \frac{1}{2}\text{sign}\left(x - \frac{1}{2}\right)\sqrt{|2x - 1|}$$

[78] An der entsprechenden Stelle findet sich in dem Paper von A. Y. Govan, C. D. Meyer, R. Albright (2008) ein schöner Lapsus. Man kann dort lesen: Some teams in our small NFL example play each other more than once. In this case we add the corresponding game scores for each of the games between the same two teams to produce one cumulative score. For example, Dallas (Dal) and Washington (Was) played twice where Dallas won 28-23 and Washington won 27-6. The cumulative score between Dallas and Washington is therefore Washington won 60-34. This is the simplest approach for dealing with multiple games between two teams. ...

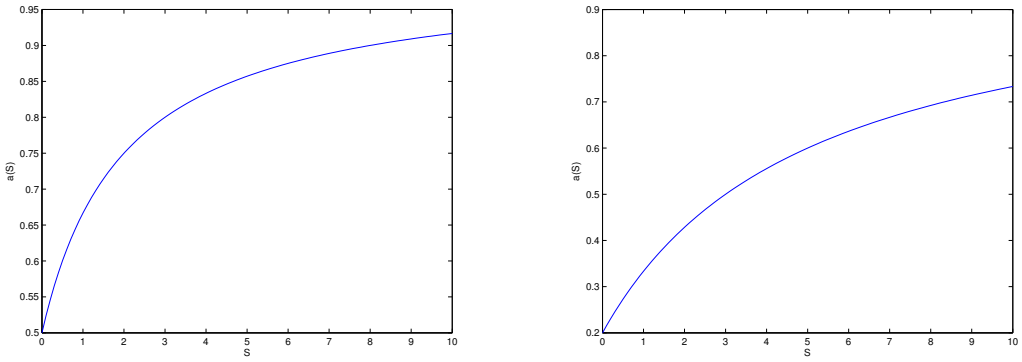


Abbildung 82: Die Funktion $a(S) := (S+1)/(S+S_{ji}+2)$ für $S_{ji} = 0$ und $S_{ji} = 3$

zu definieren und anschließend die Einträge a_{ij} der Bewertungsmatrix A durch

$$a_{ij} := \begin{cases} \dfrac{1}{n_i} \cdot h\Big(\dfrac{S_{ij}+1}{S_{ij}+S_{ji}+2}\Big), & \text{wenn Team } i \text{ gegen Team } j \text{ spielte,} \\ 0, & \text{sonst.} \end{cases}$$

Die Funktion h ist auf $[0,1]$ stetig, es gilt $h(\frac{1}{2}) = \frac{1}{2}$ und h geht schnell gegen 0 bzw. 1 für $x \to 0+$ bzw. $x \to 1-$. In Abbildung 83 skizzieren wir die Funktion h. Zum Vergleich ist gestrichelt auch $y(x) := x$ auf $[0,1]$ angegeben. Als Bewertungsmatrix zu

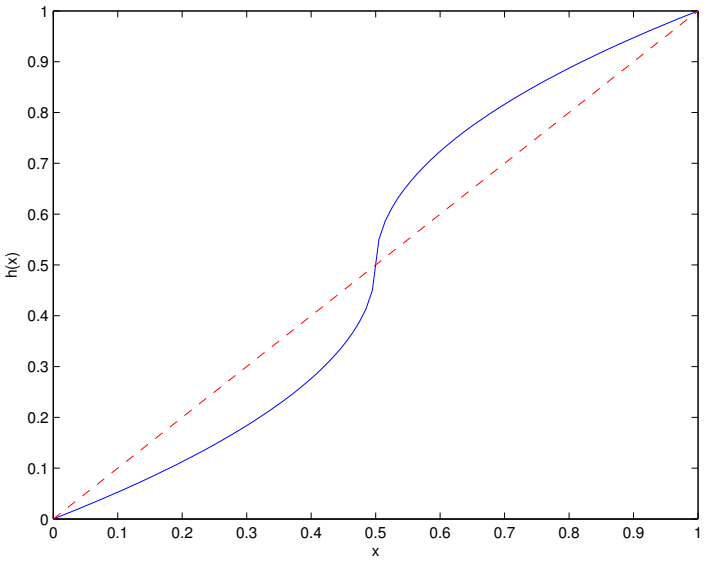


Abbildung 83: Die Funktion $h(x) := \frac{1}{2} + \frac{1}{2}\text{sign}\,(x-\frac{1}{2})\sqrt{|2x-1|}$ auf $[0,1]$

obigen kleinen NFL-Problem erhalten wir

$$A := \begin{pmatrix} 0 & 0.0921 & 0.0871 & 0.0719 & 0 & 0 \\ 0.2412 & 0 & 0 & 0 & 0.2471 & 0.0948 \\ 0.3694 & 0 & 0 & 0.4024 & 0 & 0 \\ 0.2615 & 0 & 0.0651 & 0 & 0.0853 & 0 \\ 0 & 0.0862 & 0 & 0.2480 & 0 & 0.1667 \\ 0 & 0.3578 & 0 & 0 & 0.2500 & 0 \end{pmatrix}.$$

Ein Unterschied zur Arbeit von A. Govan, C. D. Meyer, R. Albright (2008) ergibt sich daraus, dass dort die unterschiedliche Zahl der Spiele der verschiedenen nicht berücksichtigt wird und nicht alle Spiele der ausgewählten sechs NFL-Teams untereinander berücksichtigt wurden (z. B. spielte Dallas gegen Carolina und gewann $20 - 13$). Spielen Mannschaften mehrfach gegeneinander, so werden die Spielstände aufaddiert. Insgesamt haben wir daher nur acht Spiele, die wir in Tabelle 4 auflisten. Als Perron-Vektor zu A erhalten wir

Spielpaarung	Ergebnis
Hou-Car	$34 - 21$
Car-NO	$22 - 44$
Dal-Car	$20 - 13$
Dal-Phi	$44 - 27$
Was-Dal	$50 - 34$
NO-Hou	$10 - 23$
Phi-NO	$38 - 23$
Was-Phi	$45 - 45$

Tabelle 4: Spielergebnisse in der NFL-Saison 2007

$$x = (0.2025, 0.4608, 0.3795, 0.2489, 0.4345, 0.5932)^T.$$

Dies ergibt das Ranking

Washington Dallas Philadelphia Houston New Orleans Carolina.

Dieses Ranking stimmt mit dem bei A. Y. Govan, C. D. Meyer, R. Albright (2008) angegebenen überein. □

61 Konvexe, quadratisch restringierte quadratische Optimierungsaufgaben

Wir betrachten die Optimierungsaufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0,\ i = 1, \ldots, l,\ Ax = b\}. \end{cases}$$

Hierbei seien die Matrizen $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, also (P) eine konvexe Optimierungsaufgabe, bei der sowohl die Zielfunktion als auch die Restriktionabbildungen der Ungleichungen konvexe, quadratische Funktionen sind. Ferner seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c_0, c_1, \ldots, c_l \in \mathbb{R}^n$ und $\beta_1, \ldots, \beta_l \in \mathbb{R}$ gegeben.

Unser Ziel ist es, einen *Existenzsatz* für die konvexe, quadratisch restringierte quadratische Optimierungsaufgabe (P) zu beweisen. Er stammt von E. L. Peterson, J. G. Ecker (1969, 1970).

Wir präsentieren einen wesentlich einfacheren Beweis[79].

Satz *Die konvexe, quadratisch restringierte quadratische Optimierungsaufgabe (P) sei zulässig, d. h. es sei $M \neq \emptyset$. Ferner sei*

$$\inf(\mathrm{P}) := \inf_{x\in M} f(x) > -\infty,$$

die Zielfunktion f sei also auf der Menge M der zulässigen Lösungen nach unten beschränkt. Dann besitzt (P) eine Lösung[80].

Beweis: Wir können annehmen, dass in (P) keine (linearen) Gleichungen als Restriktionen auftreten, da man eine Gleichung als zwei Ungleichungen schreiben kann. Wir gehen daher o. B. d. A. von der Aufgabe

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0,\ i = 1, \ldots, l\} \end{cases}$$

aus. Wir definieren die konvexe, quadratische Funktion $g_0 : \mathbb{R}^n \longrightarrow \mathbb{R}$ durch $g_0(x) := f(x) - \inf(\mathrm{P})$, weiter die konvexe Funktion $G : \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$G(x) := \max_{i=0,\ldots,l} g_i(x).$$

Dann ist $\inf_{x\in\mathbb{R}^n} G(x) = 0$. Wir werden die Existenz eines $x^* \in \mathbb{R}^n$ mit $G(x^*) = 0$ zeigen. Offenbar ist dann $x^* \in M$ eine Lösung von (P).

Wir nennen eine (nicht notwendig nichtleere) Indexmenge $I \subset \{0, \ldots, l\}$ *kanonisch*, wenn die Implikation

$$p \in \mathbb{R}^n, \quad c_i^T p \leq 0, \quad Q_i p = 0 \quad (i \in I) \Longrightarrow c_i^T p = 0 \quad (i \in I)$$

gilt. In einem ersten Schritt zeigen wir:

[79] Die Geschichte hierzu ist die folgende: Die Arbeiten von Peterson-Ecker waren vor vielen Jahren Gegenstand eines Seminars. Angesichts der einfachen Formulierung insbesondere des Existenzsatzes empfand ich die Beweise als unbefriedigend. Ich verkürzte diese und schickte eine entsprechende Arbeit an eine Zeitschrift, ich weiß nicht mehr an welche. Von einem mir (natürlich) unbekannten Gutachter erhielt ich etliche Verbesserungsvorschläge. Da ich aber unter Termindruck beim Korrekturlesen meiner Bände Numerische Mathematik 1 und 2 war, kam ich nicht dazu, diese Verbesserungsvorschläge einzuarbeiten. Der hier angegebene Beweis des Existenzsatzes stammt daher zu großen Teilen von einem Gutachter meiner nie veröffentlichten Arbeit.

[80] Allgemein folgt für eine Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x), \quad x \in M,$$

aus der Zulässigkeit von (P) bzw. $M \neq \emptyset$ und $\inf(\mathrm{P}) > -\infty$ *nicht* die Existenz einer Lösung von (P). Hierzu betrachte man etwa $f(x) := \exp(-x)$ und $M := \{x \in \mathbb{R} : x \geq 0\}$.

- Ist $I \subset \{0,\dots,m\}$ kanonisch, so existiert ein $x \in \mathbb{R}^n$ mit $g_i(x) \le 0$, $i \in I$.

Denn: Die Aussage ist trivial, wenn $I = \emptyset$ oder $\inf_{x\in\mathbb{R}^n} \max_{i\in I} g_i(x) < 0$. Wir können also annehmen, daß $I \neq \emptyset$ und $\inf_{x\in\mathbb{R}^n} \max_{i\in I} g_i(x) = 0$. Mit $B[0;k]$ bezeichnen wir die euklidische Kugel um den Nullpunkt mit dem Radius $k \in \mathbb{N}$, ferner sei $x_k \in B[0;k]$ die Lösung minimaler euklidischer Norm[81] der Optimierungsaufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad G_I(x) := \max_{i\in I} g_i(x), \quad x \in B[0;k].$$

Offenbar ist dann

$$\lim_{k\to\infty} G_I(x_k) = \lim_{k\to\infty} \min_{x\in B[0;k]} G_I(x) = \inf_{x\in\mathbb{R}^n} G_I(x) = 0.$$

Besitzt daher $\{x_k\}$ eine Häufungspunkt x, so ist $G_I(x) = 0$, also x der gesuchte Punkt. Andernfalls ist $\|x_k\| \to \infty$, o. B. d. A. gilt $x_k/\|x_k\| \to p$, wobei natürlich $\|p\| = 1$, insbesondere also $p \neq 0$. Wegen

$$g_i(x_k) = \beta_i + c_i^T x_k + \tfrac{1}{2} x_k^T Q_i x_k \le G_I(x_k) \to 0, \qquad i \in I,$$

folgt

$$c_i^T p \le 0, \quad Q_i p = 0 \quad (i \in I).$$

Da $I \subset \{0,\dots,m\}$ nach Voraussetzung kanonisch ist, folgt $c_i^T p = 0$, $i \in I$. Für alle $t \in \mathbb{R}$ ist daher $g_i(x_k) = g_i(x_k - tp)$, $i \in I$, insbesondere $G_I(x_k) = G_I(x_k - tp)$ für alle $t \in \mathbb{R}$ und alle $k \in \mathbb{N}$. Andererseits ist

$$\lim_{t\to 0+} \frac{\|x_k - tp\|^2 - \|x_k\|^2}{t} = -2x_k^T p < 0$$

für alle hinreichend großen k. Für diese k und alle hinreichend kleinen $t > 0$ ist daher $x_k - tp$ eine Lösung von (P_k) mit einer kleineren euklidischen Norm als der von x_k, ein Widerspruch zu der Definition von x_k.

Nun kommen wir zum entscheidenden Schritt und zeigen:

- Sei $I^* \subset \{0,\dots,m\}$ unter allen kanonischen Teilmengen von $\{0,\dots,m\}$ maximal. Wegen der gerade eben bewiesenen Aussage existiert ein $x \in \mathbb{R}^n$ mit $g_i(x) \le 0$, $i \in I^*$. Dann existiert ein $x^* \in \mathbb{R}^n$ mit $g_i(x^*) = g_i(x)$, $i \in I^*$, und $g_i(x^*) \le 0$, $i \in \{0,\dots,m\} \setminus I^*$. Dieses x^* ist eine Lösung von (P).

Denn: Wir können annehmen, daß I^* eine echte Teilmenge von $I_0 := \{0,\dots,m\}$ ist, da man andernfalls $x^* := x$ wählen kann. Alle Teilmengen I von I_0, die I^* echt enthalten, sind nicht kanonisch, d. h. das Gleichungs-Ungleichungssystem

$$(I) \qquad c_i^T p \le 0, \quad Q_i p = 0 \quad (i \in I), \quad \Big(\sum_{i\in I} c_i\Big)^T p < 0$$

besitzt eine Lösung. Auf die folgende Weise bestimmen wir strikt absteigende Indexmengen $I_0 \supset I_1 \supset \cdots \supset I_r \supset I^*$, welche mit der maximalen kanonischen Indexmenge I^* enden.

[81] Die Menge der Lösungen von (P_k) ist nichtleer, abgeschlossen und konvex. Daher gibt es in dieser Menge genau ein Element mit minimaler euklidischer Norm bzw. eines, welches minimalen Abstand zum Nullpunkt besitzt. Dies ist eine Folgerung aus dem Projektionssatz für konvexe Mengen, siehe Abschnitt 37 über den Brouwerschen Fixpunktsatz.

Für $k = 0, 1, \ldots$:

- Sei p_k eine Lösung von

$$(I_k) \qquad c_i^T p \le 0, \quad Q_i p = 0 \quad (i \in I_k), \quad \Big(\sum_{i \in I_k} c_i\Big)^T p < 0$$

und definiere die (nichtleere) Indexmenge

$$J_k := \{i \in I_k : c_i^T p_k < 0\}.$$

- Falls $I_k \setminus J_k = I^*$, dann: $r := k$, STOP.
- Andernfalls: Setze $I_{k+1} := I_k \setminus J_k$.

Nun setze man

$$x^* := x + \sum_{k=0}^{r} \alpha_k p_k$$

mit noch unbestimmten Konstanten $\alpha_0, \ldots, \alpha_r \ge 0$. Wegen

$$c_i^T p_k = 0, \quad Q_i p_k = 0 \quad (i \in I^*), \quad k = 0, \ldots, r,$$

ist

$$g_i(x^*) = g_i(x) \quad (i \in I^*).$$

Für $i \in J_r = I_r \setminus I^*$ ist

$$c_i^T p_r < 0, \quad c_i^T p_k \le 0 \quad (k = 0, \ldots, r-1), \quad Q_i p_k = 0 \quad (k = 0, \ldots, r).$$

Nun wähle man $\alpha_r \ge 0$ so groß, dass (bei noch unbestimmten $\alpha_0, \ldots, \alpha_{r-1}$) gilt:

$$g_i(x^*) = g_i(x) + \alpha_r \underbrace{c_i^T p_r}_{<0} + \sum_{k=0}^{r-1} \alpha_k \underbrace{c_i^T p_k}_{\le 0} \le g_i(x) + \alpha_r c_i^T p_r \le 0, \quad i \in J_r.$$

Für $i \in J_{r-1} = I_{r-1} \setminus I_r$ ist entsprechend

$$c_i^T p_{r-1} < 0, \quad c_i^T p_k \le 0 \quad (k = 0, \ldots, r-2), \quad Q_i p_k = 0 \quad (k = 0, \ldots, r-1).$$

Durch Wahl eines hinreichend großen $\alpha_{r-1} \ge 0$ (bei noch unbestimmten $\alpha_0, \ldots, \alpha_{r-2}$) ist

$$g_i(x^*) \le g_i(x + \alpha_r p_r) + \alpha_{r-1} c_i^T p_{r-1} \le 0, \quad i \in J_{r-1}.$$

In dieser Weise kann man fortfahren. Nach endlich vielen Schritten hat man nichtnegative Zahlen $\alpha_r, \ldots, \alpha_0$ so bestimmt, dass für $x^* := x + \sum_{k=0}^{r} \alpha_k p_k$ nicht nur $g_i(x^*) = g_i(x) \le 0$, $i \in I^*$,Situation" sondern auch $g_i(x^*) \le 0$, $i \notin I^*$. Dann ist

$$0 = \inf_{z \in \mathbb{R}^n} G(z) \le G(x^*) \le 0,$$

also $G(x^*) = 0$ bzw. $x^* \in M$ und $f(x^*) = \inf(\mathrm{P})$. Damit´ ist die obige Behauptung und folglich der ganze Satz bewiesen. □

Bemerkung: Als Spezialfall (setze $Q_i := 0$, $i = 1, \ldots, l$) erhält man: Ist eine konvexe, quadratische Funktion auf einem nichtleeren Polyeder nach unten beschränkt, so nimmt sie auf diesem Polyeder ihr Minimum an. Dies ist ein Ergebnis, das zuerst von E. Barankin, R. Dorfman (1958) bewiesen wurde. Erst Recht ist natürlich der Existenzsatz der linearen Optimierung enthalten, der aussagt, dass eine zulässige lineare (Minimierungs-) Optimierungsaufgabe, deren Zielfunktion auf der Menge der zulässigen Lösungen nach unten beschränkt ist, eine Lösung besitzt. □

62 Das Minimum Covering Sphere Problem

J. J. Sylvester (1857) publizierte eine Note, welche nur aus einem Satz besteht:

- It is required to find the least circle which shall contain a given system of points in the plane.

Die gegebenen paarweise verschiedenen Punkte seien $a_1, \ldots, a_m \in \mathbb{R}^n$ mit $n \geq 2$ (wir befreien uns also von der Annahme, die gegebenen Punkte würden in der Ebene liegen[82]). Wir nehmen an[83], es sei $m \geq 3$. Man spricht auch von dem *smallest bounding sphere problem* oder dem *minimum covering sphere problem*. Eine minimax-Formulierung des Sylvester-Problems ist

$$\text{Minimiere} \quad \max_{i=1,\ldots,m} \|x - a_i\|_2, \quad x \in \mathbb{R}^n.$$

Diese unrestringierte Optimierungsaufgabe hat den Nachteil, dass ihre Zielfunktion nicht differenzierbar ist. Eine mögliche Umformulierung ist dann

$$\text{Minimiere} \quad r \quad \text{unter der Nebenbedingung} \quad \|x - a_i\|_2 \leq r, \quad i = 1, \ldots, m,$$

oder (um zu einer differenzierbaren Restriktionsabbildung zu kommen)

$$\text{Minimiere} \quad \frac{1}{2}r^2 \quad \text{unter der Nebenbedingung} \quad \frac{1}{2}\|x - a_i\|_2^2 \leq \frac{1}{2}r^2, \quad i = 1, \ldots, m.$$

Setzt man hier $q := r^2/2$, so kommt man zu der konvexen, quadratisch restringierten quadratischen Optimierungsaufgabe

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad q \quad \text{unter der Nebenbedingung} \\ \dfrac{1}{2}\|a_i\|_2^2 - a_i^T x - q + \dfrac{1}{2}\|x\|_2^2 \leq 0, \quad i = 1, \ldots, m. \end{cases}$$

[82] Ist $n = 1$, so kann die Lösung offenbar leicht geschlossen angegeben werden. Sei nämlich in diesem Falle

$$l := \min_{i=1,\ldots,m} a_i, \qquad u := \max_{i=1,\ldots,m} a_i.$$

Mit $x^* := (l + u)/2$ hat man den Mittelpunkt und mit $r^* := (u - l)/2$ den Radius des gesuchten Kreises bzw. Intervalles gefunden.

[83] Ist $m = 2$, so ist offenbar $x^* := (a_1 + a_2)/2$ der Mittelpunkt und $r^* := \|a_1 - a_2\|/2$ der Radius der gesuchten Kugel.

Diese Aufgabe besitzt wegen des gerade eben in Abschnitt 61 bewiesenen Existenzsatzes eine Lösung (x^*, q^*). Denn sie ist zulässig ist (in einer hinreichend großen Kugel sind alle m Punkte $a_1, \ldots, a_m$ enthalten) und die Zielfunktion nach unten (durch 0) beschränkt. Die gesuchte minimale Kugel hat den Mittelpunkt x^* und den Radius $r^* := \sqrt{2q^*}$. Durch eine Variablentransformation kann man (P) sogar in eine konvexe quadratische Optimierungsaufgabe überführen, also eine Aufgabe, eine konvexe quadratische Zielfunktion unter linearen (Ungleichungs-) Nebenbedingungen zu minimieren. Hierzu setze man $q = v + \frac{1}{2}\|x\|_2^2$, wodurch man die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \dfrac{1}{2}x^T x + v \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad \dfrac{1}{2}\|a_i\|_2^2 - a_i^T x - v \le 0, \quad i = 1, \ldots, m, \end{cases}$$

erhält. Ausführlich geschrieben lautet diese Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \dfrac{1}{2}\begin{pmatrix} x \\ v \end{pmatrix}^T \begin{pmatrix} I & 0 \\ 0^T & 0 \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} x \\ v \end{pmatrix} \\ \text{unter den Nebenbedingungen} \\ -\begin{pmatrix} a_1^T & 1 \\ \vdots & \vdots \\ a_m^T & 1 \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} \le -\dfrac{1}{2}\begin{pmatrix} \|a_1\|_2^2 \\ \vdots \\ \|a_m\|_2^2 \end{pmatrix}. \end{cases}$$

Diese Optimierungsaufgabe kann z. B. mit Hilfe der MATLAB-Funktion `quadprog` gelöst werden. Wir wollen uns überlegen, dass (P) *eindeutig* lösbar ist und nehmen hierzu an, (x^*, v^*) und (x^{**}, v^{**}) seien zwei Lösungen. Da (P) eine konvexe Optimierungsaufgabe ist, ist auch $(\frac{1}{2}(x^* + x^{**}), \frac{1}{2}(v^* + v^{**}))$ eine Lösung. Da der Optimalwert von (P) eindeutig ist, ist ferner

$$(*) \qquad \frac{1}{2}\|x^*\|_2^2 + v^* = \frac{1}{2}\|x^{**}\|_2^2 + v^{**} = \frac{1}{8}\|x^* + x^{**}\|_2^2 + \frac{1}{2}(v^* + v^{**}).$$

Eine einfache Manipulation dieser Gleichungen ergibt

$$\|x^*\|_2^2 + \|x^{**}\|_2^2 = \frac{1}{2}\|x^* + x^{**}\|_2^2.$$

Aus der leicht zu beweisenden Parallelogrammgleichung

$$\|x^*\|_2^2 + \|x^{**}\|_2^2 = \frac{1}{2}\|x^* + x^{**}\|_2^2 + \frac{1}{2}\|x^* - x^{**}\|_2^2$$

folgt $\|x^* - x^{**}\|_2 = 0$ und damit $x^* = x^{**}$. Aus $(*)$ erhält man auch $v^* = v^{**}$ und damit die Eindeutigkeit einer Lösung von (P). Ist (x^*, v^*) die Lösung von (P), so ist x^* der Mittelpunkt und $r^* := \sqrt{2(v^* + \|x^*\|_2^2/2)}$ der Radius der gesuchten Kugel.

Eine Anwendung von Satz 1 in Abschnitt 53 (Satz von Kuhn-Tucker für linear restringierte Optimierungsaufgaben) auf die Optimierungsaufgabe (P) liefert zu der Lösung (x^*, v^*) von (P) die Existenz eines Vektors $u^* \in \mathbb{R}^m$ mit

$$u^* \ge 0, \qquad \begin{pmatrix} x^* \\ 1 \end{pmatrix} - \sum_{i=1}^m u_i^* \begin{pmatrix} a_i \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

sowie

$$(*) \qquad u_i^*\Big(\frac{1}{2}\|a_i\|_2^2 - a_i^T x^* - v^*\Big) = 0, \qquad i = 1, \ldots, m.$$

Aus den ersten beiden Beziehungen erhalten wir, dass

$$x^* = \sum_{i=1}^m u_i^* a_i, \qquad u^* \in \Sigma_m := \{u \in \mathbb{R}^m : u \geq 0,\ e^T u = 1\},$$

wobei e der Vektor des $\mathbb{R}^m$ ist, dessen Komponenten alle gleich 1 sind. Also liegt x^* in der *konvexen Hülle* $\mathrm{co}\,(\{a_1, \ldots, a_m\})$ von $\{a_1, \ldots, a_m\}$, d. h. der kleinsten konvexen Menge, welche $\{a_1, \ldots, a_m\}$ enthält, bzw. dem Durchschnitt aller konvexen Mengen des $\mathbb{R}^n$, welche $\{a_1, \ldots, a_m\}$ enthalten. Allgemein gilt der

Satz *Sei $S \subset \mathbb{R}^n$. Dann ist die konvexe Hülle $\mathrm{co}\,(S)$, d. h. der Durchschnitt aller konvexen Mengen, die S enthalten, gegeben durch*

$$\mathrm{co}\,(S) = \Big\{\sum_{i=1}^m \lambda_i x_i : x_i \in S,\ \lambda_i \geq 0\ (i = 1, \ldots, m),\ \sum_{i=1}^m \lambda_i = 1,\ m \in \mathbb{N}\Big\}.$$

Beweis: Wir setzen zur Abkürzung

$$K := \Big\{\sum_{i=1}^m \lambda_i x_i : x_i \in S,\ \lambda_i \geq 0\ (i = 1, \ldots, m),\ \sum_{i=1}^m \lambda_i = 1,\ m \in \mathbb{N}\Big\}.$$

Dann ist K konvex (Beweis?) und $S \subset K$, und daher nach Definition der konvexen Hülle $\mathrm{co}\,(S) \subset K$. Zum Nachweis von $K \subset \mathrm{co}\,(S)$ haben wir zu zeigen: Für jedes $m \in \mathbb{N}$ gilt die Implikation

$$\lambda_i \geq 0,\ x_i \in S\ \ (i = 1, \ldots, m),\ \ \sum_{i=1}^m \lambda_i = 1 \Longrightarrow \sum_{i=1}^m \lambda_i x_i \in \mathrm{co}\,(S).$$

Dies zeigt man leicht durch vollständige Induktion nach m. Die Aussage ist für $m = 1$ offensichtlich richtig. Für den Induktionsschluss von m nach $m+1$ gehen wir von $\lambda_i \geq 0$, $x_i \in S$, $i = 1, \ldots, m$, und $\sum_{i=1}^{m+1} \lambda_i = 1$ aus. Wir setzen $\Lambda_m := \sum_{i=1}^m \lambda_i$, o. B. d. A. ist $\Lambda_m > 0$. Dann ist unter Berücksichtigung der Induktionssannahme

$$\sum_{i=1}^{m+1} \lambda_i x_i = \sum_{i=1}^m \lambda_i x_i + \lambda_{m+1} x_{m+1} = \Lambda_m \underbrace{\sum_{i=1}^m \frac{\lambda_i}{\Lambda_m} x_i}_{\in \mathrm{co}\,(S)} + (1 - \Lambda_m)\ \underbrace{x_{m+1}}_{\in S \subset \mathrm{co}\,(S)} \in \mathrm{co}\,(S),$$

wobei wir am Schluss die Konvexität von $\mathrm{co}\,(S)$ benutzt haben. Damit ist die Aussage bewiesen. □

Der Satz von Carathéodory[84] liefert, dass x^* in der konvexen Hülle von höchstens $n+1$ der Punkte $a_1, \ldots, a_m$ liegt. Im Fall der Ebene ($n = 2$) liegt x^* also in der konvexen Hülle von höchstens drei der Punkte $a_1, \ldots, a_m$. Wegen $r^* = \sqrt{2(v^* + \|x^*\|_2^2/2)}$ ist $v^* + \|x^*\|_2^2/2 = (r^*)^2/2$ und folglich

$$\begin{aligned} \frac{1}{2}\|a_i\|_2^2 - a_i^T x^* - v^* &= \frac{1}{2}\|a_i\|_2^2 - a_i^T x^* - \frac{1}{2}\|x^*\|_2^2 - \frac{1}{2}(r^*)^2 \\ &= \frac{1}{2}\|x^* - a_i\|_2^2 - \frac{1}{2}(r^*)^2. \end{aligned}$$

Wir definieren die Menge der in x^* aktiven Ungleichungsrestriktionen durch

$$I^* := \{i \in \{1, \ldots, m\} : \|x^* - a_i\|_2 = r^*\}.$$

Dann ist $u_i^* = 0$, $i \in \{1, \ldots, m\} \setminus I^*$. Wir wollen uns überlegen, dass i. Allg. $|I^*| \geq 3$, also mindestens drei Restriktionen aktiv sind bzw. mindestens drei der vorgegebenen Punkte auf der gesuchten kleinsten Kugel liegen. Wäre $I^* = \emptyset$, so wäre $u_i^* = 0$, $i = 1, \ldots, m$, was ein Widerspruch zu $\sum_{i=1}^m u_i^* = 1$ ist. Angenommen, $I^* = \{i\}$, es sei also $\|x^* - a_i\|_2 = r^*$ und $\|x^* - a_j\|_2 < r^*$ für alle $j \in \{1, \ldots, m\} \setminus \{i\}$. Dann ist $u_i^* = 1$, $u_j^* = 0$ für $j \neq i$. Dann ist $x^* = a_i$ und $r^* = 0$. Dies ist nur möglich, wenn nur ein Punkt gegeben ist. Da wir aber $m \geq 3$ vorausgesetzt haben, haben wir einen Widerspruch. Jetzt nehmen wir an, es sei $I^* = \{i, j\}$. Dann ist $u_j^* = 1 - u_i^*$ und $x^* = u_i^* a_i + (1 - u_i^*) a_j$. Dann ist

$$x^* - a_i = (1 - u_i^*)(a_j - a_i), \qquad x^* - a_j = u_i^*(a_i - a_j).$$

---

[84] **Satz von Carathéodory** *Sei $S \subset \mathbb{R}^n$ und $x \in \mathrm{co}\,(S)$. Dann ist x eine Konvexkombination von höchstens $n+1$ Punkten aus S. Genauer existiert ein $m \in \mathbb{N}$ mit $m \leq n+1$ sowie $x_i \in S$, $\lambda_i \geq 0$, $i = 1, \ldots, m$, mit $\sum_{i=1}^m \lambda_i = 1$ sowie $x = \sum_{i=1}^m \lambda_i x_i$.*

Beweis: Wegen der gerade eben bewiesenen Aussage lässt sich $x \in \mathrm{co}\,(S)$ als Konvexkombination von $m \in \mathbb{N}$ Punkten aus S darstellen, d. h. es existieren $\lambda_i \geq 0$, $x_i \in S$, $i = 1, \ldots, m$, mit $\sum_{i=1}^m \lambda_i = 1$ sowie $x = \sum_{i=1}^m \lambda_i x_i$. Wir zeigen: Ist $m > n+1$, so lässt sich x als Konvexkombination von $m-1$ Punkten aus S darstellen. Hieraus folgt die Behauptung.

O. B. d. A. ist $\lambda_i > 0$, $i = 1, \ldots, m$. Wegen $m - 1 > n$ sind $x_1 - x_m, \ldots, x_{m-1} - x_m$ linear abhängig, sodass $r_1, \ldots, r_{m-1}$, nicht alle gleich Null, existieren mit $\sum_{i=1}^{m-1} r_i(x_i - x_m) = 0$. Definiert man $r_m := -\sum_{i=1}^m r_i$, so ist $\sum_{i=1}^m r_i = 0$ und $\sum_{i=1}^m r_i x_i = 0$. Nun definiere man

$$\mu_i := \lambda_i - \alpha r_i, \quad i = 1, \ldots, m,$$

wobei $\alpha > 0$ aus

$$\frac{1}{\alpha} = \max_{i=1,\ldots,m} \frac{r_i}{\lambda_i} = \frac{r_j}{\lambda_j}$$

bestimmt ist. Dann ist

$$\mu_i \geq 0, \;\; i = 1, \ldots, m, \qquad \sum_{i=1}^m \mu_i = 1, \qquad \mu_j = 0.$$

Folglich ist

$$x = \sum_{i=1}^m \lambda_i x_i = \sum_{i=1}^m \mu_i x_i + \alpha \underbrace{\sum_{i=1}^m r_i x_i}_{=0} = \sum_{\substack{i=1 \\ i \neq j}}^m \mu_i x_i,$$

womit die gewünschte Darstellung von x als Konvexkombination von $m-1$ Elementen aus S erhalten ist. □

Aus

$$\|x^* - a_i\|_2 = \|x^* - a_k\|_2 = r^*$$

erhalten wir $u_i^* = \frac{1}{2}$ und damit $x^* = \frac{1}{2}(a_i + a_k)$. Wenn also genau zwei der Punkte aus $\{a_1, \ldots, a_m\}$ auf dem Rand der kleinsten diese Punkte enthaltenden Kugel liegen, so ist der Mittelpunkt dieser Kugel der Mittelpunkt dieser zwei Punkte und der Durchmesser der Kugel ist gleich dem (euklidischen) Abstand dieser zwei Punkte[85]. In Abbildung 84 geben wir ein Beispiel an, in welchem genau zwei der gegebenen Punkte auf dem

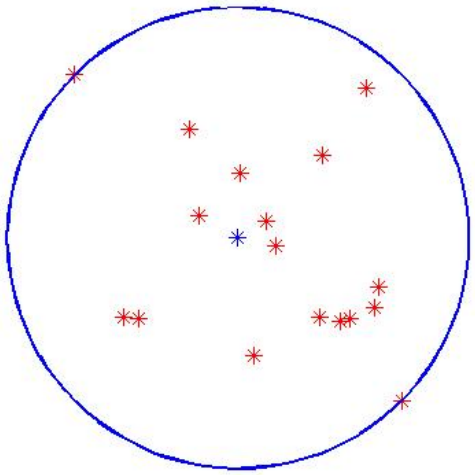


Abbildung 84: Beispiel für genau zwei aktive Restriktionen

Rand des gesuchten minimalen Kreises liegen. Hier haben wir die Punkte $(0, 1)$ und $(1, 0)$ sowie 15 zufällig in $(0, 1) \times (0, 1)$ gewählte Punkte durch $*$ gekennzeichnet.

Zum Schluss geben wir in Abbildung 85 ein Beispiel an, in dem 15 Punkte in $[0, 1] \times [0, 1]$ zufällig gewählt und durch $*$ markiert wurden und anschließend der diese Punkte enthaltende minimale Kreis mit dem Mittelpunkt $*$ gezeichnet wurde.

[85] Umgekehrt gilt: Sei

$$\|a_i - a_j\|_2 = \max_{\substack{k,l=1,\ldots,m \\ k \neq l}} \|a_k - a_l\|$$

der Durchmesser der Menge $\{a_1, \ldots, a_m\}$. Ist dann $x^* := (a_i + a_j)/2$ und $\|x^* - a_k\| \leq \|a_i - a_j\|/2 =: r^*$, $k \in \{1, \ldots, m\} \setminus \{i, j\}$, so ist die Kugel um x^* mit dem Radius r^* die kleinste Kugel (d. h. die mit dem kleinsten Radius), welche die Punkte $a_1, \ldots, a_m$ enthält. Denn sei $\|x - a_k\|_2 \leq r$, $k = 1, \ldots, m$. Wir haben zeigen, dass $r^* \leq r$. Dies ist aber einfach. Denn aus $\|x - a_i\|_2 \leq r$ und $\|x - a_j\|_2 \leq r$ folgt mit Hilfe der Dreiecksungleichung

$$\|a_i - a_j\| = \|(a_i - x) + (x - a_j)\|_2 \leq 2r,$$

also $r \geq r^*$ und das war zu zeigen.

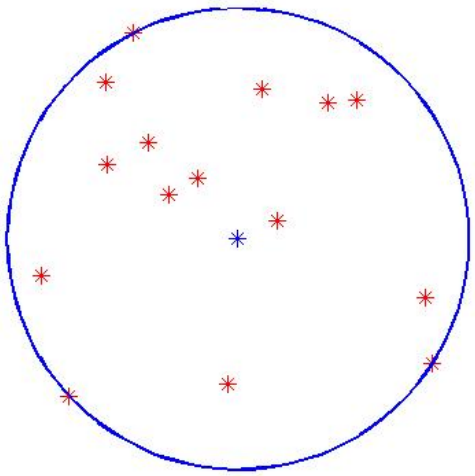


Abbildung 85: 15 Punkte und der zugehörige Sylvester-Kreis

63 Die Ungleichungen von Steinhagen und Jung

Für gewisse Teilmengen des $\mathbb{R}^n$ werden durch die Ungleichungen von P. Steinhagen (1921) und H. W.Jung (1901) die Größen *Inkugelradius, Breite, Durchmesser* und *Umkugelradius* zueinander in Beziehung gesetzt. Interessant ist, wie zum Beweis dieser Ungleichungen Ergebnisse der Optimierung eingesetzt werden.

Mit $B[x;r] \subset \mathbb{R}^n$ bezeichnen wir die abgeschlossene, euklidische Kugel um $x \in \mathbb{R}^n$ mit dem Radius $r > 0$. Weiter definieren wir:

Definition Sei $P \subset \mathbb{R}^n$ nichtleer, konvex und kompakt.

1. $$r^*(P) := \sup\{r > 0 : \text{Es existiert } x \in P \text{ mit } B[x;r] \subset P\}$$

   heißt der *Inkugelradius* von P.

2. $$w^*(P) := \inf_{c \neq 0} \frac{1}{\|c\|_2} \Big\{ \sup_{y \in P} c^T y - \inf_{y \in P} c^T y \Big\}$$

   heißt die *Breite* (engl.: width) von P. Die Breite von P ist offensichtlich der minimale Abstand zweier paralleler Hyperebenen, von denen eine P im zugehörigen nichtpositiven, die andere P im nichtnegativen Halbraum enthält[86].

3. $$R^*(P) := \inf\{R > 0 : \text{Mit } x \in \mathbb{R}^n \text{ ist } P \subset B[x;R]\}$$

   heißt der *Umkugelradius* von P.

[86] Bei Steinhagen heißen solche Hyperebenen *Stützgebilde*.

4.

$$D^*(P) := \sup_{y,z\in P} \|y-z\|_2$$

heißt der *Durchmesser* von P.

Durch die Ungleichung von Steinhagen wird die Breite eines Polytops, also eines beschränkten Polyeders, durch den Inkugelradius abgeschätzt.

Satz von Steinhagen *Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $P := \{x \in \mathbb{R}^n : Ax \le b\}$ nichtleer und beschränkt. Ist $r^*(P)$ der Inkugelradius und $w^*(P)$ die Breite von (P), so gilt*

$$w^*(P) \le r^*(P) \cdot \begin{cases} 2\sqrt{n}, & \text{falls } n \text{ ungerade,} \\ \dfrac{2(n+1)}{\sqrt{n+2}}, & \text{falls } n \text{ gerade.} \end{cases}$$

Beispiel: Für ein gleichseitiges Dreieck in der Ebene gilt in der Ungleichung von Steinhagen sogar Gleichheit. Ist nämlich die Seitenlänge des gleichseitigen Dreiecks a, so stimmen die Höhe im Dreieck und die Breite überein und es ist $w^* = (a/2)\sqrt{3}$. Der Inkreisradius r^* ist gegeben durch $r^* = (a/6)\sqrt{3}$. In diesem Fall ist also $w^* = r^* \cdot 3$, während die Ungleichung von Steinhagen $w^* \le r^* \cdot 3$ für $n = 2$ lautet. □

Beweis des Satzes von Steinhagen Sei

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

O. B. d. A. können wir annehmen, dass $\|a_i\|_2 = 1$, $i = 1, \ldots, m$. Nun überlegen wir uns:

- Es ist $B[x; r] \subset P$ genau dann, wenn $Ax - b + re \le 0$, $r \ge 0$, wobei $e \in \mathbb{R}^m$ der Vektor ist, dessen Komponenten alle gleich 1 sind.

Denn: Sei $B[x; r] \subset P$. Wegen $\|a_k\|_2 = 1$, $k = 1, \ldots, m$, ist speziell $x + ra_k \in P$, $k = 1, \ldots, m$, und daher $a_k^T(x + ra_k) = a_k^T x + r \le b_k$, $k = 1, \ldots, m$. Also ist $Ax - b + re \le 0$. Ist umgekehrt $Ax - b + re \le 0$, $r \ge 0$, und $y \in B[x; r]$, so ist

$$a_k^T y = a_k^T x + a_k^T(y - x) \le a_k^T x + \|y - x\| \le a_k^T x + r \le b_k, \quad k = 1, \ldots, m,$$

und damit $y \in P$.

Damit ist der Mittelpunkt $x^* = x^*(P)$ der Inkugel und der Inkugelradius $r^* = r^*(P)$ Lösung der linearen Optimierungsaufgabe

$$\begin{cases} \text{Maximiere} \quad r \quad \text{unter den Nebenbedingungen} \\ Ax - b + re \le 0, \quad r \ge 0 \end{cases}$$

bzw.

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \begin{pmatrix} 0 \\ -1 \end{pmatrix}^T \begin{pmatrix} x \\ r \end{pmatrix} \quad \text{unter den Nebenbedingungen} \\ \begin{pmatrix} A & e \\ 0^T & -1 \end{pmatrix} \begin{pmatrix} x \\ r \end{pmatrix} \le \begin{pmatrix} b \\ 0 \end{pmatrix}. \end{cases}$$

Diese lineare Optimierungsaufgabe besitzt eine Lösung (x^*, r^*), da sie zulässig ist (für jedes $x \in P$ ist $(x, 0)$ zulässig für (P)) und ihre Zielfunktion auf der Menge zulässiger Lösungen nach unten beschränkt ist[87], siehe die Bemerkung im Anschluss an den Existenzsatz für konvexe, quadratisch restringierte quadratische Optimierungsaufgaben in Abschnitt 61. Nun wenden wir den Satz von Kuhn-Tucker für linear restringierte Optimierungsaufgaben (siehe Satz 1 in Abschnitt 53) an. Hiernach existiert zu der Lösung (x^*, r^*) von (P) ein Paar $(y^*, q^*) \in \mathbb{R}^m \times \mathbb{R}$ mit

$$\begin{pmatrix} y^* \\ q^* \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} A^T & 0 \\ e^T & -1 \end{pmatrix} \begin{pmatrix} y^* \\ q^* \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

sowie

$$\begin{pmatrix} Ax^* - b + r^* \\ -r^* \end{pmatrix}^T \begin{pmatrix} y^* \\ q^* \end{pmatrix} = 0.$$

Zuerst nehmen wir an, es sei $r^* = 0$. Insbesondere gibt es dann keine in P enthaltene Kugel, d. h. das Innere $\operatorname{int}(P)$ von P ist leer. Ist das Innere einer konvexen Teilmenge P des $\mathbb{R}^n$ aber leer, so ist P in einer Hyperebenen enthalten und hat damit die Breite $w^* = 0$. In diesem Fall ist die Ungleichung von Steinhagen also trivialerweise richtig.

Wir können jetzt also $r^* > 0$ annehmen. Dann ist notwendigerweise $q^* = 0$ und zu der Lösung (x^*, r^*) von (P) erhalten wir die Existenz von $y^* \in \mathbb{R}^m$ mit

$$y^* \geq 0, \qquad \begin{pmatrix} A^T \\ e^T \end{pmatrix} y^* = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad (Ax^* - b + r^*)^T y^* = 0.$$

Also ist

$$y^* \in \Sigma_m := \{y \geq 0,\ e^T y = 1\}$$

und

$$\sum_{i=1}^{m} y_i^* a_i = 0, \qquad y_i^*(a_i^T x^* - b_i + r) = 0, \quad (i = 1, \ldots, m).$$

Wir definieren

$$I^* := \{i \in \{1, \ldots, m\} : y_i^* > 0\}.$$

Dann ist

$$\sum_{i \in I^*} y_i^* = 1, \qquad \sum_{i \in I^*} y_i^* a_i = 0, \qquad r^* = b_i - a_i^T x^*, \quad (i \in I^*).$$

Insbesondere liegt 0 in der konvexen Hülle (siehe S. 245) $\operatorname{co}(\{a_i\}_{i \in I^*})$ von $\{a_i\}_{i \in I^*}$. Wegen des Satzes von Carathéodory (siehe S. 246) existiert eine Indexmenge $I \subset I^*$ mit $p := |I| \leq n + 1$ und $0 \in \operatorname{co}(\{a_i\}_{i \in I})$. Folglich existieren z_i^*, $i \in I$, mit

$$z_i^* > 0 \quad (i \in I), \qquad \sum_{i \in I} z_i^* = 1, \qquad \sum_{i \in I} z_i^* a_i = 0.$$

[87] Denn ist (x, r) zulässig für (P), so ist insbesondere $x \in P$. Da P als beschränkt vorausgesetzt ist, existiert eine nur von P abhängende Konstante mit $0 \leq r \leq C$ bzw. $-C \leq -r \leq 0$.

Weiter ist

$$r^* = \Big(\underbrace{\sum_{i\in I} z_i^*}_{=1}\Big) r^* = \sum_{i\in I} z_i^* b_i - \Big(\underbrace{\sum_{i\in I} z_i^* a_i}_{=0}\Big)^T x^* = \sum_{i\in I} z_i^* b_i.$$

Nun schätzen wir die Breite w^* von P nach oben ab. Ist $\alpha \le c^T y \le \beta$ für alle $y \in P$, so ist $w^* \le (\beta-\alpha)/\|c\|_2$ (wobei diese Abschätzung für $c = 0$ trivial ist). Für $k = 1, \ldots, p-1$ sei

$$I(k) := \{J \subset I : |J| = k\}$$

die Menge der Teilmengen von I, die aus genau k Elementen bestehen. Für beliebige $J \in I(k)$ und $y \in P$ ist

$$\sum_{j\in J} z_j^* b_j \ge \Big(\sum_{j\in J} z_j^* a_j\Big)^T y = -\Big(\sum_{j\in I\setminus J} z_j^* a_j\Big)^T y \ge -\sum_{j\in I\setminus J} z_j^* b_j.$$

Daher kann für alle $J \in I(k)$, $k = 1, \ldots, p-1$, die Breite w^* von P durch

$$w^* \le \frac{1}{\|\sum_{j\in J} z_j^* a_j\|_2}\Big(\sum_{j\in J} z_j^* b_j + \sum_{j\in I\setminus J} z_j^* b_j\Big) = \frac{\sum_{j\in I} z_i^* b_i}{\|\sum_{j\in J} z_j^* a_j\|_2} = r^* \cdot \frac{1}{\|\sum_{j\in J} z_j^* a_j\|_2}$$

abgeschätzt werden. Eine bestmögliche Abschätzung erhalten wir, indem wir über $J \in I(k)$ und $k\{1, \ldots, p-1\}$ variieren, durch

$$w^* \le r^* \cdot \min_{k=1,\ldots,p-1}\Big(\max_{J\in I(k)}\Big\|\sum_{j\in J} z_j^* a_j\Big\|_2\Big)^{-1}.$$

Zunächst schätzen wir $\max_{J\in I(k)} \|\sum_{j\in J} z_j^* a_j\|_2$ bei festem $k \in \{1, \ldots, p-1\}$ nach unten ab. Hierbei berücksichtigen wir, dass $I(k)$ so viele Elemente hat, wie man k Elemente aus einer p-elementigen Menge auswählen kann, und diese Anzahl ist bekanntlich durch den Binomialkoeffizienten

$$\binom{p}{k} = \frac{p!}{k!\,(p-k)!}$$

gegeben. Daher ist

$$(*) \qquad \max_{J\in I(k)}\Big\|\sum_{j\in J} z_j^* a_j\Big\|_2 \ge \Big(\sum_{J\in I(k)}\Big\|\sum_{j\in J} z_j^* a_j\Big\|_2^2 \Big/ \binom{p}{k}\Big)^{1/2}.$$

Unter Benutzung von

$$0 = \Big\|\sum_{i\in I} z_i^* a_i\Big\|_2^2 = \sum_{i\in I}(z_i^*)^2 \underbrace{\|a_i\|_2^2}_{=1} + 2\sum_{\substack{i,j\in I\\ i<j}} z_i^* z_j^* a_i^T a_j$$

erhalten wir

$$\sum_{J\in I(k)}\Big\|\sum_{j\in J} z_j^* a_j\Big\|_2^2 = \sum_{J\in I(k)}\Big(\sum_{j\in J}(z_j^*)^2 + 2\sum_{\substack{i,j\in J\\ i<j}} z_i^* z_j^* a_i^T a_j^T\Big)$$

$$
\begin{aligned}
& \qquad\qquad (\text{wegen } \|a_i\|_2 = 1) \\
&= \binom{p-1}{k-1} \sum_{i\in I} (z_i^*)^2 + \binom{p-2}{k-2} \cdot 2 \sum_{\substack{i,j\in I \\ i<j}} z_i^* z_j^* a_i^T a_j \\
&= \Big[\binom{p-1}{k-1} - \binom{p-2}{k-2} \Big] \sum_{i\in I} (z_i^*)^2 \\
&= \binom{p-2}{k-1} \sum_{i\in I} (z_i^*)^2 \\
&\geq \binom{p-2}{k-1} \cdot p^{-1}.
\end{aligned}
$$

Hierbei haben wir am Schluss ausgenutzt, dass

$$
1 = \sum_{i\in I} z_i^* \leq p^{1/2} \Big(\sum_{i\in I} (z_i^*)^2 \Big)^{1/2}
$$

wegen der Cauchy-Schwarzschen Ungleichung. Setzt man dies in $(*)$ ein, so erhält man

$$
\max_{J\in I(k)} \Big\| \sum_{j\in J} z_j^* a_j \Big\|_2 \geq \Big[p^{-1} \binom{p-2}{k-1} \Big/ \binom{p}{k} \Big]^{1/2} = \Big[\frac{k(p-k)}{(p-1)p^2} \Big]^{1/2}.
$$

Damit erhalten wir schließlich

$$
w^* \leq r^* \cdot \min_{k=1,\ldots,p-1} \Big[\frac{(p-1)p^2}{k(p-k)} \Big]^{1/2} = \begin{cases} 2(p-1)^{1/2}, & \text{falls } p \text{ gerade,} \\ \dfrac{2p}{(p+1)^{1/2}}, & \text{falls } p \text{ ungerade.} \end{cases}
$$

Hieraus folgt nach leichter Argumentation die Behauptung, da $p \leq n+1$. □

Beispiel: Wir haben im Beweis der Ungleichung von Steinhagen gesehen, dass die Berechnung der Inkugel zu einem gegebenen Polytop durch Lösen einer linearen Optimierungsaufgabe erfolgen kann. In Abbildung 86 geben wir ein einfaches Beispiel an. Bei der Berechnung der Inkugel haben wir es uns einfach gemacht, indem wir die in der Optimization-Toolbox von MATLAB enthalte Funktion `linprog` benutzten. □

Jetzt kommen wir zur Jungschen Ungleichung bzw. dem Satz von Jung. Dieser gibt an, wie der Umkreisradius einer beschränkten Menge durch ihren Durchmesser abgeschätzt werden kann. Wir machen es uns etwas einfacher, indem wir nur endliche Mengen betrachten. Sei also $S := \{a_1, \ldots, a_m\}$ mit vorgegebenen paarweise verschiedenen Punkten $a_1, \ldots, a_m \in \mathbb{R}^n$, mit $m \geq 2$ Dann ist offenbar $R^*(S) = R^*(\mathrm{co}\,(S))$, d. h. die Umkugelradien von S und der konvexen Hülle $\mathrm{co}\,(S)$ von S stimmen überein. Da nämlich $S \subset \mathrm{co}\,(S)$ ist trivialerweise $R^*(S) \leq R^*(\mathrm{co}\,(S))$. Da eine Kugel mit S auch $\mathrm{co}\,(S)$ enthält, gilt auch die umgekehrte Ungleichung. Weiter gilt auch $D^*(S) = D^*(\mathrm{co}\,(S))$. Wegen $S \subset \mathrm{co}\,(S)$ ist trivialerweise $D^*(S) \leq D^*(\mathrm{co}\,(S)$. Zum Beweis der umgekehrten Ungleichung geben wir uns $a, \hat{a} \in \mathrm{co}\,(S)$ vor. Als Elemente der konvexen Hülle von S besitzen a und $\hat{a}$ Darstellungen der Form

$$
a = \sum_{k=1}^{m} \lambda_k a_k, \qquad \hat{a} = \sum_{l=1}^{m} \hat{\lambda}_l a_l
$$

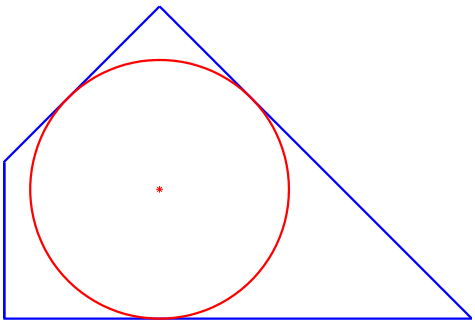


Abbildung 86: Inkugel zu einem ebenen Polyeder

mit

$$\lambda = (\lambda_k),\ \hat{\lambda} = (\hat{\lambda}_l) \in \Sigma_m := \{\lambda \in \mathbb{R}^m : \lambda \geq 0,\ e^T\lambda = 1\},$$

wobei $e \in \mathbb{R}^m$ einmal wieder der Vektor ist, dessen Komponenten sämtlich gleich 1 sind. Für $l = 1, \ldots, m$ ist

$$\|a - a_l\|_2 = \Big\|\sum_{k=1} \lambda_k - a_l\Big\|_2 = \Big\|\sum_{k=1}^{m} \lambda_k(a_k - a_l)\Big\|_2 \leq \sum_{k=1} \lambda_k \underbrace{\|a_k - a_l\|_2}_{\leq D^*(S)} \leq D^*(S).$$

Daher ist

$$\|a - \hat{a}\|_2 = \Big\|a - \sum_{l=1}^{m} \hat{\lambda}_l a_l\Big\|_2 = \Big\|\sum_{l=1}^{m} \hat{\lambda}_l(a - a_l)\Big\|_2 \leq \sum_{l=1}^{m} \hat{\lambda}_l \underbrace{\|a - a_l\|_2}_{\leq D^*(S)} \leq D^*(S).$$

Folglich ist auch $D^*(\mathrm{co}\,(S)) \leq D^*(S)$ und daher insgesamt $D^*(S) = D^*(\mathrm{co}\,(S))$.

Satz von Jung *Sei $S := \{a_1, \ldots, a_m\}$ mit paarweise verschiedenen $a_1, \ldots, a_m \in \mathbb{R}^n$ und $m \geq 2$. Ist $R^*(S)$ der Umkugelradius und $D^*(S)$ der Durchmesser von S, so ist*

$$R^*(S) \leq D^*(S) \cdot \Big(\frac{n}{2(n+1)}\Big)^{1/2}.$$

Beispiel: Die Seitenlänge a eines gleichseitigen Dreiecks in der Ebene ist der Durchmesser D^*, der Umkreisradius ist $R^* = a \cdot (\sqrt{3}/3)$. In diesem Fall ist also

$$R^*(S) = a\frac{\sqrt{3}}{3} = D^*(S)\frac{\sqrt{3}}{3},$$

in der Jungschen Ungleichung gilt in diesem Fall also sogar Gleichheit. □

Beweis des Satzes von Jung Der Mittelpunkt x^* und der Radius R^* der Umkugel zu S sind Lösung der Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x,R) := \dfrac{1}{2}R^2 \quad \text{unter der Nebenbedingung} \\ \qquad g(x,R) := \left(\begin{array}{c} \dfrac{1}{2}\|x-a_1\|_2^2 - \dfrac{1}{2}R^2 \\ \vdots \\ \dfrac{1}{2}\|x-a_m\|_2^2 - \dfrac{1}{2}R^2 \end{array} \right) \leq \left(\begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right). \end{array} \right.$$

Der Satz von Kuhn-Tucker (Satz 2 in Abschnitt 53) liefert die Existenz von $u^* \in \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \left(\begin{array}{c} 0 \\ R^* \end{array} \right) + \sum_{i=1}^m u_i^* \left(\begin{array}{c} x^* - a_i \\ -R^* \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \end{array} \right)$$

sowie

$$u_i^*(\|x^* - a_i\|_2 - R^*) = 0, \quad i = 1, \ldots, m.$$

Wegen $R^* > 0$ schließen wir hieraus, dass

$$u^* \in \Sigma_m := \{u \in \mathbb{R}^m : u \geq 0,\ e^T u = 1\}$$

und

$$x^* = \sum_{i=1}^m u_i^* a_i, \qquad u_i^*(\|x^* - a_i\|_2 - R^*) = 0, \quad (i = 1, \ldots, m).$$

Wir definieren

$$I^* := \{i \in \{1, \ldots, m\} : u_i^* > 0\}.$$

Dann ist

$$\sum_{i \in I^*} u_i^* = 1, \qquad x^* = \sum_{i \in I^*} u_i^* a_i, \qquad \|x^* - a_i\|_2 = R^*, \quad (i \in I^*).$$

Fast wörtlich wie im entsprechenden Teil im Beweis der Ungleichung von Steinhagen haben wir: Insbesondere liegt x^* in der konvexen Hülle (siehe S. 245) $\operatorname{co}(\{a_i\}_{i \in I^*})$ von $\{a_i\}_{i \in I^*}$. Wegen des Satzes von Carathéodory (siehe S. 246) existiert eine Indexmenge $I \subset I^*$ mit $p := |I| \leq n+1$ und $x^* \in \operatorname{co}(\{a_i\}_{i \in I})$. Folglich existieren v_i^*, $i \in I$, mit

$$v_i^* > 0 \quad (i \in I), \qquad \sum_{i \in I} v_i^* = 1, \qquad x^* = \sum_{i \in I} v_i^* a_i.$$

Für ein beliebiges $j \in I$ ist

$$\begin{aligned} \sum_{i \in I} v_i^* \|a_i - a_j\|_2^2 &= \sum_{i \in I} v_i^* \|(x^* - a_j) - (x^* - a_i)\|_2^2 \\ &= \sum_{i \in I} v_i^* (\underbrace{\|x^* - a_j\|_2^2}_{=(R^*)^2} - 2(x^* - a_j)^T (x^* - a_i) + \underbrace{\|x^* - a_i\|_2^2}_{=(R^*)^2}) \\ &= 2(R^*)^2 - 2\underbrace{\Big(\sum_{i \in I} v_i^* (x^* - a_i)\Big)^T}_{=0} (x^* - a_j) \\ &= 2(R^*)^2. \end{aligned}$$

Folglich ist

$$\begin{aligned}
2(R^*)^2 &= \sum_{i,j\in I} v_i^* v_j^* \|a_i - a_j\|_2^2 \\
&\le D^*(S)^2 \sum_{\substack{i,j\in I \\ i\neq j}} v_i^* v_j^* \\
&= D^*(S)^2 \Big[\Big(\underbrace{\sum_{i\in I} v_i^*}_{=1} \Big)^2 - \sum_{i\in I} (v_i^*)^2 \Big] \\
&\le D^*(S)^2 \Big(1 - \frac{1}{p} \Big) \\
&\le D^*(S)^2 \frac{n}{n+1}.
\end{aligned}$$

Hierbei haben wir am Schluss ausgenutzt, genau wie beim Beweis des Satzes von Steinhagen, dass

$$1 = \sum_{i\in I} v_i^* = \sum_{i\in I} 1 \cdot v_i^* \le \Big(\sum_{i\in I} 1^2 \Big)^{1/2} \sum_{i\in I} (v_i^*)^2 \Big)^{1/2} = p^{1/2} \Big(\sum_{i\in I} (v_i^*)^2 \Big)$$

wegen der Cauchy-Schwarzschen Ungleichung. Insgesamt haben wir die Jungsche Ungleichung bewiesen. □

64 Magische Quadrate

Die übliche Definition eines magischen Quadrates ist:

- Ein *magisches Quadrat* der *Kantenlänge* n ist eine quadratische Anordnung der Zahlen $1, 2, \ldots, n^2$ mit der Eigenschaft, dass die Summe der Zahlen aller Zeilen, Spalten und der beiden Diagonalen gleich ist.

Es gibt viele sehr gut gemachte Internet-Seiten über magische Quadrate. Wir erwähnen `http://www.hp-gramatke.de/magic_sq/german/page0020.htm` zur Geschichte magischer Quadrate, `http://www.hp-gramatke.de/magic_sq/` und die englischsprachige Seite `http://mathforum.org/alejandre/magic.square.html`. Hierauf greifen wir immer wieder zurück.

Gelegentlich heißt ein magisches Quadrat nach obiger Definition ein *normales* oder *natürliches* magisches Quadrat. Bei Varianten sind nicht alle Bedingungen erfüllt oder es werden zusätzliche Einschränkungen gefordert. So heißt z. B. eine quadratische Anordnung der Zahlen $1, \ldots, n^2$, bei der zwar die Zeilen- und Spaltensummen sämtlich gleich sind, nicht aber notwendig die Diagonalsummen, ein *semi-magisches Quadrat*. Eines der bekanntesten (normalen) magischen Quadrate findet sich im Kupferstich *Melencolia I* von Albrecht Dürer, siehe z. B. `http://de.wikipedia.org/wiki/Melencolia_`

I. Dieses magische Quadrat ist gegeben durch

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

Drehungen dieses magischen Quadrats um 90°, 180° und 270° ergibt der Reihe nach weitere magische Quadrate, nämlich

13	8	12	1
2	11	7	14
3	10	6	15
16	5	9	4

1	14	15	4
12	7	6	9
8	11	10	5
13	2	3	16

4	9	5	16
15	6	10	3
14	7	11	2
1	12	8	13

Spiegelung an den beiden Hauptdiagonalen und horizontalen und vertikalen Achsen durch den Mittelpunkt ergibt noch einmal vier weitere magische Quadrate:

16	5	9	4
3	10	6	15
2	11	7	14
13	8	12	1

1	12	8	13
14	7	11	2
15	6	10	3
4	9	5	16

4	15	14	1
9	6	7	12
5	10	11	8
16	3	2	13

13	2	3	16
8	11	10	5
12	7	6	9
1	14	15	4

Alle diese 8 magischen Quadrate werden als im wesentlichen gleich angesehen.

Bei einem magischen Quadrat der Kantenlänge n ist die Summe S_n der Zahlen in jeder Zeile (und jeder Spalte und den beiden Diagonalen), die sogenannte *magische Zahl*, gleich der Summe aller Zahlen von 1 bis n^2 geteilt durch n. Also ist

$$S_n = \frac{1}{n}\sum_{j=1}^{n^2} j = \frac{1}{n}\cdot\frac{n^2(n^2+1)}{2} = \frac{n(n^2+1)}{2}.$$

Offenbar gibt es kein magisches Quadrat der Kantenlänge 2. Wir betrachten jetzt den Fall $n = 3$. Ist

a	b	c
d	e	f
g	h	i

ein magisches Quadrat, so sind auch

c	f	i
b	e	h
a	d	g

i	h	g
f	e	d
c	b	a

g	d	a
h	e	b
i	f	c

g	h	i
d	e	f
a	b	c

c	b	a
f	e	d
i	h	g

a	d	g
b	e	h
c	f	i

i	f	c
h	e	b
g	d	a

im wesentlichen gleiche magische Quadrate. Addiert man die Zahlen in den beiden Diagonalen sowie in der Zeile sowie der Spalte, welche das mittlere Element e enthalten, so erhält man unter Berücksichtigung von $S_3 = 15$, dass

$$\begin{aligned} 60 &= (a+e+i)+(g+e+c)+(d+e+f)+(b+e+h) \\ &= 3e+\underbrace{(g+h+i)}_{=15}+\underbrace{(a+c+b)}_{=15}+\underbrace{(d+f+e)}_{=15}, \end{aligned}$$

woraus $e = 5$ folgt. Hieraus wiederum ergibt sich

$$a + i = g + c = d + f = b + h = 10.$$

Es können nicht a und i kleiner als 5 sein, da ihre Summe 10 ergibt. Da wir notfalls an der Nebendiagonalen spiegeln können, kann $a > 5$ angenommen werden. Also ist $a = 5 + x$, $i = 5 - x$ mit positivem x. Ebenso ist c oder g größer als 5. Da wir notfalls an der Hauptdiagonalen spiegeln können (diese bleibt dabei natürlich fest), können wir annehmen, dass $c = 5+y$, $g = 5-y$ mit positivem y. Wenn also ein magisches Quadrat der Kantenlänge 3 existiert, dann existiert auch eines der Form

$5 + x$	$5 - (x + y)$	$5 + y$
$5 - (x - y)$	5	$5 + (x - y)$
$5 - y$	$5 + (x + y)$	$5 - x$

Hier können wir annehmen, dass $x > y$ (beachte: es ist notwendig $x \neq y$), da wir notfalls an der senkrechten Mittellinie spiegeln können. Man überlegt sich leicht, dass nur $x = 3$, $y = 1$ mit den gestelLten Bedingungen vereinbar ist. Für $n = 3$ gibt es also im wesentlichen nur ein magisches Quadrat, nämlich

8	1	6
3	5	7
4	9	2

Bemerkung: Bei W. Sierpinski (1988, S. 436) findet man einen Hinweis auf eine Note von A. Makowski (1962), durch die die Existenz magischer Quadrate für beliebig große Kantenlänge n bewiesen werden kann. Hierzu nehmen wir an, Q_n und Q_m seien magische Quadrate der Kantenlänge n bzw. m. Hieraus kann man ein magisches Quadrat Q_{nm} der Kantenlänge nm erhalten. Dies erreicht man, indem man Q_n für jede Zahl i in Q_m substituiert, wobei man $n^2(i-1)$ zu jeder Zahl in Q_n addiert. Bevor wir uns überlegen, dass diese Aussage richtig ist, geben wir ein Beispiel an. Wir betrachten die magischen Quadrate Q_3 und Q_4, die durch

$$Q_3 = \begin{array}{|c|c|c|}\hline 8 & 1 & 6 \\ \hline 3 & 5 & 7 \\ \hline 4 & 9 & 2 \\ \hline \end{array}, \qquad Q_4 = \begin{array}{|c|c|c|c|}\hline 16 & 3 & 2 & 13 \\ \hline 5 & 10 & 11 & 8 \\ \hline 9 & 6 & 7 & 12 \\ \hline 4 & 15 & 14 & 1 \\ \hline \end{array}$$

gegeben sind. Es sei $n = 3$ und $m = 4$. Nach der obigen Vorschrift erhalten wir Q_{12}

durch

$Q_{12} =$

143	136	141	26	19	24	17	10	15	116	109	114
138	140	142	21	23	25	12	14	16	111	113	115
139	144	137	22	27	20	13	18	11	112	117	110
44	37	42	89	82	87	98	91	96	71	64	69
39	41	43	84	86	88	93	95	97	66	68	70
40	45	38	85	90	83	94	99	92	67	72	65
80	73	78	53	46	51	62	55	60	107	100	105
75	77	79	48	50	52	57	59	61	102	104	106
76	81	74	49	54	47	58	63	56	103	108	101
35	28	33	134	127	132	125	118	123	8	1	6
30	32	34	129	131	133	120	122	124	3	5	7
31	36	29	130	135	128	121	126	118	4	9	2

Wir beachten, dass durch die angegebene Vorschrift alle Zahlen von 1 bis n^2m^2 in ein $nm \times nm$-Feld geschrieben werden. Nämlich die Zahlen $1, \ldots, n^2$ dort, wo in Q_m die Zahl 1 steht bis $n^2(m^2-1)+1, \ldots, n^2m^2$ dort, wo in Q_m die Zahl m^2 steht. Die Summe über die Zeilen, Spalten und die beiden Diagonalen ist jeweils

$$\begin{aligned} mS_n + n^2(nS_m - nm) &= m\frac{n(n^2+1)}{2} + n^2\Big(n\frac{m(m^2+1)}{2} - nm\Big) \\ &= \frac{nm(n^2m^2+1)}{2} \\ &= S_{nm}, \end{aligned}$$

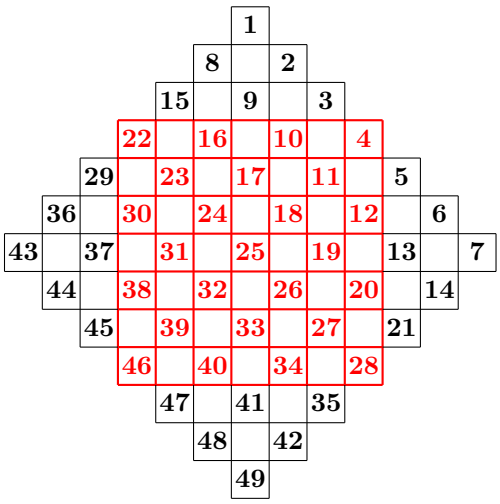
also ist das konstruierte Quadrat Q_{nm} ebenfalls magisch.Insbesondere existieren (und können auf die angegebene Weise konstruiert werden) magische Quadrate der Kantenlänge 3^k, $k = 1, 2 \ldots$, also magische Quadrate beliebig großer Kantenlänge. □

Als einzigen Satz in diesem Abschnitt formulieren und beweisen wir:

Satz *Für jede natürliche Zahl $n \geq 3$ gibt es mindestens ein (normales) magisches Quadrat der Kantenlänge n.*

Beweis: Der Beweis ist konstruktiv (mir ist kein von Konstruktionen unabhängiger Existenzbeweis bekannt) und unterscheidet insgesamt drei Fälle. Im ersten Fall ist n *ungerade*, d. h. $n = 2k + 1$ mit $k \in \mathbb{N}$. Wir schildern die sogenannte *Terrassenmethode* von Claude Gaspard Bachet de Méziriac (1581–1638). Wir veranschaulichen das Verfahren zunächst für $n = 7$ an und betrachten anschließend den allgemeinen Fall.

In Abbildung 87 links geben wir ein 7×7-Feld an, bei welchem wir an den vier Seiten eine dreieckige Anordnung von $5 + 3 + 1 = 9$ Feldern angesetzt haben. In den Nebendiagonalen haben wir, rechts oben beginnend, der Reihe nach die Zahlen von 1 bis 49 eingetragen. Diejenigen Zahlen, die in das 7×7-Feld fallen, werden rot gekennzeichnet. Das sind Zahlen, die schon an ihrem richtigen Platz sind und nicht verändert werden. Danach werden die an den Seiten angesetzten Teile (schwarze Zahlen) auf die jeweils gegenüberliegende Seite verschoben, wobei wir sie an der entsprechenden Stelle blau eintragen, siehe Abbildung 87 rechts. Man weist leicht nach, dass



22	47	16	41	10	35	4
5	23	48	17	42	11	29
30	6	24	49	18	36	12
13	31	7	25	43	19	37
38	14	32	1	26	44	20
21	39	8	33	2	27	45
46	15	40	9	34	3	28

Abbildung 87: Das Verfahren von Bachet für $n = 7$

hierdurch ein magisches Quadrat erzeugt wurde. Nach diesem Beispiel müssen wir zeigen, dass diese Methode nicht nur für $n = 7$ funktioniert. In der Hauptdiagonale des erzeugten $n \times n$-Feldes stehen die Zahlen $k \cdot n + 1, \ldots, k \cdot n + n$, ihre Summe ist $\sum_{j=1}^{n}(kn + j) = n(n^2 + 1)/2$, die magische Zahl S_n. Die Nebendiagonale wird besetzt durch die Zahlen $k + 1, k + 1 + 1 \cdot n, \ldots, k + 1 + (n - 1) \cdot n$, diese haben als Summe $(k + 1) \cdot n + n \cdot \sum_{j=1}^{n-1} j$ wieder die magische Zahl S_n. Scharfes Hinschauen ergibt auch für die Zeilen und Spalten des $n \times n$-Feldes jeweils als Summe die magische Zahl S_n.

Zum Vergleich geben wir auch noch das Verfahren von Simon de la Loubère (1642–1729) an, bei welchem sukzessive die Zahlen von 1 bis n^2 im $n \times n$-Feld abgelegt werden. Man beginnt in der Position $(i, j) := (1, (n + 1)/2)$, also in der Mitte der ersten Zeile, und setzt dort die 1. I. Allg. geht man von der Position (i, j) zur Position $(i + 1, j + 1)$ (also in Nord-Ost-Richtung), es sei denn, man verlässt dabei das $n \times n$-Feld oder man trifft auf ein schon besetztes Feld. Das $n \times n$-Feld kann nach oben (also Norden) oder nach rechts (also Osten) verlassen werden. Nachfolger von $(1, j)$ ist $(n, j + 1)$ bzw. $(2, n)$ für $j = n$, Nachfolger von (i, n) ist $(i - 1, 1)$. Ist der geplante Nachfolger schon besetzt, so nimmt man $(i - 1, j)$ als Nachfolger. In Abbildung 88 geben wir das Ergebnis für $n = 7$ an. Dasselbe Ergebnis erhält man übrigens, wenn man in MATLAB

30	39	48	1	10	19	28
38	47	7	9	18	27	29
46	6	8	17	26	35	37
5	14	16	25	34	36	45
13	15	24	33	42	44	4
21	23	32	41	43	3	12
22	31	40	49	2	11	20

Abbildung 88: Das Verfahren von de la Loubère für $n = 7$

`magic(7)` berechnet, denn für ungerade Kantenlänge wird das Verfahren von de la Loubere benutzt.

Nun nehmen wir an, es sei $n \geq 4$ gerade und unterscheiden zwei Fälle. Im ersten Fall nehmen wir an, die Kantenlänge n des gesuchten magischen Quadrats sei durch 4 teilbar. Wir sprechen dann vom *doppelt-geraden* Fall. Es sei also $n = 4k$ mit $k \in \mathbb{N}$. Das $n \times n$-Feld denken wir uns in sechzehn $k \times k$-Felder unterteilt. Von links oben beginnend tragen wir in Schritt 1 die Zahlen von 1 bis n^2 in das $n \times n$-Feld ein, wobei allerdings nur die 4 Eckblöcke und die 4 Mittelblöcke mit den entsprechenden Zahlen beschrieben werden. Danach trage man in Schritt 2 die fehlenden Zahlen in die noch freien Felder ein, wobei man rechts unten beginnt. Als Beispiel betrachten wir den Fall $k = 3$ bzw. $n = 12$. Der Schritt 1 wird in Abbildung 89 dargestellt. Danach trägt man in Schritt

1	2	3							10	11	12
13	14	15							22	23	24
25	26	27							34	35	36
			40	41	42	43	44	45			
			52	53	54	55	56	57			
			64	65	66	67	68	69			
			76	77	78	79	80	81			
			88	89	90	91	92	93			
			100	101	102	103	104	105			
109	110	111							118	119	120
121	122	123							130	131	132
133	134	135							142	143	144

Abbildung 89: Konstruktion eines magischen Quadrats der Kantenlänge 12, Schritt 1

2 die fehlenden Zahlen blau in die noch freien Felder ein, wobei man rechts unten beginnt. Dies geben wir in Abbildung 90 an. Man weist leicht nach, dass hierdurch ein 12×12-magisches Quadrat gegeben ist. Nun müssen wir uns überlegen, dass die angegebene Vorgehensweise auch im allgemeinen Fall zum Ziel führt. Wir denken uns das $n \times n$-Feld in sechzehn $k \times k$-Felder zerlegt. Beim Durchnummerieren von links oben nach rechts unten bleiben die vier Eckfelder sowie vier $k \times k$-Felder in der Mitte unverändert, die entsprechenden Zahlen werden schwarz in die Felder eingetragen. Die zunächst noch freien Felder werden mit den restlichen Zahlen von rechts unten nach

1	2	3	141	140	139	138	137	136	10	11	12
13	14	15	129	128	127	126	125	124	22	23	24
25	26	27	117	116	115	114	113	112	34	35	36
108	107	106	40	41	42	43	44	45	99	98	97
96	95	94	52	53	54	55	56	57	87	86	85
84	83	82	64	65	66	67	68	69	75	74	73
72	71	70	76	77	78	79	80	81	63	62	61
60	59	58	88	89	90	91	92	93	51	50	49
48	47	46	100	101	102	103	104	105	39	38	37
109	110	111	33	32	31	30	29	28	118	119	120
121	122	123	21	20	19	18	17	16	130	131	132
133	134	135	9	8	7	6	5	4	142	143	144

Abbildung 90: Konstruktion eines magischen Quadrats der Kantenlänge 12, Schritt 2

links oben gefüllt. Man erhält ein Feld der Form

A_{11}	A_{12}	A_{13}	A_{14}
A_{21}	A_{22}	A_{23}	A_{24}
A_{31}	A_{32}	A_{33}	A_{34}
A_{41}	A_{42}	A_{43}	A_{44}

Mit E_k bezeichnen wir die $k \times k$-Matrix, deren Einträge sämtlich gleich 1 sind. Man weist leicht nach, dass

$$\begin{aligned}
(A_{11} + A_{12}) + (A_{13} + A_{14}) &= (n^2 - k + 1)E_k + (n^2 + k + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{21} + A_{22}) + (A_{23} + A_{24}) &= (n^2 + k + 1)E_k + (n^2 - k + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{31} + A_{32}) + (A_{33} + A_{34}) &= (n^2 + k + 1)E_k + (n^2 - k + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{41} + A_{42}) + (A_{43} + A_{44}) &= (n^2 - k + 1)E_k + (n^2 + k + 1)E_k = 2(n^2 + 1)E_k.
\end{aligned}$$

Daher ist die Summe der Zahlen in den Zeilen des $n \times n$-Feldes jeweils gleich $2(n^2+1)k = n(n^2 + 1)/2 = S_n$. Ebenso ist

$$\begin{aligned}
(A_{11} + A_{21}) + (A_{31} + A_{41}) &= (n^2 - kn + 1)E_k + (n^2 + kn + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{12} + A_{22}) + (A_{32} + A_{42}) &= (n^2 + kn + 1)E_k + (n^2 - kn + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{13} + A_{23}) + (A_{33} + A_{43}) &= (n^2 + kn + 1)E_k + (n^2 - kn + 1)E_k = 2(n^2 + 1)E_k, \\
(A_{14} + A_{24}) + (A_{34} + A_{44}) &= (n^2 - kn + 1)E_k + (n^2 + kn + 1)E_k = 2(n^2 + 1)E_k.
\end{aligned}$$

Hieraus folgt, dass auch die Summe der Zahlen in den Spalten des $n \times n$-Feldes jeweils gleich $2(n^2+1)k = n(n^2+1)/2 = S_n$ ist. Die Summe der Zahlen in der Hauptdiagonale ist

$$\begin{aligned} 1 + (1 \cdot (n+1) + 1) + \cdots + ((n-1) \cdot (n+1) + 1) &= n + (n+1) \sum_{j=1}^{n-1} j \\ &= n + (n+1) \frac{n(n-1)}{2} \\ &= \frac{n(n^2+1)}{2} \\ &= S_n. \end{aligned}$$

Entsprechend ist die Summe der Zahlen in der Nebendiagonale

$$\begin{aligned} (1 \cdot (n-1) + 1) + \cdots + (n \cdot (n-1) + 1) &= n + (n-1) \sum_{j=1}^{n} j \\ &= n + (n-1) \frac{n(n+1)}{2} \\ &= \frac{n(n^2+1)}{2} \\ &= S_n. \end{aligned}$$

Also ist die Summe der Zahlen in den Spalten, Zeilen und den beiden Diagonalen gleich der magischen Zahl S_n, womit die angegebene Vorgehensweise gerechtfertigt ist.

Nun kommen wir zum schwierigsten der drei Fälle, dass nämlich die Kantenlänge n des gesuchten magischen Quadrats zwar gerade, aber nicht durch 4 teilbar ist. Wir sprechen dann von dem *einfach-geraden* Fall. Hier ist also $n = 2(2k+1)$ mit $k \in \mathbb{N}$. Die Idee[88] besteht darin, sich das gesuchte magische $n \times n$-Feld in vier $(2k+1) \times (2k+1)$-Felder aufgeteilt zu denken und einen Ansatz der Form

A_{11}	A_{12}	Q	$Q + 2(2k+1)^2 E$
A_{21}	A_{22}	$Q + 3(2k+1)^2 E$	$Q + (2k+1)^2 E$

zu machen. Hierbei ist Q ein magisches Quadrate der Kantenlänge $2k+1$, welches mit dem Verfahren von Bachet oder de la Loubère berechnet werden kann, und E ein $(2k+1) \times (2k+1)$-Feld bzw. Matrix, dessen bzw. deren Einträge sämtlich gleich 1 sind. Klar ist, dass in diesem Ansatzfeld alle Zahlen von 1 bis $4(2k+1)^2 = n^2$ genau einmal auftreten. Die Spaltensummen sind offenbar $2S_{k+1} + 3(2k+1)^3 = S_n$, was für ein magisches Quadrat der Kantenlänge n schon die richtige Größe ist. Diese

[88] Wir schildern im Prinzip die Methode, die in der MATLAB-Funktion `magic` realisiert ist. Diese Funktion ist selbst in MATLAB geschrieben, man kann sie sich durch den Befehl `type magic` ansehen.

Eigenschaft bleibt beim Vertauschen von Komponenten innerhalb der Spalten erhalten. Die Zeilensumme der oberen $2k+1$ Zeilen ist jeweils gleich $2S_{2k+1}+2(2k+1)^3$, die der unteren $2k+1$ Zeilen jeweils gleich $2S_{2k+1}+4(2k+1)^3$. Daher werden wir versuchen, innerhalb von Spalten Vertauschungen vorzunehmen, welche die unteren Zeilensummen um $(2k+1)^3$ vermindern und die oberen um $(2k+1)^3$ vergrößern. Vertauscht man in der j-ten Spalte mit $j \in \{1,\ldots,2k+1\}$ des $n \times n$-Ansatzfeldes die Komponenten i und $i+(2k+1)$ (wobei sich die Spaltensumme nicht verändert), so vergrößert sich die i-te Zeilensumme um $3(2k+1)^2$ und die $(i+2k+1)$-te Zeilensumme vermindert sich um denselben Wert. Hier sprechen wir von einer Vertauschung vom Typ I. Ist dagegen $j \in \{2k+2,\ldots,2(2k+1)\}$, so vermindert sich beim Vertauschen der Komponenten i und $i+(2k+1)$ in der j-ten Spalte die i-te Zeilensumme um $(2k+1)^2$, während sich die $(i+2k+1)$-te Zeilensumme um den gleichen Wert vergrößert. Dies nennen wir eine Vertauschung vom Typ II. Macht man daher p Vertauschungen vom Typ I und q Vertauschungen vom Typ II, so vergrößert sich die i-te Zeilensumme um $(3p-q)(2k+1)^2$, während sich die $(i+2k+1)$-te Zeilensumme um $(3p-q)(2k+1)^2$ vermindert. Es liegt daher nahe, $p=k$ und $q=k-1$ zu wählen, und in den Spalten $1,\ldots,k$ (Vertauschungen vom Typ I) bzw. $n-k+2,\ldots,n$ (Vertauschungen vom Typ II) k bzw. $k-1$ Vertauschungen in den Zeilen i und $i+(2k+1)$, $i=1,\ldots,2k+1$, vorzunehmen. Wenn man diesen Schritt gemacht hat, sind alle Zeilensummen gleich S_n und die Spaltensummen haben ihren richtigen Wert S_n behalten.

Die Diagonalsummen sind noch nicht ganz richtig. *Vor* den eben geschilderten Vertauschungen vom Typ I und Typ II ist die Hauptdiagonalsumme $2S_{2k+1}+(2k+1)^3$, die Nebendiagonalsumme ist $2S_{2k+1}+5(2k+1)^3$. Daher muss die Hauptdiagonalsumme um $2(2k+1)^3$ vergrößert und die Nebendiagonalsumme um denselben Wert vermindert werden. *Nach* den Vertauschungen (k Vertauschungen vom Typ I, $k-1$ Vertauschungen vom Typ II) hat sich die Hauptdiagonalsumme um $(4k-1)(2k+1)^2$ vergrößert und die Nebendiagonalsumme um denselben Wert vermindert. Also muss die Hauptdiagonalsumme noch um $3(2k+1)^2$ vergrößert, die Nebendiagonalsumme um denselben Wert vermindert werden. Dies muss natürlich geschehen, ohne dass die schon richtigen Spalten- und Zeilensummen verändert werden. Hierzu vertauschen wir in der ersten und der $(k+1)$-ten Spalte (jeweils also eine Vertauschung vom Typ I) die Komponenten $k+1$ und $k+1+(2k+1)$. Die Spaltensummen bleiben dabei selbstverständlich unverändert. Das (Haupt-) Diagonalelement in der Position $(k+1,k+1)$ (in der $(k+1)$-ten Spalte hat noch keine Vertauschung stattgefunden) vergrößert sich um den Wert $3(2k+1)^2$, alle anderen Haupdiagonalelemente bleiben unverändert. Das Nebendiagonalelement in der Position $(k+1,k+1+(2k+1)$ vermindert sich um $3(2k+1)^2$, alle anderen Nebendiagonalelemente bleiben unverändert.Also haben die Spalten- und die beiden Diagonalsummen den richtigen Wert. Jetzt müssen wir uns nur noch davon überzeugen, dass die Zeilensummen den richtigen Wert behalten haben. Das ist aber klar, denn verändert werden nur die $(k+1)$-te und die $(3k+2)$-te Zeile und diese auch nur in den Komponenten 1 und $k+1$. Die Vergrößerungen und Verminderungen um $3(2k+1)^2$ heben sich jeweils gegenseitig auf, die Zeilensummen behalten den richtigen Wert. Damit ist die Konstruktion auch im letzten Fall abgeschlossen und insbesondere der Satz bewiesen. □

Bemerkung: Sei $n>2$ eine gerade Zahl, die nicht durch 4 teilbar ist. Ein MATLAB-

Programm zur Berechnung eines magischen Quadrats Q der Kantenlänge n könnte folgendermaßen aussehen (und so ist es auch in `magic` realisiert), wobei davon ausgegangen wird, dass `magic(p)` für ungerades p ein magisches Quadrat der Kantenlänge p liefert.

```
p=n/2; Q=magic(p); Q=[Q Q+2*p^2;Q+3*p^2 Q+p^2];
i=(1:p)'; k=(n-2)/4; j=[1:k (n-k+2):n]; Q([i;i+p],j)=Q([i+p;i],j);
i=k+1;j=[1 i]; Q([i;i+p],j)=Q([i+p;i],j);
```

Nach Abschluss ist Q ein magisches Quadrat der Kantenlänge n. □

Beispiel: Wir wollen das Verfahren im letzten Teil des Satzes, also die Bestimmung eines magischen Quadrats im einfach-geraden Fall, für $n = 10$ bzw. $k = 2$ (es ist $n = 2(2k+1)$) demonstrieren. Das Ansatzfeld (der linke obere Block ist mit dem Verfahren von de la Loubère berechnet) ist links angegeben. Um die richtigen Zeilensummen zu erhalten, finden Vertauschungen in den ersten beiden und der letzten Spalte statt. Das Ergebnis geben wir rechts an.

17	24	1	8	15	67	74	51	58	65
23	5	7	14	16	73	55	57	64	66
4	6	13	20	22	54	56	63	70	72
10	12	19	21	3	60	62	69	71	53
11	18	25	2	9	61	68	75	52	59
92	99	76	83	90	42	49	26	33	40
98	80	82	89	91	48	30	32	39	41
79	81	88	95	97	29	31	38	45	47
85	87	94	96	78	35	37	44	46	28
86	93	100	77	84	36	43	50	27	34

92	99	1	8	15	67	74	51	58	40
98	80	7	14	16	73	55	57	64	41
79	81	13	20	22	54	56	63	70	47
85	87	19	21	3	60	62	69	71	28
86	93	25	2	9	61	68	75	52	34
17	24	76	83	90	42	49	26	33	65
23	5	82	89	91	48	30	32	39	66
4	6	88	95	97	29	31	38	45	72
10	12	94	96	78	35	37	44	46	53
11	18	100	77	84	36	43	50	27	59

Jetzt ergeben die Zeilen- und Spaltensummen jeweils die magische Zahl $S_{10} = 505$. Um auch noch die richtigen Diagonalsummen zu erhalten, werden in der ersten und dritten Spalte die dritte und die achte Komponente miteinander vertauscht. Das Ergebnis ist:

92	99	1	8	15	67	74	51	58	40
98	80	7	14	16	73	55	57	64	41
79	81	88	20	22	54	56	63	70	47
85	87	19	21	3	60	62	69	71	28
86	93	25	2	9	61	68	75	52	34
17	24	76	83	90	42	49	26	33	65
23	5	82	89	91	48	30	32	39	66
79	6	13	95	97	29	31	38	45	72
10	12	94	96	78	35	37	44	46	53
11	18	100	77	84	36	43	50	27	59

Das gesuchte magische Quadrat der Kantenlänge 10 ist damit gefunden. □

Über die Anzahl magischer Quadrate einer gegebenen Kantenlänge n gibt es nur für kleine n gesicherte Aussagen. Hierauf und auf spezielle magische Quadrate wollen wir nicht mehr eingehen, sondern nur erwähnen, dass in Tabelle 1 rechts in Abschnitt 16 ein semi-magisches 8×8-Rösselsprung-Quadrat (zur Erinnerung: Bei einem semi-magischen Quadrat stimmen zwar die Zeilen- und die Spaltensummen überein, nicht

notwendig sind diese aber gleich den Diagonalsummen) angegeben wurde. Erst seit 2003 weiß man, dass es kein magisches 8×8-Rösselsprung-Quadrat gibt und dass die Anzahl (im wesentlichen) verschiedener semi-magischer 8×8-Rösselsprung-Quadrate genau 140 ist, siehe `http://mathworld.wolfram.com/news/2003-08-06/magictours/`. Das in Tabelle 5 angegebene magische Quadrat ist *fast* ein magisches Rösselsprung-Quadrat. Nur der Übergang von 32 zu 33 ist kein Rösselsprung. Dieses magische Quadrat soll von M. A. Feisthamel stammen, siehe `http://plus.maths.org/content/anything-square-magic-squares-sudoku`.

5	14	53	62	3	12	51	60
54	63	4	13	52	61	2	11
15	6	55	24	41	10	59	50
64	25	16	7	58	49	40	1
17	56	33	42	23	32	9	48
34	43	26	57	8	39	22	31
27	18	45	36	29	20	47	38
44	35	28	19	46	37	30	21

Tabelle 5: Ein fast magisches Rösselsprung-Quadrat

65 Lateinische Quadrate

65.1 Orthogonale lateinische Quadrate

Die übliche Definition eines lateinischen Quadrats ist:

- Ein *lateinisches Quadrat der Ordnung* n ist eine quadratische Anordnung von n verschiedenen Symbolen, z. B. den Zahlen $1, \ldots, n$, oder n Buchstaben oder Farben, mit der Eigenschaft, dass jedes Symbol in jeder Zeile und in jeder Spalte jeweils genau einmal auftritt.

Bei magischen Quadraten mussten wir uns ein wenig anstrengen, um die Existenz magischer Quadrate jeder Ordnung $n \geq 3$ nachzuweisen. Bei lateinischen Quadraten ist das ganz einfach, denn durch

1	2	3	$\cdots$	n
2	3	4	$\cdots$	1
3	4	5	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$n-1$	n	1	$\cdots$	$n-2$
n	1	2	$\cdots$	$n-1$

ist ein lateinisches Quadrat der Ordnung n gegeben. Lateinische Quadrate wurden von Leonhard Euler eingeführt, der als Symbole lateinische Buchstaben benutzte, was zu dem Namen führte. Eine etwas genauer gefasste und offensichtlich äquivalente Definition eines lateinischen Quadrats der Ordnung n ist die folgende:

- Sei S eine n-elementige Menge. Ein *lateinisches Quadrat der Ordnung* oder *Kantenlänge* n über S ist eine Abbildung $L\colon \{1,\dots,n\} \times \{1,\dots,n\} \longrightarrow S$ mit der Eigenschaft, dass

$$L(i,j) = L(i',j) \Longrightarrow i = i', \qquad L(i,j) = L(i,j') \Longrightarrow j = j'.$$

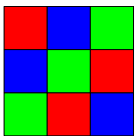
Diese Definition ist natürlich unabhängig von der Menge S (der Symbole). Ob die Elemente von S Zahlen, lateinische Buchstaben, griechische Buchstaben oder Farben sind, spielt keine Rolle. Durch

1	2	3
2	3	1
3	1	2

a	b	c
b	c	a
c	a	b

α	β	γ
β	γ	α
γ	α	β

und



sind vier äquivalente lateinische Quadrate der Ordnung 3 gegeben. *Äquivalenz* bedeutet hierbei, dass die Quadrate durch die Umbenennung $1 \leftrightarrow \text{a} \leftrightarrow \alpha \leftrightarrow$ rot usw. ineinander übergehen. Dagegen sind

1	2	3
2	3	1
3	1	2

1	2	3
3	1	2
2	3	1

zwei lateinische Quadrate der Ordnung 3, die nicht äquivalent sind. Sie sind aber *orthogonal*, d. h. alle $3^2 = 9$ geordneten Paare

$(1,1)$	$(2,2)$	$(3,3)$
$(2,3)$	$(3,1)$	$(1,2)$
$(3,2)$	$(1,3)$	$(2,1)$

die man erhält, indem man die Einträge der beiden Quadrate nebeneinander in ein neues Quadrat schreibt, sind paarweise verschieden. Genauer definieren wir:

- Sei S eine n-elementige Menge. Zwei lateinische Quadrate L_1, L_2 der Ordnung n über S heißen *orthogonal*, wenn es zu jedem Paar $(a,b) \in S \times S$ ein Paar (i,j) mit $(L_1(i,j), L_2(i,j)) = (a,b)$ gibt. Denkt man sich also zwei lateinische Quadrate „übereinandergelegt“, so gibt es für jedes Paar $(a,b) \in S \times S$ eine Position (i,j), in der a und b in dieser Reihenfolge übereinanderliegen.

Statt von einem Paar orthogonaler lateinischer Quadrate spricht man auch von einem *lateinisch-griechischen* oder einem *Euler-Quadrat*. So ist z. B.

(a,α)	(b,β)	(c,γ)	(d,δ)	(e,ϵ)
(b,ϵ)	(c,α)	(d,β)	(e,γ)	(a,δ)
(c,δ)	(d,ϵ)	(e,α)	(a,β)	(b,γ)
(d,γ)	(e,δ)	(a,ϵ)	(b,α)	(c,β)
(e,β)	(a,γ)	(b,δ)	(c,ϵ)	(d,α)

ein Euler-Quadrat der Ordnung 5, da

a	b	c	d	e
b	c	d	e	a
c	d	e	a	b
d	e	a	b	c
e	a	b	c	d

α	β	γ	δ	ϵ
ϵ	α	β	γ	δ
δ	ϵ	α	β	γ
γ	δ	ϵ	α	β
β	γ	δ	ϵ	α

jeweils lateinische Quadrate der Ordnung 5 sind. Insgesamt gibt es sogar vier paarweise orthogonale lateinische Quadrate der Ordnung 5. Die beiden anderen (wir nehmen jetzt die Ziffern $1, \ldots, 5$ als Symbolmenge) sind

1	2	3	4	5
3	4	5	1	2
5	1	2	3	4
2	3	4	5	1
4	5	1	2	3

1	2	3	4	5
4	5	1	2	3
2	3	4	5	1
5	1	2	3	4
3	4	5	1	2

Es gibt kein Euler-Quadrat der Ordnung 6 und daher ist das Problem der 36 Offiziere, das angeblich (?) Leonhard Euler 1779 von der Zarin Katharina der Großen erhalten hat, *nicht* lösbar. Es lautet:

- Beim Divisionsball ordnet jedes der sechs anwesenden Regimenter für jeden der sechs Dienstgrade je einen Offizier für eine besondere Aufgabe ab. Die sechsunddreißig Offiziere sollen zur Feier des Tages so im Quadrat aufgestellt werden, dass in jeder Zeile und jeder Spalte genau ein Offizier jeden Regiments und jeden Dienstgrades steht.

Dass eine Lösung des Problems der 36 Offiziere nicht möglich ist, ist erst 1901 von Gaston Tarry bewiesen worden.

Über die Maximalzahl paarweise orthogonaler lateinischer Quadrate der Ordnung n beweisen wir die folgende Aussage.

Satz *Sei $n \in \mathbb{N}$ und $N(n)$ die Maximalzahl paarweise orthogonaler lateinischer Quadrate der Ordnung n. Dann gilt:*

1. *Es ist $N(n) \leq n - 1$.*
2. *Ist $n = p^m$ eine Primzahlpotenz, also $p \in \mathbb{N}$ eine Primzahl und $m \in \mathbb{N}$, so ist $N(n) = n - 1$.*
3. *Ist $n = n_1 n_2$ mit $n_1, n_2 \in \mathbb{N}$, so ist $N(n_1 n_2) \geq \min(N(n_1), N(n_2))$.*
4. *Ist $n = p_1^{m_1} \cdots p_k^{m_k}$ mit Primzahlen $p_1, \ldots, p_k$ und $m_1, \ldots, m_k \in \mathbb{N}$, so ist*

$$N(n) \geq \min_{i=1,\ldots,k} (p_i^{m_i} - 1).$$

Insbesondere ist $N(n) \geq 2$ für alle $n \not\equiv 2 \pmod 4$.

Beweis: Zum Beweis der ersten Aussage nehmen wir an, $L_1, \ldots, L_k$ seien paarweise orthogonale lateinische Quadrate der Ordnung n, wobei wir $S = \{1, \ldots, n\}$ annehmen können. Die Orthogonalität bleibt erhalten, wenn wir die Spalten in jedem Quadrat L_i so permutieren, dass in der ersten Zeile die Elemente in der Reihenfolge $1, 2, \ldots, n$ erscheinen, dass also

$$L_i(1,1) = 1,\ L_i(1,2) = 2,\ \ldots, L_i(1,n) = n, \qquad i = 1, \ldots, k.$$

Nun betrachten wir die Elemente an der Position $(2,1)$. Es ist $L_i(2,1) \neq 1$, $i = 1, \ldots, m$, da die 1 bei jedem der L_i schon in der Position $(1,1)$ vorkommt. Weiter ist $L_i(2,1) \neq L_j(2,1)$ für $i \neq j$. Denn wäre $q := L_i(2,1) = L_j(2,1)$, so würde beim Übereinanderlegen der lateinischen Quadrate L_i und L_j das Paar (q,q) sowohl in der Position $(1,q)$ als auch der Position $(2,1)$ erscheinen, was wegen der Orthogonalität von L_i und L_j nicht möglich ist. Also sind $L_i(2,1) \in \{2, \ldots, n\}$, $i = 1, \ldots, m$, paarweise verschieden, damit $m \leq n-1$ und folglich $N(n) \leq n-1$.

Zum Beweis der zweiten Aussage benutzen wir eine aus der Algebra bekannte Aussage, von der bei K. Jacobs, D. Jungnickel (2004, S. 69) gesagt wird, dass sie zur Allgemeinbildung gehört.

- Ist $n = p^m$ eine Primzahlpotenz, so gibt es (bis auf Isomorphie genau) einen Körper $\mathbb{F}_n$ mit n Elementen. D. h. $\mathbb{F}_n = \{a_1 = 0, a_2, \ldots, a_n\}$ ist eine n-elementige Menge mit zwei Verknüpfungen $+\colon \mathbb{F}_n \times \mathbb{F}_n \longrightarrow \mathbb{F}_n$ und $\cdot\colon \mathbb{F}_n \times \mathbb{F}_n \longrightarrow \mathbb{F}_n$ derart, dass $(\mathbb{F}_n, +)$ und $(\mathbb{F}_n \setminus \{0\}, \cdot)$ abelsche Gruppen sind und das Distributivgesetz gilt.

Wir definieren $L_k\colon \{1, \ldots, n\} \times \{1, \ldots, n\} \longrightarrow \mathbb{F}_n$, $k = 2, \ldots, n$, durch

$$L_k(i,j) = a_k \cdot a_i + a_j \qquad (i,j = 1, \ldots, n).$$

Hierdurch sind $n-1$ paarweise orthogonale lateinische Quadrate über $\mathbb{F}_n$ definiert. Denn zunächst sind die L_k, $k = 2, \ldots, n$, lateinische Quadrate. Ist nämlich $L_k(i,j) = L_k(i',j)$ bzw. $a_k \cdot a_i + a_j = a_k \cdot a_{i'} + a_j$, so ist $a_k \cdot a_i = a_k \cdot a_{i'}$ und wegen $a_k \neq 0$ damit $a_i = a_{i'}$ bzw. $i = i'$. Entsprechend folgt aus $L_k(i,j) = L_k(i,j')$ bzw. $a_k \cdot a_i + a_j = a_k \cdot a_i + a_{j'}$, dass $a_j = a_{j'}$ bzw. $j = j'$. Nun zeigen wir, dass die L_k, $k = 2, \ldots, n$, paarweise orthogonal sind. Seien L_k, L_l mit $k, l \in \{2, \ldots, n\}$ und $k \neq l$ zwei lateinische Quadrate. Wir geben uns $(a_r, a_s) \in \mathbb{F}_n \times \mathbb{F}_n$ vor und haben zu zeigen, dass es genau ein Paar $(i,j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ mit $(L_k(i,j), L_l(i,j)) = (a_r, a_s)$ gibt. Dies ergibt für (i,j) die Gleichungen

$$a_k \cdot a_i + a_j = a_r, \qquad a_l \cdot a_i + a_j = a_s.$$

Wegen $(a_k - a_l) \neq 0$ erhält man i aus $(a_k - a_l) \cdot a_i = a_r - a_s$, anschließend j aus $a_j = a_r - a_k \cdot a_i$. Damit ist nachgewiesen, dass es für eine Primzahlpotenz $n = p^m$ mindestens $n-1$ paarweise orthogonale lateinische Quadrate der Ordnung n gibt. Unter Berücksichtigung des ersten Teiles des Satzes folgt $N(n) = n-1$, wenn n eine Primzahlpotenz ist. Damit ist auch der zweite Teil des Satzes bewiesen.

Sei $n = n_1 n_2$ mit $n_1, n_2 \in \mathbb{N}$. Sei $k := \min(N(n_1), N(n_2))$. Dann gibt es paarweise orthogonale lateinische Quadrate $L_1, \ldots, L_k$ der Ordnung n_1 auf $S_1 := \{1, \ldots, n_1\}$

sowie paarweise orthogonale lateinische Quadrate $L'_1, \ldots, L'_k$ der Ordnung n_2 auf $S_2 := \{1, \ldots, n_2\}$. Wir definieren die $n_1 n_2$-elementige Menge $S := S_1 \times S_2$ sowie $L^*_h \colon S \times S \longrightarrow S$, $h = 1, \ldots, k$, durch

$$L^*_h((i, i'), (j, j')) := (L_h(i, j), L'_h(i', j')).$$

Wir wollen uns überlegen, dass $L^*_1, \ldots, L^*_k$ paarweise orthogonale lateinische Quadrate der Ordnung $n_1 n_2$ über S sind. Zunächst sind hierdurch wirklich lateinische Quadrate über S definiert. Denn

$$L^*_h((i, i'), (j, j')) = L^*_h((i_1, i'_1), (j, j'))$$

bedeutet nach Definition von L^*_h, dass

$$(L_h(i, j), L'_h(i', j')) = (L_h(i_1, j), L'_h(i'_1, j'))$$

bzw.

$$L_h(i, j) = L_h(i_1, j), \quad L'_h(i', j') = L'_h(i'_1, j').$$

Da aber L_h und L'_h lateinische Quadrate sind, ist $(i, i') = (i_1, i'_1)$. Da entsprechend aus $L^*_h((i, i'), (j, j')) = L^*_h((i, i'), (j_1, j'_1))$ folgt, dass $(j, j') = (j_1, j'_1)$, sind L^*_h, $h = 1, \ldots, k$, lateinische Quadrate über S. Wir zeigen nun, dass diese paarweise orthogonal sind. Hierzu seien $((r, r'), (s, s')) \in S \times S$ und $h \neq l$ mit $h, l \in \{1, \ldots, k\}$ gegeben. Da L_h und L_l orthogonale lateinische Quadrate sind, gibt es ein Paar $(i, j) \in S_1 \times S_1$ mit $(L_h(i, j), L_l(i, j)) = (r, s)$. Entsprechend existiert ein Paar $(i', j') \in S_2 \times S_2$ mit $(L'_h(i', j'), L'_l(i', j')) = (r', s')$. Dann ist aber

$$\begin{aligned}(L^*_h((i, i'), (j, j')), L^*_l((i, i'), (j, j')) &= ((L_h(i, j), L'_h(i', j')), (L_l(i, j), L'_l(i', j'))) \\ &= ((r, r'), (s, s')).\end{aligned}$$

Damit haben wir die Existenz von $k = \min(N(n_1), N(n_2))$ paarweise orthogonalen lateinischen Quadraten der Ordnung $n_1 n_2$ bewiesen. Damit ist auch der dritte Teil des Satzes bewiesen.

Sei[89] $n = p_1^{m_1} \cdots p_k^{m_k}$ mit Primzahlen $p_1, \ldots, p_k$ und $m_1, \ldots, m_k \in \mathbb{N}$. Mit Hilfe des dritten und des zweiten Teils des Satzes erhalten wir

$$\begin{aligned}N(n) &\geq \min(N(p_1^{m_1}), N(p_2^{m_2} \cdots p_k^{m_k})) \\ &= \min(p_1^{m_1} - 1, N(p_2^{m_2} \cdots p_k^{m_k})). \\ &\geq \min_{i=1,\ldots,k}(p_i^{m_i} - 1).\end{aligned}$$

Ist $n \not\equiv 2 \pmod 4$, ist also n nicht von der Form $n = 2 + 4m$ mit $m \in \mathbb{N}$, so ist n ungerade oder n besitzt 4 als Teiler. In beiden Fällen ist $N(n) \geq 2$, es existiert also mindestens ein Paar orthogonaler lateinischer Quadrate der Ordnung n. □

Bemerkung: Wegen des letzten Teiles des Satzes ist nur für $n = 6, 10, 14, \ldots$ ungeklärt, ob $N(n) \geq 2$ ist bzw. mindestens ein Paar orthogonaler lateinischer Quadrate der

[89] Jede natürliche Zahl n besitzt eine bis auf die Reihenfolge der Faktoren eindeutige *Primfaktorzerlegung*, siehe z. B. `http://de.wikipedia.org/wiki/Primfaktorzerlegung`.

Ordnung n existiert. In der Tat vermutete Euler, dass für $n \equiv 2 (\mathrm{mod}\, 4)$ kein Paar orthogonaler lateinischer Quadrate der Ordnung n existiert. Diese Vermutung ist nur für $n = 6$ richtig, d. h. das Problem der 36 Offiziere ist nicht lösbar (TARRY um 1900). Für alle anderen Fälle ist die Vermutung falsch, wie 1960 von R. C. BOSE, S. S. SHRIKHANDE, E. T. PARKER gezeigt wurde, siehe `http://mathworld.wolfram.com/EulersGraeco-RomanSquaresConjecture.html`. Werte von $N(n)$ sind bisher nur für Primzahlpotenzen n bekannt. □

Bemerkung: Lateinische und lateinisch-griechische (bzw. griechisch-lateinische) Quadrate spielen in der *statistischen Versuchsplanung* (design of experiments, DOE) eine Rolle. Bei einem (mehrfaktoriellen) Versuchsplan hat man eine bestimmte Anzahl von *Faktoren*, die unabhängigen Variablen (an denen man sozusagen drehen kann). Diese liegen in verschiedenen *Stufungen* vor, die sich i. Allg. in Abhängigkeit vom jeweiligen Faktor unterscheiden können. Interessiert ist man an einer abhängigen Variablen, welche in Abhängigkeit von den Faktoren gemessen bzw. bestimmt wird. Hierbei kann es sich z. B. um die Konzentrationsfähigkeit von Versuchspersonen in Abhängigkeit ihrer Ernährung oder ihrer körperlichen Aktivitäten handeln oder die Arbeitsleistung in Abhängigkeit von Umweltvariaben (z. B. Lärm, Temperatur, Beleuchtung und Luftfeuchtigkeit). Man spricht von einem *vollständigen Versuchsplan*, wenn alle Faktorkombinationen (in allen Stufungen) berücksichtigt werden. Hat man z. B. vier Faktoren, die jeweils in zwei Stufungen vorkommen, so hat man insgesamt $2^4 = 16$ Kombinationen. Wenn bei jeder Kombination n Versuchspersonen beteiligt sind, so benötigt man also insgesamt $16 \cdot n$ Personen. Das kann für große n ein Problem werden, so dass man eventuell aus Kostengründen auf einen *unvollständigen Versuchsplan* ausweicht, bei dem nicht alle Stufen jedes Faktors miteinander kombiniert werden. Spezielle unvollständige Versuchspläne sind die *quadratischen Versuchspläne* mit drei oder vier Faktoren. Hier haben alle Faktoren die gleiche Anzahl von Faktorstufen und es werden nicht alle Stufen miteinander kombiniert. Bei einem Versuchsplan nach der Bauart lateinischer Quadrate hat man z. B. drei Faktoren A, B, C in jeweils drei Stufungen. Zwei mögliche Versuchspläne, die auf lateinischen Quadraten basieren, sind in Tabelle 6 angegeben. Jede Faktorstufenkombination $A_i \times B_j$ mit $1 \leq i, j \leq 3$ wird genau einmal

	B_1	B_2	B_3
A_1	C_1	C_2	C_3
A_2	C_2	C_3	C_1
A_3	C_3	C_1	C_2

	B_1	B_2	B_3
A_1	C_1	C_2	C_3
A_2	C_3	C_1	C_2
A_3	C_2	C_3	C_1

Tabelle 6: Zwei auf lateinischen Quadraten basierende Versuchspläne

mit einer Faktorstufe C_k mit $1 \leq k \leq 3$ kombiniert. Da jede Stufe des Faktors C nur mit 3 der möglichen 9 Kombinationen von $A \times B$ auftritt, hat man gegenüber einem vollständigen Versuchsplan nur ein Drittel der benötigten Kombinationen, nämlich 9 statt $3^3 = 27$. Noch signifikanter wird die Verringerung der Kombinationsmöglichkeiten bei vier Faktoren A, B, C, D in jeweils drei Abstufungen und einem Versuchsplan nach

der Bauart eines griechisch-lateinischen Quadrats. Ein solcher sieht z. B. wie in Abbildung 7 aus. Bei einem vollständigen Versuchsplan mit vier Faktoren und jeweils drei

	B_1	B_2	B_3
A_1	C_1D_1	C_2D_2	C_3D_3
A_2	C_2D_3	C_3D_1	C_1D_2
A_3	C_3D_2	C_1D_3	C_2D_1

Tabelle 7: Auf griechisch-lateinischem Quadrat basierender Versuchsplan

Faktorstufen hat man insgesamt 81 Kombinationsmöglichkeiten, bei einem auf einem griechisch-lateinischen Quadrat basierenden Versuchsplan sind es dagegen nur 9. □

65.2 Die Vervollständigung lateinischer Quadrate: Der Satz von Evans-Smetaniuk

Angenommen, ein $n \times n$-Quadrat sei schon zum Teil mit Zahlen aus $\{1, \ldots, n\}$ besetzt. Unter welchen Bedingungen können die noch nicht besetzten Felder so gefüllt werden, dass man ein lateinisches Quadrat der Ordnung n erhält? Hierbei muss natürlich vorausgesetzt werden, dass zu Beginn jede der schon festgelegten Zahlen aus $\{1, \ldots, n\}$ höchstens einmal in jeder Zeile und Spalte vorkommt. Ein so zum Teil mit Zahlen aus $\{1, \ldots, n\}$ gefülltes $n \times n$-Quadrat nennen wir ein *partielles lateinisches Quadrat* der Ordnung n. Unter welchen Voraussetzungen kann ein partielles lateinisches Quadrat zu einem lateinischen Quadrat vervollständigt werden? Klar ist, dass dies nicht immer möglich ist. Denn z. B. können die partiellen lateinischen Quadrate

1	2	$\cdots$	$n-1$	
				n

1				
	1			
		$\ddots$		
			1	
				2

nicht zu einem lateinischen Quadrat vervollständigt werden. Wir folgen der Darstellung von M. Aigner, G. M. Ziegler (2002, 203 ff.), verweisen aber insbesondere auch auf `http://www.nerdburrow.com/smetaniuklatin/` und J. H. van Lint, R. M. Wilson (1992). Ziel in diesem Abschnitt ist es, die folgende Aussage zu beweisen, deren Gültigkeit von T. Evans 1960 vermutet und von B. Smetaniuk 1981 bewiesen wurde. Genauere Literaturangaben findet man bei M. Aigner, G. M. Ziegler (2002, S. 210).

Satz (Evans-Smetaniuk) *Jedes partielle lateinische Quadrat der Ordnung n, in dem höchstens $n-1$ Felder gefüllt sind, kann zu einem lateinischen Quadrat derselben Ordnung vervollständigt werden.*

Bevor wir in den Beweis einsteigen, sind einige Vorbereitungen nötig. Zunächst notieren wir, dass wir ein lateinisches Quadrat der Ordnung n auch als eine $(3 \times n^2)$-Matrix auffassen können, die *Zeilenmatrix* oder der *line array* des lateinischen Quadrats. In jeder Spalte dieser Matrix stehen drei Komponenten, nämlich der Reihe nach ein Zeilenindex i, ein Spaltenindex j und das Element in der Position (i, j). Es folgt ein lateinisches Quadrat und eine zugehörige Zeilenmatrix:

$(*)$

3	4	2	1
4	3	1	2
1	2	4	3
2	1	3	4

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 & 4 & 3 & 1 & 2 & 1 & 2 & 4 & 3 & 2 & 1 & 3 & 4 \end{pmatrix}$$

Eine $(3 \times n^2)$-Matrix, deren Einträge die Zahlen von 1 bis n sind, ist die Zeilenmatrix eines lateinischen Quadrats genau dann, wenn in je zwei Zeilen der Zeilenmatrix alle n^2 geordneten Paare von Zahlen aus $\{1, \ldots, n\}$ (genau einmal) auftreten. Hieraus können wir zweierlei ableiten:

- In jeder der drei Zeilen der Zeilenmatrix eines lateinischen Quadrats können wir die Symbole, also die Zahlen $1, \ldots, n$, beliebig permutieren und erhalten wieder die Zeilenmatrix eines lateinischen Quadrat. Permutationen in der ersten Zeile entsprechen Permutationen der Zeilen, in der zweiten Permutationen der Spalten und in der dritten Permutationen der Elemente.

Geht man vom obigen Beispiel $(*)$ aus und vertauscht (in der Zeilenmatrix) in der ersten Zeile 1 und 2, in der zweiten Zeile 1 und 3 sowie 2 und 4 und in der dritten Zeile 1 und 4, so erhält man die folgende Zeilenmatrix samt zugehörigem lateinischen Quadrat:

$$\begin{pmatrix} 2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \\ 3 & 2 & 1 & 4 & 3 & 2 & 1 & 4 & 3 & 2 & 1 & 4 & 3 & 2 & 1 & 4 \\ 3 & 1 & 2 & 4 & 1 & 3 & 4 & 2 & 4 & 2 & 1 & 3 & 2 & 4 & 3 & 1 \end{pmatrix}$$

4	3	1	2
2	1	3	4
1	2	4	3
3	4	2	1

- Vertauscht man zwei Zeilen in der Zeilenmatrix eines lateinischen Quadrats, so erhält man wieder die Zeilenmatrix eines lateinischen Quadrats.

Vertauscht man in der Zeilenmatrix in $(*)$ z. B. die erste und die dritte Zeile, so erhält man eine neue Zeilenmatrix und ein zugehöriges lateinisches Quadrat:

$$\begin{pmatrix} 3 & 4 & 2 & 1 & 4 & 3 & 1 & 2 & 1 & 2 & 4 & 3 & 2 & 1 & 3 & 4 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \end{pmatrix}$$

3	4	2	1
4	3	1	2
1	2	4	3
2	1	3	4

Lateinische Quadrate, die durch Permutationen dieser Art auseinander hervorgehen, heißen *konjugierte* lateinische Quadrate. Einem partiellen lateinischen Quadrat entspricht eine partielle Zeilenmatrix (jedes geordnete Paar von Zahlen aus $\{1, \ldots, n\}$

tritt in einem Paar von Zeilen höchstens einmal auf). Es folgt ein Beispiel eines partiellen lateinischen Quadrats und einer zugehörigen partiellen Zeilenmatrix:

3	4		
		1	2
	1		

$$\begin{pmatrix} 1 & 1 & 2 & 2 & 4 \\ 1 & 2 & 3 & 4 & 2 \\ 3 & 4 & 1 & 2 & 1 \end{pmatrix}$$

Permutationen in den Zeilen und Permutationen der Zeilen einer partiellen Zeilenmatrix liefern wieder partielle Zeilenmatrizen, die partiellen lateinischen Quadraten entsprechen. Daher können wir von *konjugierten* partiellen lateinischen Quadraten sprechen. Wichtig ist nun die folgende Erkenntnis:

- Ein partielles lateinisches Quadrat kann genau dann vervollständigt werden, wenn irgendein hierzu konjugiertes partielles lateinisches Quadrat vervollständigt werden kann.

Denn hierzu brauchen nur die Permutationen in den bzw. unter den Zeilen der Zeilenmatrix rückgängig gemacht zu werden. Vertauschen wir z. B. in der obigen partiellen Zeilenmatrix in der ersten Zeile die 2 und die 4 und anschließend die erste und die dritte Zeile, so erhalten wir die folgende partielle Zeilenmatrix samt zugehörigem konjugierten partiellen lateinischen Quadrat:

$$\begin{pmatrix} 3 & 4 & 1 & 2 & 1 \\ 1 & 2 & 3 & 4 & 2 \\ 1 & 1 & 4 & 4 & 2 \end{pmatrix}$$

	2	4	
			4
1			
	1		

Dieses konjugierte partielle lateinische Quadrat kann leicht vervollständigt werden. Eine mögliche Vervollständigung samt zugehöriger Zeilenmatrix geben wir nun an:

3	2	4	1
2	3	1	4
1	4	2	3
4	1	3	2

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 & 2 & 3 & 1 & 4 & 1 & 4 & 2 & 3 & 4 & 1 & 3 & 2 \end{pmatrix}$$

Vertauscht man in dieser Zeilenmatrix die erste und die dritte Zeile und anschließend in der ersten Zeile die 2 und die 4, so erhält man die folgende Zeilenmatrix samt zugehörigem lateinischen Quadrat:

$$\begin{pmatrix} 3 & 4 & 2 & 1 & 4 & 3 & 1 & 2 & 1 & 2 & 4 & 3 & 2 & 1 & 3 & 4 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \end{pmatrix}$$

3	4	2	1
4	3	1	2
1	2	4	3
2	1	3	4

Wir haben damit eine Vervollständigung des am Anfang gegebenen partiellen lateinischen Quadrats erhalten. Die ergänzten Zahlen haben wir rot kenntlich gemacht.

Ein partielles lateinisches Quadrat, bei dem die die ersten r Zeilen gefüllt und alle übrigen Felder leer sind, heißt ein $(r \times n)$-*lateinisches Rechteck*.

Lemma 1 *Ein $(r \times n)$-lateinisches Rechteck mit $r < n$ kann zu einem $((r+1) \times n)$-lateinischen Rechteck erweitert werden. Insbesondere kann ein lateinisches Rechteck zu einem lateinischen Quadrat vervollständigt werden.*

Beweis: Der Beweis verwendet den Heiratssatz, siehe Abschnitt 29. Hierzu sei $U := \{1, \ldots, n\}$ die Menge der Damen und $W := \{1, \ldots, n\}$ die Menge der Herren. Wir definieren A_j als die Menge der $i \in \{1, \ldots, n\}$, die *nicht* in der j-ten Spalte des gegebenen $(r \times n)$-lateinischen Rechtecks auftreten, $j = 1, \ldots, n$. Wir sagen, die Dame $j \in U$ sei mit dem Herrn $i \in W$ befreundet, wenn $i \in A_j$ bzw. i in der j-ten Spalte des lateinischen Rechtecks nicht enthalten ist. Wir wollen jetzt die sogenannte Partybedingung des Heiratssatzes nachweisen. Diese besagt, dass je k Damen mit mindestens k Herren befreundet sind bzw. in k Spalten des lateinischen Rechtecks mindestens k der Zahlen aus $\{1, \ldots, n\}$ nicht enthalten sind, $k = 1, \ldots, n$. Die Anzahl der Elemente in A_j ist $n-r$, da in der j-ten Spalte des $(r \times n)$-lateinischen Quadrats genau r (paarweise verschiedene) Zahlen aus $\{1, \ldots, n\}$ stehen. Daher ist jede Dame mit genau $n - r$ Herren befreundet. Andererseits ist jedes $i \in \{1, \ldots, n\}$ genau r Mal im lateinischen Rechteck vertreten, da es in jeder der r Zeilen genau einmal auftritt. Diese r-vielen Vorkommen von i müssen über r verschiedene Spalten verteilt sein. Also bleiben genau $n - r$ Spalten übrig, in denen i nicht vertreten ist. Dies bedeutet, dass jeder Herr mit genau $n-r$ Damen befreundet ist. Nun betrachten wir eine Gruppe von k Damen. Jede dieser Damen ist mit $n - r$ Herren befreundet. Daher gibt es genau $k(n - r)$ Freundschaftsbeziehungen[90] zwischen den k Damen und den Herren. Wären die k Damen nur mit $l < k$ Herren befreundet, so gäbe es nur $l(n - r)$ Freundschaftsbeziehungen zwischen den k Damen und den Herren, da jeder Herr mit genau $n - r$ Damen befreundet ist. Damit haben wir einen Widerspruch erreicht und die Partybedingung ist nachgewiesen. Der Heiratssatz liefert, dass alle n Damen in einer Massenhochzeit jeweils mit einem befreundeten Herren verheiratet werden können, und zwar so, dass weder eine Dame noch ein Herr Bigamisten werden. Hierdurch hat man nun die $(r + 1)$-te Zeile des gesuchten $((r+1) \times n)$-lateinischen Rechtecks bestimmt. Genauer steht in der j-ten Komponente dieser Zeile die Nummer des Herrn, mit dem die j-te Dame verheiratet wird, $j = 1, \ldots, n$. Damit ist Lemma 1 bewiesen. □

Beispiel: Gegeben sei das (3×5)-lateinische Rechteck

1	2	3	4	5
2	4	1	5	3
3	5	2	1	4

Wir wollen den Beweis von Lemma 1 anhand dieses Beispiels verfolgen Wegen

$$A_1 = \{4, 5\}, \quad A_2 = \{1, 3\}, \quad A_3 = \{4, 5\}, \quad A_4 = \{2, 3\}, \quad A_5 = \{1, 2\}$$

[90] Dies entspricht Kanten in dem bipartiten Graphen, dessen Eckenmengen aus den Damen einerseits und den Herren andererseits gegeben sind.

können wir diesem lateinischen Rechteck den in Abbildung 91 links angegebenen bipartiten Graphen zuordnen: Die Damen sind blau durch •, die Herren rot durch •

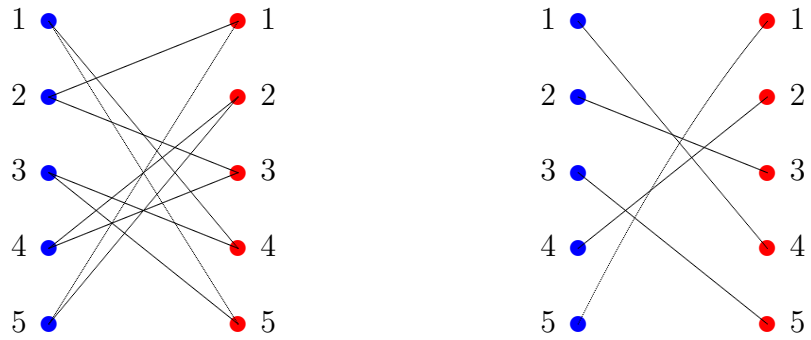


Abbildung 91: Erläuterung zum Beweis von Lemma 1

gekennzeichnet. Eine Kante verbindet eine Dame j und einen Herrn i, wenn diese miteinander befreundet sind bzw. $i \in A_j$ gilt. In Abbildung 91 rechts geben wir einen Heiratsplan an. Hierdurch erhalten wir das (4×5)-lateinische Rechteck

1	2	3	4	5
2	4	1	5	3
3	5	2	1	4
4	3	5	2	1

Dieses kann in einem weiteren Schritt zu einem lateinischen Quadrat vervollständigt werden. □

Beim Beweis des folgenden Lemmas benutzen wir neben der Darstellung von Aigner-Ziegler auch J. H. van Lint, R. M. Wilson (1992, S. 164 ff.)

Lemma 2 *Ein partielles lateinisches Quadrat der Ordnung n mit höchstens $n-1$ gefüllten Feldern und höchstens $\frac{n}{2}$ verschiedenen Elementen kann zu einem lateinischen Quadrat der Ordnung n vervollständigt werden.*

Beweis: Indem man notfalls zu einem konjugierten partiellen lateinischen Quadrat übergeht, können wir die Bedingung „höchstens $\frac{n}{2}$ verschiedene Elemente“ durch die Bedingung ersetzen, dass die Einträge in höchstens $\frac{n}{2}$ Zeilen auftreten (z. B. durch Vertauschen der ersten und der dritten Zeile der partiellen Zeilenmatrix).

Ist z. B. das partielle lateinische Quadrat der Ordnung $n = 6$ mit zugehöriger partieller Zeilenmatrix

	1				
		1			
		3			
			2		
					3

$$\begin{pmatrix} 1 & 3 & 4 & 5 & 6 \\ 2 & 3 & 3 & 4 & 6 \\ 1 & 1 & 3 & 2 & 3 \end{pmatrix}$$

mit $3 = \frac{n}{2}$ verschiedenen Elementen gegeben, so erhalten wir durch Vertauschen der ersten und der dritten Zeile in der partiellen Zeilenmatrix ein konjugiertes partielles lateinisches Quadrat:

$$\begin{pmatrix} 1 & 1 & 3 & 2 & 3 \\ 2 & 3 & 3 & 4 & 6 \\ 1 & 3 & 4 & 5 & 6 \end{pmatrix}$$

	1	3			
			5		
		4			6

Die Einträge treten hier nur in $3 = \frac{n}{2}$ verschiedenen Zeilen auf.

Durch Zeilenvertauschungen können wir weiter annehmen, dass nur die ersten $r \leq \frac{n}{2}$ Zeilen gefüllte Felder haben, und dass

$$f_1 \geq f_2 \geq \cdots \geq f_r > 0$$

gilt, wobei f_i die Anzahl der gefüllten Felder in Zeile i ist. Im obigen Beispiel erreichen wir dies z. B. durch Vertauschen der zweiten und dritten Zeile:

$$\begin{pmatrix} 1 & 1 & 2 & 3 & 2 \\ 2 & 3 & 3 & 4 & 6 \\ 1 & 3 & 4 & 5 & 6 \end{pmatrix}$$

	1	3			
		4			6
			5		

Da die Anzahl gefüllter Felder höchstens $n-1$ ist, ist $\sum_{i=1}^{r} f_i \leq n-1$.

Die Zeilen $1, \ldots, r$ werden sukzessive gefüllt bis wir ein $(r \times n)$-lateinisches Rechteck erhalten. Dieses kann wegen Lemma 1 zu einem lateinischen Quadrat der Ordnung n erweitert werden.

Wir nehmen an, die Zeilen $1, \ldots, l-1$ seien bereits gefüllt. In Zeile l gibt es f_l gefüllte Felder. Die Vervollständigung der l-ten Zeile erfolgt durch eine erneute Anwendung des Heiratssatzes. Hierzu sei die Menge U der Damen die Menge derjenigen Elemente $j \in \{1, \ldots, n\}$, für die das Feld (l, j) nicht gefüllt ist. Die Menge W der Herren bestehe aus der Menge der $i \in \{1, \ldots, n\}$, die *nicht* in Zeile l auftreten. Dann ist $|U| = |W| = n - f_l$. Für $j \in U$ sei weiter A_j die Menge der $i \in W$, die nicht in Spalte j auftreten (weder ober- noch unterhalb der Zeile l). Wir sagen, die Dame $j \in U$ und der Herr $i \in W$ seien miteinander befreundet, wenn $i \in A_j$ bzw. i weder in der l-ten Zeile noch in der j-ten Spalte vorkommt. Die l-te Zeile kann aufgefüllt werden, indem man einen zulässigen Heiratsplan bestimmt. Ein solcher existiert wegen des Heiratssatzes genau dann, wenn die Partybedingung erfüllt ist.

Im obigen Beispiel ist $l = 1$, $U = \{1, 4, 5, 6\}$ und $W = \{2, 4, 5, 6\}$. Welche der Damen mit welchem Herrn befreundet ist, wird durch eine Kante in einem bipartiten Graphen mit der Eckenmenge $U \cup W$ verdeutlicht, siehe Abbildung 92 links. Rechts haben wir in dieser Abbildung einen zulässigen Heiratsplan angegeben (was wegen der

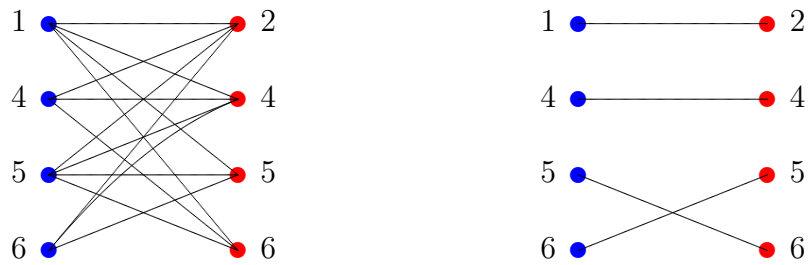


Abbildung 92: Erläuterung zum Beweis von Lemma 2

vielen Freundschaften nicht schwierig ist). Das neue partielle lateinische Quadrat mit aufgefüllter l-ter Zeile ist also

2	1	3	4	6	5
		4			6
			5		

Dieses Verfahren kann offenbar fortgesetzt werden.

Bevor wir die Partybedingung nachweisen, zeigen wir, dass

$$(*) \qquad n - f_l - l + 1 > l - 1 + f_{l+1} + \cdots + f_r.$$

Zum Nachweis werden drei Fälle unterschieden. Ist $l = 1$, so ist $(*)$ wegen $\sum_{i=1}^{r} f_i < n$ richtig. Daher kann jetzt $l \geq 2$ vorausgesetzt werden. Ist $f_{l-1} \geq 2$, so ist

$$f_1 + \cdots + f_{l-1} \geq (l-1) f_{l-1} \geq 2(l-1)$$

und daher

$$\begin{aligned} l - 1 + f_{l+1} + \cdots + f_r &= 2(l-1) + f_{l+1} + \cdots + f_r - l + 1 \\ &\leq \sum_{i=1}^{r} f_i - f_l - l + 1 \\ &< n - f_l - l + 1. \end{aligned}$$

Also gilt $(*)$ in diesem Falle. Ist schließlich $f_{l-1} = 1$, so folgt $f_l = \cdots = f_r = 1$. Daher ist $(*)$ gleichwertig mit $n - l > r - 1$ bzw. $n > r + l - 1$, was wegen $l \leq r \leq \frac{n}{2}$ richtig ist. Damit ist $(*)$ nachgewiesen.

Nun kommen wir zum Nachweis der Partybedingung. Mit einem $k \in \{1, \ldots, n - f_l\}$ sei $A \subset U$ mit $|A| = k$ eine Menge von k Damen und $B \subset W$ die Menge der Herren, die mit einer der Damen aus A befreundet sind. Wir haben zu zeigen, dass $|B| \geq k$. Sei c die Anzahl der Felder in den k zu A gehörenden Spalten, die ein Element aus W

enthalten. Es gibt höchstens $(l-1)k$ solche Felder oberhalb der Zeile l, kein solches Feld in Zeile l und höchstens $f_{l+1}+\cdots+f_r$ unterhalb der Zeile l. Daher ist

$$c \le (l-1)k + f_{l+1} + \cdots + f_r.$$

Andererseits ist $i \in W \setminus B$ ein Herr, der mit keiner der Damen aus A befreundet ist, folglich tritt i in jeder der Spalten j mit $j \in A$ auf. Daher ist $c \ge k(|W| - |B|)$. Wegen $|W| = n - f_l$ folgt

$$|B| \ge |W| - \frac{c}{k} \ge n - f_l - (l-1) - \frac{f_{l+1}+\cdots+f_r}{k}.$$

Es ist $|B| \ge k$, falls

$$n - f_l - (l-1) - \frac{f_{l+1}+\cdots+f_r}{k} > k-1$$

bzw.

$$(**) \qquad k(n - f_l - l + 2 - k) > f_{l+1} + \cdots + f_r.$$

Für $k = 1$ und $k = n - f_l - l + 1$ ist

$$k(n - f_l - l + 2 - k) = n - f_l - l + 1 > l - 1 + f_{l+1} + \cdots f_r \ge f_{l+1} + \cdots + f_r$$

wegen $(*)$, für diese k gilt also $(**)$ und dann auch $|B| \ge k$. Die linke Seite von $(**)$ ist ein konkaves quadratisches Polynom in k, daher gilt $(**)$ bzw. $|B| \ge k$ nicht nur für $k = 1$ und $k = n - f_l - l + 1$, sondern auch für alle Werte von k dazwischen. Sei daher jetzt $k > n - f_l - l + 1$ und wegen $(*)$ auch $k > l - 1 + f_{l+1} + \cdots + f_r$. Ein $i \in W$ kann nicht in Zeile l, und daher in höchstens $r-1$ Zeilen und in höchstens ebenso vielen Spalten auftreten. Erst Recht kann ein $i \in W$ in höchstens $l - 1 + f_{l+1} + \cdots + f_r \ge r - 1$ Spalten auftreten. Wegen $k > l - 1 + f_{l+1} + \cdots + f_r$ fehlt $i \in W$ in wenigstens einer der k Spalten $j \in A$. Dies bedeutet, dass jeder Herr $i \in W$ mit wenigstens einer Dame $j \in A$ befreundet ist. Folglich ist $|B| = |W| = n - f_l \ge k = |A|$, die Partybedingung ist erfüllt. Das Lemma ist bewiesen. □

Beispiel: Wir geben ein partielles lateinisches Quadrat (siehe oben) sowie eine Vervollständigung zu einem lateinischen Rechteck (Verfahren von Lemma 2) und einem lateinischen Quadrat (Verfahren von Lemma 1) an:

	1	3			
		4			6
			5		

2	1	3	4	6	5
		4			6
			5		

2	1	3	4	6	5
1	2	4	3	5	6
3	4	6	5	2	1

2	1	3	4	6	5
1	2	4	3	5	6
3	4	6	5	2	1
6	3	5	2	1	4
4	5	1	6	3	2
5	6	2	1	4	3

In dem den Beweis von Lemma 2 begleitenden Beispiel waren wir von einem partiellen lateinischen Quadrat der Ordnung 6 mit drei verschiedenen Einträgen ausgegangen.

Durch eine Vertauschung der Zeilen und Elementen erreichten wir, dass nur in drei Zeilen Einträge stehen. Durch anschließende Zeilenvertauschungen sicherten wir, dass nur die ersten Zeilen besetzt sind und die Füllung der Zeilen von Zeile zu Zeile abnimmt (oder zumindest nicht zunimmt). Diese Vertauschungen müssen rückgängig gemacht werden, um eine Vervollständigung des ursprünglich gegebenen partiellen lateinischen Quadrats zu erhalten. In unserem Falle hat man in der zu dem gerade berechneten lateinischen Quadrat gehörenden Zeilenmatrix in der ersten Zeile 2 und 3 und anschließend die erste und die dritte Zeile zu vertauschen. Wir geben das ursprüngliche partielle lateinische Quadrat sowie die gewonnene Vervollständigung an:

	1				
		1			
		3			
			2		
					3

3	1	5	6	4	2
1	3	6	4	2	5
2	4	1	3	5	6
5	2	3	1	6	4
6	5	4	2	3	1
4	6	2	5	1	3

□

Beweis des Satzes von Evans-Smetaniuk Der Beweis erfolgt durch Induktion nach n. Für $n \leq 2$ ist die Aussage trivialerweise richtig. Wir betrachten jetzt ein partielles lateinisches Quadrat der Ordnung $n \geq 3$ mit höchstens $n-1$ gefüllten Feldern. Diese sind in höchstens $r \leq n-1$ verschiedenen Zeilen $s_1, \ldots, s_r$ mit $f_1, \ldots, f_r$ gefüllten Feldern enthalten und es ist $\sum_{i=1}^{r} f_i < n$. Wegen Lemma 2 können wir annehmen, dass es mehr als $\frac{n}{2}$ verschiedene Elemente gibt. Daher gibt es ein Element, welches nur einmal auftritt. Durch Permutieren der Elemente können wir annehmen, dass das Element n nur einmal auftritt. Indem man notfalls Zeilen vertauscht können wir erreichen, dass das Element n in Zeile s_1 auftritt.

Als den Beweis begleitendes Beispiel, siehe auch Aigner-Ziegler, betrachten wir ein partielles lateinisches Quadrat der Ordnung $n = 7$. Bei diesem hat man vier verschiedene Elemente in vier verschiedenen Zeilen sowie Spalten. Daher ist das obige Lemma 2 hier nicht anwendbar.

$s_1 = 2$		2			5			$f_1 = 2$
$s_2 = 3$			7		4			$f_2 = 2$
$s_3 = 5$				7				$f_3 = 1$
$s_4 = 7$		4						$f_4 = 1$

	2			7		
		5		4		
			5			
	4					

In einem ersten Schritt haben wir durch ein Vertauschen der Elemente 5 und 7 dafür gesorgt, dass das Element $n = 7$ nur einmal auftritt. Hierdurch haben wir auch erreicht, dass $n = 7$ in Zeile s_1 vorkommt.

Im nächsten Schritt werden die Zeilen und Spalten des partiellen lateinischen Quadrats so vertauscht, dass danach alle gefüllten Felder oberhalb der Nebendiagonalen (das sind die Felder $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ mit $i + j = n + 1$) stehen, mit Ausnahme

des mit n gefüllten Feldes, das auf der Nebendiagonalen liegen soll. Dies erreichen wir folgendermaßen: Vertausche Zeile s_1 mit Zeile $n-f_1$ wird, Zeile s_2 mit Zeile $n-(f_1+f_2)$ und allgemein Zeile s_i mit Zeile $n-(f_1+\cdots+f_i)$, $i=1,\ldots,r$.

In unserem Beispiel ergibt sich das folgende, links angegebene, (konjugierte) partielle lateinische Quadrat:

	4					
			5			
		5		4		
	2			7		

	4					
			5			
4		5				
7	2					

	4					•
			5		•	
5		4		•		
			•			
	2	7				
	•					
•						

Anschließend permutiere man die Spalten so, dass die Spalten, die jetzt die f_1 gefüllten Felder in Zeile $n-f_1$ enthalten, die ersten f_1 Spalten werden; die Spalten, die die verbleibenden gefüllten Felder in Zeile $n-(f_1+f_2)$ (es sind höchstens f_2) enthalten, werden die nächsten Spalten. Dies setzt man fort. Zum Abschluss sind alle gefüllten Felder oberhalb oder auf der Nebendiagonalen und auf oder oberhalb der Zeile $n-f_1$.

Das Ergebnis in unserem Beispiel haben wir oben in der Mitte angegeben. Hierbei brauchen nur die Spalten 1 und 5 vertauscht zu werden.

Das nur einmal auftretende Element n stehe in der Position $(n-f_1,c)$. Man vertausche nun die Spalten f_1+1 und c. Dadurch erhält man ein (konjugiertes) partielles lateinisches Quadrat M, bei dem das Element n auf der Nebendiagonalen und alle anderen Elemente oberhalb der Nebendiagonalen liegen. Wir werden zeigen, dass man M zu einem lateinischen Quadrat der Ordnung n erweitern kann. Indem man die Vertauschungen in umgekehrter Reihenfolge rückgängig macht, erhält man eine Vervollständigung des ursprünglich gegebenen partiellen lateinischen Quadrats.

In unserem begleitenden Beispiel geben wir das Ergebnis, also das partielle lateinische Quadrat M, oben rechts an. Die außer mit $n=7$ nichtgefüllten Nebendiagonalelemente kennzeichnen wir mit •.

Nun entferne man das Element n in der Nebendiagonale und streiche die n-te Zeile und die n-te Spalte. Man erhält ein partielles lateinisches Quadrat der Ordnung $n-1$ mit höchstens $n-2$ gefüllten Feldern (da ja n gestrichen wurde). Nach Induktionsvoraussetzung lässt sich dieses zu einem lateinischen Quadrat L_{n-1} der Ordnung $n-1$ vervollständigen.

In unserem Beispiel geben wir das so entstandene partielle lateinische Quadrat der Ordnung $n-1=6$ und eine mögliche Vervollständigung zu einem lateinischen Quadrat L_6 der Ordnung 6 an.

	4				
			5		
5		4			
	2				

1	**4**	2	3	5	6
2	6	1	**5**	4	3
5	3	**4**	2	6	1
4	5	6	1	3	2
6	**2**	3	4	1	5
3	1	5	6	2	4

Sei A die obere Hälfte von L_{n-1} einschließlich der Nebendiagonalen. Ist also

$$L_{n-1} = \begin{array}{|c|c|c|}\hline a_{11} & \cdots & a_{1,n-1} \\ \hline \vdots & \vdots & \vdots \\ \hline a_{n-1,1} & \cdots & a_{n-1,n-1} \\ \hline \end{array},$$

so ist

$$A = \begin{array}{|c|c|c|c|c|c|}\hline a_{11} & \cdots & a_{1,n-k} & \cdots & a_{1,n-2} & a_{1,n-1} \\ \hline a_{21} & \cdots & a_{2,n-k} & \cdots & a_{2,n-2} & \\ \hline \vdots & \vdots & \vdots & \iddots & & \\ \hline a_{k,1} & \cdots & a_{k,n-k} & & & \\ \hline \vdots & \iddots & & & & \\ \hline a_{n-1,1} & & & & & \\ \hline \end{array}$$

Mit L'_n bezeichnen wir das partielle lateinische Quadrat der Ordnung n, welches A oberhalb der Nebendiagonalen als Einträge besitzt und bei dem jedes Feld der Nebendiagonalen mit dem Element n gefüllt ist. L'_n hat also die Form

$$L'_n = \begin{array}{|c|c|c|c|c|c||c|}\hline a_{11} & \cdots & a_{1j} & \cdots & a_{1,n-2} & a_{1,n-1} & n \\ \hline a_{21} & \cdots & a_{2j} & \cdots & a_{2,n-2} & n & \\ \hline \vdots & \vdots & \vdots & \iddots & \iddots & & \\ \hline a_{n-j,1} & \cdots & a_{n-j,j} & n & & & \\ \hline \vdots & \iddots & \iddots & & & & \\ \hline a_{n-1,1} & \iddots & & & & & \\ \hline\hline n & & & & & & \\ \hline \end{array}$$

Die letzte Spalte und die letzte Zeile von L'_n spielen eine Sonderrolle, was wir durch einen horizontalen bzw. waagerechten Doppelstrich deutlich machen. Denn es genügt natürlich, die noch freien Felder im oberen $(n-1) \times (n-1)$-Block so mit Zahlen aus $\{1, \ldots, n-1\}$ zu füllen, dass in jeder Zeile und jeder Spalte (paarweise) verschiedene Elemente stehen. Die dann noch freien Felder in der letzten Spalte und der letzten Zeile werden durch die jeweiligen fehlenden Elemente gefüllt. Wir werden zeigen, dass L'_n zu einem lateinischen Quadrat L_n der Ordnung n erweitert werden kann.

In unserem Beispiel geben wir in Abbildung 93 noch einmal das lateinische Quadrat L_6 sowie das zu einem lateinischen Quadrat L_7 zu vervollständigende partielle lateinische Quadrat L'_7 an:

Die Vervollständigung von L'_n zu einem lateinischen Quadrat L_n der Ordnung n erfolgt Zeile für Zeile unter Benutzung des lateinischen Quadrats L_{n-1}, dessen oberer Teil (einschließlich der Nebendiagonalen) mit dem oberen Teil von L'_n (ausschließlich der Nebendiagonalen) übereinstimmt. Wie schon gesagt, wird nur der linke obere $(n-1) \times (n-1)$-Block gefüllt, und hier natürlich auch nur die freien Felder. Das Füllen der dann noch freien Felder in der letzten Spalte und der letzten Zeile ist trivial. Wir nehmen an, es sei $k \geq 3$ und die ersten $k-1$ Zeilen von L'_n seien schon so gefüllt, dass die folgenden beiden Bedingungen gelten:

$L_6 =$

1	4	2	3	5	6
2	6	1	5	4	3
5	3	4	2	6	1
4	5	6	1	3	2
6	2	3	4	1	5
3	1	5	6	2	4

$L'_7 =$

1	4	2	3	5	6	7
2	6	1	5	4	7	
5	3	4	2	7		
4	5	6	7			
6	2	7				
3	7					
7						

Abbildung 93: Das lateinische Quadrat L_6 und das partielle lateinische Quadrat L'_7

- Mit $j \in \{n-k+2, \ldots, n-1\}$ gibt es unter den $k-1$ ersten Komponenten der j-ten Spalte von L_n genau ein Element x_j, das "fehlende Element", welches nicht in der j-ten Spalte von L'_n vorkommt. Sei $X_{k-1} := \{x_{n-k+2}, \ldots, x_{n-1}\}$.
- Die "fehlenden Elemente" x_j, $j \in \{n-k+2, \ldots, n-1\}$, sind (paarweise) verschieden bzw. $|X_{k-1}| = k-2$.

Für $k = 1, 2$ sind diese beiden Bedingungen leer bzw. erfüllt. Wir nehmen also an, dass die ersten k Zeilen von L'_n vor dem Füllen der k-ten Zeile die folgende Form haben:

a_{11}	$\cdots$	$a_{1,n-k}$	$a_{1,n-k+1}$	$\cdots$	$a_{1,n-2}$	$a_{1,n-1}$	n
a_{21}	$\cdots$	$a_{2,n-k}$	$a_{2,n-k+1}$	$\cdots$	$a_{2,n-2}$	n	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
$a_{k-1,1}$	$\cdots$	$a_{k-1,n-k}$	$a_{k-1,n-k+1}$	$\ddots$	$*$	$*$	$*$
$a_{k,1}$	$\cdots$	$a_{k,n-k}$	n				

In der k-ten Zeile von L'_n müssen die Felder in den Positionen $(k, n-k+2), \ldots, (k, n-1)$ gefüllt werden. Es seien die "fehlenden Elemente" $x_{n-k+2}, \ldots, x_{n-1}$ in den Spalten $n-k+2, \ldots, n-1$ von L'_n bekannt, ferner seien $y_{n-k} := a_{k,n-k}$, $\ldots$, $y_{n-1} := a_{k,n-1}$ beim Nebendiagonalelement beginnend Elemente in der k-ten Zeile des lateinischen Quadrats L_{n-1}. Folglich ist $y_{n-k+1} = a_{k,n-k+1}$ das neue "fehlende Element" in Spalte $n-k+1$. Die neue Menge der "fehlenden Elemente" ist also $X_k := \{y_{n-k+1}, x_{n-k+2}, \ldots, x_{n-1}\}$. Als einen ersten Versuch fülle man die leeren Felder in der k-ten Zeile (bis auf die n-te Komponente) mit den Elementen $y_{n-k+2}, \ldots, y_{n-1}$ bzw. $a_{k,n-k+2}, \ldots, a_{k,n-1}$. Diese Zeile ist *zulässig*, d. h. in ihr kommt bis auf $a_{k,n-k+1}$ jede der Zahlen von 1 bis n genau einmal vor. Weiter ist $x_j \neq y_j$, $j = n-k+2, \ldots, n-1$. Denn sowohl x_j als auch y_j stammen aus der j-ten Spalte des lateinischen Quadrats L_n, aber x_j aus den ersten $k-1$ Komponenten und y_j aus der k-ten Komponente.

Ist $y_{n-k+1} \notin \{x_{n-k+2}, \ldots, x_{n-1}\}$, so haben wir die k-te Zeile von L'_n (bis auf die letzte Komponente) erfolgreich gefüllt, und zwar so, dass mit $x_{n-k+1} := y_{n-k+1}$ und $X_k := \{x_{n-k+1}, \ldots, x_{n-1}\}$ wieder die beiden obigen Bedingungen gelten.

Ist andernfalls $y_{n-k+1} = x_j$ mit einem $j \in \{n-k+2, \ldots, n-1\}$, so vertausche man x_j und y_j. Die neue k-te Zeile, in der jetzt x_j statt y_j steht, ist wieder zulässig, da $x_j = y_{n-k+1}$ und y_{n-k+1} in der k-ten Zeile war. Die neue Menge der "fehlenden

Elemente" ist $X_k := \{x_{n-k+1}, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_{n-1}\}$. Man ist fertig, falls nicht $y_j = x_l$ mit einem $l \in \{n-k+2, \ldots, n-1\} \setminus \{j\}$. In diesem Falle vertausche man x_l und y_l. Auf diese Weise kann man fortfahren. Dieser Prozess endet spätestens dann, wenn alle "fehlenden Elemente" in die k-te Zeile gebracht wurden. Das Füllen der letzten Zeile und der letzten Spalte ist trivial. Der Satz von Evans-Smetaniuk ist bewiesen. □

Beispiel: Wir wollen das den Beweis des Satzes von Evans-Smetaniuk begleitende Beispiel zu Ende führen bzw. das in Abbildung 93 angegebene partielle lateinische Quadrat L_7' mit Hilfe des lateinischen Quadrats L_6 vervollständigen. Beim Füllen der dritten und der vierten Zeile ist man jeweils mit dem ersten Versuch erfolgreich:

1	4	2	3	5	6	7
2	6	1	5	4	7	
5	3	4	2	7	1	
4	5	6	7	3	2	
6	2	7				
3	7					
7						

1	4	2	3	5	6
2	6	1	5	4	**3**
5	3	4	2	**6**	1
4	5	6	**1**	3	2
6	2	3	4	1	5
3	1	5	6	2	4

Rechts haben wir im lateinischen Quadrat die fehlenden Elemente in den letzten drei Spalten rot eingetragen, es ist also $X_4 = \{x_4, x_5, x_6\} = \{1, 6, 3\}$. Das neue "fehlende Element" in der dritten Spalte ist $y_3 = 3$, die neue fünfte Zeile ist

6	2	7	4	1	5	

$=$

6	2	7	y_4	y_5	y_6	

Wegen $y_3 = x_6$ werden x_6 und $y_6 = 5$ vertauscht. Die neue Menge "fehlender Elemente" ist also $X_5 = \{x_3, x_4, x_5, x_6\} = \{3, 1, 6, 5\}$, das Ergebnis bis zum Füllen der fünften Zeile ist also

1	4	2	3	5	6	7
2	6	1	5	4	7	
5	3	4	2	7	1	
4	5	6	7	3	2	
6	2	7	4	1	3	
3	7					
7						

1	4	2	3	5	6
2	6	1	5	4	3
5	3	4	2	**6**	1
4	5	6	**1**	3	2
6	2	**3**	4	1	**5**
3	1	5	6	2	4

Das neue "fehlende Element" in der zweiten Spalte ist $y_2 = 1$, die neue sechste Zeile ist

3	7	5	6	2	4	

$=$

3	7	y_3	y_4	y_5	y_6	

Wegen $y_2 = x_4$ ist man noch nicht fertig mit dem Füllen der sechsten Zeile, sondern es werden x_4 und $y_4 = 6$ miteinander vertauscht. Wegen $y_4 = 6 = x_5$ ist man leider immer noch nicht fertig, sondern hat auch noch x_5 und $y_5 = 2$ miteinander zu vertauschen.

Als Ergebnis nach dem Füllen der sechsten Zeile erhält man daher

1	4	2	3	5	6	7
2	6	1	5	4	7	**3**
5	3	4	2	7	1	**6**
4	5	6	7	3	2	**1**
6	2	7	4	1	3	**5**
3	7	5	1	6	4	**2**
7	**1**	**3**	**6**	**2**	**5**	**4**

1	4	2	3	5	6
2	6	1	5	4	3
5	3	4	2	6	1
4	5	6	1	3	2
6	2	**3**	4	1	**5**
3	**1**	5	**6**	**2**	4

Hierbei haben wir die letzte Zeile schon mit den roten "fehlenden Elementen" aufgefüllt und auch die letzte Spalte blau ergänzt. Nun war die ursprüngliche Aufgabe, das in Abbildung 94 links angegebene partielle lateinische Quadrat zu vervollständigen. Durch

	2			5		
		7		4		
			7			
	4					

	4					
			5			
5		4				
	2	7				

Abbildung 94: Ein gegebenes und ein dazu konjugiertes partielles lateinisches Quadrat

mehrere Vertauschungen überführten wir dieses gegebene partielle lateinische Quadrat in das in Abbildung 94 rechts stehende konjugierte partielle lateinische Quadrat, von dem wir gerade eben eine Vervollständigung zu einem lateinischen Quadrat berechnet haben. Diese Vertauschungen waren der Reihe nach:

1. Vertausche die Elemente 5 und 7.
2. Vertausche Zeile 2 mit Zeile 5 und vertausche Zeile 1 mit Zeile 7.
3. Vertausche Spalte 1 und Spalte 5.
4. Vertausche Spalte 1 und Spalte 3 (um 7 als Nebendiagonalelement zu erhalten).

Diese Vertauschungen müssen nun in umgekehrter Reihenfolge wieder rückgängig gemacht werden, um eine Vervollständigung des ursprünglich gegebenen partielle lateinischen Quadrat zu erhalten. Im ersten Schritt werden die Vertauschungen der Spalten rückgängig gemacht, danach die der Zeilen und schließlich die der Elemente:

5	4	1	3	2	6	7
4	6	2	5	1	7	3
7	3	5	2	4	1	6
3	5	4	7	6	2	1
1	2	6	4	7	3	5
6	7	3	1	5	4	2
2	1	7	6	3	5	4

2	1	7	6	3	5	4
1	2	6	4	7	3	5
7	3	5	2	4	1	6
3	5	4	7	6	2	1
4	6	2	5	1	7	3
6	7	3	1	5	4	2
5	4	1	3	2	6	7

2	1	5	6	3	7	4
1	**2**	6	4	**5**	3	7
5	3	**7**	2	**4**	1	6
3	7	4	5	6	2	1
4	6	2	**7**	1	5	3
6	5	3	1	7	4	2
7	**4**	1	3	2	6	5

Rechts steht die Vervollständigung des ursprünglich gegebenen partiellen lateinischen Quadrats, dessen Einträge noch einmal in rot angegeben sind. □

66 Sudoku

Ziel beim Logikrätsel Sudoku ist es, ein 9×9-Gitter mit den Ziffern 1 bis 9 so zu füllen, dass jede Ziffer in jeder Spalte, in jeder Zeile und in jedem Block (3×3-Unterquadrat) genau einmal vorkommt. Ausgangspunkt ist ein Gitter, in dem bereits mehrere Ziffern vorgegeben sind, und zwar natürlich so, dass jede der Ziffern von 1 bis 9 in jeder Spalte, jeder Zeile und in jedem Block *höchstens* einmal vorkommt. Dieses Gitter gilt es zu vervollständigen. In Abbildung 95 links geben wir ein mittelschweres Sudoku aus der (heutigen) Süddeutschen Zeitung vom 2. Juli 2012 an, daneben seine Lösung. Wir wollen hier nichts über die Geschichte und die Ursprünge von Sudoku sagen und

			6					5
	2	4			8	9		
	5	1		9	3		8	
1				8		6		
		3	1	6		7		2
		9		3			5	
		5						9
7			3			8		
			2	4	6			3

8	9	7	6	2	1	3	4	5
3	2	4	5	7	8	9	1	6
6	5	1	4	9	3	2	8	7
1	7	2	9	8	5	6	3	4
5	8	3	1	6	4	7	9	2
4	6	9	7	3	2	1	5	8
2	3	5	8	1	7	4	6	9
7	4	6	3	5	9	8	2	1
9	1	8	2	4	6	5	7	3

Abbildung 95: Ein Sudoku-Rätsel und seine Lösung

verweisen hierzu nur auf `http://de.wikipedia.org/wiki/Sudoku`. Vielmehr wollen wir auf einige mathematische Fragen im Zusammenhang mit Sudoku und auf Verfahren zur Lösung eines Sudoku-Rätsels eingehen.

66.1 Mathematische Aussagen zu Sudoku

Ein 9×9-Gitter oder Gitterfeld mit 3×3 Blöcken, welches mit den Ziffern 1 bis 9 so gefüllt ist, dass jede Ziffer in jeder Spalte, jeder Zeile und jedem Block genau einmal vorkommt, nennen wir ein *vollständiges Sudoku*. Ist ein solches Gitter nur *teilweise* mit Ziffern von 1 bis 9 gefüllt, und zwar so, dass jede der Ziffern in jeder Spalte, jeder Zeile und jedem Block höchstens einmal vorkommt, so sprechen wir von einem *partiellen Sudoku*, wenn eine Ergänzung zu einem vollständigen Sudoku möglich ist. Ein partielles Sudoku entsteht also aus einem vollständigen Sudoku durch Weglassen gewisser Einträge. Das Sudoku-Rätsel besteht darin, ein gegebenes partielles Sudoku zu einem vollständigen Sudoku zu ergänzen. Die in den Zeitungen, Zeitschriften und Rätselheften veröffentlichten Sudoku-Rätsel sind i. Allg. *eindeutig* lösbar, d. h. zu dem

vorgegebenen partiellen Sudoku gibt es genau eine Ergänzung zu einem vollständigen Sudoku.

Wir wollen wenigstens flüchtig und häufig nur mit Verweisen auf die Literatur auf die folgenden Fragen eingehen.

1. Wie viele verschiedene vollständige Sudokus gibt es?

2. Wie viele *im wesentlichen verschiedene* vollständige Sudokus gibt es? Hierzu müssen wir erklären, wann wir zwei vollständige Sudokus *im wesentlichen gleich* oder *äquivalent* nennen. So werden wir z. B. zwei vollständige Sudokus im wesentlichen gleich nennen, wenn das eine aus dem anderen dadurch hervor geht, dass die Ziffern permutiert, also z. B. die Ziffern 1 und 2 vertauscht werden.

3. Wie viele Ziffern müssen in einem partiellen Sudoku mindestens vorgegeben sein, dass es hierzu eine eindeutige Vervollständigung gibt? Die Lösung dieses Problems im Jahre 2011, dass nämlich mindestens 17 Ziffern vorgegeben sein müssen, wurde auch in deutschsprachigen Zeitungen gemeldet, siehe z. B. FAZ vom 17.03.2012 `http://www.faz.net/-gx7-6yel5`, Neue Zürcher Zeitung vom 18.01.2012, Spektrum der Wissenschaft vom 16.03.2012 oder Wissenschaft Aktuell vom 30.01.2012. Die entsprechenden Web-Adressen erhält man leicht, indem man nach "sudoku eindeutig lösbar" googelt. Unter `http://www.math.ie/checker.html` findet man links zu weiteren Reaktionen in der internationalen Presse. Übrigens sind von Gordon Royle von der University of Western Australia, siehe `http://mapleta.maths.uwa.edu.au/~gordon/sudokumin.php`, 49151 im wesentlichen verschiedene partielle Sudokus mit 17 Einträgen gefunden worden, die auf eindeutige Weise zu einem vollständigen Sudoku ergänzt werden können.

66.1.1 Die Anzahl vollständiger Sudokus

Nun gehen wir auf die erste Frage ein, nämlich die Anzahl N_0 vollständiger Sudokus anzugeben[91]. Grundlage der folgenden Bemerkungen ist ein Aufsatz von B. Felgen-

[91] Relativ einfach ist es, die Anzahl vollständiger 4×4-Sudokus anzugeben, siehe `http://www.math.cornell.edu/~mec/Summer2009/Mahmood/Four.html`. Hier hat man ein 4×4-Gitter so mit den Zahlen 1 bis 4 zu füllen, dass jede der Ziffern in allen Zeilen, Spalten und 2×2-Unterblöcken genau einmal auftritt. Die gesuchte Anzahl ist 288. Denn man denke sich das 4×4-Gitter zerlegt in der Form

B_1	B_2
B_3	B_4

mit 2×2-Blöcken $B_1, \ldots, B_4$. Der linke obere Block B_1 kann auf 4!-Weisen mit den Zahlen 1 bis 4 gefüllt werden. Nach einer Umnummerierung können wir annehmen, dass die beiden oberen Blöcke die Form

1	2	$\{3,4\}$
3	4	$\{1,2\}$

haben. Hierbei bedeutet z. B. $\{3,4\}$ die Zahlen 3 und 4 in irgendeiner Reihenfolge. Zum Füllen der ersten beiden Zeilen hat man also $4! \cdot 2^2$ Möglichkeiten, zum Füllen auch des linken unteren Blockes B_3 folglich $4! \cdot 2^4$ Möglichkeiten. Von den $2^4 = 16$ Möglichkeiten, die Blöcke B_2 und B_3 zu füllen,

HAUER, F. JARVIS (2006), dem wir eng folgen und in dem mit Hilfe von Computern nachgewiesen wird, dass die gesuchte Anzahl durch

$$N_0 = 6670903752021072936960 \approx 6,671 \cdot 10^{21}$$

gegeben ist. Eine obere Schranke für die gesuchte Anzahl erhält man durch die Beobachtung, dass ein vollständiges Sudoku ein spezielles lateinisches Quadrat der Ordnung 9 ist. Deren Anzahl ist

$$5524751496156892842531225600 \approx 5.525 \cdot 10^{27},$$

siehe `http://oeis.org/A002860` und S. F. BAMMEL, J. ROTHSTEIN (1975). Elementar erzielbare, schlechtere obere Schranken für die Anzahl vollständiger Sudokus findet man unter `http://www.math.cornell.edu/~mec/Summer2009/meerkamp/Site/Counting_Sudokus_2.html`. Die 3×3-Unterblöcke eines vollständigen Sudokus werden mit $B_1, \ldots, B_9$ bezeichnet, siehe Abbildung 96. Der linke obere Block B_1 kann

B_1	B_2	B_3
B_4	B_5	B_6
B_7	B_8	B_9

Abbildung 96: Unterblöcke eines vollständigen Sudoku

auf $9! = 362880$ verschiedene Weise gefüllt werden. Nach einer Umnummerierung der Ziffern können wir annehmen, dass B_1 durch

1	2	3
4	5	6
7	8	9

gegeben ist. Es genügt daher, die Anzahl $N_1 = N_0/9!$ der vollständigen Sudokus zu berechnen, deren linker oberer Block B_1 die angegebene Standardform hat, wovon im

können aber nur 12 zu einem 4×4-Sudoku vervollständigt werden. Ist nämlich eine Spalte von B_2 gleich einer Zeile von B_3, so ist dies nicht möglich. Dies ist z. B. für

1	2	3	4
3	4	2	1
2	3	$*$	$*$
4	1	$*$	$*$

der Fall. Insgesamt gibt es daher $4! \cdot 12 = 288$ verschiedene vollständige 4×4-Sudokus.

Folgenden ausgegangen wird. Jetzt überlegen wir uns, wie viele Möglichkeiten es gibt, die Blöcke B_2 und B_3 zulässig zu füllen. Zunächst betrachten wir die erste Zeile von Block B_2. Sie besteht in einer gewissen Anordnung aus den Zeilen 2 und 3 von Block B_1 bzw. einer Mischung aus beiden. Wir nennen diese *reine* bzw. *gemischte* Zeilen. In Tabelle 8 geben wir die möglichen ersten Zeilen der Blöcke B_2 und B_3 wieder. Hierbei

$\{4,5,6\}$	$\{7,8,9\}$
$\{4,5,7\}$	$\{6,8,9\}$
$\{4,5,8\}$	$\{6,7,9\}$
$\{4,5,9\}$	$\{6,7,8\}$
$\{4,6,7\}$	$\{5,8,9\}$
$\{4,6,8\}$	$\{5,7,9\}$
$\{4,6,9\}$	$\{5,7,8\}$
$\{5,6,7\}$	$\{4,8,9\}$
$\{5,6,8\}$	$\{4,7,9\}$
$\{5,6,9\}$	$\{4,7,8\}$

$\{7,8,9\}$	$\{4,5,6\}$
$\{6,8,9\}$	$\{4,5,7\}$
$\{6,7,9\}$	$\{4,5,8\}$
$\{6,7,8\}$	$\{4,5,9\}$
$\{5,8,9\}$	$\{4,6,7\}$
$\{5,7,9\}$	$\{4,6,8\}$
$\{5,7,8\}$	$\{4,6,9\}$
$\{4,8,9\}$	$\{5,6,7\}$
$\{4,7,9\}$	$\{5,6,8\}$
$\{4,7,8\}$	$\{5,6,9\}$

Tabelle 8: Mögliche erste Zeilen in den Blöcken $B_1|B_2$

bedeutet $\{a,b,c\}$ die Ziffern a,b,c in *irgendeiner* Reihenfolge. Die reine erste Zeile

$$1 \quad 2 \quad 3 \parallel \{4,5,6\} \parallel \{7,8,9\}$$

kann auf die folgenden Weisen vervollständigt werden:

1	2	3	$\{4,5,6\}$	$\{7,8,9\}$
4	5	6	$\{7,8,9\}$	$\{1,2,3\}$
7	8	9	$\{1,2,3\}$	$\{4,5,6\}$

Dies ergibt insgesamt $(3!)^6$ mögliche Konfigurationen, da sich jede Menge aus drei Zahlen auf $3! = 6$ verschiedene Weisen anordnen lässt. Das entsprechende gilt für die andere reine erste Zeile

$$1 \quad 2 \quad 3 \parallel \{7,8,9\} \parallel \{4,5,6\} \ .$$

Es bleiben noch 18 gemischte erste Zeilen. Bei diesen besteht die erste Zeile des Blocks B_2 aus Zahlen sowohl der zweiten als auch der dritten Zeile von B_1. Z. B. kann die gemischte erste Zeile

$$1 \quad 2 \quad 3 \parallel \{4,5,7\} \parallel \{6,8,9\}$$

auf die folgenden Weisen vervollständigt werden:

1	2	3	$\{4,5,7\}$	$\{6,8,9\}$
4	5	6	$\{8,9,a\}$	$\{7,b,c\}$
7	8	9	$\{6,b,c\}$	$\{4,5,a\}$

wobei a, b und c für 1, 2 und 3 in beliebiger Reihenfolge stehen, was $3 \cdot (3!)^6$ mögliche Konfigurationen ergibt (b und c können vertauscht werden). Insgesamt haben wir damit

$$2 \cdot (3!)^6 + 18 \cdot 3 \cdot (3!)^6 = 56 \cdot (3!)^6 = 2612736$$

mögliche Vervollständigungen der ersten drei Zeilen, wenn wir von einem Block B_1 in Standardform ausgehen.

Jetzt müssen wir uns überlegen, auf wie viele Weise die möglichen Blöcke B_2 und B_3 (immer nehmen wir an, der Block B_1 sei in Standardform) zu einem vollständigen Sudoku ergänzt werden können. Ein Durchprobieren aller 2612736 Möglichkeiten wäre zu zeitaufwendig. Daher ist es wichtig, die Anzahl der Möglichkeiten zu reduzieren. Hierzu sagen wir, die Blöcke (B_2, B_3) und (B_2', B_3') seien *äquivalent*, wenn die Anzahl möglicher Vervollständigungen von (B_2, B_3) und (B_2', B_3') zu einem vollständigen Sudoku gleich ist. Z. B. sind (B_2, B_3) und (B_3, B_2) äquivalent, denn man erhält eine Vervollständigung von (B_3, B_2) aus einer von (B_2, B_3), indem man die Blöcke B_5 und B_6 sowie B_8 und B_9 miteinander vertauscht. Allgemeiner können wir sogar B_1, B_2 und B_3 beliebig vertauschen. Nach eventueller Umnummerierung ist der Block B_1 wieder in Standardform. Weiter kann man in jedem der Blöcke B_1, B_2 und B_3 die Spalten beliebig permutieren. Nachdem man durch eine Umnummerierung den ersten Block wieder in Standardform gebracht hat, hat man äquivalente Blöcke (B_2', B_3') erhalten. Entsprechend kann man die drei Zeilen von B_1, B_2 und B_3 permutieren und den ersten Block durch eine Umnummerierung in Standardform überführen. Aus jeder Klasse äquivalenter Blöcke braucht man dann nur für einen Vertreter die Anzahl möglicher Vervollständigungen zu berechnen.

Unter *lexikographischer Reduktion* von einem der 2612736 möglichen (B_2, B_3) Blöcke verstehen wir das Ergebnis der folgenden beiden Operationen:

1. Permutiere die Spalten in B_2 und B_3 so, dass die ersten Zeilen jeweils aufsteigend angeordnet sind.

2. Wenn nötig vertausche B_2 und B_3 so, dass der linke obere Eintrag von B_2 kleiner als der entsprechende von B_3 ist.

Ohne lexikographische Reduktion hat man für die erste Zeile von (B_2, B_3) genau $6! = 720$ Möglichkeiten. Nach lexikographischer Reduktion gibt es nur noch 10 mögliche erste Zeilen von (B_1, B_2), nämlich

$$456\,789,\ 457\,689,\ 458\,679,\ 459\,678,\ 467\,589,\ 468\,579,\ 469\,578,\ 478\,569,\ 479\,568$$

und 489 567, also "nur noch" $2612736/72 = 36288$ Möglichkeiten für die Blöcke (B_2, B_3). Es gibt aber weitere Möglichkeiten, äquivalente Blöcke zu bestimmen zu gegebenen Blöcken (B_2, B_3) zu bestimmen, von denen wir annehmen, dass sie lexikographisch reduziert sind. Wir können alle sechs Permutationen der drei Blöcke B_1, B_2 und B_3 betrachten, ferner alle sechs Permutationen von Spalten innerhalb der drei Blöcke, was insgesamt $6^4 = 1296$ Möglichkeiten ergibt. Der neue erste Block wird i. Allg. nicht in Standardform sein, das wird notfalls durch eine Umnummerierung erreicht. Auf das Ergebnis wende man anschließend eine lexikographische Reduktion an. Bei Felgenhauer-Jarvis wird angegeben, dass Computer-Rechnungen ergeben haben, dass nur noch 2051 Paare (B_2, B_3) getestet zu werden brauchen. Man kann aber auch die drei Zeilen der drei Blöcke permutieren, dann den ersten Block wieder in Standardform bringen und lexikographische Reduktion anwenden. Stets erhält man äquivalente

neue Blöcke (B_2', B_3'). Dies ergibt nach Felgenhauer-Jarvis eine weitere Reduktion auf 416 auf Vervollständigung zu testende Paare (B_2, B_3).

Beispiel: Die ersten drei Blöcke seien gegeben durch

$(*)$

1	2	3	4	7	8	5	6	9
4	5	6	1	3	9	2	7	8
7	8	9	2	5	6	1	3	4

Hier hat der linke obere Block Standardform und die beiden Blöcke daneben sind lexikographisch reduziert. Wir vertauschen die ersten beiden Spalten des ersten Blocks und gehen anschließend wieder zur Standardform über:

2	1	3	4	7	8	5	6	9
5	4	6	1	3	9	2	7	8
8	7	9	2	5	6	1	3	4

1	2	3	5	8	7	4	6	9
4	5	6	2	3	9	1	8	7
7	8	9	1	4	6	2	3	5

Jetzt muss man in zwei Schritten noch die lexikographische Reduktion anwenden:

1	2	3	5	7	8	4	6	9
4	5	6	2	9	3	1	8	7
7	8	9	1	6	4	2	3	5

1	2	3	4	6	9	5	7	8
4	5	6	1	8	7	2	9	3
7	8	9	2	3	5	1	6	4

Zum Vergleich vertauschen wir in $(*)$ die zweite und die dritte Zeile und führen das Ergebnis anschließend in Standardform über:

1	2	3	4	7	8	5	6	9
7	8	9	2	5	6	1	3	4
4	5	6	1	3	9	2	7	8

1	2	3	7	4	5	8	9	6
4	5	6	2	8	9	1	3	7
7	8	9	1	3	6	2	4	5

Die lexikographische Reduktion ergibt

1	2	3	4	5	7	6	8	9
4	5	6	8	9	2	7	1	3
7	8	9	3	6	1	5	2	4

Wir haben erhalten, dass

1	2	3	4	6	9	5	7	8
4	5	6	1	8	7	2	9	3
7	8	9	2	3	5	1	6	4

1	2	3	4	5	7	6	8	9
4	5	6	8	9	2	7	1	3
7	8	9	3	6	1	5	2	4

äquivalent zu den Ausgangsblöcken in $(*)$ sind. □

Es gibt aber weitere Möglichkeiten, äquivalente Blöcke zu bestimmen. So sind z. B. die Blöcke

1	2	3	4	6	9	5	7	8
4	5	6	7	8	2	3	9	1
7	8	9	5	3	1	6	4	2

1	2	3	4	6	9	5	7	8
4	5	6	7	8	1	3	9	2
7	8	9	5	3	2	6	4	1

äquivalent, denn eine Vervollständigung des einen ist auch eine Vervollständigung des anderen. Allgemeiner besitzen zulässige (es dürfte klar sein, was wir darunter verstehen) Blöcke

B_1	B_2	B_3

B'_1	B'_2	B'_3

dieselbe Anzahl von Vervollständigungen, wenn bis auf die Reihenfolge entsprechende Spalten der Blöcke B_i und B'_i, $i = 1, 2, 3$, dieselben Einträge besitzen. Denn jede Vervollständigung des einen ist auch eine Vervollständigung des anderen. So haben z. B.

1	2	3	4	6	7	5	8	9
4	5	6	8	1	9	3	2	7
7	8	9	2	5	3	1	4	6

4	5	3	2	6	7	1	8	9
1	2	6	8	5	9	3	4	7
7	8	9	4	1	3	5	2	6

dieselbe Anzahl von Vervollständigungen. Die letzteren Blöcke ergeben nach Überführung des linken Blockes in Standardform durch Umnummerierung und lexikographischer Reduktion

1	2	3	5	6	7	4	8	9
4	5	6	8	2	9	3	1	7
7	8	9	1	4	3	2	5	6

1	2	3	4	8	9	5	6	7
4	5	6	3	1	7	8	2	9
7	8	9	2	5	6	1	4	3

Felgenhauer-Jarvis berichten in Ihrer Arbeit (dies müssen wir hinnehmen, denn ohne eigene Rechnungen oder Überlegungen, die diese überflüssig machen, kann dies nicht überprüft werden), dass die 36288 möglichen Blöcke B_2, B_3 (der Block B_1 wird in Standardform angenommen) sich in nur 44 Klassen (Blöcke in einer Klasse haben dieselbe Anzahl von Vervollständigungen) zusammenfassen lassen. Für die 44 (B_2, B_3) Konfigurationen (diese sind natürlich lexikographisch reduziert) haben wir zu bestimmen, auf wie viele Weise (B_1, B_2, B_3) zu einem vollständigen Sudoku ergänzt werden kann. Hierbei können wir annehmen, dass die ersten Spalten der Blöcke B_4 und B_7 lexikographisch reduziert sind, was die Rechnung um einen Faktor 72 beschleunigt. In der ersten Spalte von B_4 und B_7 brauchen also nur noch eine Füllung mit

235 689, 236 589, 238 579, 239 568, 256 389, 258 369, 259 378, 268 359, 269 358

und 289 356 auf Vervollständigung untersucht zu werden. Einer der insgesamt $44 \cdot 10$ auf Vervollständigung zu untersuchenden Fälle wäre dann z. B.

1	2	3	4	6	9	5	7	8
4	5	6	7	8	2	3	9	1
7	8	9	5	3	1	6	4	2
2								
3								
5								
6								
8								
9								

Dies geschieht, wie Felgenhauer-Jarvis schreiben, mit "brutaler Gewalt" (brute force) indem alle Möglichkeiten, die fehlenden Felder zu füllen ausprobiert werden. Das Gesamtergebnis ist bemerkenswert. Die Anzahl verschiedener vollständiger Sudokus ist

$$N_0 = 9! \cdot 72^2 \cdot 2^7 \cdot 27704267971 = 6670903752021072936960 \approx 6.671 \cdot 10^{21}.$$

Hierbei stammt der Faktor 9! aus der Überführung des Blockes B_1 in Standardform, der Faktor 72^2 stammt von den lexikographischen Reduktionen der ersten Zeile von B_2, B_3 sowie der ersten Spalte von B_4, B_7. Der letzte Faktor 27704267971 ist eine Primzahl!

66.1.2 Die Anzahl im wesentlichen verschiedener vollständiger Sudokus

Nun gehen wir auf das zweite gestellte Problem ein, nämlich wie viele im wesentlichen verschiedene vollständige Sudokus es gibt. Hierzu müssen wir natürlich erklären, wann zwei vollständige Sudokus als *im wesentlichen verschieden* angesehen werden. Die 9 Blöcke eines vollständigen Sudokus werden wieder mit $B_1, \ldots, B_9$ bezeichnet. Wir nennen zwei vollständige Sudokus *äquivalent* oder *im wesentlichen gleich*, wenn sie durch eine Folge von Operationen, auch *Symmetrien* genannt, der folgenden Art in einander übergeführt werden können, andernfalls heißen sie *im wesentlichen verschieden*.

1. Umnummerierung der 9 Ziffern.
2. Permutation der drei Spaltenblöcke

$$\begin{array}{|c|}\hline B_1 \\ \hline B_4 \\ \hline B_7 \\ \hline \end{array}, \qquad \begin{array}{|c|}\hline B_2 \\ \hline B_5 \\ \hline B_8 \\ \hline \end{array}, \qquad \begin{array}{|c|}\hline B_3 \\ \hline B_6 \\ \hline B_9 \\ \hline \end{array}.$$

3. Permutation der drei Zeilenblöcke

$$\begin{array}{|c|c|c|}\hline B_1 & B_2 & B_3 \\ \hline \end{array}, \qquad \begin{array}{|c|c|c|}\hline B_4 & B_5 & B_6 \\ \hline \end{array}, \qquad \begin{array}{|c|c|c|}\hline B_6 & B_7 & B_8 \\ \hline \end{array}.$$

4. Permutation der drei Spalten innerhalb eines der drei Spaltenblöcke.
5. Permutation der drei Zeilen innerhalb eines der drei Zeilenblöcke.
6. Rotation und Spiegelung.
   (a) Rotiere um 0° (tue also nichts).
   (b) Rotiere um 90° im Uhrzeigersinn.
   (c) Rotiere um 180° im Uhrzeigersinn.
   (d) Rotiere um 270° im Uhrzeigersinn bzw. um 90° entgegen dem Uhrzeigersinn.
   (e) Spiegele an der horizontalen Mittelachse, also der 5. Zeile (von unten oder oben).

8	9	7	6	2	1	3	4	5
3	2	4	5	7	8	9	1	6
6	5	1	4	9	3	2	8	7
1	7	2	9	8	5	6	3	4
5	8	3	1	6	4	7	9	2
4	6	9	7	3	2	1	5	8
2	3	5	8	1	7	4	6	9
7	4	6	3	5	9	8	2	1
9	1	8	2	4	6	5	7	3

9	7	2	4	5	1	6	3	8
1	4	3	6	8	7	5	2	9
8	6	5	9	3	2	1	4	7
2	3	8	7	1	9	4	5	6
4	5	1	3	6	8	9	7	2
6	9	7	2	4	5	3	8	1
5	8	4	1	7	6	2	9	3
7	2	6	5	9	3	8	1	4
3	1	9	8	2	4	7	6	5

Abbildung 97: Ein vollständiges und das um 90° rotierte Sudoku

3	7	5	6	4	2	8	1	9
1	2	8	9	5	3	6	4	7
9	6	4	7	1	8	5	3	2
8	5	1	2	3	7	9	6	4
2	9	7	4	6	1	3	8	5
4	3	6	5	8	9	2	7	1
7	8	2	3	9	4	1	5	6
6	1	9	8	7	5	4	2	3
5	4	3	1	2	6	7	9	8

5	6	7	4	2	8	9	1	3
4	1	8	3	9	5	6	2	7
3	9	2	6	7	1	4	8	5
1	8	3	5	4	2	7	9	6
2	7	9	8	6	3	1	5	4
6	5	4	9	1	7	8	3	2
7	4	1	2	3	9	5	6	8
9	2	5	7	8	6	3	4	1
8	3	6	1	5	4	2	7	9

Abbildung 98: Rotation um 180° und 270°

(f) Spiegele an der vertikalen Mittelachse, also der 5. Spalte (von links oder rechts).

(g) Spiegele an der Hauptdiagonalen (diese verläuft von oben links nach unten rechts).

(h) Spiegele an der Nebendiagonalen (diese verläuft von oben rechts nach unten links).

Offensichtlich transformieren die Symmetrien ein vollständiges Sudoku in ein weiteres.

Es ist nicht schwierig, die Anzahl im wesentlichen verschiedener vollständiger 4×4-Sudokus anzugeben. Durch Umnummerierung im linken oberen Block und eventuelle

9	1	8	2	4	6	5	7	3
7	4	6	3	5	9	8	2	1
2	3	5	8	1	7	4	6	9
4	6	9	7	3	2	1	5	8
5	8	3	1	6	4	7	9	2
1	7	2	9	8	5	6	3	4
6	5	1	4	9	3	2	8	7
3	2	4	5	7	8	9	1	6
8	9	7	6	2	1	3	4	5

5	4	3	1	2	6	7	9	8
6	1	9	8	7	5	4	2	3
7	8	2	3	9	4	1	5	6
4	3	6	5	8	9	2	7	1
2	9	7	4	6	1	3	8	5
8	5	1	2	3	7	9	6	4
9	6	4	7	1	8	5	3	2
1	2	8	9	5	3	6	4	7
3	7	5	6	4	2	8	1	9

Abbildung 99: An der horizontalen bzw. vertikalen Mittelachse gespiegelte Sudokus

8	3	6	1	5	4	2	7	9
9	2	5	7	8	6	3	4	1
7	4	1	2	3	9	5	6	8
6	5	4	9	1	7	8	3	2
2	7	9	8	6	3	1	5	4
1	8	3	5	4	2	7	9	6
3	9	2	6	7	1	4	8	5
4	1	8	3	9	5	6	2	7
5	6	7	4	2	8	9	1	3

3	1	9	8	2	4	7	6	5
7	2	6	5	9	3	8	1	4
5	8	4	1	7	6	2	9	3
6	9	7	2	4	5	3	8	1
4	5	1	3	6	8	9	7	2
2	3	8	7	1	9	4	5	6
8	6	5	9	3	2	1	4	7
1	4	3	6	8	7	5	2	9
9	7	2	4	5	1	6	3	8

Abbildung 100: An der Haupt- bzw. Nebendiagonale gespiegelte Sudokus

Vertauschung der beiden Zeilen im unteren Zeilenblock sowie eine eventuelle Vertauschung der beiden Spalten im rechten Spaltenblock können wir annehmen, dass das 4×4-Sudoku die folgenden Einträge besitzt, siehe Abbildung 101 links. In der Position

1	2	3	4
3	4		
2			
4			

1	2	3	4
3	4		
2		4	
4			1

1	2	3	4
3	4		
2		4	
4			2

1	2	3	4
3	4		
2		4	
4			3

Abbildung 101: Konstruktion im wesentlichen verschiedener 4×4-Sudokus

$(3, 3)$ muss die 4 stehen, weil 2 und 3, da sie in der dritten Zeile bzw. Spalte vorkommen, verboten sind und die 1 nicht möglich ist, da andernfalls für die 4 in der dritten Zeile kein Platz wäre. In der Position $(4, 4)$ sind 1, 2 und 3 möglich. Das Ergebnis tragen wir in Abbildung 101 rechts ein. Die rechten drei partiellen 4×4-Sudokus können leicht vervollständigt werden, siehe Abbildung 102. Das zweite und das dritte vollständige

1	2	3	4
3	4	1	2
2	1	4	3
4	3	2	1

1	2	3	4
3	4	2	1
2	1	4	3
4	3	1	2

1	2	3	4
3	4	1	2
2	3	4	1
4	1	2	3

Abbildung 102: Vervollständigung partieller 4×4-Sudokus

Sudoku sind aber äquivalent, da das dritte aus dem zweiten dadurch hervorgeht, dass man an der Hauptdiagonalen spiegelt und anschließend 2 und 3 vertauscht. Es gibt also nur 2 im wesentlichen verschiedene 4×4-Sudokus.

Von Ed Russell und Frazer Jarvis (siehe `http://www.afjarvis.staff.shef.ac.uk/sudoku/` und `http://www.afjarvis.staff.shef.ac.uk/sudoku/sudgroup.html`) ist 2006 mit einem Computerbeweis gezeigt worden, dass es 5472730538 im wesentlichen verschiedene vollständige Sudokus gibt. Wir wollen wenigstens einige Hinweise dazu geben, wie dieses Resultat erhalten werden kann. Hierzu müssen einige Begriffe der *Gruppentheorie* eingeführt werden.

Eine *Gruppe* ist eine Menge G, für die ein Element $e \in G$ und eine binäre Operation $\cdot : G \times G \longrightarrow G$ existieren mit

- Sind g_1, $g_2 \in G$, so ist auch $g_1 \cdot g_2 \in G$.
- Für alle g_1, g_2, $g_3 \in G$ ist $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$.
- Für alle $g \in G$ ist $g \cdot e = e \cdot g = g$
- Für jedes $g \in G$ gibt es ein $g^{-1} \in G$ mit $g \cdot g^{-1} = g^{-1} \cdot g = e$.

Die obigen Symmetrien 2.–6. können als Permutationen der Zahlen $\{1, \ldots, 81\}$ und damit als Elemente der symmetrischen Gruppe S_{81} aufgefasst werden. Als Beispiel einer auf einer Permutation der Spaltenblöcke (erster Spaltenblock wird dritter, der zweite wird erster und der dritte wird zweiter) basierenden Symmetrie geben wir an:

1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54
55	56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81

$\longrightarrow$

4	5	6	7	8	9	1	2	3
13	14	15	16	17	18	10	11	12
22	23	24	25	26	27	19	20	21
31	32	33	34	35	36	28	29	30
40	41	42	43	44	45	37	38	39
49	50	51	52	53	54	46	47	48
58	59	60	61	62	63	55	56	57
67	68	69	70	71	72	64	65	66
76	77	78	79	80	81	73	74	75

Die von den Symmetrien 2.–6. *erzeugte Gruppe* G, ist nach Definition die kleinste in S_{81} enthaltene Untergruppe, die diese Symmetrien enthält. Es ist leicht einzusehen, dass G wohldefiniert ist. Wir wollen uns jetzt überlegen, dass G schon von den Symmetrien 2.–5. und 6.(g) (Spiegelung an der Hauptdiagonalen bzw. Transposition) erzeugt wird. Die Rotation um 90° ist die Permutation

1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54
55	56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81

$\longrightarrow$

73	64	55	46	37	28	19	10	1
74	65	56	47	38	29	20	11	2
75	66	57	48	39	30	21	12	3
76	67	58	49	40	31	22	13	4
77	68	59	50	41	32	23	14	5
78	69	60	51	42	33	24	15	6
79	70	61	52	43	34	25	16	7
80	71	62	53	44	35	26	17	8
81	72	63	54	45	36	27	18	9

Diese Permutation erreicht man auch, indem man zunächst zur transponierten Permutation übergeht, danach den ersten und den dritten Spaltenblock und dann innerhalb der Spaltenblöcke die erste und die dritte Spalte vertauscht:

1	10	19	28	37	46	55	64	73
2	11	20	29	38	47	56	65	74
3	12	21	30	39	48	57	66	73
4	13	22	31	40	49	58	67	76
5	14	23	32	41	50	59	68	77
6	15	24	33	42	51	60	69	78
7	16	25	34	43	52	61	70	79
8	17	26	35	44	53	62	71	80
9	18	27	36	45	54	63	72	81

$\longrightarrow$

55	64	73	28	37	46	1	10	19
56	65	74	29	38	47	2	11	20
57	66	73	30	39	48	3	12	21
76	67	58	49	40	31	4	13	22
77	68	59	50	41	32	5	14	23
60	69	78	33	42	51	6	15	24
661	70	79	34	43	52	7	16	25
62	71	80	35	44	53	8	17	26
63	72	81	36	45	54	9	18	27

Damit können auch die Rotationen um 180° und 270° aus den Symmetrien 2., 4. und 6.(g) gewonnen werden. Dass eine Spiegelung an der horizontalen bzw. vertikalen Mittelachse mit Hilfe der Vertauschung von Zeilen- bzw. Spaltenblöcken und Zeilen- bzw. Spaltenvertauschungen innerhalb der Zeilen- bzw. Spaltenblöcke erreicht werden können, dürfte klar sein. Die Spiegelung an der Nebendiagonalen kann man erhalten, indem man zunächst um 90° im Uhrzeigersinn rotiert und anschließend an der horizontalen Mittelachse spiegelt. Damit haben wir gezeigt, dass die sogenannte *Sudoku-Gruppe* G von den Symmetrien 2.–5. und 6.(g) erzeugt wird. Die *Ordnung* von G, also die Anzahl $|G|$ der Elemente von G, ist $(3!)^8 \cdot 2 = 3359232$. Dies liegt daran, dass die Elemente von G sich als Produkte von Permutationen (in der oben angegebenen Reihenfolge) und der Identität bzw. der Spiegelung an der Hauptdiagonalen darstellen lassen. Der Faktor $(3!)^8$ rührt von der Tatsache her, dass es jeweils 3! Permutationen der Spaltenblöcke, der Zeilenblöcke, der Spalten innerhalb jedes der drei Spaltenblöcke sowie der Zeilen innerhalb jedes der drei Zeilenblöcke gibt. Die Spiegelung kann nicht durch ein Produkt von Permutationen der eben genannten Art erzeugt werden, daher der Faktor 2.

Mit X bezeichnen wir die Menge der vollständigen Sudokus. Wir haben oben schon bemerkt, dass die Symmetrien und damit die Elemente g der Sudoku-Gruppe G, einem vollständigen Sudoku $S \in X$ ein (bezüglich der Gruppe G) äquivalentes Sudoku $g(S) \in X$ zuordnen. Bei gegebenem $S \in X$ nennen wir

$$G(S) := \{g(S) \in X : g \in G\}$$

die *G-Bahn von S* und bezeichnen mit X/G die Menge der G-Bahnen in X. Die Anzahl $|X/G|$ ist die Anzahl paarweise verschiedener G-Bahnen bzw. von Sudokus, die bezüglich der Gruppe G verschieden sind. Diese Anzahl kann mit Hilfe des sogenannten *Burnside Lemmas*, das eigentlich schon auf Cauchy und Frobenius zurückgeht[92], ausgerechnet werden. Wir formulieren es für die hier vorliegende Situation, notieren aber vorher die sogenannte *Bahnformel*:

- *Für jedes $S \in X$ ist $|G| = |G(S)| \cdot |G_S|$, wobei*

  $$G_S := \{g \in G : g(S) = S\}$$

  der sogenannte Stabilisator von S ist.

Denn: Bei vorgegebenem $S \in X$ betrachten wir die Menge

$$M := \{(g, T) \in G \times X : T = g(S)\}$$

und zählen die Anzahl der Elemente von M auf zweierlei Weise. Einerseits gibt es zu jedem $g \in G$ genau ein $T \in X$ mit $T = g(S)$ bzw. $(g, T) \in M$, was $|M| = |G|$ ergibt. Für jedes $T = g_0(S) \in G(S)$ mit $g_0 \in G$ gibt es genau $|G_S|$ Elemente $h \in G$ mit $(h, T) \in M$ bzw. $h(S) = T$, nämlich Elemente $h = g_0 g$ mit $g \in G_S$. $(g_0 g, T) \in M$ für

[92] Deswegen heißt das Burnside Lemma auch *Nicht Burnsides Lemma*, siehe `http://www.math.rwth-aachen.de/homes/neunhoef/Publications/pdf/sudoku.pdf`

jedes $g \in G_S$ und folglich $|G(S)| \cdot |G_S| \leq |M|$. Für $T \notin G(S)$ ist $(g,T) \notin M$ für alle $g \in G$. Daher ist die Anzahl der Elemente von M andererseits

$$|M| = \sum_{T \in G(S)} |G_S| = |G(S)| \cdot |G_S|.$$

Insgesamt ist die Bahnformel bewiesen. □

Burnside's Lemma *Es ist*

$$|X/G| = \frac{1}{|G|} \sum_{g \in G} |\mathrm{Fix}_X(g)|,$$

wobei

$$\mathrm{Fix}_X(g) := \{S \in X : g(S) = S\}$$

die Menge der unter $g \in G$ invarianten Elemente S aus X bedeutet.

Beweis: Für $S \in X$ sei $G_S := \{g \in G : g(S) = S\}$ wieder der Stabilisator von S und $G(S)$ die G-Bahn von S. Zunächst zeigen wir:

- Ist $S \in X$ und $T \in G(S)$, so ist $|G_T| = |G_S|$.

Denn: $T \in g(S)$ hat eine Darstellung $T = g_0(S)$ mit $g_0 \in G$. Es ist $g \in G_S$, also $g \in G$ und $g(S) = S$, genau dann, wenn $g_0 g g_0^{-1} \in G_T$, wie man leicht nachrechnet. Hieraus folgt $|G_T| = |G_S|$ für alle $T \in G(S)$. Weiter benutzen wir:

- Sind $S, T \in X$ und $G(S) \cap G(T) \neq \emptyset$, so ist $G(S) = G(T)$. Insbesondere ist X disjunkte Vereinigung von G-Bahnen sind. Mit $m := |X/G|$ existieren daher $S_1, \ldots, S_m \in X$ mit $X = \bigcup_{i=1}^m G(S_i)$ und $G(S_i) \cap G(S_j) = \emptyset$ für $i \neq j$.

Denn: Wegen $G(S) \cap G(T) \neq \emptyset$ existieren $g_1, g_2 \in G$ mit $g_1(S) = g_2(T)$. Ist $U \in G(S)$, so existiert $g \in G$ mit $U = g(S)$. Dann ist aber $U = (g g_1^{-1} g_2)(T)$, also $U \in G(T)$ bzw. $G(S) \subset G(T)$. Aus Symmetriegründen ist auch $G(T) \subset G(T)$ und damit $G(S) = G(T)$.

Es ist

$$\begin{aligned}
\sum_{g \in G} |\mathrm{Fix}_X(g)| &= |\{(g,S) \in G \times X : g(S) = S\}| \\
&= \sum_{S \in X} |G_S| \\
&= \sum_{i=1}^{m} \sum_{S \in G(S_i)} |G_S| \\
&\qquad \text{(da } X \text{ disjunkte Vereinigung von } G(S_i),\ i = 1, \ldots, m) \\
&= \sum_{i=1}^{m} |G(S_i)| \cdot |G_{S_i}| \\
&\qquad \text{(Elemente einer Bahn haben gleich große Stabilisatoren)} \\
&= \sum_{i=1}^{m} |G| \\
&\qquad \text{(wegen der Bahnformel)} \\
&= |X/G| \cdot |G|.
\end{aligned}$$

Damit ist Burnside's Lemma, das nicht von Burnside stammt, bewiesen.. □

Im Prinzip haben wir eine Möglichkeit gefunden, die Anzahl der vollständigen Sudokus, die bezüglich der Gruppe G im wesentlichen verschieden sind bzw. in unterschiedlichen Bahnen liegen, zu bestimmen. Für alle $g \in G$ müssen wir die Anzahl der unter g invarianten vollständigen Sodukus bestimmen, diese Zahlen aufaddieren und am Schluss durch die Gruppenordnung $|G| = 3359232$ dividieren. Das ist natürlich ein sehr großer Aufwand. Daher ist die Bemerkung wichtig, dass Elemente aus einer *Konjugiertenklasse* dieselbe Anzahl von invarianten Elementen besitzen. Genauer gilt:

- *Sei $g \in G$. Dann ist $|\mathrm{Fix}_X(g)| = |\mathrm{Fix}_X(hgh^{-1})|$ für alle $h \in G$, d. h. $|\mathrm{Fix}_X(\cdot)|$ ist auf der zu $g \in G$ gehörenden* Konjugiertenklasse $\{hgh^{-1} : h \in G\}$ *konstant.*

Denn: Bei gegebenem $h \in G$ ist $S \in \mathrm{Fix}_X(hgh^{-1})$ genau dann, wenn $h^{-1}(S) \in \mathrm{Fix}_X(g)$. Daher ist die Aussage trivialerweise richtig.

Es bleibt daher die Aufgabe, die Konjugiertenklassen von G und jeweils für einen Repräsentanten aus der Konjugiertenklasse die Anzahl der nach einer eventuellen Umnummerierung fest bleibenden vollständigen Sudokus zu bestimmen. Dies ist von E. Russell-F. Jarvis mit Hilfe des Computeralgebra-Systems GAP (**G**roups, **A**lgorithms, **P**rogramming) geschehen, welches sich insbesondere für gruppen-theoretische Berechnungen eignet. Es stellt sich heraus, dass es in G genau 275 Konjugiertenklassen gibt. Nur in 27 von diesen gibt es (nach eventueller Umnummerierung) Fixpunkte. Die Ergebnisse findet man unter `http://www.afjarvis.staff.shef.ac.uk/sudoku/sudgroup.html`. Als Resultat erhalten Russell-Jarvis, dass es 5472730538 im wesentlichen verschiedene vollständige Sudokus gibt.

Bemerkung: Es gibt (viele) Symmetrien $g \in G$, zu denen kein $S \in X$ existiert derart, dass S und $g(S)$ bis auf eine Umnummerierung übereinstimmen. Ist z. B. g die Spiegelung an der horizontalen Mittelachse, so stimmen die mittlere, also die fünfte Zeile, von $S \in X$ und $g(S)$ überein. Wie man sich leicht überlegt, können S und $g(S)$ nicht bis auf eine Umnummerierung gleich sein. Bei anderen Symmetrien kann das aber anders sein. Wir betrachten z. B. eine Rotation um 90° im Uhrzeigersinn. In Abbildung 103 geben wir links ein vollständiges Sudoku und rechts daneben das um 90° gedrehte Sudoku an. Mit der Umnummerierung $1 \to 3 \to 9 \to 7 \to 1$, $2 \to 6 \to 8 \to 4 \to 2$ und

1	2	4	5	6	7	8	9	3
3	7	8	2	9	4	5	1	6
6	5	9	8	3	1	7	4	2
9	8	7	1	2	3	4	6	5
2	3	1	4	5	6	9	7	8
5	4	6	7	8	9	3	2	1
8	6	3	9	7	2	1	5	2
4	9	5	6	1	8	2	3	7
7	1	2	3	4	5	6	8	9

7	4	8	5	2	9	6	3	1
1	9	6	4	3	8	5	7	2
2	5	3	6	1	7	9	8	4
3	6	9	7	4	1	8	2	5
4	1	7	8	5	2	3	9	6
5	8	2	9	6	3	1	4	7
6	2	1	3	9	4	7	5	8
8	3	5	2	7	6	4	1	9
9	7	2	1	8	5	2	6	3

Abbildung 103: Bis auf Umnummerierung gleiche Sudokus

$5 \to 5$ erhält man wieder das Ausgangssudoku. □

66.1.3 Minimale Anzahl von Ziffern in einem eindeutigen Sudoku-Rätsel

Wir haben schon erwähnt, dass im Jahre 2011 nachgewiesen wurde, dass mindestens 17 Ziffern (auch Hinweise oder Clues genannt) in einem partiellen Sudoku vorgegeben sein müssen, damit eine *eindeutige* Ergänzung zu einem vollständigen Sudoku möglich ist. Dies ist durch einen Computerbeweis von G. McGuire et. al. (2012) erbracht worden. Im Prinzip wurden von G. McGuire und seinen Mitarbeitern alle 5472730538 im wesentlichen verschiedenen vollständigen Sudokus daraufhin durchsucht, ob es in ihnen partielle Sudokus mit 16 Hinweisen gibt, die sich auf eindeutige Weise vervollständigen lassen. In ihrem Aufsatz schreiben die Autoren:

- The strategy we used to finally solve this problem is an obvious one — exhaustively search through all possible solution grids, one by one, for a 16-clue puzzle. So we took the point of view of considering each particular completed sudoku grid one at a time, and then looking for puzzles whose solution is that particular grid. We think of these puzzles as being "contained in" that particular grid. Our search turned up no proper 16-clue puzzles, but had one existed, then we would have found it.

Weiter kann man in der Einleitung des Aufsatzes lesen:

- It is worth noting that there have been attempts to solve the minimum number of clues problem using mathematics only, i.e., not using a computer. However, nobody has made any serious progress. In fact, while it is very easy to see that a sudoku puzzle with seven clues will always have multiple completions, because the two missing digits can be interchanged in any solution, finding a theoretical reason why eight clues are not enough for a unique solution already seems hard. This is far from the conjectured answer of 17, so a purely mathematical solution of the minimum number of clues problem is a long way off.

Im Aufsatz von G. McGuire et al. kann man nachlesen, dass beim Computerbeweis einige mathematische Ideen umgesetzt wurden, damit das Problem überhaupt in einigermaßen angemessener Zeit behandelt werden konnte. Trotzdem war ein enormer Aufwand an Computerleistung nötig. Auf Einzelheiten wollen wir nicht mehr eingehen, sondern nur darauf hinweisen, dass von G. McGuire et al. ein bestimmtes vollständiges Sudoku angegeben wurde, dass genau 29 partielle Sudokus mit jeweils 17 Hinweisen enthält, was ein Rekord zu sein scheint.

Dagegen ist das entsprechende Problem für ein 4×4-Sudoku (manchmal auch *Shidoku* genannt) relativ einfach zu lösen. Zunächst bemerken wir, dass es partielle 4×4-Sudokus mit vier Hinweisen gibt, die sich eindeutig zu einem vollständigen 4×4-Sudoku ergänzen lassen. Ein Beispiel ist etwa

1			
		3	
	4		
			2

1	3	2	4
4	2	3	1
2	4	1	3
3	1	4	2

Rechts haben wir die eindeutige Vervollständigung angegeben. Ein partielles 4×4-Sudoku mit nur drei Hinweisen kann entweder nicht eindeutig oder überhaupt nicht vervollständigt werden. Dies kann man folgendermaßen einsehen, wobei wir die Argumentation von L. Taalman (2007) übernehmen. Wie wir uns in 66.1.2 überlegten, gibt es nur 2 im wesentlichen verschiedene vollständige 4×4-Sudokus. Repräsentanten dieser beiden sind z. B. die ersten beiden 4×4-Sudokus in Abbildung 102. In Abbildung 104 links betrachten wir das erste der Sudokus aus Abbildung 102. Die Ziffern

1	2	3	4
3	4	1	2
2	1	4	3
4	3	2	1

1	2	3	4
3	4	2	1
2	1	4	3
4	3	1	2

Abbildung 104: Es müssen vier Hinweise gegeben sein

in den Positionen $(1,1)$, $((1,3)$, $(2,1)$, $(2,3)$ sind gelb gefärbt, die in den Positionen $(1,2)$, $(1,4)$, $(2,2)$, $(2,4)$ rot, während die in den Positionen $(3,1)$, $(3,3)$, $(4,1)$, $(4,3)$ grün bzw. $(3,2)$, $(3,4)$, $(4,2)$, $(4,4)$ blau gefärbt sind. Aus jeder Farbgruppe muss wenigstens ein Hinweis gegeben sein, damit eine eindeutige Vervollständigung möglich ist. Das zeigt, dass mindestens vier Hinweise gegeben sein müssen. Entsprechend kann man auch für das andere, im wesentlichen verschiedene Sudoku argumentieren. Die entsprechenden Farben haben wir in Abbildung 104 rechts eingetragen.

Während ein Sudoku-Rätsel mit vielen vorgegebenen Ziffern bzw. Hinweisen i. Allg. einfach ist, ist eines mit wenigen Hinweisen, z. B. mit 17, nicht notwendig schwierig. Ein einfaches haben wir bei Gordon Royle[93] gefunden, siehe Abbildung 105. Es gibt Sudoku-

			7					
1								
			4	3		2		
								6
			5		9			
						4	1	8
				8	1			
		2					5	
	4					3		

2	6	4	7	1	5	8	3	9
1	3	7	8	9	2	6	4	5
5	9	8	4	3	6	2	7	1
4	2	3	1	7	8	5	9	6
8	1	6	5	4	9	7	2	3
7	5	9	6	2	3	4	1	8
3	7	5	2	8	1	9	6	4
9	8	2	3	6	4	1	5	7
6	4	1	9	5	7	3	8	2

Abbildung 105: Ein leichtes Sudoku-Rätsel mit 17 Hinweisen und seine Lösung

Rätsel mit 77 Hinweisen, welche zwei Lösungen haben. In Abbildung 106 geben wir ein solches an, in rot findet man die beiden möglichen Ergänzungen. In Abbildung 107 links geben wir ein partielles Sudoku mit einem leeren 5×6-Rechteck, rechts ein partielles Sudoku mit jeweils 3 leeren Quadraten, Zeilen und Spalten. Diese haben wir auf der

[93] http://theconversation.edu.au/good-at-sudoku-heres-some-youll-never-complete-5234

1	**2**	3	4	5	6	7	8	9
6	5	4	9	8	7	3	2	1
7	8	9	1	2	3	4	5	6
2	**1**	5	6	7	8	9	3	4
4	3	6	2	9	5	1	7	8
8	9	7	3	1	4	5	6	2
5	4	1	7	6	2	8	9	3
9	7	2	8	3	1	6	4	5
3	6	8	5	4	9	2	1	7

2	**1**	3	4	5	6	7	8	9
6	5	4	9	8	7	3	2	1
7	8	9	1	2	3	4	5	6
1	**2**	5	6	7	8	9	3	4
4	3	6	2	9	5	1	7	8
8	9	7	3	1	4	5	6	2
5	4	1	7	6	2	8	9	3
9	7	2	8	3	1	6	4	5
3	6	8	5	4	9	2	1	7

Abbildung 106: Ein Sudoku-Rätsel mit 77 Hinweisen und zwei Lösungen

		6	7		3	5		
				4				
5								2
9								7
	3						4	
8								1
1								4
	5	9	2	6	7	3	1	

1	4		9	3				
9	3		1					
5	6							4
8	9						2	7
			2	5			1	8
			7	4			3	2

Abbildung 107: Partielle Sudokus mit vielen leeren, zusammenhängenden Feldern

Wikipedia-Seite *Mathematics of Sudoku* gefunden, siehe `http://en.wikipedia.org/wiki/Mathematics_of_Sudoku`.

66.2 Verfahren zur Lösung eines Sudoku-Rätsels

Wir stellen zwei Verfahren vor, ein Sudoku-Rätsel mit einem Computer zu lösen. Es handelt sich hier jeweils um "brute force"-Methoden, die die Power eines Computers ausnutzen und keineswegs scharfsinnige Methoden anwenden, wie wir sie bei der Lösung eines Sudoku-Rätsels anwenden.

66.2.1 Sudoku und (binäre) lineare Optimierung

Wir stellen hier eine sehr einfache Methode vor, ein Sudoku-Rätsel als eine ganzzahlige bzw. binäre lineare Optimierungsaufgabe zu formulieren, genauer eine binäre Lösung (die Variablen sind nur 0 oder 1) eines Systems von linearen Gleichungen zu bestimmen. Daher ist eigentlich nur eine zulässige Lösung einer binären linearen Optimierungsaufgabe zu finden. Dies kann mit der in der Optimization Toolbox von MATLAB enthaltenen Funktion `bintprog` geschehen. Diese löst (wenn erfolgreich) binäre lineare Optimierungsaufgen also lineare Optimierungsaufgaben mit der Zusatzbedingung,

dass alle Variablen 0 oder 1 sind. Wir benutzen im Folgenden einen Aufsatz von A. C. Bartlett et al. (2008). Wir wollen Shidoku- und Sudoku-Rätsel gleichzeitig behandeln. Im Folgenden sei daher $n = 4$ (Shidoku) oder $n = 9$ (Sudoku). Wir benutzen die Variablen

$$x_{ijk} = \begin{cases} 1, & \text{falls Position } (i,j) \text{ im } n \times n\text{-Gitter mit } k \in \{1,\ldots,n\} \text{ besetzt ist,} \\ 0, & \text{sonst,} \end{cases}$$

wobei

$$(i,j,k) \in \{1,\ldots,n\} \times \{1,\ldots,n\} \times \{1,\ldots,n\}.$$

Als Vektor geschrieben hat x also n^3 Komponenten. Gegeben ist ferner eine Menge G von Tripeln (i,j,k), die Anzahl der Elemente von G ist gleich der Anzahl der Hinweise. Daher bedeutet $(i,j,k) \in G$, dass die Position (i,j) im $n \times n$-Gitter mit $k \in \{1,\ldots,n\}$ belegt ist. Alternativ können wir auch eine Matrix G vorgeben, die so viele Zeilen wie Hinweise und 3 Spalten besitzt. Für das Sudoku-Rätsel in Abbildung 95 ist z. B.

$$G = \{(1,4,6),(1,9,5),(2,2,2),\ldots,(9,5,4),(9,6,6),(9,9,3)\}$$

bzw.

$$G = \begin{pmatrix} 1 & 4 & 6 \\ 1 & 9 & 5 \\ 2 & 2 & 2 \\ \vdots & \vdots & \vdots \\ 9 & 5 & 4 \\ 9 & 6 & 6 \\ 9 & 9 & 3 \end{pmatrix}.$$

Als zu erfüllende Bedingungen an $x = (x_{ijk})$ haben wir:

- Es ist

$$x_{ijk} \in \{0,1\}, \qquad (i,j,k) \in \{1,\ldots,n\} \times \{1,\ldots,n\} \times \{1,\ldots,n\}.$$

- Es ist

$$\sum_{i=1}^{n} x_{ijk} = 1, \qquad (j,k) \in \{1,\ldots,n\} \times \{1,\ldots,n\}.$$

  Dies bedeutet, dass es für jede Spalte $j \in \{1,\ldots,n\}$ und jede Ziffer $k \in \{1,\ldots,n\}$ genau eine Zeile $i \in \{1,\ldots,n\}$ mit $x_{ijk} = 1$ gibt, die Position (i,j) also mit k besetzt ist.

- Es ist

$$\sum_{j=1}^{n} x_{ijk} = 1, \qquad (i,k) \in \{1,\ldots,n\} \times \{1,\ldots,n\}.$$

  Dies bedeutet, dass es für jede Zeile $i \in \{1,\ldots,n\}$ und jede Ziffer $k \in \{1,\ldots,n\}$ genau eine Spalte $j \in \{1,\ldots,n\}$ mit $x_{ijk} = 1$ gibt, die Position (i,j) also mit k besetzt ist.

- Jetzt wird es etwas komplizierter. Wir haben n Teilquadrate der Größe $\sqrt{n} \times \sqrt{n}$ und in jedem dieser Teilquadrate muss jede Ziffer $k \in \{1, \ldots, n\}$ genau einmal vertreten sein. Wir setzen $m := \sqrt{n}$. Das $n \times n$-Gitter besitzt m Zeilen- und m Spaltenblöcke. Mit $(p, q) \in \{1, \ldots, m\} \times \{1, \ldots, m\}$ besitzt das $m \times m$-Unterquadrat im p-ten Zeilen- und q-ten Spaltenblock die Positionen (i, j) mit $i = (m-1)p + 1, \ldots, mp$, $j = (m-1)q + 1, \ldots, mq$. Die Forderung, dass jedes $k \in \{1, \ldots, n\}$ in jedem der Unterquadrate genau einmal vorkommt, kann also formuliert werden als

$$\sum_{i=(m-1)p+1}^{mp} \sum_{j=(m-1)q+1}^{mq} x_{ijk} = 1, \qquad (p, q, k) \in \{1, \ldots, m\} \times \{1, \ldots, m\} \times \{1, \ldots, n\}.$$

- Es ist

$$\sum_{k=1}^{n} x_{ijk} = 1, \qquad (i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}.$$

  Dies bedeutet, dass jede Position (i, j) im $n \times n$-Gitter mit einer Zahl $k \in \{1, \ldots, n\}$ besetzt ist.

- Es ist

$$x_{ijk} = 1, \qquad (i, j, k) \in G.$$

  Hierdurch werden die vorgegebenen Hinweise berücksichtigt.

Diese Bedingungen können als ein lineares Gleichungssystem $A_0 x = b_0$ geschrieben werden, zu dem eine binäre Lösung gesucht wird, also eine Lösung mit Komponenten aus $\{0, 1\}$. Bei dieser ganz naiven Vorgehensweise ist $x \in \{0, 1\}^{n^3}$ und A_0 eine Matrix mit $4n^2 + |G|$ Zeilen und n^3 Spalten, ferner b_0 ein (Spalten-) Vektor, dessen $4n^2 + |G|$ Komponenten sämtlich gleich 1 sind. Die Funktion `bintprog` aus der Optimization Toolbox von MATLAB löst lineare, binäre Optimierungsaufgaben der Form

$$\text{Minimiere} \quad f^T x \quad \text{unter den Nebenbedingungen} \quad \begin{cases} Ax \leq b, \\ A_0 x = b_0, \\ x \text{ binär.} \end{cases}$$

Der entsprechende Aufruf geschieht z. B. durch

```
x=bintprog(f,A,b,A_0,b_0)
```

In unserem Falle (wir haben nur Gleichungen und die Zielfunktion ist irrelevant bzw. kann als Nullfunktion gewählt werden) ist der Aufruf also z. B.

```
x=bintprog(zeros(n^3,1),[],[],A_0,b_0)
```

nachdem A_0 und b_0 entsprechend besetzt sind. Hiermit haben wir ohne Schwierigkeiten das schwere Sudoku-Rätsel aus der (heutigen) Süddeutschen Zeitung vom 22. August 2012 lösen können, auch das Sudoku-Rätsel aus Abbildung 95 war kein Problem.

Man kann hiermit sogar sogenannte *diabolische Sudokus* lösen, bei welchen man zu einem bestimmten Zeitpunkt raten muss und die gewählte Entscheidung gegebenenfalls wieder rückgängig machen muss. Einige dieser "gemeinen" Sudoku-Rätsel stammen von Michael Mepham und wurden in London's *Daily Telegraph* veröffentlicht, so auch das in Abbildung 108 (siehe `http://entertainment.howstuffworks.com/leisure/brain-games/sudoku4.htm`). Ein Nachteil der hier geschilderten Vorgehensweise ist

			9	2				
		6	8		3			
1	9			7				6
2	3			4		1		
		1				7		
		8		3			2	9
7				8			9	1
			5		7	2		
				6	4			

3	8	7	9	2	6	4	1	5
5	4	6	8	1	3	9	7	2
1	9	2	4	7	5	8	3	6
2	3	5	7	4	9	1	6	8
9	6	1	2	5	8	7	4	3
4	7	8	6	3	1	5	2	9
7	5	4	3	8	2	6	9	1
6	1	3	5	9	7	2	8	4
8	2	9	1	6	4	3	5	7

Abbildung 108: Ein diabolisches Sudoku von Michael Mepham und seine Lösung

natürlich, dass die MATLAB-Funktion `bintprog` benutzt wird. Der hierbei verwendete Algorithmus wird in `http://www.mathworks.de/help/toolbox/optim/ug/brnox9y.html#f786287` geschildert, man kann ihn sich sogar in MATLAB mit Hilfe von `type bintprog` ansehen.

Bemerkung: Wir haben uns oben überlegt, dass wir die Lösung eines Sudoku-Rätsels auf die Bestimmung einer binären Lösung eines überbestimmten linearen Gleichungssystems $A_0x = b_0$ zurückführen können. Statt eine binäre Lösung von $A_0x = b_0$ (z. B. mit `bintprog`) zu berechnen, wird von P. Babu, K. Pelckmans, P. Stoica, J. Li (2010) vorgeschlagen, die bezüglich der L_1-Norm kleinste Lösung des linearen Gleichungssystems zu bestimmen, also die Optimierungsaufgabe

$$(P_1) \qquad \text{Minimiere} \quad \|x\|_1 \quad \text{unter der Nebenbedingung} \quad A_0x = b_0$$

zu lösen. Hierbei ist $\|\cdot\|_1$ die Betragssummen-Norm. Von Babu et al. wird nicht behauptet, dass eine Lösung von (P_1) auch Lösung des gegebenen Sudokus ist oder Bedingungen dafür angegeben, wann dies der Fall ist, sondern lediglich gesagt: "(P_1) solves most Sudoku puzzles". Diese Aussage wollen wir überprüfen. Hierzu wird (P_1) zunächst eine äquivalente lineare Optimierungsaufgabe zugeordnet, nämlich

$$(Q_1) \qquad \begin{cases} \text{Minimiere} \quad \begin{pmatrix} e \\ e \end{pmatrix}^T \begin{pmatrix} u \\ v \end{pmatrix} \quad \text{unter den Nebenbedingungen} \\ \qquad \begin{pmatrix} u \\ v \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} A_0 & -A_0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = b_0. \end{cases}$$

Hierbei ist e einmal wieder der Vektor, dessen Komponenten sämtlich gleich 1 sind. Die Probleme (P_1) und (Q_1) sind im folgenden Sinne äquivalent:

- Ist x^* eine Lösung von (P_1) und definiert man (u^*, v^*) durch $u^* := \max(x^*, 0)$ und $v^* := \max(-x^*, 0)$, so ist (u^*, v^*) eine Lösung von (Q_1).

Denn: Es ist $(u^*, v^*) \geq (0, 0)$, $x^* = u^* - v^*$ und daher (u^*, v^*) zulässig für (Q_1). Weiter ist $|x^*| = u^* + v^*$ und daher $\|x^*\|_1 = e^T(u^* + v^*)$. Sei ein beliebiges für (Q_1) zulässiges Element (u, v) vorgegeben. Dann ist $x := u - v$ zulässig für (P_1), daher

$$e^T(u^* + v^*) = \|x^*\|_1 \leq \|x\|_1 \leq e^T(u + v)$$

und folglich (u^*, v^*) eine Lösung von (Q_1).

- Ist (u^*, v^*) eine Lösung von (Q_1) und definiert man $x^* := u^* - v^*$, so ist x^* eine Lösung von (P_1).

Denn: Es ist $A_0x^* = A_0(u^* - v^*) = b_0$, also x^* zulässig für (P_1). Sei ein beliebiges für (P_1) zulässiges Element x vorgegeben. Man definiere $u := \max(x, 0)$, $v := \max(-x, 0)$. Dann ist (u, v) für (Q_1) zulässig, $x = u - v$, $|x| = u + v$ und $\|x\|_1 = e^T(u + v)$. Folglich ist

$$\|x\|_1 = e^T(u + v) \geq e^T(u^* + v^*) \geq \|x^*\|_1,$$

also x^* eine Lösung von (P_1).
Daher können wir (P_1) dadurch lösen, dass wir z. B. mit der MATLAB-Funktion `linprog` eine Lösung (u^*, v^*) der linearen Optimierungsaufgabe (P_2) bestimmen und anschließend $x^* := u^* - v^*$ setzen. Sind A_0 und b_0 wie oben besetzt, so lösen wir (P_1) also z. B. durch

```
c=[ones(n^3,1);ones(n^3,1)];
l=[zeros(n^3,1);zeros(n^3,1)];A=[A_0  -A_0];
y=linprog(c,[],[],A,b_0,l,[]);u=y(1:n^3);v=y(n^3+1:2*n^3);
x=u-v;x=round(x);
```

Das Sudoku-Rätsel aus Abbildung 95 wird hierdurch erfolgreich gelöst. Auch das (leichte) Sudoku-Rätsel mit nur 17 Hinweisen aus Abbildung 105 konnte auf die selbe Weise gelöst werden, ferner auch das schwere Suduko-Rätsel in der Süddeutschen Zeitung vom 25./26. August 2012, siehe Abbildung 109. Erfolgreich (und zwar wesentlich schneller als bei der Anwendung von `bintprog`) waren wir bei diesen drei Sudoku-Rätseln auch mit einer noch einfacheren linearen Optimierungsaufgabe, nämlich

(P_2) Minimiere e^Tx unter den Nebenbedingungen $x \geq 0,\ A_0x = b_0$.

Zur Motivation von (P_2) beachten wir, dass $\|x\|_1 = e^Tx$ für nichtnegative x. Keinen Erfolg hatten wir aber bei der Anwendung von (P_1) und (P_2) auf das diabolische Sudoku aus Abbildung 108. Es wäre interessant, Bedingungen dafür anzugeben, dass eine Lösung von (P_1) oder (P_2) binär ist und daher eine Lösung des gegebenen Sudoku-Rätsels liefert. □

	2		8		9	5		
			4			3		
4							1	
			9	4				
	7				5	2		6
					8	9		
1				5		4		
	5	8						
					6			9

7	2	1	8	3	9	5	6	4
6	8	5	4	2	1	3	9	7
4	9	3	5	6	7	8	1	2
8	1	6	9	4	2	7	3	5
9	7	4	3	1	5	2	8	6
5	3	2	6	7	8	9	4	1
1	6	9	7	5	3	4	2	8
2	5	8	1	9	4	6	7	3
3	4	7	2	8	6	1	5	9

Abbildung 109: Ein schweres Sudoku und seine Lösung

66.2.2 Lösung eines Sudoku-Rätsels durch rekursives Backtracking

Von Cleve Moler, dem Entwickler von MATLAB, stammt eine MATLAB-Implementation eines Sudoku-Lösers[94], der auf *rekursivem Backtracking* basiert. Diesen Zugang wollen wir hier schildern. Wir illustrieren das Vorgehen anhand eines leichten Sudoku-Rätsels aus der Süddeutschen Zeitung vom 30.8. 2012, siehe Abbildung 110. Hierbei

				5		6	9	3
6	4				1			
	2							
		7	5					
1			9		4	8	2	
9	8	2	1	3				
		6				9	3	4
	3		8			2		
								7

0	0	0	0	5	0	6	9	3
6	4	0	0	0	1	0	0	0
0	2	0	0	0	0	0	0	0
0	0	7	5	0	0	0	0	0
1	0	0	9	0	4	8	2	0
9	8	2	1	3	0	0	0	0
0	0	6	0	0	0	9	3	4
0	3	0	8	0	0	2	0	0
0	0	0	0	0	0	0	0	7

Abbildung 110: Ein leichtes Sudoku-Rätsel

haben wir in Abbildung 110 rechts im partiellen Sudoku die Zellen ohne (roten) Hinweis mit einer 0 gefüllt. Die entscheidende Funktion in der Implementation von Cleve Moler ist die Funktion

```
function [C,N]=candidates(X)
```

Der Input `X` ist ein partielles Sudoku bzw. ein 9×9-Gitter, bei welchem nicht besetzte Zellen mit 0 gefüllt sind. Der Output `[C,N]` besteht zum einen aus dem 9×9-*Cell Array*

[94] Siehe `http://www.mathworks.de/company/newsletters/news_notes/2009/clevescorner.html` und `http://www.mathworks.de/moler/exm/chapters.html` (hier gehe man auf *Individual Files* und sehe sich die sudoku-Files an).

C und dem Zeilenvektor N mit 81 Komponenten. Hierbei gibt C{i,j} für $X(i,j) = 0$ die Menge der Kandidaten für die Position (i,j) an, während C{i,j}=[], wenn $X(i,j) > 0$ (der Wert in der Position (i,j) also schon bestimmt ist). Man erhält als Cell Array der möglichen Kandidaten die in Abbildung 111 angegebenen Werte. Hierbei ist die Menge

7 8	1 7	1 8	2 4 7	[ ]	2 7 8	[ ]	[ ]	[ ]
[ ]	[ ]	3 5 8 9	2 3 7	2 7 8 9	[ ]	5 7	5 7 8	2 5 8
3 5 7 8	[ ]	1 3 5 8 9	3 4 6 7	4 6 7 8 9	3 6 7 8 9	1 4 5 7	1 4 5 7 8	1 5 8
3 4	6	[ ]	[ ]	2 6 8	2 6 8	1 3 4	1 4 6	1 6 9
[ ]	5 6	3 5	[ ]	6 7	[ ]	[ ]	[ ]	5 6
[ ]	[ ]	[ ]	[ ]	[ ]	6 7	4 5 7	4 5 6 7	5 6
2 5 7 8	1 5 7	[ ]	2 7	1 2 7	2 5 7	[ ]	[ ]	[ ]
4 5 7	[ ]	1 4 5 9	[ ]	1 4 6 7 9	5 6 7 9	[ ]	1 5 6	1 5 6
2 4 5 8	1 5 9	1 4 5 8 9	2 3 4 6	1 2 4 6 9	2 3 5 6 9	1 5	1 5 6 8	[ ]

Abbildung 111: Cell Array der Kandidaten

der Kandidaten für die Position $(i,j) \in \{1,\ldots,9\} \times \{1,\ldots,9\}$ relativ einfach berechnet worden, wobei dies natürlich nur dann geschehen muss, wenn $X(i,j) = 0$ (andernfalls ist schon der richtige Wert berechnet). Und zwar werden als Kandidaten für die Position (i,j) alle Zahlen von 1 bis 9 angegeben, die nicht schon in der i-ten Zeile, der j-ten Spalte bzw. dem entsprechenden Unterquadrat vorkommen, eingetragen. Dies geschieht dadurch, dass eine leere Zelle mit z=1:9 startet und alle Elemente in z auf Null gesetzt werden, deren Zahlenwert in der entsprechenden Zeile, Spalte bzw. Unterquadrat vorkommen. Die verbleibenden Nicht-Nullwerte sind die Kandidaten. Wir verzichten darauf, eine feinere Strategie zu entwickeln. Z. B. ist in unserem Beispiel die 5 eigentlich kein Kandidat für die Position $(5,9)$ und die 8 kein Kandidat für die Positionen $(9,1)$ und $(9,3)$, wie man unschwer erkennt. N erhält man aus C dadurch, dass man die Matrix $(|C\{i,j\}|)$ der Anzahl der Elemente von C bildet, dann die Einträge mit $C\{i,j\} = 0$ bzw. $X(i,j) > 0$ (hier steht in X schon die richtige Ziffer) gleich Inf (gleich ∞) setzt, anschließend zu einem Vektor übergeht, indem man die Spalten1 bis 9 hintereinander schreibt und zum Schluss zu einem Zeilenvektor übergeht. Bei Cleve Moler sieht die Implementation der Funktion candidates daher folgendermaßen aus:

```
function [C,N]=candidates(X);
% Input:     X ist ein partielles Sudoku, bei dem nicht
%               besetzte Zellen mit 0 gefüllt sind
% Output:    C ist ein 9x9-Cell Array, wobei C{i,j} die Menge
%                der Kandidaten für X(i,j) ist.
%          N ist ein Zeilenvektor mit 81 Komponenten, wobei
%                          |C{k}|,          wenn X(k)=0,
%             N(k)=
%                          Inf,             wenn X(k)>0.
     tri=@(k) 3*ceil(k/3-1)+(1:3);
```

```
% tri(1)=tri(2)=tri(3)=1 2 3
% tri(4)=tri(5)=tri(6)=4 5 6
% tri(7)=tri(8)=tri(9)=7 8 9
     C=cell(9,9);  %C ist 9x9-Cell Array mit leeren Zellen
     for j=1:9
         for i=1:9
            if X(i,j)==0
                 z=1:9;
                 z(nonzeros(X(i,:)))=0;
                 z(nonzeros(X(:,j)))=0;
                 z(nonzeros(X(tri(i),tri(j))))=0;
                 C{i,j}=nonzeros(z)';
             end
          end
      end
      N=cellfun(@length,C);  N(X>0)=Inf;    N=N(:)';
 end  %candidates
```

Jetzt kommen wir zu der eigentlichen Funktion `sudoku`. Solange man beim Aufruf von `[C,N]=candidates(X)` die Information erhält, dass es für jede noch nicht besetzte Zelle Kandidaten gibt (`all(N>0)`) und es Zellen mit genau einem Element, sogenannte *singletons*, gibt (`any(N==1)`), wird man `X` entsprechend verändern. Hier ist der einzige Unterschied unserer Version gegenüber dem Original von Cleve Moler. Während Moler immer nur *ein* singleton aufnimmt und anschließend die Kandidaten neu berechnet, nehmen wir *alle* singletons auf, wobei wir mit Hilfe der Funktion `cell2mat` Cell Arrays in Matrizen umwandeln. Gibt es dagegen für jede noch nicht besetzte Zelle mehr als einen Kandidaten, so wählt man eine Zelle mit einer minimalen Anzahl von Kandidaten und iteriert anschließend über die Kandidaten. Insgesamt erhalten wir (bzw. Cleve Moler):

```
function [X,steps]=sudoku(X,steps)
%Input:      X ist ein partielles Sudoku, bei dem nicht
%             besetzte Zellen mit 0 gefüllt sind
%Output:    Wenn erfolgreich, ist X ein vollständiges
%               Sudoku, und zwar eine Ergänzung des Inputs.
%               steps gibt die benötigten Schritte an
%Benutzt wird rekursives Backtracking
%Die Funktion stammt von Cleve Moler
if nargin<2
      steps=0;
end;
[C,N]=candidates(X);
while all(N>0) & any(N==1)
   s=find(N==1);                    %Finde alle singletons
   X(s)=cell2mat(C(s));             %Trage diese in X ein
   steps=step+1;
```

```
   [C,N]=candidates(X);
end;
if all(N>0)
   Y=X;
   s=find(N==min(N),1);
   for t=[C{s}]                    %Iteriere über die Kandidaten
       X=Y;
       X(s)=t;                     %Trage Probewert ein
       steps=steps+1;
       X=sudoku(X);
       if all(X(:)>0)              %Lösung gefunden
          break
       end
   end
end %sudoku
```

Da die Funktion `candidates` wohl nur in `sudoku` auftreten wird, hätte man sie auch an diese "anhängen" können. So geschieht es bei Cleve Moler. Das diabolische Sudoku-Rätsel aus Abbildung 108 wurde mit 47 Schritten in 0.4221 Sekunden gelöst.

67 Die schnelle Fourier-Transformation

Die schnelle Fourier-Transformation ist eine sehr clevere Methode, ein bestimmtes $n \times n$-Matrix mal n-Vektor Produkt, welches naiv durchgeführt n^2 Multiplikationen benötigt, wesentlich effizienter zu berechnen, was an der speziellen Struktur der Matrix liegt. Im Vorwort von C. Van Loan (1992) kann man die folgenden bemerkenswerten Sätze über die schnelle Fourier-Transformation (Fast Fourier Transform, FFT) finden:

- The fast Fourier transform (FFT) is one of the truly great computational developments of this century. It has changed the face of science and engineering so much that it is not an exaggeration to say that *life as we know it would be very different without the FFT*.

Uns kommt es hier nicht darauf an, die Einsatzmöglichkeiten der schnellen Fourier-Transformation (wir benutzen im Folgenden die Abkürzung FFT) z. B. in der Signalverarbeitung oder bei der Bildkompression zu schildern und damit die gerade eben zitierte Bedeutung der FFT zu untermauern, sondern wenigstens in dem Spezialfall, dass n eine Zweierpotenz ist, eine möglichst einfache Darstellung der FFT zu geben. Wir zitieren wieder aus dem Vorwort von C. Van Loan (1992):

- Unfortunately, the simplicity and intrinsic beauty of many FFT ideas are buried in research papers that are rampant with vectors of subscripts, multiple summations, and poorly specified recursions. The poor mathematical and algorithmic notation has retarded progress and has led to a literature of duplicated results.

Bei gegebenem $n \in \mathbb{N}$ sei

$$\omega_n := \exp(-2\pi i/n).$$

Für das Folgende ist es wichtig, dass ω_n^k, $k = 0, \ldots, n-1$, die n-ten Einheitwurzeln sind. Beginnend mit $\omega_n^0 = 1$ sind die weiteren Einheitswurzeln $\omega_n^1, \ldots, \omega_n^{n-1}$ auf dem Einheitskreis im Uhrzeigersinn angeordnet. Die *diskrete Fourier-Transformation der Länge n* (kurz: DFT) $F_n \colon \mathbb{C}^n \longrightarrow \mathbb{C}^n$ ordnet einem Vektor $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$ den Vektor $y = F_n x$ zu, dessen Komponenten durch

$$y_j := \sum_{k=0}^{n-1} \omega_n^{jk} x_k, \qquad j = 0, \ldots, n-1,$$

gegeben sind. Der Vektor y heißt die *diskrete Fourier-Transformierte* von x. Wir werden nicht unterscheiden zwischen der gerade eben definierten Abbildung

$$F_n \colon \mathbb{C}^n \longrightarrow \mathbb{C}^n$$

und der Matrix

$$F_n := (\omega_n^{jk})_{0 \le j,k \le n-1} \in \mathbb{C}^{n \times n}.$$

Eine naive Berechnung der diskreten Fourier-Transformierten $y = F_n x$ eines gegebenen Vektors $x \in \mathbb{C}^n$ erfordert n^2 komplexe Multiplikationen[95]. Mit Hilfe der FFT ist es möglich, jedenfalls für den Fall, dass $n = 2^p$ eine Zweierpotenz ist, diese Anzahl auf $\frac{1}{2}np = \frac{1}{2}n \log_2 n$ zu drücken, was schon für moderat große n eine wesentliche Verbesserung bedeutet, siehe Tabelle 9. Da ein Tag aus 86400 Sekunden besteht, bedeutet dies:

p	$n = 2^p$	$n^2/(\frac{1}{2}n \log_2 n)$
5	32	≈ 13
10	1024	≈ 205
15	32768	≈ 4369
20	2048576	≈ 104858

Tabelle 9: Effizienzgewinn durch die FFT

Benötigt man eine Sekunde zur Berechnung der FFT einer Länge $n = 2^{20}$, so benötigt man mehr als einen Tag um die diskrete Fourier-Transformierte $F_n x$ auf konventionelle Weise zu berechnen. Die FFT ist von J. W. Cooley, J. W. Tukey (1965) angegeben worden. Sie gehört ganz sicher zu den anwendungsreichsten Verfahren der numerischen Mathematik.

Bemerkung: Mit $n \in \mathbb{N}$ sei $t_k := 2\pi k/n$, $k = 0, \ldots, n-1$. Bei gegebenem $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$, $c := (1/n) F_n x$ und

$$p(t) := \sum_{j=0}^{n-1} c_j \exp(ijt)$$

ist $p(t_k) = x_k$, $k = 0, \ldots, n-1$, also $p(\cdot)$ ein *komplexes trigonometrisches Polynom vom Grad* $\le n-1$, welches an den Stützstellen t_k den Wert x_k annimmt, $k = 0, \ldots, n-1$.

[95] Man könnte noch berücksichtigen, dass die erste Zeile und die erste Spalte von F_n aus Einsen besteht. Das macht den Kohl aber auch nicht fett.

Dies liegt daran, dass

$$(*) \qquad F_n^H F_n = nI_n,$$

wobei I_n die $n \times n$-Einheitsmatrix und F_n^H die konjugierte, transponierte Matrix zu F_n bezeichnet[96]. Zum Nachweis von $(*)$ beachten wir, dass

$$(F_n^H F_n)_{pq} = \sum_{k=0}^{n-1} (F_n^H)_{pk}(F_n)_{kq} = \sum_{k=0}^{n-1} \overline{\omega_n}^{kp}\omega_n^{kq} = \sum_{k=0}^{n-1} \omega_n^{(q-p)k}.$$

Ist $p = q$, so ist folglich $(F_n^H F_n)_{pq} = n$. Für $0 \le p, q \le n-1$ und $p \ne q$ ist $\omega_n^{q-p} \ne 1$ und daher

$$(F_n^H F_n)_{pq} = \sum_{k=0}^{n-1} \omega_n^{(q-p)k} = \sum_{k=0}^{n-1} (\omega_n^{q-p})^k = \frac{1-\omega_n^{(q-p)n}}{1-\omega_n^{q-p}} = 0.$$

Damit ist $(*)$ bewiesen. Mit $x \in \mathbb{C}^n$ und $c := (1/n)F_n x$ ist

$$F_n^H c = \frac{1}{n} F_n^H F_n x = x,$$

was gleichbedeutend mit der oben angegebenen Interpolationseigenschaft ist. □

Die grundlegende Idee der FFT besteht in der Anwendung einer "Teile und Herrsche"- bzw. "Divide et impera"-Strategie, welche darin besteht, eine große Aufgabe in kleinere zu zerlegen und die Lösungen dieser kleineren Probleme zu einer Lösung des Gesamtproblems zusammenzusetzen. Wir werden zeigen, dass wir die Berechnung der diskreten Fourier-Transformierten eines Vektors $x \in \mathbb{C}^n$ in einem ersten Schritt auf die Berechnung von zwei diskreten Fourier-Transformierten der halben Länge zurückführen kann. Ist $n = 2^p$ eine Zweierpotenz, und davon werden wir ausgehen, so kann dieser Prozess fortgesetzt werden, bis man im p-ten Schritt n diskrete Fourier-Transformationen der Länge 1 durchführt. Entscheidend ist (siehe Theorem 1.2.1 bei C. Van Loan (1992))

Lemma *Sei $n = 2m$ eine gerade natürliche Zahl, $\omega_n := \exp(-2\pi i/n)$ und*

$$F_n := (\omega_n^{jk})_{0 \le j,k \le n-1}.$$

Sei weiter $\Pi_n \in \mathbb{R}^{n \times n}$ die Permutationsmatrix

$$\Pi_n := \left(\begin{array}{cccc|cccc} e_1 & e_3 & \cdots & e_{n-1} & e_2 & e_4 & \cdots & e_n \end{array} \right),$$

wobei e_j der j-te Einheitsvektor im $\mathbb{R}^n$ ist. Mit $\Omega_m := \operatorname{diag}(\omega_{2m}^0, \omega_{2m}^1, \cdots, \omega_{2m}^{m-1})$ ist

$$F_n \Pi_n = \left(\begin{array}{c|c} F_m & \Omega_m F_m \\ \hline F_m & -\Omega_m F_m \end{array} \right) = \left(\begin{array}{c|c} I_m & \Omega_m \\ \hline I_m & -\Omega_m \end{array} \right) \left(\begin{array}{c|c} F_m & 0_m \\ \hline 0_m & F_m \end{array} \right).$$

Hierbei bedeuten I_m bzw. 0_m die $m \times m$-Einheits- bzw. Nullmatrix.

Beweis: Durch Multiplikation von F_n von rechts mit der Permutationsmatrix Π_n werden die Spalten von F_n permutiert. Zunächst kommt die erste Spalte und dann die

[96] Ist allgemein $A = (a_{jk}) \in \mathbb{C}^{n \times n}$, so ist $A^H := (\overline{a_{kj}})$. Also geht A^H dadurch aus A hervor, dass man A transponiert und zu den konjugiert komplexen Einträgen übergeht.

weiteren ungeraden Spalten, danach die zweite und die weiteren geradzahligen Spalten. Wegen

$$F_n = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n^1 & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)(n-1)} \end{pmatrix}$$

ist

$$F_n\Pi_n = \left(\begin{array}{cccc|cccc} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n^2 & \cdots & \omega_n^{n-2} & \omega_n^1 & \omega_n^3 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^4 & \cdots & \omega_n^{2(n-2)} & \omega_n^2 & \omega_n^6 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)(n-2)} & \omega_n^{n-1} & \omega_n^{(n-1)3} & \cdots & \omega_n^{(n-1)(n-1)} \end{array}\right).$$

Ist $0 \le j, k < m$, so ist

$$\begin{array}{lcllcll} (F_n\Pi_n)_{jk} & = & \omega_n^{j(2k)} & = & \omega_m^{jk} & = & (F_m)_{jk}, \\ (F_n\Pi_n)_{j+m,k} & = & \omega_n^{(j+m)(2k)} & = & \omega_m^{(j+m)k} & = & (F_m)_{jk}, \\ (F_n\Pi_n)_{j,k+m} & = & \omega_n^{j(2k+1)} & = & \omega_n^j\omega_m^{jk} & = & (\Omega_m F_m)_{jk}, \\ (F_n\Pi_n)_{j+m,k+m} & = & \omega_n^{(j+m)(2k+1)} & = & -\omega_n^{j(2k+1)} & = & (-\Omega_m F_m)_{jk}. \end{array}$$

Hierbei haben wir ausgenutzt, dass $\omega_n^2 = \omega_m$ und $\omega_n^m = -1$. Diese Gleichungen bestätigen, dass die vier $m \times m$-Blocks von $F_n\Pi_n$ die angebene Form haben. Das Lemma ist bewiesen. □

Wir benutzen im Folgenden, wie C. Van Loan (1992), eine MATLAB-ähnliche[97] Notation. Ist $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$ mit geradem $n = 2m$, so sei z. B.

$$x(0:2:n-1) := (x_0, x_2, \ldots, x_{n-2})^T, \qquad x(1:2:n-1) := (x_1, x_3, \ldots, x_{n-1})^T.$$

Wie in MATLAB soll weitgehend mit Matrizen und Vektoren, nicht so sehr mit ihren Einträgen bzw. Komponenten operiert werden. Für $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$ ist dann

$$\Pi_n^T x = \begin{pmatrix} x(0:2:n-1) \\ x(1:2:n-1) \end{pmatrix}.$$

Mit den Bezeichnungen des letzten Lemmas und geradem $n = 2m$ ist unter Berücksichtigung von $\Pi_n\Pi_n^T = I_n$ daher

$$F_n x = \begin{pmatrix} I_m & \Omega_m \\ I_m & -\Omega_m \end{pmatrix} \begin{pmatrix} F_m & 0_m \\ 0_m & F_m \end{pmatrix} \Pi_n^T x$$

[97] Diese Notation ist nur MATLAB-*ähnlich*, da in MATLAB nur *positive* Indices vorkommen dürfen. In einer Realisierung der FFT in MATLAB müssen daher die Indices von 1 bis n und nicht von 0 bis $n-1$ laufen.

$$
\begin{aligned}
&= \begin{pmatrix} I_m & \Omega_m \\ I_m & -\Omega_m \end{pmatrix} \begin{pmatrix} F_m & 0_m \\ 0_m & F_m \end{pmatrix} \begin{pmatrix} x(0:2:n-1) \\ x(1:2:n-1) \end{pmatrix} \\
&= \begin{pmatrix} I_m & \Omega_m \\ I_m & -\Omega_m \end{pmatrix} \begin{pmatrix} F_m x(0:2:n-1) \\ F_m x(1:2:n-1) \end{pmatrix} \\
&= \begin{pmatrix} F_m x(0:2:n-1) + \Omega_m F_m x(1:2:n-1) \\ F_m x(0:2:n-1) - \Omega_m F_m x(1:2:n-1) \end{pmatrix}.
\end{aligned}
$$

Man erkennt hieran, dass man eine diskrete Fourier-Transformierte der Länge $n = 2m$ durch die Berechnung von zwei diskreten Fourier-Transformierten der halben Länge m erhalten kann. Berücksichtigt man noch, dass eine DFT der Länge 1 die Identität ist, so kann man leicht eine *rekursive* Fassung der FFT angeben, z. B. in MATLAB. Eine sehr einfache Version könnte z. B. folgendermaßen aussehen:

```
function y=RekFFT(x);
%Input:  x ist ein n=2^p-Spaltenvektor
%Output: y ist die diskrete Fourier-Transformierte von x
n=length(x); m=n/2;
if n==1
   y=x;
else
   ung_ind=1:2:n;x_ung=x(ung_ind);
   ger_ind=2:2:n;x_ger=x(ger_ind);
   x_ung=RekFFT(x_ung); x_ger=RekFFT(x_ger);
   Omega_m=(exp(-2*pi*i*[0:m-1]'/n));
   z=Omega_m.*x_ger;   y=[x_ung+z;x_ung-z];
end;
```

Unser Ziel wird es sein, eine nichtrekursive und damit wesentlich effizientere Formulierung der FFT darzustellen. Die Anzahl $f(n)$ der benötigten Multiplikationen zur Berechnung von $F_n x$ kann leicht abgeleitet werden, wobei davon ausgegangen wird, dass die Potenzen $\omega_n^0, \ldots, \omega_n^{n-1}$ im Vorfeld berechnet wurden. Denn wegen obiger Darstellung von $F_n x$ ist offenbar

$$f(n) = 2f\Big(\frac{n}{2}\Big) + \frac{n}{2}.$$

Wegen

$$F_2 \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & \omega_2^1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} x_0 + x_1 \\ x_0 + \omega_2^1 \cdot x_1 \end{pmatrix}$$

ist $f(2) = 1$, wobei wir *nicht* ausnutzen, dass $\omega_1^1 = -1$. Für alle $p \in \mathbb{N}$ ist folglich

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

wie man leicht durch vollständige Induktion nach p zeigt. Denn für $p = 1$ ist die Behauptung richtig. Sie sei auch für $p - 1$ richtig. Dann ist

$$f(2^p) = 2f(2^{p-1}) + 2^{p-1} = 2\Big[\frac{1}{2} 2^{p-1}(p-1)\Big] + 2^{p-1} = \frac{1}{2} 2^p p,$$

die Behauptung ist bewiesen. Ist also $n = 2^p$ eine Zweierpotenz, so ist die Anzahl benötigter Multiplikationen zur Berechnung der diskreten Fourier-Transformierten $F_n x$ mit der obigen rekursiven Fassung der FFT gleich $\frac{1}{2}2^p p = \frac{1}{2} n \log_2 n$.

Beispiel: Sei $n = 2^4 = 16$ und $x = x(0 : 1 : 15) \in \mathbb{C}^{16}$, zu berechnen sei $F_{16}x$. Wir berechnen $F_{16}x$ mit $F_8 x(0 : 2 : 15)$ und $F_8 x(1 : 2 : 15)$. Mit $F_4 x(0 : 4 : 15)$ und $F(2 : 4 : 15)$ erhält man den ersten Wert, den zweiten mit $F_4 x(1 : 4 : 15)$ sowie $F_4 x(3 : 4 : 15)$. Dies kann man fortsetzen, bis man im letzten Schritt triviale diskrete Fourier-Transformationen der Länge 1 durchführt. In Abbildung 112 wird der erhaltene Baum angegeben. Man erkennt, dass die "Blätter" dieses Baumes die Komponenten des

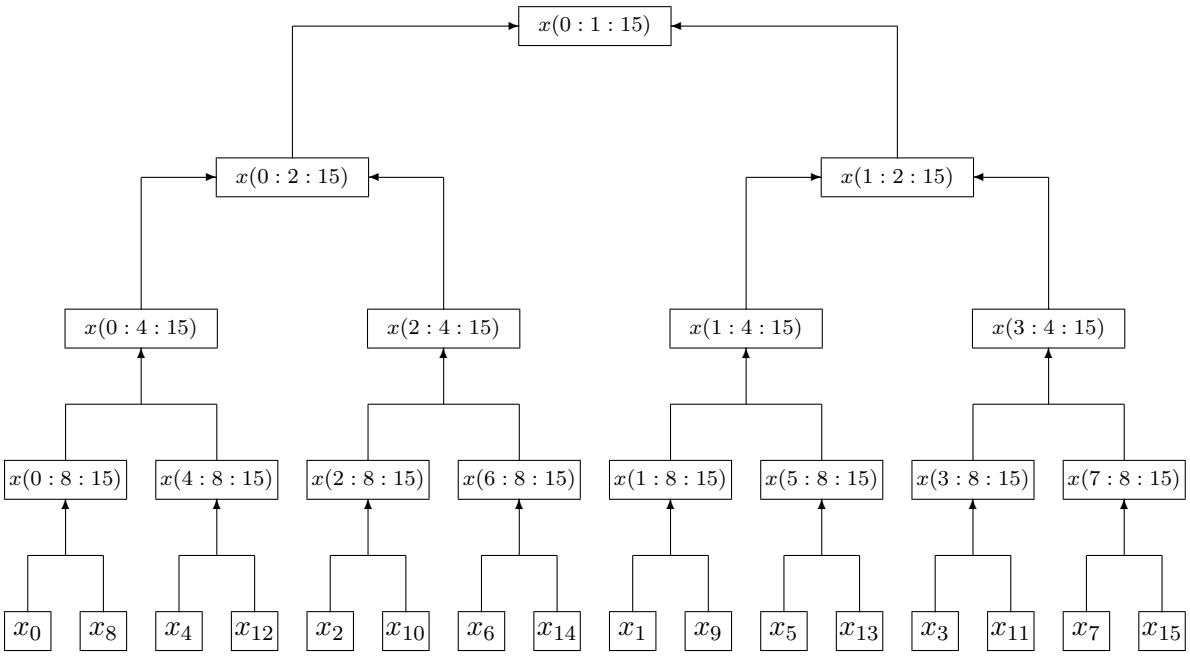


Abbildung 112: Berechnung von $F_{16}x$ mit der FFT

zu transformierenden Vektors x in einer permutierten Reihenfolge sind. Wir definieren

$$y = (y_0, \ldots, y_{15})^T := (x_0, x_8, x_4, x_{12}, x_2, x_{10}, x_6, x_{14}, x_1, x_9, x_5, x_{13}, x_3, x_{11}, x_7, x_{15})^T.$$

Natürlich erfolgt die Berechnung von $F_{16}x$ von unten nach oben. Aus den (trivialen) diskreten Fourier-Transformierten der Länge 1 erhält man die entsprechenden der Länge 2, aus diesen die der Länge 4 usw. Wodurch ist die Permutation $\pi : \{0, \ldots, 15\} \longrightarrow \{0, \ldots 15\}$ mit $y_j = x_{\pi(j)}$, $j = 0, \ldots, 15$ gegeben? In Tabelle 10 geben wir außen j und $\pi(j)$ (in Dezimaldarstellung) und innen die Binärdarstellungen j_2 bzw. $\pi(j)_2$ von j bzw. $\pi(j)$ an. Man erkennt, dass man $\pi(j)_2$ dadurch erhält, dass man die Binärdarstellung j_2 von j in umgekehrter Reihenfolge aufschreibt. Dass dies auch allgemein richtig ist, werden wir uns gleich durch einen Beweis des folgenden Satzes überlegen. □

Mit etwas anderer Notation findet man den folgenden Satz bei C. Van Loan (1992, S. 37).

j	j_2	$\pi(j)_2$	$\pi(j)$
0	0000	0000	0
1	0001	1000	8
2	0010	0100	4
3	0011	1100	12
4	0100	0010	2
5	0101	1010	10
6	0110	0110	6
7	0111	1110	14
8	1000	0001	1
9	1001	1001	9
10	1010	0101	5
11	1011	1101	13
12	1100	0011	3
13	1101	1011	11
14	1110	0111	7
15	1111	1111	15

Tabelle 10: Die Binärdarstellungen von j und $\pi(j)$, $j = 0, \ldots, 15$

Satz *Für eine Zweierpotenz $n = 2^p$ sei die Permutationsmatrix P_n induktiv definiert, indem $P_2 := I_2$ und*

$$P_n := \begin{pmatrix} P_{n/2} & 0_{n/2} \\ 0_{n/2} & P_{n/2} \end{pmatrix} \Pi_n^T$$

gesetzt wird. Hierbei ist Π_n die im obigen Lemma definierte Permutationsmatrix. Mit $\pi_n : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, n-1\}$ bezeichnen wir die Bit-Reversal-Permutation. *Diese ist dadurch definiert, dass $j \in \{0, \ldots, n-1\}$ mit der Binärdarstellung*

$$j = \alpha_0 + \alpha_1 \cdot 2 + \cdots + \alpha_{p-1} \cdot 2^{p-1}, \qquad \alpha_0, \alpha_1, \cdots, \alpha_{p-1} \in \{0, 1\},$$

der Wert

$$\pi_n(j) := \alpha_{p-1} + \alpha_{p-2} \cdot 2 + \cdots + \alpha_0 \cdot 2^{p-1}$$

zugeordnet wird. Für $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$ ist dann

$$(P_n x)_j = x_{\pi_n(j)}, \qquad j = 0, \ldots, n-1.$$

Beweis: Der Beweis erfolgt durch vollständige Induktion nach p. Für $p = 1$ bzw. $n = 2$ ist die Behauptung offenbar richtig, da $P_2 = I_2$ und π_2 die Identität ist. Wir nehmen an, die Behauptung sei für $p-1$ bzw. $n/2 = 2^{p-1}$ richtig. Mit $x = (x_0, \ldots, x_{n-1})^T \in \mathbb{C}^n$ ist dann

$$\begin{aligned} P_n x &= \begin{pmatrix} P_{n/2} & 0_{n/2} \\ 0_{n/2} & P_{n/2} \end{pmatrix} \Pi_n^T x \\ &= \begin{pmatrix} P_{n/2} & 0_{n/2} \\ 0_{n/2} & P_{n/2} \end{pmatrix} \begin{pmatrix} x(0:2:n-1) \\ x(1:2:n-1) \end{pmatrix} \\ &= \begin{pmatrix} P_{n/2} x(0:2:n-1) \\ P_{n/2} x(1:2:n-1) \end{pmatrix}. \end{aligned}$$

Nach Induktionsvoraussetzung gilt mit der Bit-Reversal-Permutation

$$\pi_{n/2} : \{0, \ldots, n/2-1\} \longrightarrow \{0, \ldots, n/2-1\},$$

dass $(P_{n/2}z)_k = z_{\pi_{n/2}(k)}$, $k = 0, \ldots, n/2-1$, für $z = (z_0, \ldots, z_{n/2-1})^T \in \mathbb{C}^{n/2}$. Dann ist

$$(P_n x)_j = \begin{cases} (P_{n/2}x(0:2:n-1))_j & = & x_{2\pi_{n/2}(j)}, & j \in \{0, \ldots, n/2-1\}, \\ (P_{n/2}x(1:2:n-1))_{j-n/2} & = & x_{2\pi_{n/2}(j-n/2)+1}, & j \in \{n/2, \ldots, n-1\}. \end{cases}$$

Zu zeigen bleibt daher, dass

$$\pi_n(j) = \begin{cases} 2\pi_{n/2}(j), & j \in \{0, \ldots, n/2-1\}, \\ 2\pi_{n/2}(j-n/2)+1, & j \in \{n/2, \ldots, n-1\}. \end{cases}$$

Sei zunächst $j \in \{0, \ldots, n/2-1\}$. Die Binärdarstellung von j sei

$$j = \alpha_0 + \alpha_1 \cdot 2 + \cdots + \alpha_{p-2} \cdot 2^{p-2} + 0 \cdot 2^{p-1}.$$

Daher ist

$$\begin{aligned} \pi_n(j) &= 0 + \alpha_{p-2} \cdot 2 + \cdots + \alpha_1 \cdot 2^{p-2} + \alpha_0 \cdot 2^{p-1} \\ &= 2 \cdot (\alpha_{p-2} + \cdots \alpha_1 \cdot 2^{p-3} + \alpha_0 \cdot 2^{p-2}) \\ &= 2\pi_{n/2}(j). \end{aligned}$$

Nun sei $j \in \{n/2, \ldots, n-1\}$ mit der Binärdarstellung

$$j = \alpha_0 + \alpha_1 \cdot 2 + \cdots + \alpha_{p-2} \cdot 2^{p-2} + 1 \cdot 2^{p-1}$$

gegeben. Dann ist

$$\pi_{n/2}(j-n/2) = \pi_{n/2}(\alpha_0 + \alpha_1 \cdot 2 + \cdots \alpha_{p-2} \cdot 2^{p-2}) = \alpha_{p-2} + \cdots \alpha_1 \cdot 2^{p-3} + \alpha_0 \cdot 2^{p-2}$$

und daher

$$\begin{aligned} \pi_n(j) &= 1 + \alpha_{p-2} \cdot 2 + \cdots + \alpha_1 \cdot 2^{p-2} + \alpha_0 \cdot 2^{p-1} \\ &= 1 + 2 \cdot (\alpha_{p-2} + \cdots + \alpha_1 \cdot 2^{p-3} + \alpha_0 \cdot 2^{p-2}) \\ &= 2\pi_{n/2}(j-n/2) + 1. \end{aligned}$$

Der Satz ist damit bewiesen. □

Durch $P_n x$ sind bei gegebenem $x \in \mathbb{C}^n$ offenbar die Blätter des im obigen Beispiel für $n = 2^4$ angegebenen Baumes.

Die Bit-Reversal-Permutation kann in MATLAB mit der Funktion `bitrevorder` berechnet werden. So erhält man z. B. durch `x=[0:1:7]';y=bitrevorder(x);` die Bit-Reversal-Permutation von `x`. Hierbei muss die Länge von `x` eine Zweierpotenz sein. Es ist z. B. `y(2)=4`, da die Feldindizierung in MATLAB mit 1 beginnt. Für einen Vektor $x \in \mathbb{C}^n$ mit einer Zweierpotenz n liefert `bitrevorder(x)` den durch die Bit-Reversal-Permutation permutierten Vektor `x`. Im Kern sieht die MATLAB-Funktion `bitrevorder` (wir geben ihr gleich den Namen `BitReversal`) folgendermaßen aus[98]:

[98] Hierbei haben wir ausgenutzt, dass wir uns durch `type digitrevorder` eine Funktion ansehen können, die beim Aufruf von `bitrevorder` benutzt wird.

```
function y=BitReversal(x);
%*********************************************************
%Input:   x ist ein n-Vektor, n eine Zweierpotenz,
%         was nicht geprüft wird.
%Output:  y ist der durch die Bit-Reversal-Permutation
%         permutierte Vektor x.
%*********************************************************
n=length(x);
idx=base2dec(fliplr(dec2base(0:n-1,2)),2)+1;
y=x(idx);
```

Hierbei werden durch `dec2base` Dezimalzahlen in ihre Binärdarstellung umgewandelt, durch `fliplr` wird die Binärdarstellung von rechts nach links gelesen und anschließend durch `base2dec` wieder in die Dezimaldarstellung umgewandelt. Da Felder in MATLAB immer mit dem Index 1 beginnen, wird 1 addiert. Durch `y=x(idx)` wird der permutierte Vektor bestimmt.

Nun wollen wir eine MATLAB-Funktion zur Berechnung der FFT angeben, wobei wir uns diesmal nicht an der "built in function" `fft` orientieren können. Wir geben zunächst eine Version an, die noch stark auf Komponenten von Vektoren zugreift und die wir im Anschluss versuchen werden zu vektorisieren, siehe Algorithm 1.6.1 bei C. Van Loan (1992, S. 44).

```
function y=MyFFT(x)
%*****************************************************************
% Mit FFT wird die DFT von x berechnet. Hierbei wird vorausgesetzt
% und nicht getestet, dass die Länge von x eine Zweierpotenz ist.
%*****************************************************************
n=length(x); p=floor(log2(n)); %Es ist n=2^p
y=BitReversal(x); L=1;
for q=1:p
      L=2*L;  %Es ist L=2^q
      r=n/L;  %Es ist r=2^(p-q)
      m=L/2;  %Es ist $m=2^(q-1)
      omega=exp(-2*pi*i/L);
      for k=0:r-1
            omega_m=1;
            for j=1:m
                 z=omega_m*y(k*L+m+j);
                 y(k*L+m+j)=y(k*L+j)-z;
                 y(k*L+j)=y(k*L+j)+z;
                 omega_m=omega*omega_m;
            end;
        end;
end;
```

Hierbei zählt die äußerste Schleife mit dem Laufindex q die p *Ebenen* (wobei $n = 2^p$ die Länge des zu transformierenden Vektors x ist), in denen diskrete Fourier-Transformierte

$F_{2^1}, \ldots, F_{2^p}$ berechnet werden. In der q-ten Ebene sind $r = 2^{p-q}$ diskrete Fourier-Transformierte der Länge 2^q zu berechnen. Der Laufindex in dieser mittleren Schleife ist k, er zählt die *Abschnitte* innerhalb der Ebenen. In der innersten Schleife mit dem Laufindex j wird die diskrete Fourier-Transformierte des k-ten Abschnitts in der q-ten Ebene berechnet, wobei insgesamt $m = 2^q$-mal eine sogenannte Butterfly-Operation gemacht wird. In Abbildung 113 veranschaulichen wir uns diesen Schritt. Außerdem

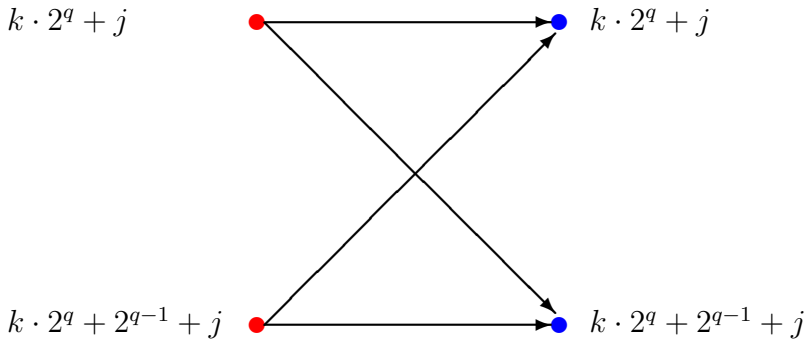


Abbildung 113: Die Butterfly-Operation in der innersten Schleife der FFT

können wir noch einmal leicht die Anzahl der benötigten Multiplikationen $f(n)$ zur Berechnung von $F_n x$ ausrechnen. Mit $n = 2^p$ ist nämlich

$$f(n) = \sum_{q=1}^{p} \sum_{k=0}^{2^{p-q}-1} \sum_{j=1}^{2^{q-1}} 1 = \sum_{q=1}^{p} \sum_{k=0}^{2^{p-q}-1} 2^{q-1} = \sum_{q=1}^{p} 2^{p-1} = \frac{1}{2} 2^p p = \frac{1}{2} n \log_2 n.$$

Es ist einfach, die innere Schleife einzusparen, wobei wir es uns bei der Berechnung des Vektors `Omega_m` allerdings einfach machen:

```
function y=MyVekFFT(x)
%*****************************************************************
% Mit FFT wird die DFT von x berechnet. Hierbei wird vorausgesetzt
% und nicht getestet, dass die Länge von x eine Zweierpotenz ist.
% Es werden die Möglichkeiten von MATLAB ausgenutzt und wenig auf
% Komponenten zurückgegriffen
%*****************************************************************
n=length(x); p=floor(log2(n)); y=BitReversal(x); L=1;
for q=1:p
     L=2*L; r=n/L; m=L/2;  Omega_m=exp(-2*pi*i*[0:m-1]'/L);
     for k=0:r-1
          z=Omega_m.*y(k*L+[1:m]+m);
          y(k*L+[1:m]+m)=y(k*L+[1:m])-z;
          y(k*L+[1:m])=y(k*L+[1:m])+z;
     end;
end;
```

Hiermit wollen wir uns zufriedengeben und auf die Literatur zur FFT verweisen.

68 Geraden in der Ebene: Problem von Sylvester

Wir orientieren uns in diesem Abschnitt stark an der Darstellung bei M. Aigner, G. M. Ziegler (2002, S. 61 ff.), siehe auch H. S. M. Coxeter (1969, S. 65–66). Von J. J. Sylvester[99] wurde 1893 das folgende Problem formuliert:

- Prove that it is not possible to arrange any finite number of real points so that a right line through every two of them shall pass through a third, unless they all lie in the same right line.

Diese "negative" Aussage von Sylvester wurde von T. Motzkin (1951) in eine "positive" Aussage umformuliert:

- Consider a finite number of distinct points in the real Euclidean plane; these points are collinear or there exists a straight line through exactly 2 of them.

Der folgende Satz wird auch Satz von Sylvester-Gallai genannt, da Tibor Gallai 1933 die Vermutung von Sylvester bewies. Siehe auch `http://en.wikipedia.org/wiki/Sylvester-Gallai_theorem`. Der Beweis stammt von L. M. Kelly, er kann bei H. S. M. Coxeter (1948) nachgelesen werden.

Satz *Gegeben seien n paarweise verschiedene Punkte in der Ebene, die nicht alle auf einer Geraden liegen. Dann gibt es eine Gerade, die genau zwei der n Punkte enthält.*

Beweis: Die n paarweise verschiedenen Punkte seien mit $P_1, \ldots, P_n$ bezeichnet. Durch zwei Punkte P_j, P_k mit $j \neq k$ ist eine Gerade $\overline{P_jP_k}$ bestimmt. Daher bestimmen die Punkte $P_1, \ldots, P_n$ höchstens $\frac{1}{2}n(n-1)$ Geraden

$$\overline{P_1P_2}, \ldots, \overline{P_1P_n}, \overline{P_2P_3}, \ldots, \overline{P_2P_n}, \ldots, \overline{P_{n-1}P_n}.$$

Nun betrachte man Paare von Punkten und Geraden $(P_i, \overline{P_jP_k})$ mit $P_i \notin \overline{P_jP_k}$. Es gibt höchstens $\frac{1}{2}n(n-1)(n-2)$ solcher Paare und wenigstens ein Paar, da die gegebenen Punkte als nicht kollinear vorausgesetzt sind. . Unter diesen endlich vielen Paaren gibt es ein Paar aus Punkt und Gerade mit der Eigenschaft, dass der Punkt minimalen (euklidischen) Abstand zu der Geraden hat. O. B. d. A. sei $(P_1, \overline{P_2P_3})$ dieses Paar. Sei Q der Lotpunkt von P_1 auf $\overline{P_2P_3}$, also die Projektion von P_1 auf $\overline{P_2P_3}$. Wir veranschaulichen dies in Abbildung 114. Wir zeigen, dass $\overline{P_2P_3}$ die gesuchte Gerade ist, also außer P_2 und P_3 keinen weiteren der gegebenen Punkte enthält. Angenommen, dies sei nicht der Fall und $\overline{P_2P_3}$ enthielte einen weiteren Punkt P_4. Wenigstens zwei der drei Punkte P_2, P_3, P_4 müssen auf einer Seite von Q (sozusagen rechts oder links) liegen[100], wobei einer der Punkte mit Q zusammenfallen kann. Wir nehmen an, dies seien P_2 und P_3, wobei wir weiter annehmen können, dass P_2 näher zu Q liegt (oder mit Q zusammenfällt). Dann ist aber $(P_2, \overline{P_1P_3})$ ein Paar aus Punkt und Gerade mit der Eigenschaft, dass der

[99] In Abschnitt 62 hatten wir schon ein Problem von Sylvester kennengelernt. Mein Geburtstag ist der 31. 12., mag sein, dass dies der Grund ist, dass mich Probleme von Sylvester interessieren. Vielleicht folgt deswegen noch ein Abschnitt über die Sylvestersche Trägheitsformel oder Sylvester's Vierpunktproblem.

[100] Dies ist wieder ein "Sockenargument": Hat man drei Socken, die schwarz oder weiß sind, so hat man wenigstens zwei von derselben Farbe.

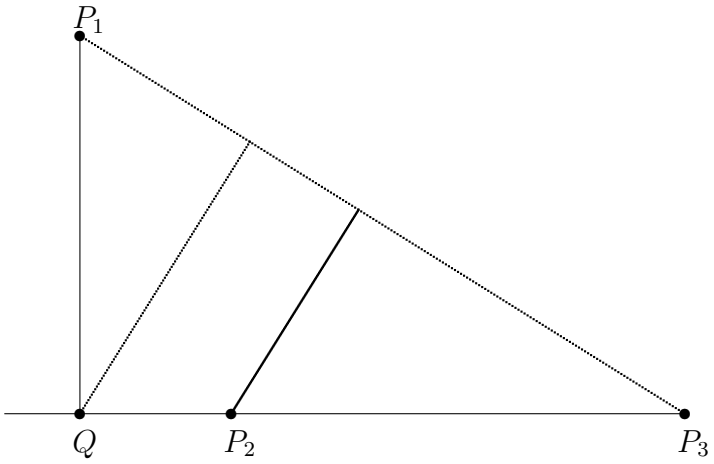


Abbildung 114: Veranschaulichung des Beweises zum Satz von Sylvester-Gallai

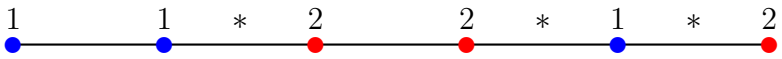


Abbildung 115: Sperners Lemma für $n = 1$

Abstand von P_2 zu der Geraden $\overline{P_1P_3}$ kleiner als der Abstand von P_1 zu $\overline{P_2P_3}$ ist. Dies ist ein Widerspruch zur Definition des Paares $(P_1, \overline{P_2P_3})$. Der Satz von Sylvester-Gallai ist bewiesen. □

69 Sperners Lemma und Brouwers Fixpunktsatz

Mit Hilfe des später nach ihm benannten Lemmas hat E. Sperner (1928) einen verblüffend einfachen Beweis des Brouwerschen Fixpunktsatzes (siehe Abschnitt 37) gefunden. Im folgenden benutzen wir online-Aufzeichnungen von Jacob Fox. Im Falle $n = 1$ sagt Sperners Lemma aus:

- Färbt man die Endpunkte eines Intervalls auf unterschiedliche Weise, etwa mit den Farben blau (1) und rot (2), unterteilt man das Intervall ferner in Teilintervalle, deren Endpunkte beliebig blau oder rot gefärbt sind, so ist die Anzahl der Teilintervalle mit unterschiedlich gefärbten Endpunkten ungerade, insbesondere von Null verschieden.

Wir veranschaulichen uns die Aussage in Abbildung 115. In diesem Beispiel gibt es drei Teilintervalle, deren Endpunkte unterschiedlich gefärbt sind. Die Gültigkeit der Aussage des Sperner-Lemmas für $n = 1$ ist einfach einzusehen. Geht man nämlich vom linken blauen Endpunkt zum rechten roten Endpunkt, so muss man die Farbe eine ungerade Anzahl von Malen verändern. Daher ist die Aussage für $n = 1$ richtig.

Nun formulieren und beweisen wir Sperners Lemma für $n = 2$, auch wenn es genügen würde den Fall $n = 1$ (als Induktionsanfang) und den allgemeinen Fall zu betrachten. Eine Triangulierung eines Dreiecks $\mathbb{R}^2$ ist aber anschaulicher als eine simpliziale Zerle-

gung eines n-dimensionale Simplex. Die Beweise von Sperners Lemma für $n = 2$ und im allgemeinen Fall sind fast identisch.

Sperners Lemma (für $n = 2$) *Ein „großes“ Dreieck mit Ecken v_1, v_2 und v_3 sei trianguliert, also in eine endliche Zahl von „kleinen“ Dreiecken zerlegt, bei denen je zwei entweder disjunkt sind oder eine gemeinasame Kante haben. Die Eckpunkte v_1, v_2, v_3 seien mit drei verschiedenen Farben, die wir mit 1 (blau), 2 (rot) und 3 (grün) bezeichnen, gefärbt. Alle Knoten auf der (v_i, v_j)-Seite des Dreiecks, also der Seite mit den Eckpunkten v_i und v_j, seien mit den Farben i oder j gefärbt, während alle inneren Knoten beliebig gefärbt seien. Dann ist die Anzahl der „kleinen“ Dreiecke, deren Eckpunkte mit den drei verschiedenen Farben gefärbt sind, ungerade und daher insbesondere von Null verschieden.*

In Abbildung 116 wird die Aussage veranschaulicht. Wenn man sie sich genau ansieht,

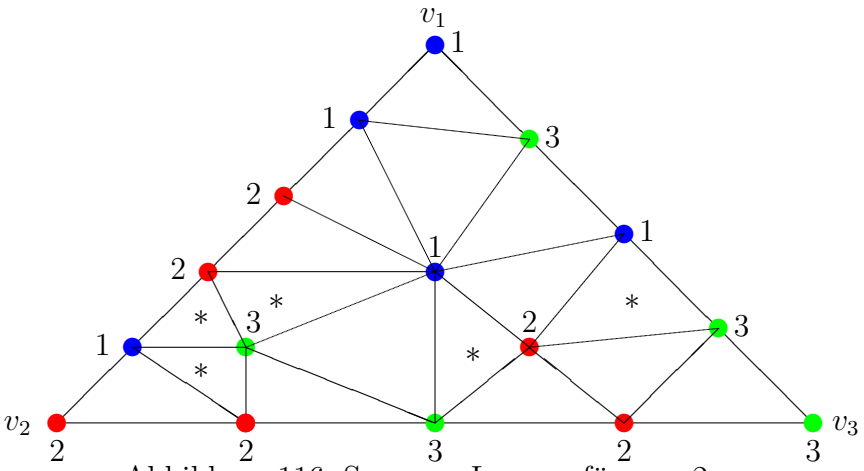


Abbildung 116: Sperners Lemma für $n = 2$

so erkennt man, dass es genau fünf „kleine“ Dreiecke gibt, deren Eckpunkte mit den drei unterschiedlichen Farben gefärbt sind. Wir haben sie durch ein $*$ gekennzeichnet. In diesem Beispiel ist die Aussage von Sperners Lemma also richtig.

Beweis: Sei Q die Anzahl der Zellen bzw. „kleinen“ Dreiecke, deren Eckpunkte $(1, 1, 2)$ oder $(1, 2, 2)$ gefärbt sind, und R die Anzahl der Zellen, deren Eckpunkte gemäß $(1, 2, 3)$ unterschiedlich gefärbt sind. Wir haben zu zeigen, dass R ungerade ist. In dem Beispiel in Abbildung 116 ist $Q = 4$ und $R = 5$. Nun betrachten wir die Kanten bzw. Dreiecksseiten der Triangulierung, die $(1, 2)$ bzw. (blau, rot) gefärbt sind. Diese sind entweder auf dem Rand des gegebenen Dreiecks oder im Innern. Mit X bezeichnen wir die Anzahl der Kanten auf dem Rand, die $(1, 2)$ gefärbt sind, mit Y die Anzahl der ebenso gefärbten Kanten im Inneren. In unserem Beispiel in Abbildung 116 ist $X = 3$ und $Y = 5$. Wir erhalten:

- Randkanten, die $(1, 2)$ gefärbt sind, können nur auf der Seite des Dreiecks mit den Eckpunkten v_1 und v_2 auftreten. Da v_1 die Farbe 1 und v_2 die Farbe 2 hat, ist die Anzahl X der $(1, 2)$ gefärbten Randkanten ungerade (siehe das Argument für $n = 1$).

- Es gibt Q Zellen, deren Eckpunkte mit $(1,1,2)$ oder $(1,2,2)$ gefärbt sind. Jede dieser Zellen hat also zwei Kanten, die mit $(1,2)$ gefärbt sind. Dagegen haben die R Zellen mit $(1,2,3)$ gefärbten Eckpunkten genau eine $(1,2)$ gefärbte Kante. Hierbei werden innere $(1,2)$-Kanten doppelt gezählt, Randkanten nur einfach. Daher ist $2Q + R = X + 2Y$.

Insgesamt ist $R = X + 2(Y - Q)$ eine ungerade Zahl. Sperners Lemma ist für $n = 2$ bewiesen. □

Nun kommen wir zum allgemeinen Fall. Statt eines Dreiecks in der Ebene ist jetzt ein *n-dimensionales Simplex* S gegeben, also die *konvexe Hülle* $\operatorname{co}(\{v_1, \ldots, v_{n+1}\})$ (siehe Seite 245) bzw. die Menge aller Konvexkombinationen von gegebenen Punkten $v_1, \ldots, v_{n+1} \in \mathbb{R}^n$ in *allgemeiner Lage* (d. h. $\{v_2 - v_1, \ldots, v_{n+1} - v_1\}$ sind linear unabhängig). Ein n-dimensionales Simplex S mit Ecken $v_1, \ldots, v_{n+1}$ ist also gegeben durch

$$S = \Big\{ \sum_{i=1}^{n+1} \alpha_i v_i : \alpha_i \geq 0 \ (i = 1, \ldots, n+1), \ \sum_{i=1}^{n+1} \alpha_i = 1 \Big\}.$$

In der Darstellung eines Punktes $x \in S$ in der Form

$$x = \sum_{i=1}^{n+1} \alpha_i v_i$$

mit $\alpha_i \geq 0$, $i = 1, \ldots n+1$, und $\sum_{i=1}^{n+1} \alpha_i = 1$, heißen $\alpha_i = \alpha_i(x)$, $i = 1, \ldots, n+1$, die *baryzentrischen Koordinaten* von x. Einer Kante eines Dreiecks entspricht eine *Seite* eines Simplexes, d. h. der konvexen Hülle von n oder weniger der Ecken $v_1, \ldots, v_{n+1}$. Der Triangulierung eines Dreiecks entspricht eine *simpliziale Zerlegung* des Simplex S, d. h. eine Zerlegung von S in kleine n-dimensionale Simplexe bzw. Zellen derart, dass zwei Zellen entweder disjunkt sind oder eine Seite einer gewissen Dimension gemein haben.

Nun können wir Sperners Lemma im allgemeinen Fall formulieren und anschließend beweisen.

Sperners Lemma *Gegeben sei ein n-dimensionales Simplex S mit Ecken $v_1, \ldots, v_{n+1}$ und hierzu eine simpliziale Zerlegung von S. Es mögen $n+1$ verschiedene Farben zur Verfügung stehen, die wir mit $1, \ldots, n+1$ bezeichnen. Die Eckpunkte der simplizialen Zerlegung seien so mit jeweils einer Farbe aus $\{1, \ldots, n+1\}$ gefärbt, dass die folgenden Regeln gelten:*

1. *Die Ecke v_i ist mit der Farbe i gefärbt, $i = 1, \ldots, n+1$.*

2. *Ein Eckpunkt auf einer Seite von S ist mit einer diese Seite definierenden Eckpunkte gefärbt. Genauer: Sei I eine echte Teilmenge von $\{1, \ldots, n+1\}$, so ist ein Eckpunkt auf der durch $\{v_i\}_{i \in I}$ erzeugten Seite von S mit einer der Farben aus I gefärbt.*

3. *Eckpunkte im Inneren von S sind beliebig mit einer der Farben aus $\{1, \ldots, n+1\}$ gefärbt.*

Dann gibt es eine ungerade Anzahl von Zellen in der simplizialen Zerlegung von S, deren Ecken mit $n+1$ verschiedenen Farben aus $\{1,\ldots,n+1\}$ gefärbt sind.

Beweis: Der Beweis benutzt Induktion nach n. Die Aussage ist richtig für $n = 2$ (oder für $n = 1$, dann hätten wir uns Sperners Lemma für $n = 2$ sparen können). Wir nehmen an, die Aussage sei für $(n-1)$-dimensionale Simplexe richtig. Sei Q die Anzahl der Zellen in der simplizialen Zerlegung von S, deren Ecken mit allen Farben außer der Farbe $n+1$ gefärbt sind. Die $n+1$ Ecken dieser Zellen sind also mit den Farben $1,\ldots,n$ gefärbt. D. h. genau eine dieser Farben wird zweimal, die anderen einmal benutzt. Mit R bezeichnen wir die Anzahl der Zellen, deren Ecken alle verschieden gefärbt sind, also mit den Farben $1,\ldots,n+1$. Ferner betrachten wir in der simplizialen Zerlegung $(n-1)$-dimensionale Seiten, deren Eckpunkte genau mit den Farben $1,\ldots,n$ gefärbt sind. Sei X die Anzahl solcher Seiten auf dem Rand von S und Y die Anzahl im Inneren von S. Wir erhalten:

- Seiten auf dem Rand von S, die genau mit den Farben $1,\ldots,n$ gefärbt sind, können nur auf der Seite $F \subset S$ liegen, deren Ecken $v_1,\ldots,v_n$ sind bzw. mit den Farben $1,\ldots,n$ gefärbt sind. Auf den $(n-1)$-dimensionalen Simplex F können wir die Induktionsvoraussetzung anwenden und erhalten, dass X ungerade ist.

- Es gibt R Zellen, deren Ecken verschieden, also mit den Farben $1,\ldots,n+1$ gefärbt sind. Jede dieser R Zellen hat genau eine Seite mit durch $1,\ldots,n$ gefärbten Ecken. Es gibt Q Zellen, deren Ecken mit den Farben $1,\ldots,n$ gefärbt sind. Jede dieser Q Zellen hat zwei Seiten, deren Ecken mit $1,\ldots,n$ gefärbt sind. Randseiten treten nur in einer, innere Seiten in zwei Zellen auf. Daher ist $2Q + R = X + 2Y$.

Da X ungerade ist, ist es auch R. Sperners Lemma ist bewiesen. □

Nun kommen wir zum Brouwerschen Fixpunktsatz. Beim Beweis benötigen wir, dass es zu einem n-dimensionalen Simplex S eine Folge $\{\mathcal{S}_k\}$ simplizialer Zerlegungen von S gibt mit der Eigenschaft, dass $\lim_{k\to\infty} \delta(\mathcal{S}_k) = 0$, wobei $\delta(\mathcal{S}_k)$ der maximale Durchmesser einer Zelle von $\mathcal{S}_k$ ist. Hierzu definieren wir:

Definition Sei $S := \mathrm{co}\,(\{v_1,\ldots,v_{n+1}\})$ ein n-dimensionales Simplex mit den Ecken $v_1,\ldots,v_{n+1}$. Für eine Permutation $(i_1,\ldots,i_{n+1})$ von $(1,\ldots,n+1)$ sei

$$y_p := \frac{1}{p}\sum_{j=1}^{p} v_{i_j}, \qquad p = 1,\ldots,n+1$$

und

$$S(i_1,\ldots,i_{n+1}) := \mathrm{co}\,(\{y_1,\ldots,y_{n+1}\}).$$

Die Menge aller $(n+1)!$ Simplexe $S(i_1,\ldots,i_{n+1})$ bildet die *baryzentrische Unterteilung* von S.

Beispiel: Für den Fall eines Dreiecks als Ausgangssimplex veranschaulichen wir uns die baryzentrische Unterteilung in Abbildung 117. Mit der Permutation $(1,3,2)$ von $(1,2,3)$ erhalten wir das Simplex $S(1,3,2)$ mit den Ecken v_1, $\frac{1}{2}(v_1+v_3)$ und dem Schwerpunkt $\frac{1}{3}(v_1+v_3+v_2)$ des Dreiecks. Auch die fünf weiteren Simplexe der baryzentrischen Unterteilung haben wir in Abbildung 117 eingetragen. □

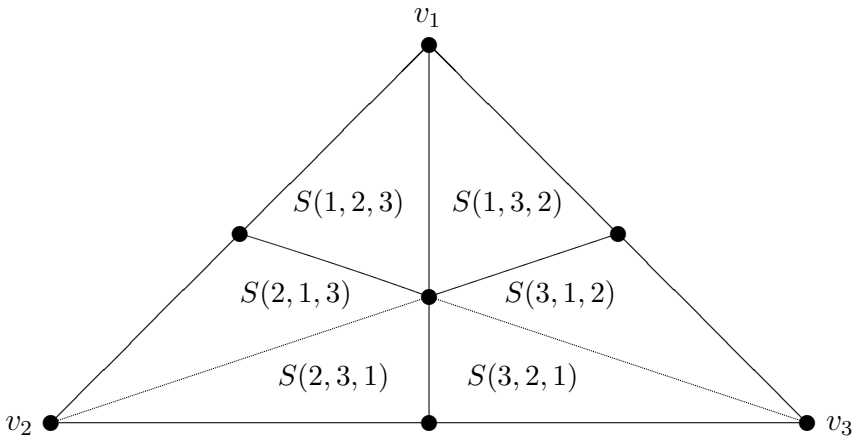


Abbildung 117: Die baryzentrische Unterteilung eines Dreiecks

Eine Folge simplizialer Zerlegungen eines gegebenen Simplex erhält man, indem man jedes Simplex einer baryzentrischen Unterteilung wieder baryzentrisch unterteilt. Der maximale Durchmesser der hierbei entstehenden Simplexe konvergiert dann gegen Null, was eine Konsequenz der folgenden Aussage ist:

- Sei $S := \operatorname{co}(\{v_1, \ldots, v_{n+1}\})$ mit $v_1, \ldots, v_{n+1} \in \mathbb{R}^n$ ein n-dimensionales Simplex. Mit einer Permutation $(i_1, \ldots, i_{n+1})$ von $(1, \ldots, n+1)$ sei

$$S(i_1, \ldots, i_{n+1}) := \operatorname{co}(\{y_1, \ldots, y_{n+1}\})$$

mit

$$y_p := \frac{1}{p} \sum_{j=1}^{p} v_{i_j}, \qquad p = 1, \ldots, n+1.$$

Dann ist

$$\delta(S(i_1, \ldots, i_{n+1})) \leq \frac{n}{n+1} \delta(S).$$

Denn: Der Durchmesser eines Simplex ist gleich dem maximalen Abstand von je zwei seiner Ecken (Beweis?). Daher ist

$$\delta(S(i_1, \ldots, i_{n+1})) = \max_{1 \leq p < q \leq n+1} \|y_p - y_q\|,$$

wobei $\|\cdot\|$ z. B. die euklidische Norm bedeutet. Für $1 \leq p < q \leq n+1$ ist

$$\begin{aligned} y_p - y_q &= \frac{1}{p} \sum_{j=1}^{p} v_{i_j} - \frac{1}{q} \sum_{j=1}^{q} v_{i_j} \\ &= \frac{1}{pq} \Big[(q-p) \sum_{j=1}^{p} v_{i_j} - p \sum_{j=p+1}^{q} v_{i_j} \Big]. \end{aligned}$$

Rechts stehen $(q-p)p$ Differenzen der Form $v_{i_j} - v_{i_k}$. Für $1 \leq p < q \leq n+1$ ist daher

$$\|y_p - y_q\| \leq \frac{q-p)p}{pq} \delta(S) = \frac{q-p}{q} \delta(S) \leq \frac{n}{n+1} \delta(S).$$

Damit ist obige Behauptung bewiesen. Insbesondere ist gezeigt, dass

$$\delta(\mathcal{S}_k) \leq \Big(\frac{n}{n+1}\Big)^k \delta(S)$$

und folglich $\lim_{k\to\infty} \delta(\mathcal{S}_k) = 0$, wenn $\mathcal{S}_0 := S$ und $\mathcal{S}_k$ durch eine baryzentrische Unterteilung von $\mathcal{S}_{k-1}$ entsteht.

Wir formulieren den Brouwerschen Fixpunktsatz in der folgenden Weise.

Brouwers Fixpunktsatz *Sei S ein im $\mathbb{R}^{n+1}$ eingebettetes Simplex mit den Ecken $v_i := e_i$, $i = 1, \ldots, n+1$, wobei e_i der i-te Einheitsvektor im $\mathbb{R}^{n+1}$ ist*[101]*. Ist $f\colon S \longrightarrow S$ stetig, so gibt es ein $x \in S$ mit $f(x) = x$.*

Beweis: Wir machen einen Widerspruchsbeweis und nehmen an, dass f keinen Fixpunkt besitzt. Wie wir eben gesehen haben, gibt es eine Folge $\{\mathcal{S}_k\}$ simplizialer Zerlegungen von S mit $\lim_{k\to\infty} \delta(\mathcal{S}_k) = 0$. Für jede simpliziale Zerlegung $\mathcal{S}_k$ definieren wir eine Färbung der Ecken, welche den Regeln in Sperners Lemma genügen. Eine Ecke $x \in \mathcal{S}_k$ färben wir mit der Farbe $\lambda(x) \in \{1, \ldots, n+1\}$, wobei

$$\lambda(x) := \min\{j \in \{1, \ldots, n+1\} : f(x)_j < x_j\}.$$

Also ist $\lambda(x)$ der kleinste Index j, für den die j-te Koordinate von $f(x) - x$ negativ ist. Eine solche Koordinate gibt es, so dass $\lambda(x)$ wohldefiniert ist. Denn wäre $f(x)_j \geq x_j$, $j = 1, \ldots, n+1$, so folgt aus $x \in S$ und $f(x) \in S$, dass

$$0 = \underbrace{\sum_{j=1}^{n+1} f(x)_j}_{=1} - \underbrace{\sum_{j=1}^{n+1} x_j}_{=1} = \sum_{j=1}^{n+1} (\underbrace{f(x)_j - x_j}_{\geq 0}),$$

somit $f(x)_j - x_j = 0$, $j = 1, \ldots, n+1$, und damit $f(x) = x$ im Widerspruch zu der Annahme, dass f keinen Fixpunkt besitzt. Damit ist gezeigt, dass bei der Definition von $\lambda(x)$ das Minimum über eine nichtleere Menge gebildet wird und damit $\lambda(x)$ wohldefiniert ist. Nun ist zu zeigen, dass hierdurch die Färbungsregeln in Sperners Lemma erfüllt sind. Es ist $\lambda(v_i) = \lambda(e_i) = i$, die i-te Ecke i also mit der Farbe i gefärbt, $i = 1, \ldots, n+1$. Denn es ist $f(e_i)_j - (e_i)_j = f(e_i)_j \geq 0$ für $j < i$ und $f(e_i)_i - (e_i)_i = f(e_i)_i - 1 < 0$, da aus $f(e_i) = 1$ folgen würde, dass e_i ein Fixpunkt von f ist, was wir ausgeschlossen hatten. Damit ist die erste Färbungsregel für eine simpliziale Zerlegung von S erfüllt. Zum Nachweis der zweiten Färbungsregel nehmen wir an, die Ecke x liege auf einer von den Ecken $\{v_i\}_{i\in I}$ erzeugten Seite von S, es sei also I eine echte Teilmenge von $\{1, \ldots, n+1\}$ und $x = \sum_{i\in I} \alpha_i v_i$ mit $\alpha_i > 0$, $i \in I$, und $\sum_{i\in I} \alpha_i = 1$. Ist $f(x)_j < x_j$ und damit

$$0 \leq f(x)_j < x_j = \sum_{i\in I} \alpha_i (e_i)_j = \begin{cases} \alpha_j, & j \in I, \\ 0, & j \notin I, \end{cases}$$

[101] Es ist also

$$S := \{\alpha \in \mathbb{R}^{n+1} : \alpha \geq 0,\ e^T\alpha = 1\},$$

wobei $e \in \mathbb{R}^{n+1}$ der Vektor ist, dessen Komponenten sämtlich gleich 1 sind.

so folgt $j \in I$. Eine Ecke auf der von $\{v_i\}_{i\in I}$ erzeugten Seite muss also mit einer Farbe aus I gefärbt sein. Damit ist auch die zweite Färbungsregel erfüllt. Sperners Lemma impliziert, dass es in $\mathcal{S}_k$ ein Simplex mit Ecken $x^{(k,1)}, \ldots, x^{(k,n+1)}$ gibt, die mit den Farben $1, \ldots, n+1$ gefärbt sind, d. h. es ist $f(x^{(k,j)})_j < x_j^{(k,j)}$, $j = 1, \ldots, n+1$. Da die Folgen $\{x^{(k,j)}\}_{k\in\mathbb{N}}$ in der kompakten Teilmenge S liegen, konvergieren jeweils Teilfolgen gegen einen Punkt aus S. Da man sukzessive, von $\{x^{(k,1)}\}_{k\in\mathbb{N}}$ ausgehend, zu konvergenten Teilfolgen übergehen kann, können wir annehmen, dass $\{x^{(k,j)}\}_{k\in\mathbb{N}}$, $j = 1, \ldots, n+1$, konvergiert. Wegen $\delta(\mathcal{S}_k) \to 0$ ist der Limes von j unabhängig. Wir definieren $x^* := \lim_{k\to\infty} x^{(k,j)}$. Es ist $x^* \in S$. Wegen $f(x^{(k,j)})_j < x_j^{(k,j)}$, $j = 1, \ldots, n+1$, und der Stetigkeit von f ist $f(x^*)_j \leq x_j^*$, $j = 1, \ldots, n+1$. Komponentenweise ist also $x^* - f(x^*) \geq 0$. Wegen $x^* \in S$ und $f(x^*) \in S$ ist $e^T(x^* - f(x^*)) = 0$. Insgesamt ist $f(x^*) = x^*$, ein Widerspruch dazu, dass f keinen Fixpunkt besitzt. Damit ist Brouwers Fixpunktsatz bewiesen. □

Bemerkung: Der Vollständigkeit halber wollen wir uns noch davon überzeugen, dass das im $\mathbb{R}^{n+1}$ enthaltene Simplex

$$S := \{\alpha \in \mathbb{R}^{n+1} : \alpha \geq 0,\ e^T\alpha = 1\},$$

wobei e der Vektor im $\mathbb{R}^{n+1}$ ist, dessen Komponenten sämtlich gleich 1 sind, und die abgeschlossene euklidische Einheitskugel im $\mathbb{R}^n$, also

$$B^n[0;1] := \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\},$$

homöomorph sind, also eineindeutig und umkehrbar stetig aufeinander abgebildet werden können. Hierzu zeigen wir zunächst, dass S und $B^{n+1}[0;1] \cap H$ homöomorph sind, wobei die Hyperebene H durch $H := \{y \in \mathbb{R}^{n+1} : e^T y = 0\}$ definiert ist. Am Schluss überlegen wir uns, dass $B^{n+1}[0;1] \cap H$ und $B^n[0;1]$ homöomorph sind.

Zunächst haben wir also die Existenz einer bijektiven, stetigen Abbildung $\phi\colon S \longrightarrow B^{n+1}[0;1] \cap H$ zu zeigen, für die auch $\phi^{-1}\colon B^{n+1}[0;1] \cap H \longrightarrow S$ stetig ist. Mit

$$s := \frac{1}{n+1}e$$

bezeichnen wir den Schwerpunkt des Simplex S. Für $\alpha \in S \setminus \{s\}$ sei

$$t(\alpha) := \frac{1/(n+1)}{1/(n+1) - \min_{j=1,\ldots,n+1} \alpha_j}.$$

Offenbar ist $t(\alpha) \geq 1$ und $t(\alpha) = 1$ genau dann, wenn $\min_{j=1,\ldots,n+1} \alpha_j = 0$ bzw. α auf dem Rand des Simplex S liegt. Wir definieren die Abbildung $m : \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$m(y) := \min_{j=1,\ldots,n+1} y_j$$

und beachten, dass $m(\alpha) < 1/(n+1)$ für $\alpha \in S \setminus \{s\}$. Nun sei

$$\phi(\alpha) := \begin{cases} \dfrac{1}{t(\alpha)} \cdot \dfrac{\alpha - s}{\|\alpha - s\|_2}, & \alpha \neq s, \\ 0, & \alpha = s. \end{cases}$$

Offenbar ist $\phi\colon S \longrightarrow B^{n+1}[0;1] \cap H$. Die Abbildung ϕ ist stetig, was nur im Punkte s nicht völlig trivial ist. Ist $\{\alpha_k\} \subset S \setminus \{s\}$ eine Folge mit $\alpha_k \to s$, so gilt $t(\alpha_k) \to +\infty$ und damit $\|\phi(\alpha_k)\|_2 = 1/t(\alpha_k) \to 0$ bzw. $\phi(\alpha_k) \to 0$, womit die Stetigkeit von ϕ auch im Punkt s nachgewiesen ist. Zum Nachweis der Injektivität von ϕ nehmen wir an, es sei $\phi(\alpha) = \phi(\beta)$ mit $\alpha, \beta \in S \setminus \{s\}$. Aus $\|\phi(\alpha)\|_2 = \|\phi(\beta)\|_2$ folgt $t(\alpha) = t(\beta)$ und hieraus $m(\alpha) = m(\beta)$. Aus $\phi(\alpha) = \phi(\beta)$ folgt also $\alpha - s = c(\beta - s)$ mit $c > 0$. Wegen $m(\alpha) = m(\beta)$ ist $c = 1$ und folglich $\alpha = \beta$. Um zu zeigen, dass ϕ auch surjektiv ist, geben wir uns ein von Null verschiedenes Element y aus $B^{n+1}0;1] \cap H$ beliebig vor. Wir haben die Existenz eines $\alpha \in S$ mit $\phi(\alpha) = y$ zu zeigen. Für α machen wir den Ansatz $\alpha = s + cy$ mit noch unbekanntem $c > 0$. Dann ist $m(\alpha) = 1/(n+1) + cm(y)$. Wegen $y \in H \setminus \{0\}$ ist $m(y) < 0$. Dann ist

$$\phi(\alpha) = -c(n+1)m(y) \cdot \frac{y}{\|y\|_2}.$$

Um $\phi(\alpha) = y$ zu erreichen, haben wir also

$$c := -\frac{\|y||_2}{(n+1)m(y)}$$

zu setzen. Wir beachten, dass dann $m(\alpha) = (1 - \|y\|_2)/(n+1) \geq 0$, also $\alpha \in S$ gesichert ist. Die Umkehrabbildung $\phi^{-1}\colon B^{n+1}[0;1] \cap H \longrightarrow S$ ist daher gegeben durch

$$\phi^{-1}(y) = \begin{cases} s - \dfrac{\|y\|_2}{(n+1)m(y)} \cdot y, & y \neq 0, \\ s, & y = 0. \end{cases}$$

Damit ist gezeigt, dass die Abbildung $\phi\colon S \longrightarrow B^{n+1}[0;1] \cap H$ auch surjektiv, insgesamt also bijektiv ist. Die Stetigkeit von ϕ^{-1} ist nur im Punkte $y = 0$ fraglich. Um auch in diesem Fall die Stetigkeit von ϕ^{-1} zu zeigen, geben wir uns eine Folge $\{y_k\} \subset B^{n+1}[0;1] \cap H \setminus \{0\}$ mit $y_k \to 0$ vor. Wegen

$$(*) \qquad \|\phi^{-1}(y_k) - s\|_2 = -\frac{\|y_k\|_2^2}{(n+1)m(y_k)}$$

haben wir zu zeigen, dass die rechte Seite dieser Gleichung mit $k \to \infty$ gegen Null konvergiert. Dies wird sehr einfach aus dem folgenden Ergebnis folgen.

- Man betrachte die Optimierungsaufgabe

  $$\text{Maximiere} \quad m(z) := \min_{j=1,\ldots,n+1} z_j \quad \text{auf} \quad M := \{z \in \mathbb{R}^{n+1} : e^T z = 0,\ \|z\|_2 = 1\}.$$

  Diese Aufgabe hat, da M kompakt und m stetig ist, eine Lösung $z^* \in M$. Es ist $m^* := m(z^*)$ negativ, da $m(z) < 0$ für alle $z \in M$. Dann ist

  $$-\frac{\|y\|_2^2}{m(y)} \leq -\frac{\|y\|_2}{m^*}, \qquad y \in H \setminus \{0\}.$$

Denn: Ist $y \in H \setminus \{0\}$, so ist $z := y/\|y\|_2 \in M$ und daher $m(z) = m(y)/\|y\|_2 \leq m^*$, woraus die Behauptung folgt.

Hieraus folgt die Stetigkeit von ϕ^{-1} auch im Nullpunkt, da $(*)$ und das eben bewiesene Ergebnis uns die Abschätzung

$$\|\phi^{-1}(y_k) - s\|_2 \leq -\frac{\|y_k\|_2}{(n+1)m^*}$$

liefern.

Damit ist bewiesen, dass das Simplex S und $B^{n+1}[0;1] \cap H$ homöomorph sind. Dass $B^{n+1}[0;1] \cap H$ und $B^n[0;1]$ homöomorph sind, ist einfach einzusehen. Die Hyperebene H ist ein n-dimensionaler linearer Teilraum des $\mathbb{R}^{n+1}$. Mit $\{u_1, \ldots, u_n\} \subset H$ bezeichnen wir ein Orthonormalsystem für H, d. h. es ist

$$u_i^T u_j = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Daher ist

$$B^{n+1}[0;1] \cap H = \Big\{ \sum_{i=1}^{n} \lambda_i u_i : \sum_{i=1}^{n} \lambda_i^2 \leq 1 \Big\}.$$

Die durch

$$\psi\Big(\sum_{i=1}^{n} \lambda_i u_i\Big) := \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$$

erklärte Abbildung ψ bildet offenbar $B^{n+1}[0;1] \cap H$ homöomorph auf $B^n[0;1]$ ab. Insgesamt sind das Simplex S und die Kugel $B^n[0;1]$ also homöomorph. Weiter ist damit mit Hilfe des Spernerschen Lemmas auch der Brouwersche Fixpunktsatz für die Einheitskugel und damit allgemein für eine beliebige nichtleere, konvexe und kompakte Teilmenge des $\mathbb{R}^n$ bewiesen, siehe die Bemerkung zu Beginn von Abschnitt 37. □

70 Die Gammafunktion und der Satz von Bohr-Mollerup

Ich werde animiert, diesen Abschnitt über die Gammafunktion zu schreiben, weil ich als junger Student die kleine, inzwischen klassische Abhandlung von E. Artin (1931) über die Gammafunktion mit Begeisterung gelesen habe. Zu der Zeit lebte Emil Artin noch. Ich erwähnte ihn schon in Abschnitt 36 über den Satz von Stone-Weierstraß. Wie E. Hewitt im Vorwort zur englischsprachigen Ausgabe (erschienen 1964 und online verfügbar) habe ich empfunden:

- A generation has passed since the late Emil Artin's little classic on the gamma function appeared in the *Hamburger Mathematische Einzelschriften*. Since that time, it has been read with joy and fascination by many thousands of mathematicians and students of mathematics.

Daher schreibe ich diesen Abschnitt, um noch einmal diese Faszination zu spüren. Allerdings werde ich mich auf die Definition der Gammafunktion, den Beweis von Artin des schönen Satzes von Bohr-Mollerup und den Nachweis dafür, dass die Gammafunktion auf $\mathbb{R}_+$ beliebig oft differenzierbar ist, beschränken.

Die Gammafunktion ist u. a. aus dem Bemühen entstanden, eine Interpolationsfunktion für die für natürliche Zahlen gegebene Abbildung $n \mapsto (n-1)!$ zu finden[102]. Gesucht ist also eine stetige Funktion $\Gamma\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, für die $\Gamma(n) = (n-1)!$ für alle $n \in \mathbb{N}$ gilt und die außerdem die Grundeigenschaft $n! = n \cdot (n-1)!$ der Fakultät besitzt, also $\Gamma(x+1) = x \cdot \Gamma(x)$ für alle $x > 0$ gilt.

Im folgenden Satz definieren wir die Gammafunktion auf $\mathbb{R}_+$ und zeigen, dass diese Definition einen Sinn macht. Außerdem wird die Grundeigenschaft der Fakultätfunktion sowie die *logarithmische Konvexität* der Gammafunktion nachgewiesen.

Satz *Für $x > 0$ sei die Gammafunktion an der Stelle x durch*

$$\Gamma(x) := \int_0^\infty e^{-t} t^{x-1}\, dt$$

definiert. Dann gilt:

1. *Für jedes $x > 0$ existiert das $\Gamma(x)$ definierende Integral.*
2. *Es gilt $\Gamma(x+1) = x \cdot \Gamma(x)$ für jedes $x > 0$ und $\Gamma(1) = 1$.*
3. *Die Gammafunktion $\Gamma\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ ist eine logarithmisch konvexe Funktion, d. h. $\log\Gamma\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ ist eine konvexe Funktion.*

Beweis: Zum Nachweis der ersten Aussage zeigen wir, dass

$$\int_0^1 e^{-t} t^{x-1}\, dt = \lim_{\epsilon \to 0+} \int_\epsilon^1 e^{-t} t^{x-1}\, dt, \qquad \int_1^\infty e^{-t} t^{x-1}\, dt = \lim_{\delta \to \infty} \int_1^\delta e^{-t} t^{x-1}\, dt$$

für alle $x > 0$ existieren. Für $x > 0$ und $\epsilon \in (0,1)$ ist

$$\int_\epsilon^1 e^{-t} t^{x-1}\, dt \le \int_\epsilon^1 t^{x-1}\, dt = \frac{1}{x} t^x \Big|_{t=\epsilon}^{t=1} = \frac{1}{x} - \frac{\epsilon^x}{x} \le \frac{1}{x}.$$

Bei festem $x > 0$ ist das Integral $\int_\epsilon^1 e^{-t} t^{x-1}\, dt$ also durch $1/x$ nach oben beschränkt. Mit fallendem ϵ wächst das Integral monoton. Damit existiert $\int_0^1 e^{-t} t^{x-1}\, dt$ und die erste Gleichung ist bewiesen. Um die zweite Gleichung zu beweisen, beachten wir, dass für positive t jeder Term der Reihe für e^t positiv ist, also die Ungleichung $e^t > t^n/n!$ für alle $n \in \mathbb{N}$ gilt. Daher ist $e^{-t} < n!/t^n$, so dass man den Integranden in der Definition der Gammafunktion durch

$$e^{-t} t^{x-1} < \frac{n!}{t^{n+1-x}}$$

[102] Der Shift im Argument hat historische Gründe.

abschätzen. Hält man $x > 0$ fest und wählt $n > x + 1$, so ist

$$\int_1^\delta e^{-t}t^{x-1}\,dt \le \int_1^\delta \frac{n!}{t^{n+1-x}}\,dt = \frac{n!}{x-n}t^{x-n}\,\Big|_{t=1}^{t=\delta} = \underbrace{\frac{n!}{x-n}\delta^{x-n}}_{<0} + \frac{n!}{n-x} < \frac{n!}{n-x}.$$

Da außerdem $\int_1^\delta e^{-t}t^{x-1}\,dt$ mit wachsendem δ ebenfalls wächst, ist die Existenz von $\int_1^\infty e^{-t}t^{x-1}\,dt$ und damit die erste Aussage des Satzes bewiesen.
Mit Hilfe partieller Integration erhalten wir

$$\begin{aligned}\int_\epsilon^\delta e^{-t}t^x\,dt &= \int_\epsilon^\delta \Big(-\frac{d}{dt}e^{-t}\Big)t^x\,dt \\ &= -e^{-t}t^x\,\Big|_{t=\epsilon}^{t=\delta} + \int_\epsilon^\delta e^{-t}\Big(\frac{d}{dt}t^x\Big)\,dt \\ &= e^{-\epsilon}\epsilon^x - e^{-\delta}\delta^x + x\int_\epsilon^\delta e^{-t}t^{x-1}\,dt.\end{aligned}$$

Bei festem $x > 0$ ist offenbar $\lim_{\epsilon\to 0+} e^{-\epsilon}\epsilon^x = 0$, der erste Term auf der rechten Seite konvergiert also mit $\epsilon \to 0+$ gegen Null. Der positive Term $e^{-\delta}\delta^x$ kann durch $n!/\delta^{n-x}$ nach oben abgeschätzt werden. Wählt man etwa wieder $n > x + 1$, so erkennt man, dass ausch der zweite Term auf der rechten Seite mit $\delta \to \infty$ gegen Null konvergiert. Mit $\epsilon \to 0+$ und $\delta \to \infty$ verschwinden also die beiden ersten Terme auf der rechten Seite und wir erhalten $\Gamma(x + 1) = x\Gamma(x)$. Weiter ist

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = -e^{-t}\,\Big|_{t=0}^{t=\infty} = 1.$$

Damit ist die zweite Aussage des Satzes bewiesen.
Nun zeigen wir, dass die Gammafunktion auf $\mathbb{R}_+$ logarithmisch konvex ist. Wir benutzen hierzu, dass die Logarithmusfunktion log auf $\mathbb{R}_+$ eine konkave Funktion (bzw. $-\log$ eine konvexe Funktion) ist, da die zweite Ableitung des Logarithmus auf $\mathbb{R}_+$ negativ ist. Für $a, b \in \mathbb{R}_+$ und $\lambda \in (0, 1)$ ist daher

$$\log((1-\lambda)a + \lambda b) \ge (1-\lambda)\log a + \lambda \log b = \log a^{1-\lambda} + \log b^\lambda$$

und folglich

$$(*) \qquad a^{1-\lambda}b^\lambda \le (1-\lambda)a + \lambda b.$$

Nun seien $x, y \in \mathbb{R}_+$ und $\lambda \in (0, 1)$ vorgegeben. Dann ist

$$\begin{aligned}\frac{\Gamma((1-\lambda)x + \lambda y)}{\Gamma(x)^{1-\lambda}\Gamma(y)^\lambda} &= \frac{1}{\Gamma(x)^{1-\lambda}\Gamma(y)^\lambda}\int_0^\infty e^{-t}t^{(1-\lambda)x+\lambda y-1}\,dt \\ &= \int_0^\infty \Big(\frac{e^{-t}t^{x-1}}{\Gamma(x)}\Big)^{1-\lambda}\Big(\frac{e^{-t}t^{y-1}}{\Gamma(y)}\Big)^\lambda\,dt \\ &\le \int_0^\infty \Big[(1-\lambda)\frac{e^{-t}t^{x-1}}{\Gamma(x)} + \lambda\frac{e^{-t}t^{y-1}}{\Gamma(y)}\Big]\,dt \\ &\qquad \text{(Anwendung von } (*)) \\ &= 1.\end{aligned}$$

Durch Logarithmieren dieser Ungleichung folgt die logarithmische Konvexität der Gammafunktion. Der Satz ist bewiesen. □

Bemerkung: Ist die Gammafunktion auf $(0, 1]$ bekannt, so auch wegen der Funktionalgleichung $\Gamma(x+1) = x\Gamma(x)$ auf $(1, 2]$ und den folgenden Intervallen der Länge 1. Durch wiederholte Anwendung der Funktionalgleichung erhalten wir die Gültigkeit von

$$\Gamma(x+n) = (x+n-1)(x+n-2)\cdots(x+1)x\Gamma(x)$$

für alle $x \in \mathbb{R}_+$ und alle $n \in \mathbb{N}$. Setzen wir hier $n = 1$, so erhalten wir $\Gamma(n+1) = n!$, d. h. die Gammafunktion interpoliert die Fakultätfunktion. □

In Abbildung 118 haben wir die Gammafunktion auf einem Teil von $\mathbb{R}_+$ dargestellt.

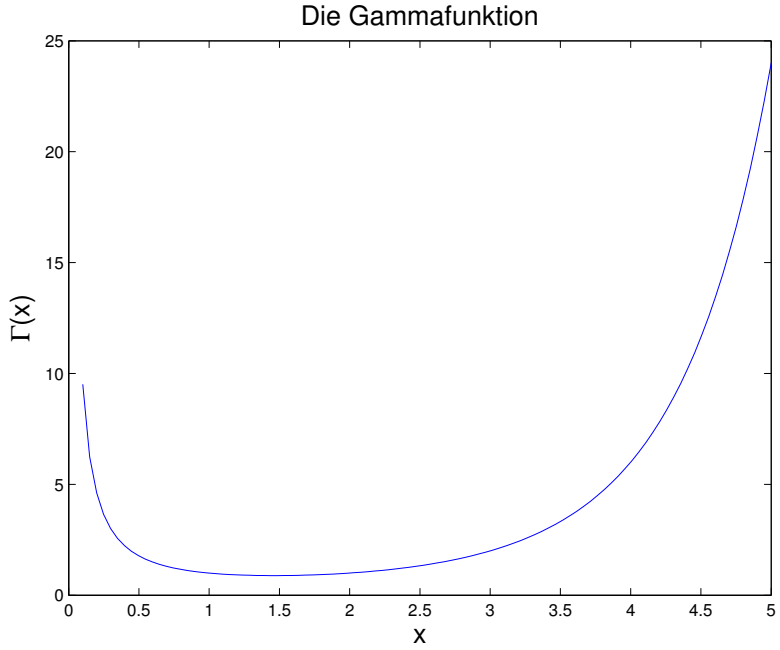


Abbildung 118: Die Gammafunktion

Durch die Integraldarstellung haben wir eine Definition der Gammafunktion nur für positive Argumente. Ist $x \in (-n, -n+1)$ mit $n \in \mathbb{N}$, so definieren wir den Wert der Gammafunktion in diesem x durch

$$\Gamma(x) := \frac{1}{x(x+1)\cdots(x+n-1)}\Gamma(x+n).$$

Damit ist die Gammafunktion auf ganz $\mathbb{R}$ mit Ausnahme von 0 und den negativen ganzen Zahlen definiert. Ferner gilt die Funktionalgleichung $\Gamma(x+1) = x\Gamma(x)$ für alle x, für die $\Gamma(x)$ definiert ist.

Jetzt kommen wir zu dem Hauptergebnis dieses Abschnitts, nämlich dem folgenden Satz von Mohr-Mollerup (1922). In ihm wird gezeigt, dass die Gammafunktion durch ihre Funktionalgleichung, ihre Anfangsbedingung und die logarithmische Konvexität

festgelegt ist. Bei N. Bourbaki (1961) wird die Gammafunktion im Kapitel "La Fonction Gamma" sogar durch diese Eigenschaften definiert! Den außerordentlich eleganten Beweis des folgenden Satzes haben wir E. Artin (1931, S. 14–15) fast wörtlich entnommen. Oder um es mit Bourbaki zu sagen: „Nous avons suivi d'assez près son exposé". Dieser Beweis hätte es meiner Meinung nach verdient, in das BUCH der Beweise aufgenommen zu werden.

Satz *Sei $f\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ eine Abbildung mit den folgenden beiden Eigenschaften.*

1. *Es ist $f(x+1) = xf(x)$ für alle $x \in \mathbb{R}_+$ und $f(1) = 1$.*
2. *Die Abbildung f ist auf $\mathbb{R}_+$ logarithmisch konvex.*

Dann stimmen f und die Gammafunktion auf $\mathbb{R}_+$ überein, d. h. es ist $f(x) = \Gamma(x)$ für alle $x > 0$.

Beweis: Dass die Gammafunktion den beiden Bedingungen genügt, ist in dem vorigen Satz bewiesen worden. Daher nehmen wir jetzt an, $f\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ sei eine Funktion, die den beiden obigen Bedingungen genügt. Für $x \in (0,1]$ und $n \in \mathbb{N}$ ist

$$(*) \qquad f(x+n) = (x+n-1)(x+n-2)\cdots(x+1)xf(x),$$

wie wir leicht durch vollständige Induktion nach n beweisen können. Aus der ersten Bedingung folgt $f(n) = (n-1)!$ für alle $n \in \mathbb{N}$, d. h. f und die Gammafunktion stimmen zumindest auf $\mathbb{N}$ überein. Es genügt zu zeigen, dass f und die Gammafunktion Γ auf $(0,1]$ gleich sind. Denn ist dies der Fall, so stimmen f und Γ wegen $(*)$ auf ganz $\mathbb{R}_+$ überein. Sei nun $x \in (0,1]$ und $n \in \mathbb{N}$ mit $n \geq 2$. Mit $g(x) := \log f(x)$ ist dann

$$(**) \qquad \frac{g(-1+n) - g(n)}{(-1+n) - n} \leq \frac{g(x+n) - g(n)}{(x+n) - n} \leq \frac{g(1+n) - g(n)}{(1+n) - n}.$$

Dies folgt aus der logarithmischen Konvexität von f bzw. der Konvexität von g. Denn die rechte Ungleichung erhält man aus

$$g(x+n) = g((1-x)n + x(1+n)) \leq (1-x)g(n) + xg(1+n),$$

die linke aus

$$\begin{aligned} g(n) &= g\Big(\Big(1 - \frac{1}{1+x}\Big)(-1+n) + \frac{1}{1+x}(x+n)\Big) \\ &\leq \Big(1 - \frac{1}{1+x}\Big)g(-1+n) + \frac{1}{1+x}g(x+n). \end{aligned}$$

Die linke und die rechte Seite kann man wegen $f(n) = (n-1)!$ vereinfachen und erhält

$$\log(n-1) \leq \frac{\log f(x+n) - \log(n-1)!}{x} \leq \log n$$

bzw.

$$\log((n-1)^x(n-1)!) \leq \log f(x+n) \leq \log(n^x(n-1)!).$$

Da die Exponentialfunktion monoton wachsend ist, ist

$$(n-1)^x(n-1)! \leq f(x+n) \leq n^x(n-1)!.$$

Mit Hilfe von Gleichung $(*)$ erhalten wir Ungleichungen für $f(x)$ selber:

$$\begin{aligned} \frac{(n-1)^x(n-1)!}{x(x+1)\cdots(x+n-1)} \leq f(x) &\leq \frac{n^x(n-1)!}{x(x+1)\cdots(x+n-1)} \\ &= = \frac{n^x n!}{x(x+1)\cdots(x+n)}\frac{x+n}{n}. \end{aligned}$$

Da diese Ungleichungen für alle $n \geq 2$ gelten, können wir auf der linken Seite n durch $n+1$ ersetzen. Daher ist

$$\frac{n^x n!}{x(x+1)\cdots(x+n)} \leq f(x) \leq \frac{n^x n!}{x(x+1)\cdots(x+n)}\frac{x+n}{n}$$

und folglich

$$f(x)\frac{n}{x+n} \leq \frac{n^x n!}{x(x+1)\cdots(x+n)} \leq f(x).$$

Mit $n \to \infty$ erhalten wir

$$f(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)\cdots(x+n)}.$$

Da die Gammafunktion aber auch den Bedingungen des Satzes genügt, ist auch

$$\Gamma(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)\cdots(x+n)}, \qquad x \in (0,1].$$

Also stimmen f und die Gammafunktion auf $(0,1]$ und damit auf ganz $\mathbb{R}_+$ überein. Der Satz ist bewiesen. □

Bemerkung: Wir wollen uns noch davon überzeugen dass

$$\Gamma(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)\cdots(x+n)}$$

nicht nur für alle $x \in (0,1]$, sondern für alle $x \in \mathbb{R}_+$ richtig ist. Hierzu definieren wir

$$\Gamma_n(x) := \frac{n^x n!}{x(x+1)\cdots(x+n)}.$$

Dann ist

$$\Gamma_n(x+1) = \frac{n n^x n!}{(x+1)(x+2)\cdots(x+n+1)} = x\Gamma_n(x)\frac{n}{x+n+1}$$

und folglich

$$\Gamma_n(x) = \frac{1}{x}\frac{x+n+1}{n}\Gamma_n(x+1).$$

Hieraus erkennt man: Existiert $\lim_{n\to\infty}\Gamma_n(x)$, so existiert auch $\lim_{n\to\infty}\Gamma_n(x+1)$. Und existiert umgekehrt $\lim_{n\to\infty}\Gamma_n(x+1)$ und ist $x\neq 0$, so existiert auch $\lim_{n\to\infty}\Gamma_n(x)$. Da $f(x)=\Gamma(x)$ für alle $x\in(0,1]$ und $f(x+1)=xf(x)$ gilt, stimmen f und Γ auf $\mathbb{R}_+$ überein. □

Die Darstellung

$$\Gamma(x)=\int_0^\infty e^{-t}t^{x-1}\,dt,\qquad x>0,$$

der Gammafunktion geht auf Euler zurück und wird auch *Eulers Integral zweiter Gattung* genannt. Dagegen stammt die Darstellung

$$\Gamma(x)=\lim_{n\to\infty}\frac{n^x n!}{x(x+1)\cdots(x+n)},\qquad x\in\mathbb{R}\setminus\{0,-1,-2,\ldots\},$$

von Gauß. Definiert man wie oben

$$\begin{aligned}\Gamma_n(x) &:= \frac{n^x n!}{x(x+1)\cdots(x+n)}\\ &= e^{x(\log n-1/1-1/2-\cdots-1/n)}\frac{1}{x}\cdot\frac{e^{x/1}}{1+x/1}\cdot\frac{e^{x/2}}{1+x/2}\cdot\cdots\cdot\frac{e^{x/n}}{1+x/n}\end{aligned}$$

und beachtet, dass der Grenzwert

$$(*)\qquad C:=\lim_{n\to\infty}\Big(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}-\log n\Big)$$

existiert (dieser wird oft *Eulersche Konstante* genannt), so erkennt man die Gültigkeit der auf Weierstraß zurückgehenden Darstellung

$$\Gamma(x)=e^{-Cx}\frac{1}{x}\lim_{n\to\infty}\prod_{i=1}^{n}\frac{e^{x/i}}{1+x/i}=e^{-Cx}\frac{1}{x}\prod_{i=1}^{\infty}\frac{e^{x/i}}{1+x/i}.$$

Dass der Grenzwert in $(*)$ existiert, sieht man folgendermaßen ein: Sei

$$C_n:=\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}-\log n,\qquad D_n:=C_n-\frac{1}{n}.$$

Dann ist

$$C_{n+1}-C_n=\frac{1}{n+1}-\log\Big(1+\frac{1}{n}\Big),\qquad D_{n+1}-D_n=\frac{1}{n}-\log\Big(1+\frac{1}{n}\Big).$$

Für alle $x\geq 0$ ist

$$\frac{x}{1+x}\leq\log(1+x)\leq x$$

und daher (setze $x:=1/n$)

$$\frac{1}{n+1}\leq\log\Big(1+\frac{1}{n}\Big)\leq\frac{1}{n}.$$

Folglich ist $\{C_n\}$ monoton fallend und $\{D_n\}$ monoton wachsend. Weiter ist $D_n < C_n$. Folglich ist $D_1 = 0$ eine untere Schranke für $\{C_n\}$. Also besitzt $\{C_n\}$ einen Grenzwert. Damit haben wir schon drei Darstellungen der Gammafunktion gewonnen, die mit den Namen von Euler, Gauß und Weierstraß verbunden sind. Aus der letzten folgern wir:

Satz *Die Gammafunktion ist auf $\mathbb{R}_+$ beliebig oft differenzierbar.*

Beweis: Wir zeigen, dass $\log\Gamma(x)$ auf $\mathbb{R}_+$ beliebig oft differenzierbar ist. Wegen $\Gamma(x) = e^{\log\Gamma(x)}$ folgt dann die Behauptung. Aus der Weierstraß-Darstellung der Gammafunktion erhalten wir

$$\begin{aligned}\log\Gamma(x) &= -Cx - \log x + \lim_{n\to\infty}\sum_{i=1}^{n}\Big(\frac{x}{i} - \log\Big(1+\frac{x}{i}\Big)\Big)\\ &= -Cx - \log x + \sum_{i=1}^{\infty}\Big(\frac{x}{i} - \log\Big(1+\frac{x}{i}\Big)\Big).\end{aligned}$$

Um nachzuweisen, dass $\log\Gamma(x)$ differenzierbar ist, benutzen wir die folgende Hilfsaussage (siehe z. B. W. Walter (1985, S. 261)):

- Sei $I \subset \mathbb{R}$ ein Intervall, $f_i\colon I \longrightarrow \mathbb{R}$ in I stetig differenzierbar, $i \in \mathbb{N}$. Ist $f(x) := \sum_{i=1}^{\infty} f_i(x)$ konvergent und $\sum_{i=1}^{\infty} f_i'(x)$ gleichmäßig konvergent in I, dann ist f stetig differenzierbar und

  $$f'(x) = \Big(\sum_{i=1}^{\infty} f_i(x)\Big)' = \sum_{i=1}^{\infty} f_i'(x)$$

  in I.

Die gliedweise differenzierte Reihe für $\log\Gamma(x)$ ist

$$-C - \frac{1}{x} + \sum_{i=1}^{\infty}\Big(\frac{1}{i} - \frac{1}{x+i}\Big) = -C - \frac{1}{x} + \sum_{i=1}^{\infty}\frac{x}{i(x+i)}.$$

Wegen des Weierstraßschen Majorantenkriteriums ist die Reihe auf $(0, r]$ für beliebiges $r > 0$ gleichmäßig konvergent, da das allgemeine Glied wegen $x > 0$ durch x/i^2 abgeschätzt werden kann und die Majorantenreihe $\sum_{i=1}^{\infty} r/i^2$ konvergiert. Auf $(0, r]$ ist also $\log\Gamma(x)$ differenzierbar und es ist

$$\frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} = -C - \frac{1}{x} + \sum_{i=1}^{\infty}\Big(\frac{1}{i} - \frac{1}{x+i}\Big)$$

auf $(0, r]$. Da $r > 0$ beliebig ist, gilt die Aussage auf ganz $\mathbb{R}_+$. Eine erneute Differentiation führt auf eine Reihe, die wegen des Majorantenkriteriums offenbar auf ganz $\mathbb{R}_+$ gleichmäßig konvergent ist. □

71 Das Problem der feindlichen Brüder bzw. dichteste Packungen von Kreisen in einem Quadrat

Von L. Moser (1960) stammt das folgende Problem:

- For small n (say $n \leq 10$) obtain sharp upper bounds for the smallest distance determined by n points in or on a unit square. Conjecture: 8 points in or on a unit square determine at least one distance $\leq \frac{1}{2}\sec 15^\circ = \frac{1}{2}(\sqrt{6}-\sqrt{2})$.

Man hat also die Aufgabe, Punkte $P_1, \ldots, P_n$ mit $n \geq 2$ in einem Quadrat P der Seitenlänge 1 so zu verteilen, dass $\min_{1\leq i<j\leq n} \|P_i - P_j\|_2$ maximal ist. Bei L. Collatz, W. Wetterling (1971, S. 174) wird vom *Problem der feindlichen Brüder* gesprochen. Der Grund ist der folgende: Auf einem quadratischen Grundstück wollen n verfeindete Brüder so ihre Häuser (diese werden als punktförmig angenommen!) bauen, dass der minimale Abstand von je zweien maximal ist. Bei gegebenem $n \in \mathbb{N}$ mit $n \geq 2$ ist also eine Lösung der Optimierungsaufgabe

$$\text{Maximiere} \quad d(P_1, \ldots, P_n) := \min_{1\leq i<j\leq n} \|P_i - P_j\|_2, \quad P_1, \ldots, P_n \in P,$$

gesucht. Den Optimalwert dieser Optimierungsaufgabe bezeichnen wir mit d_n, es ist also

$$d_n := \max_{P_1,\ldots,P_n\in P} \min_{1\leq i<j\leq n} \|P_i - P_j\|_2.$$

Ein eng mit dem Problem der feindlichen Brüder zusammenhängendes Problem besteht darin, eine *dichteste Packung* von $n \geq 2$ Kreisen mit einem gemeinsamen Radius in einem vorgegebenen Quadrat S der Seitenlänge $s > 0$ zu finden, d. h. n in S enthaltene Kreise, deren Inneres paarweise disjunkt ist, mit einem maximalen (gemeinsamen) Radius r zu bestimmen. Mit r_n bezeichnen wir den maximalen Radius der n Kreise einer Packung von S. Der Zusammenhang der beiden Probleme ist der folgende:

- Seien $P_1, \ldots, P_n$ Punkte aus P mit dem maximalen Minimalabstand d_n. Die Kreise mit den Mittelpunkten P_i, $i = 1, \ldots, n$, und dem Radius $\frac{1}{2}d_n$ bilden eine Packung für ein Quadrat der Seitenlänge $1 + d_n$. Daher gibt es n Kreise in S mit dem Radius $\frac{1}{2}d_n s/(1 + d_n)$, die eine Packung für S bilden. Folglich ist

$$r_n \geq \frac{s d_n}{2(1 + d_n)}.$$

  Seien umgekehrt $Q_1, \ldots, Q_n$ Mittelpunkte von Kreisen, die eine dichteste Packung von S liefern und daher einen Radius r_n besitzen. Diese liegen in einem Quadrat der Seitenlänge $s - 2r_n$ und haben paarweise mindestens den Abstand $2r_n$ voneinander. Daher gibt es im Einheitsquadrat n Punkte, die mindestens den Abstand $2r_n/(s - 2r_n)$ voneinander haben. Folglich ist

$$d_n \geq \frac{2r_n}{s - 2r_n}$$

bzw.

$$r_n \leq \frac{sd_n}{2(1+2d_n)}.$$

Insgesamt ist

$$\frac{r_n}{s} = \frac{d_n}{2(1+d_n)}$$

und man kann aus einer Lösung des einen Problems leicht eine des anderen gewinnen.

Eine Darstellung einer Lösung des Packungsproblems mit n gleichen Kreisen findet man z. B. in `http://www2.stetson.edu/~efriedma/cirinsqu/`, wobei hier Einheitskreise in ein Quadrat möglichst kleiner Seitenlänge gepackt werden. Man vergleiche auch Wikipedia unter `http://en.wikipedia.org/wiki/Circle_packing_in_a_square` oder `http://www.packomania.com`. Die Lösung der Probleme für $n = 2, 3, 4, 5$ ist relativ einfach. Wir geben sie in Abbildung 119 zunächst für $n = 2$ und $n = 3$ an. Es ist

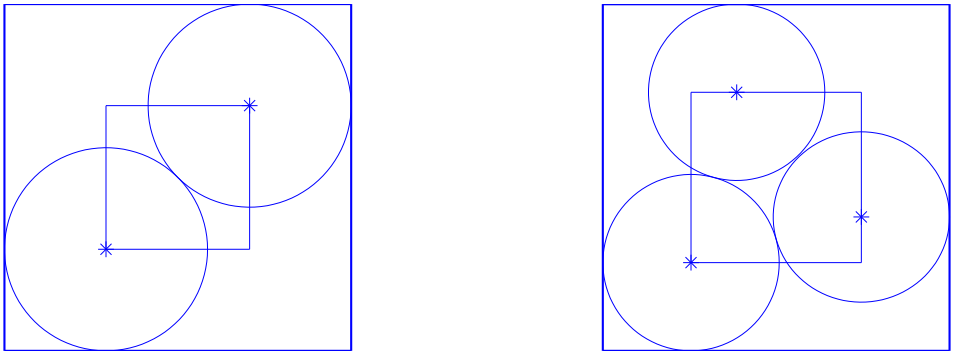


Abbildung 119: Dichteste Packungen für $n = 2$ und $n = 3$

$d_2 = \sqrt{2}$ und $d_3 = \sqrt{6} - \sqrt{2}$. Die Lösung für $n = 3$ erhält man natürlich dadurch, dass man in dem Einheitsquadrat ein gleichseitiges Dreieck mit Ecken auf dem Rand des Quadrats konstruiert. Die Lösungen für $n = 4$ und $n = 5$ sind außerordentlich naheliegend. Sie sind in Abbildung 120 angegeben. Offenbar ist $d_4 = 1$ und $d_5 = \sqrt{2}/2$. Einen Beweis dafür, dass die in Abbildung 123 links angegebene Konfiguration optimal für $n = 6$ ist, ist von J. B. M. Melissen (1994) angegeben worden. Es ist $d_6 = \frac{1}{6}\sqrt{13}$. Wir wollen diesen schönen, gar nicht so einfachen Beweis hier (wesentlich ausführlicher als der Autor) reproduzieren und formulieren den folgenden Satz.

Satz *Das Problem der sechs feindlichen Brüder hat die in Abbildung 121 angegebene Konfiguration als Lösung. Der minimale Abstand zwischen den sechs Punkten ist* $d_6 = \frac{1}{6}\sqrt{13}$.

Beweis: In der obigen Konfiguration ist der minimale Abstand zwischen zwei Punkten, wie man leicht nachweist, gleich $d_6 = \frac{1}{6}\sqrt{13}$. Wir nehmen an, wir hätten eine Konfiguration $\mathcal{N} = \{P_1, P_2, \ldots, P_6\}$ von sechs Punkten in dem Einheitsquadrat, für welche

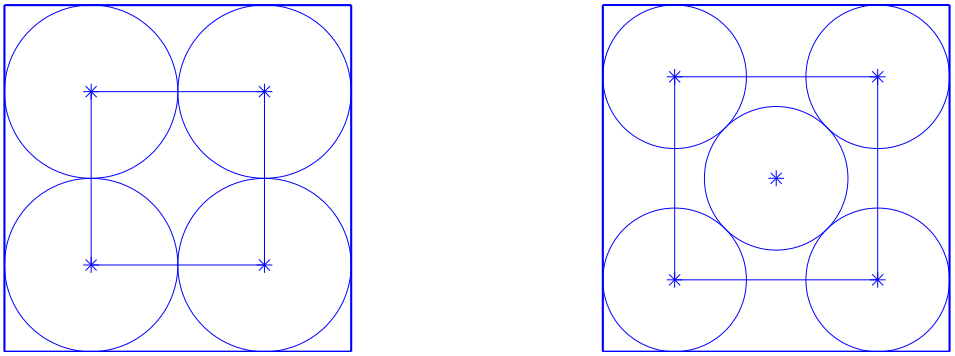


Abbildung 120: Dichteste Packungen für $n = 4$ und $n = 5$

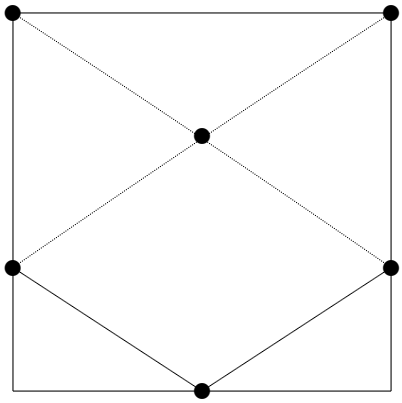


Abbildung 121: Die optimale Konfiguration für $n = 6$

der minimale Abstand der Punkte gleich $d \geq d_6$ ist und zeigen, dass notwendigerweise $d = d_6$ ist. Dann ist der Satz bewiesen.

Der Beweis basiert auf einer disjunkten Partitionierung des Einheitsquadrats $Q := \{(x, y) : 0 \leq x, y \leq 1\}$ in neun kleinere Gebiete, wie in Abbildung 122 angegeben. Genauer handelt es sich um Gebiete A_1, A_2, A_3, A_4, Gebiete B_1, B_2, B_3, B_4 und ein Gebiet C in der Mitte. Die Kanten zwischen zwei Gebieten sind i. Allg. rot und gehören zu einem A-Gebiet, nur die Kanten zwischen C und einem A-Gebiet sind grün und gehören zu C. Die Partition ist vollständig durch die Abstände $|p_8p_{10}| = |p_{10}p_{11}| = \frac{1}{3}$, $|p_5p_8| = \frac{1}{6}\sqrt{13}$ $(=d_6)$ und den offensichtlichen Symmetrien bestimmt. Der Durchmesser jedes der neun Teilgebiete ist $\leq d_6$. Dann kann jedes Teilgebiet höchstens einen Punkt der Konfiguration enthalten. Dies ist für die vier B-Gebiete und für das Gebiet C in der Mitte des Quadrats selbstverständlich, denn deren Durchmesser ist kleiner als d_6. Es ist aber auch für die A-Gebiete richtig, da diese den Durchmesser d_6 besitzen und die gemeinsame Kante mit C zu C gehört.

Nun überlegen wir uns:

- Gibt es sowohl einen Punkt aus $\mathcal{N}$ in einem B-Gebiet als auch zwei weitere Punkte aus $\mathcal{N}$ in benachbarten A-Gebieten (wir sagen dann, dass $\mathcal{N}$ die ABA-

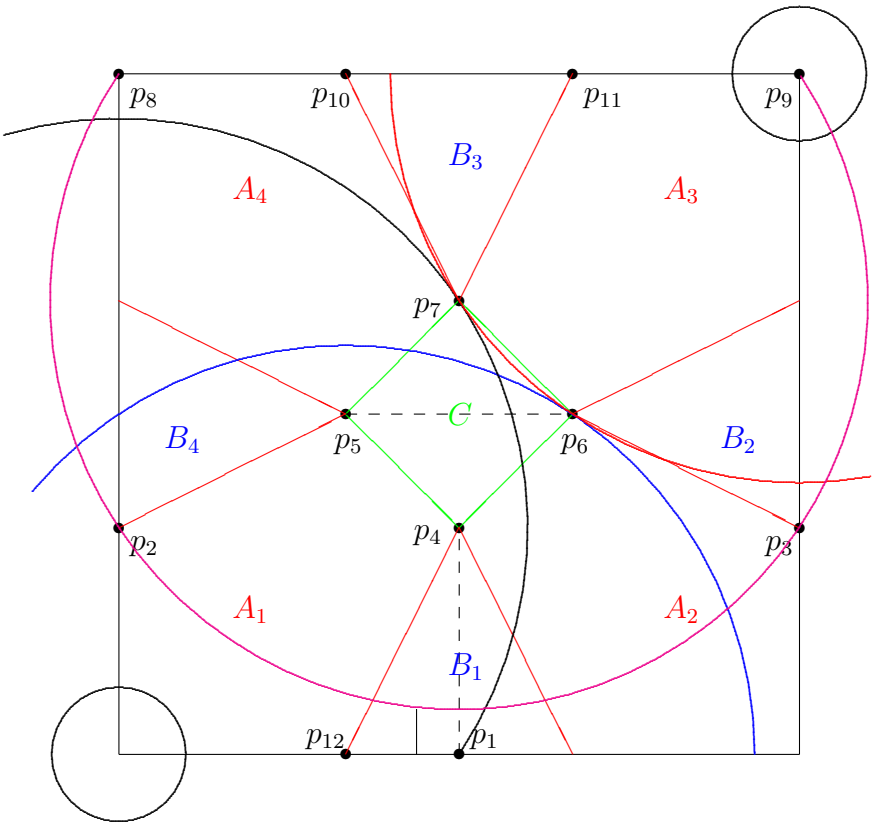


Abbildung 122: Partitionierung des Einheitsquadrats

Eigenschaft besitzt), so ist $d = d_6$ und die Aussage des Satzes ist richtig.

Denn: Wir nehmen o. B. d. A. an, in A_1, B_1 und A_2 würde je ein Punkt aus $\mathcal{N}$ liegen. Dann zerlegen wir die Vereinigung von A_1, B_2 und A_2 durch einen Schnitt von p_4 nach p_1 (in Abbildung 122 durch eine gestrichelte Linie angedeutet) in zwei Gebiete vom Durchmesser d_6. In einem der beiden Gebiete müssen zwei Punkte aus $\mathcal{N}$ liegen und daher ist $d = d_6$. Die Punkte aus $\mathcal{N}$ können offenbar nur die Punkte p_1, p_2, p_3 sein.

Jetzt machen wir eine Fallunterscheidung und nehmen zunächst an, es sei $\mathcal{N} \cap C = \emptyset$, kein Punkt der Konfiguration liege also in C. Wenn es drei oder vier A-Gebiete mit jeweils einem Punkt aus $\mathcal{N}$ gibt, so hat $\mathcal{N}$ die ABA-Eigenschaft und die Aussage ist richtig. Es bleibt der Fall zu betrachten, dass es zwei A-Gebiete mit Punkten aus $\mathcal{N}$ gibt. In den verbleibenden vier B-Gebieten liegt ebenfalls genau ein Punkt aus $\mathcal{N}$. Wir können annehmen, dass die beiden A-Gebiete bezüglich C einander gegenüber liegen, da $\mathcal{N}$ andernfalls die ABA-Eigenschaft besitzt, und wollen uns überlegen, dass diese Situation nicht eintreten kann. Es seien etwa jeweils ein Punkt aus $\mathcal{N}$ in A_1 und A_3 enthalten, die vier anderen Punkte liegen in den vier B-Gebieten. Daher nehmen wir an, es sei $P_j \in B_j$, $j = 1, \ldots, 4$, und $P_5 \in A_1$, $P_6 \in A_3$. Der offene Kreis $B(p_3; d_6)$ um $p_3 = (1, \frac{1}{3})$ mit dem Radius d_6 enthält ganz B_2 und damit insbesondere P_2. Daher ist $P_6 \notin B(p_3; d_6)$, da ja P_6 und P_2 mindestens einen Abstand d_6 von einander haben. Entsprechend enthält der offene Kreis $B(p_{10}; d_6)$ um $p_{10} = (\frac{1}{3}, 1)$ das Gebiet B_3 und damit auch den Punkt P_3. Genau wie eben folgt hieraus, dass $P_6 \notin B(p_{10}; d_6)$. Also liegt P_6 in dem Teil von A_3, der weder in $B(p_3; d_6)$ noch in $B(p_{10}; d_6)$ liegt. Das ist aber nur eine „kleine“ Umgebung von $p_9 = (1, 1)$. Denn der Schnittpunkt der beiden

abgeschlossenen Kreise $B[p_3; d_6]$ und $B[p_{10}; d_6]$ in A_3 ist in $(\frac{1}{12}\sqrt{10} + \frac{2}{3}, \frac{1}{12}\sqrt{10} + \frac{2}{3}) \approx (0.93, 0.93)$ und dieser Punkt hat zu $p_9 = (1, 1)$ einen Abstand $r := \sqrt{2}(\frac{1}{3} - \frac{1}{12}\sqrt{10}) \approx 0.0987$. Daher ist P_6 in einem Kreis um p_9 mit dem Radius r enthalten. Diesen haben wir in Abbildung 122 eingezeichnet. Ganz entsprechend ist der in A_1 enthaltene Punkt P_5 in einem Kreis um den Nullpunkt mit demselben Radius r enthalten. Auch diesen Kreis haben wir in Abbildung 122 eingetragen. Da aber $P_2 \in B_2$ von $P_6 \in B[p_9; r] \cap A_3$ mindestens einen Abstand $d \geq d_6$ besitzt, entsprechend $P_1 \in B_1$ von $P_5 \in B[0; r] \cap A_1$ mindestens einen Abstand $d \geq d_6$, erhalten wir Einschränkungen für die Lage von P_2 in B_2 bzw. P_1 in B_1. Der Abstand von $P_1 \in B_1$ zu $P_2 \in B_2$ ist höchstens so groß wie der Abstand von $(\sqrt{d_6^2 - r^2}, 0)$ (Schnittpunkt des Kreises um $(0, r)$ mit Radius d_6 und $y = 0$) zu $(1, 1 - \sqrt{d_6^2 - r^2})$ (Schnittpunkt des Kreises um $(1 - r, 1)$ mit Radius d_6 und $x = 1$), und dieser Abstand ist kleiner als d_6, wie man leicht nachrechnet. Für den Fall, dass kein Punkt aus $\mathcal{N}$ in C liegt, ist die Behauptung also bewiesen.

Jetzt sei $\mathcal{N} \cap C \neq \emptyset$, ein Punkt der Konfiguration liege also in C. Wenn zwei gegenüberliegende B-Gebiete wie z. B. B_1 und B_3 beide einen Punkt aus $\mathcal{N}$ enthalten, so ist $d = d_6$ und die Aussage ist richtig. Um dies einzusehen, denken wir uns die Vereinigung der Gebiete B_1, C und B_3 durch einen Schnitt von p_5 nach p_6 (in Abbildung 122 durch eine gestrichelte Linie angedeutet) in zwei Gebiete vom Durchmesser d_6 zerlegt. Eines der beiden Gebiete muss zwei Punkte aus $\mathcal{N}$ enthalten und es ist $d = d_6$. Gibt es also drei oder vier B-Gebiete mit einem Punkt aus $\mathcal{N}$, so sind wir fertig, denn dann gibt es mindestens ein sich gegenüberliegendes Paar von B-Gebieten, welche einen Punkt aus $\mathcal{N}$ enthalten. Es genügt also den Fall zu betrachten, in dem höchstens zwei B-Gebiete einen Punkt aus $\mathcal{N}$ enthalten. Enthält nur ein B-Gebiet einen Punkt aus $\mathcal{N}$, so hat $\mathcal{N}$ die ABA-Eigenschaft und wir sind fertig. Daher gehen wir jetzt davon aus, dass es genau zwei B-Gebiete gibt, die einen Punkt aus $\mathcal{N}$ enthalten und sich nicht gegenüber liegen. Diese beiden Gebiete seien etwa B_1 und B_4. Wir können annehmen, dass A_1 keinen Punkt aus $\mathcal{N}$ enthält, da $\mathcal{N}$ andernfalls die ABA-Eigenschaft besitzt. Es bleibt also der Fall zu untersuchen, dass außer in C noch in A_2, A_3, A_4, B_1 und B_4 ein Punkt von $\mathcal{N}$ enthalten ist. Wir werden zeigen, dass dieser Fall nicht eintreten kann.

Der Punkt aus $\mathcal{N}$ in C liegt weder in $B(p_2; d)$ (dieser Kreis enthält B_4), noch in $B(p_9; d)$ (enthält A_3) oder in $B(p_{12}; d)$ (enthält B_1), siehe Abbildung 122. Er liegt also in einem kleinen Gebiet in C, das von drei Kreissegmenten vom Radius d um p_2, p_9, p_{12} begrenzt ist. Da die Situation bezüglich der Diagonalen durch p_9 symmetrisch ist, können wir annehmen, dass der Punkt in C oberhalb der Diagonalen durch p_9 liegt. Seien $P_j = (x_j, y_j)$, $j = 1, \ldots, 4$, die Punkte von $\mathcal{N}$ in den Gebieten C, B_4, B_1 bzw. A_2. Da wir annehmen, dass P_1 oberhalb der Diagonalen liegt, ist $x_1 \leq y_1$. Weiter ist

$$1 - d < \frac{1}{2} \leq x_1 \leq \frac{7}{12} < d,$$

wobei wir berücksichtigen, dass $(\frac{7}{12}, \frac{7}{12})$ der zu C gehörende Mittelpunkt von p_6 und p_7 ist. Außerdem ist $y_2 \leq \frac{1}{2}$, da auch in A_4 ein Punkt aus $\mathcal{N}$ liegt. Da in B_1 und A_2 ein Punkt aus $\mathcal{N}$ liegt, ist $x_3 \leq \frac{1}{2}$. Da in A_2 und in C ein Punkt aus $\mathcal{N}$ liegt, ist $y_4 \leq \frac{1}{3}$.

Wir erhalten die folgenden Beziehungen, wobei wir immer wieder auf Abbildung 122 zurückgreifen

- Der Punkt $P_1 = (x_1, y_1) \in C$ liegt weder in $B(p_2; d)$ noch in $B(p_9; d)$. Daher ist

$$d^2 \le x_1^2 + \left(y_1 - \frac{1}{3}\right)^2, \qquad d^2 \le (1 - x_1)^2 + (1 - y_1)^2.$$

  Aus der zweiten Ungleichung erhalten wir wegen $0 \le 1 - y_1 \le 1 - x_1$, dass $d^2 \le 2(1 - x_1)^2$ und daher

$$\frac{1}{2} \le x_1 \le 1 - \frac{d}{\sqrt{2}}$$

  gilt. Unter Berücksichtigung von $x_1^2 \le d^2$, $(1 - x_1)^2 \le d^2$ und $y_1 \ge \frac{1}{3}$ erhalten wir ferner aus den beiden obigen Ungleichungen

$$\frac{1}{3} + \sqrt{d^2 - x_1^2} \le y_1 \le 1 - \sqrt{d^2 - (1 - x_1)^2}. \tag{1}$$

- Der Punkt $P_2 = (x_2, y_2) \in B_4$ hat mindestens den Abstand d von $P_1 \in C$, d. h. es ist

$$d^2 \le (x_1 - x_2)^2 + (y_1 - y_2)^2 \le x_1^2 + (y_1 - y_2)^2.$$

  Hierbei haben wir ausgenutzt, dass $x_2 \le \frac{1}{3} < \frac{1}{2} \le x_1$ und folglich $(x_1 - x_2)^2 \le x_1^2$. Es ist $\frac{1}{3} \le y_2 \le \frac{1}{2} \le y_1$ und daher

$$\frac{1}{3} \le y_2 \le y_1 - \sqrt{d^2 - x_1^2}. \tag{2}$$

- Der Punkt $P_3 = (x_3, y_3) \in B_1$ hat mindestens den Abstand d von $P_2 \in B_4$, d. h. es ist

$$d^2 \le (x_3 - x_2)^2 + (y_3 - y_2)^2 \le x_3^2 + y_2^2.$$

  Hierbei haben wir ausgenutzt, dass $(y_3 - y_2)^2 \le y_2^2$, da $y_3 \le \frac{1}{3} \le y_2$. Wegen $y_2 \le \frac{1}{2} < d$ und $x_3 \le \frac{1}{2}$ folgt hieraus

$$\sqrt{d^2 - y_2^2} \le x_3 \le \frac{1}{2}. \tag{3}$$

Oben haben wir notiert, dass

$$\frac{1}{2} \le x_1 \le 1 - \frac{d}{\sqrt{2}} \le 1 - \frac{d_6}{\sqrt{2}} = 1 - \frac{\sqrt{26}}{12}.$$

Schreiben wir $x_1 = \frac{1}{2} + \epsilon$, so erhalten wir

$$0 \le \epsilon \le \epsilon_0 := \frac{6 - \sqrt{26}}{12} \approx 0.0751.$$

Aus (1) erhalten wir unter Beachtung von $\epsilon \in [0, \epsilon_0]$, dass

$$y_1 \le 1 - \sqrt{d^2 - (1 - x_1)^2} \le 1 - \sqrt{d_6^2 - \left(\frac{1}{2} - \epsilon\right)^2} = 1 - \sqrt{\frac{1}{9} + \epsilon - \epsilon^2} \le \frac{2}{3} - \frac{7}{6}\epsilon,$$

wobei die letzte Ungleichung wegen

$$\Big(1+\frac{49}{36}\Big)\epsilon_0 \le \frac{2}{9}$$

richtig ist. Wieder für $\epsilon \in [0,\epsilon_0]$ erhalten wir aus (2), dass

$$y_2 \le y_1 - \sqrt{d^2-x_1^2} \le \frac{2}{3} - \frac{7}{6}\epsilon - \sqrt{\frac{13}{36} - \Big(\frac{1}{2}+\epsilon\Big)^2} = \frac{2}{3} - \frac{7}{6}\epsilon - \sqrt{\frac{1}{9} - \epsilon - \epsilon^2} \le \frac{1}{3} + \epsilon,$$

wobei die letzte Ungleichung wegen

$$\Big(1+\frac{169}{36}\Big)\epsilon_0 \le \frac{4}{9}$$

richtig ist. Aus (3) erhalten wir für $\epsilon \in [0,\epsilon_0]$, dass

$$x_3 \ge \sqrt{d^2-y_2^2} \ge \sqrt{\frac{13}{36} - \Big(\frac{1}{3}+\epsilon\Big)^2} = \sqrt{\frac{1}{4} - \frac{2}{3}\epsilon - \epsilon^2} \ge \frac{1}{2} - \frac{5}{6}\epsilon,$$

wobei wir für die letzte Ungleichung ausgenutzt haben, dass

$$\Big(1+\frac{25}{36}\Big)\epsilon_0 \le \frac{1}{6}.$$

Insbesondere ist $0.4374 \approx \frac{1}{2} - \frac{5}{6}\epsilon_0 \le x_3 \le \frac{1}{2}$. Außerdem liegt $P_3 = (x_3, y_3)$ außerhalb des (offenen) magentafarbenen Kreises $B(p_7; d)$ um $p_7 = (\frac{1}{2}, \frac{2}{3})$ mit dem Radius d (siehe Abbildung 122), da andernfalls der Abstand zwischen P_1 und P_3 kleiner als d wäre. Also ist

$$\frac{1}{2} - \frac{5}{6}\epsilon_0 \le x_3 \le \frac{1}{2}, \qquad d^2 \le \Big(\frac{1}{2}-x_3\Big)^2 + \Big(\frac{2}{3}-y_3\Big)^2.$$

Aus diesen beiden Ungleichungen erhalten wir

$$\frac{13}{36} = d_6^2 \le d^2 \le \Big(\frac{5}{6}\epsilon_0\Big)^2 + \Big(\frac{2}{3}-y_3\Big)^2$$

und hieraus

$$y_3 \le \frac{2}{3} - \sqrt{\frac{13}{36} - \Big(\frac{5}{6}\epsilon_0\Big)^2} \le 0.07.$$

Hiermit können wir zeigen:

- Der Punkt $P_4 = (x_4, y_4) \in A_2$ hat mindestens den Abstand d von $P_3 \in B_1$, d. h. es ist
$$d^2 \le (x_4-x_3)^2 + (y_4-y_3)^2 \le (1-x_3)^2 + y_4^2.$$
Hierbei haben wir ausgenutzt, dass $x_3 \le \frac{1}{2} \le x_4$ und daher $0 \le x_4 - x_3 \le 1 - x_3$ und damit $(x_4-x_3)^2 \le (1-x_3)^2$. Ferner ist $(y_4-y_3)^2 \le y_4^2$, da $y_3 \le 2y_4$. Denn andernfalls wäre
$$-0.07 \le -y_3 \le y_4 - y_3 \le y_4 < \frac{1}{2}y_3 \le 0.035$$

und daher $(y_4 - y_3)^2 \leq 0.0049$. Dann ist

$$0.3611 \leq d_6^2 \leq d^2 \leq (1 - x_3)^2 + (y_4 - y_3)^2 \leq \Big(\frac{1}{2} + \frac{5}{6}\epsilon_0\Big)^2 + 0.0049 \leq 0.3214,$$

ein Widerspruch. Unter Berücksichtigung von

$$0 \leq 1 - x_3 \leq \frac{1}{2} + \frac{5}{6}\epsilon \leq \frac{1}{2} + \frac{5}{6}\epsilon_0 \leq d$$

für $\epsilon \in [0, \epsilon_0]$ ist damit

$$\sqrt{d^2 - (1 - x_3)^2} \leq y_4 \leq \frac{1}{3} \tag{4}$$

bewiesen.

Für $\epsilon \in [0, \epsilon_0]$ erhalten wir aus (4), dass

$$y_4 \geq \sqrt{d^2 - (1 - x_3)^2} \geq \sqrt{\frac{13}{36} - \Big(\frac{1}{2} + \frac{5}{6}\epsilon\Big)^2} = \sqrt{\frac{1}{9} - \Big(\frac{5}{6}\epsilon + \frac{25}{36}\epsilon^2\Big)} \geq \frac{1}{3} - \frac{5}{3}\epsilon,$$

wobei die letzte Ungleichung wegen

$$\frac{125}{36}\epsilon_0 \leq \frac{5}{8}$$

richtig ist. Weiter zeigen wir:

- Der Punkt $P_4 = (x_4, y_4) \in A_2$ hat mindestens den Abstand d von $P_1 \in C$, d. h. es ist
$$d^2 \leq (x_4 - x_1)^2 + (y_4 - y_1)^2 \leq (1 - x_1)^2 + \Big(\frac{1 - y_4}{2}\Big)^2.$$
Hierbei haben wir zunächst ausgenutzt, dass $(x_4 - x_1)^2 \leq (1 - x_1)^2$, da
$$(1 - x_1)^2 - (x_4 - x_1)^2 = (1 - x_4^2) - 2x_1(1 - x_4) = \underbrace{(1 - x_4)}_{\geq 0}\underbrace{(1 + x_4 - 2x_1)}_{\geq 0} \geq 0.$$
Für $\epsilon \in [0, \epsilon_0]$ ist weiter
$$0 \leq \frac{1 - y_4}{2} \leq \frac{1}{2} - \frac{1}{2}\Big(\frac{1}{3} - \frac{5}{6}\epsilon\Big) = \frac{1}{3} + \frac{5}{6}\epsilon \leq \frac{1}{3} + \frac{5}{6}\epsilon_0 \leq d_6 \leq d$$
und daher
$$\frac{1}{2} \leq x_1 \leq 1 - \sqrt{d^2 - \Big(\frac{1 - y_4}{2}\Big)^2}. \tag{5}$$

Für $\epsilon \in [0, \epsilon_0]$ erhalten wir aus (5), dass für $d > d_6$ gilt

$$\frac{1}{2} + \epsilon = x_1 < 1 - \sqrt{d_6^2 - \Big(\frac{1}{3} + \frac{5}{6}\epsilon\Big)^2} = 1 - \sqrt{\frac{1}{4} - \frac{5}{9}\epsilon - \frac{25}{36}\epsilon^2} \leq \frac{1}{2} + \epsilon,$$

wobei wir für die letzte Ungleichung ausgenutzt haben, dass

$$\Big(1+\frac{25}{36}\Big)\epsilon_0 \le \frac{4}{9}.$$

Es ergibt sich also nur dann kein Widerspruch, wenn $d = d_6$ und $\epsilon = 0$. Es ist dann notwendig $y_2 = \frac{1}{3}$. In B_4 gibt es aber keinen Punkt, dessen zweite Koordinate gleich $\frac{1}{3}$ ist, denn $(0, \frac{1}{3})$ gehört vereinbarungsgemäß zu A_1 und nicht zu B_4. Damit haben wir gezeigt, dass $\mathcal{N}$ notwendigerweise die ABA-Eigenschaft besitzt und damit die Aussage des Satzes richtig ist. Seien z. B. drei Punkte von $\mathcal{N}$ in A_1, B_1 und A_2 enthalten. Diese Punkte müssen p_2, p_1 und p_3 sein. Es kann keine Punkte aus $\mathcal{N}$ in B_2 oder B_4 geben, weil diese sozusagen zu nah an p_3 bzw. p_2 sind. Weiter können die drei übrigen Punkte nicht in A_3, B_3 und A_4 liegen, da sie außerhalb eines Kreises mit dem Radius d_6 um p_2 bzw. p_3 liegen müssen. Daher muss es einen Punkt aus $\mathcal{N}$ in C geben. Dies muss notwendigerweise p_7 sein, was man daran erkennt, dass man wiederum um p_2 und p_3 einen Kreis mit dem Radius d_6 schlägt. Die beiden restlichen Punkte sind p_8 und p_9. Dies ergibt die in Abbildung 121 angegebene Konfiguration. □

Der von J. Schaer (1965) behandelte Fall $n = 7$ ist besonders, da ein Kreis frei ist, d. h. innerhalb gewisser Grenzen frei gewählt werden kann. Diesen Fall haben wir in Abbildung 123 rechts angegeben. Der rot gezeichnete Kreis rechts oben ist frei. Insbesondere ist die Lösung nicht eindeutig. Es ist $d_7 = 2(2-\sqrt{3})$. Von J. Schaer, A.

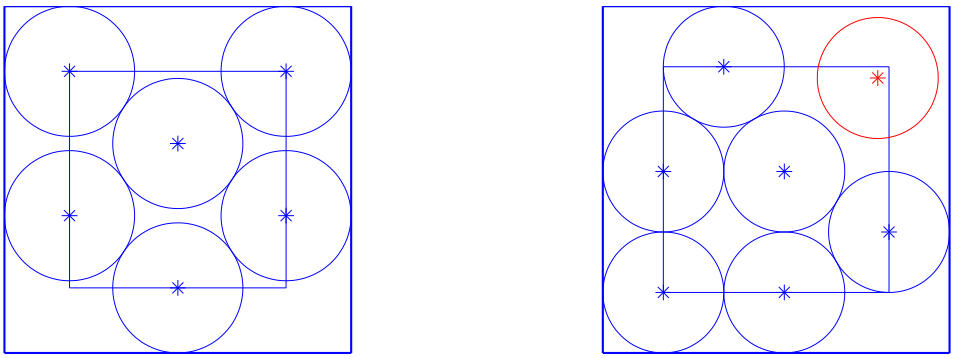


Abbildung 123: Dichteste Packungen für $n = 6$ und $n = 7$

Meir (1965) stammt ein (keineswegs einfacher) Beweis, dass die Vermutung von L. Moser (1960) für $n = 8$ richtig ist. J. Schaer (1965) löste das Problem für $n = 9$. In beiden Fällen sind die Lösungen nicht überraschend. Sie sind in Abbildung 124 angegeben. Es ist $d_8 = \frac{1}{2}(\sqrt{6}-\sqrt{2})$ und $d_9 = \frac{1}{2}$. Wir erhalten mit dem `format long` von MATLAB die in Tabelle 11 angegebenen Werte. Von C. de Groot, R. Peikert, D. Würtz (1990) und R. Peikert, D. Würtz, M. Monagan, C. de Groot (1992) sind mit Computerunterstützung Lösungen für $n = 10, \ldots, 20$ angegeben worden. Interessant ist der Fall $n = 10$. Hier wurde von M. Goldberg (1970) vermutet, dass die in Abbildung 125 links angegebene symmetrische Konfiguration optimal sei. Der maximale Minimalabstand der Goldbergschen (aus zehn Punkten bestehenden) Konfiguration ist offenbar $d_G = \frac{5}{12} \approx 0.416667$, was z. B. der Abstand von $P := (\frac{1}{4}, \frac{2}{3})$ und

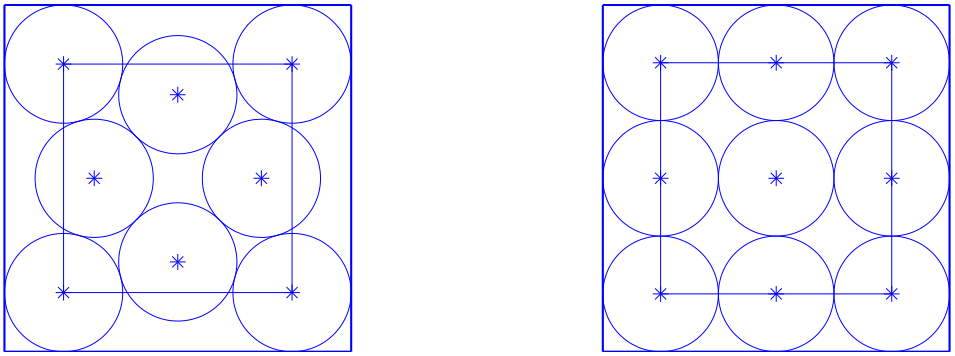


Abbildung 124: Dichteste Packungen für $n = 8$ und $n = 9$

n	d_n	$\approx d_n$	$\approx r_n$
2	$\sqrt{2}$	1.414213562373095	0.292893218813453
3	$\sqrt{6}-\sqrt{2}$	1.035276180410083	0.254333095030250
4	1	1.000000000000000	0.250000000000000
5	$\frac{1}{2}\sqrt{2}$	0.707106781186548	0.207106781186548
6	$\frac{1}{6}\sqrt{13}$	0.600925212577332	0.187680601147477
7	$4-2\sqrt{3}$	0.535898384862246	0.174457630187009
8	$\frac{1}{2}(\sqrt{6}-\sqrt{2})$	0.517638090205041	0.170540688701054
9	$\frac{1}{2}$	0.500000000000000	0.166666666666667

Tabelle 11: Maximaler Minimalabstand und Packungsradius im Einheitskreis

$Q := (\frac{1}{2}, 1)$ ist. Von J. Schaer (1971) stammt eine Konfiguration mit (geringfügig) größerem maximalen Minimalabstand d_S, siehe Abbildung 125 rechts. Bei ihr sind zwei Kreise frei, außerdem hat man Symmetrie bezüglich der beiden Diagonalen. Wir wollen uns überlegen, dass

$$d_S = (4+\sqrt{2})\frac{\sqrt{\sqrt{8}+1}-\sqrt{2}}{7} \approx 0.419542.$$

Hierzu betrachten wir Abbildung 126. Mit einem $d \in (0,1)$ liegen die Punkte der Konfiguration von Schaer in den vier Eckpunkten des Einheitsquarates, außerdem in den Punkten

$$P_1(d) := (0,d), \quad P_2(d) := (1-d,1), \quad Q_1(d) := (1,1-d), \quad Q_2(d) := (d,0)$$

auf dem Rande des Einheitsquadrates sowie den Punkten

$$P(d) := \Big(\frac{1}{2}\Big(1-\frac{d}{\sqrt{2}}\Big), \frac{1}{2}\Big(1+\frac{d}{\sqrt{2}}\Big)\Big), \qquad Q(d) := \Big(\frac{1}{2}\Big(1+\frac{d}{\sqrt{2}}\Big), \frac{1}{2}\Big(1-\frac{d}{\sqrt{2}}\Big)\Big),$$

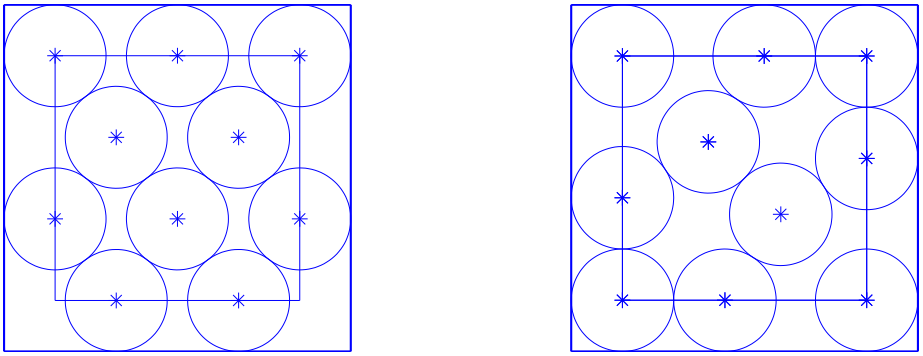


Abbildung 125: Packung von zehn Kreisen nach Goldberg und Schaer

die auf der Diagonalen im Einheitsquadrat symmetrisch zum Mittelpunkt $(\frac{1}{2},\frac{1}{2})$ liegen und den Abstand d voneinander haben. Der Parameter d wird so bestimmt, dass der (rote) Umkreis zum Dreieck $\triangle P_1(d)P_2(d)Q(d)$ bzw. der (blaue) Umkreis zum Dreieck $\triangle Q_1(d)Q_2(d)P(d)$ gerade $P(d)$ bzw. $Q(d)$ zum Umkreismittelpunkt und d jeweils als Umkreisradius besitzen. Aus Symmetriegründen ist dies genau dann der Fall, wenn

$$d = |P_1(d)P(d)| = \sqrt{\Big[\frac{1}{2}\Big(1-\frac{d}{\sqrt{2}}\Big)\Big]^2 + \Big[d-\frac{1}{2}\Big(1+\frac{d}{\sqrt{2}}\Big)\Big]^2}$$

bzw.

$$\begin{aligned} d^2 &= \frac{1}{4}\Big(1-\frac{d}{\sqrt{2}}\Big)^2 + d^2 - d\Big(1+\frac{d}{\sqrt{2}}\Big) + \frac{1}{4}\Big(1+\frac{d}{\sqrt{2}}\Big)^2 \\ &= \frac{1}{2} - d + \Big(\frac{1}{4}+1-\frac{1}{\sqrt{2}}\Big)d^2 \end{aligned}$$

oder

$$\frac{1}{2} - d + \Big(\frac{1}{4}-\frac{1}{\sqrt{2}}\Big)d^2 = 0.$$

Als positive Lösung dieser quadratischen Gleichung erhält man

$$d_S := \frac{2\sqrt{\sqrt{2}+\frac{1}{2}}-2}{2\sqrt{2}-1} = (2\sqrt{2}+1)\frac{2\sqrt{\sqrt{2}+\frac{1}{2}}-2}{7} = (4+\sqrt{2})\frac{\sqrt{\sqrt{8}+1}-\sqrt{2}}{7}$$

und genau das wollten wir zeigen.

Auch die Konfiguration von Schaer wurde verbessert, nämlich durch G. Valette (1989), siehe Abbildung 127 links. Eine ausführliche Darstellung findet man auch bei J. Werner (1992, S. 83–84). Die Konstruktion verläuft folgendermaßen, wobei $|P_iP_j|$ den euklidischen Abstand zwischen Punkten P_i und P_j bezeichne. Sei $d \in [\sqrt{2}-1, \frac{1}{2}]$ vorgegeben. Genau für diese d ist die folgende Konstruktion durchführbar. In Abbildung 128 haben wir $d = 0.45$ gewählt. Man definiere

$$P_1(d) := (2d, 0), \quad P_2(d) := (d, 0), \quad P_3(d) := (0, 0), \quad P_4(d) := (0, d), \quad P_5(d) := (0, 1)$$

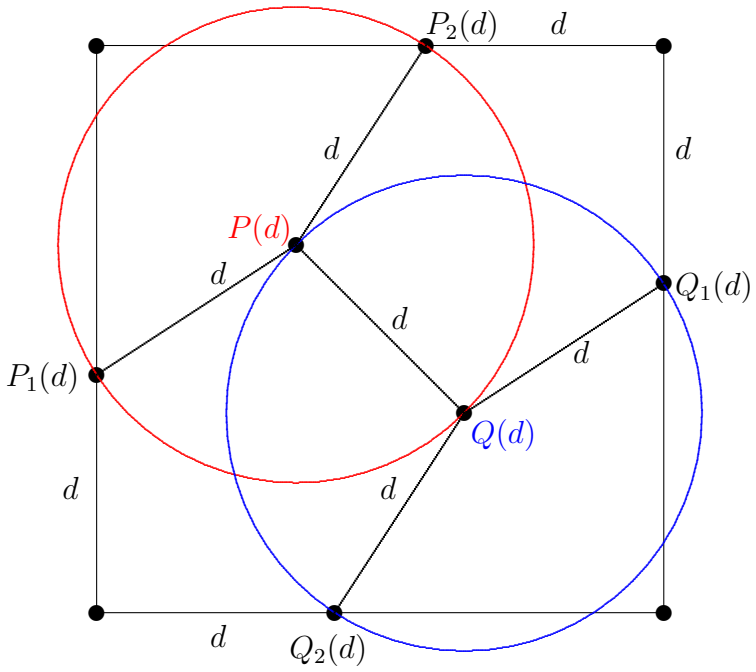


Abbildung 126: Der maximale Minimalabstand in der Konfiguration von Schaer

und anschließend $P_6(d)$ im Innern und $P_7(d)$, $P_8(d)$ und $P_9(d)$ auf dem Rande des Einheitsquadrates so, dass

$$d = |P_4(d)P_6(d)| = |P_5(d)P_6(d)| = |P_6(d)P_7(d)| = |P_7(d)P_8(d)| = |P_8(d)P_9(d)|.$$

Man setze zur Abkürzung

$$x(d) := \sqrt{d^2 - \left(\frac{1-d}{2}\right)^2}, \qquad y(d) := \sqrt{d^2 - (1-2x(d))^2}.$$

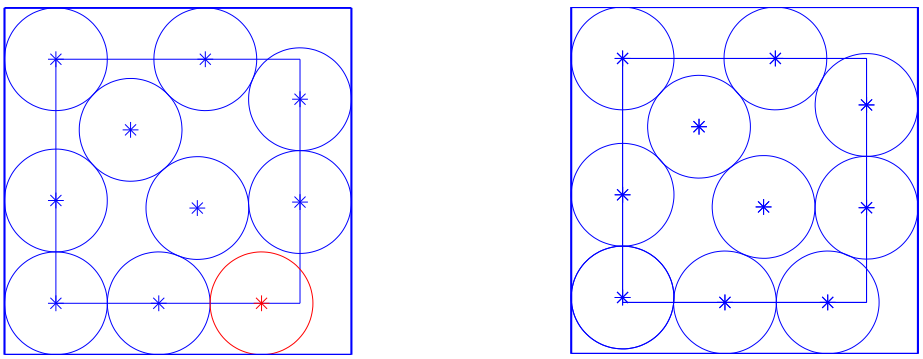


Abbildung 127: Packung von zehn Kreisen nach Valette und Schlüter

Wegen $d \in [\sqrt{2}-1, \frac{1}{2}]$ sind $x(d)$ und $y(d)$ reell. Nach einfacher Rechnung ist

$$P_6(d) = \Big(x(d), \frac{d+1}{2}\Big), \qquad P_7(d) = (2x(d), 1), \qquad P_8(d) = (1, 1-y(d))$$

sowie

$$P_9(d) = (1, 1-d-y(d)).$$

Schließlich sei $P_{10}(d)$ der Umkreismittelpunkt und $F(d)$ der Radius des Umkreises zum Dreieck $\triangle P_2(d)P_6(d)P_9(d)$. Also ist

$$F(d) = |P_2(d)P_{10}(d)| = |P_6(d)P_{10}(d)| = |P_9(d)P_{10}(d)|.$$

Bezeichnet man mit

$$a(d) := |P_2(d)P_6(d)|, \qquad b(d) := |P_6(d)P_9(d)|, \qquad c(d) := |P_9(d)P_2(d)|$$

die Seitenlängen des Dreiecks $\triangle P_2(d)P_6(d)P_9(d)$ und setzt man anschließend

$$s(d) := \frac{1}{2}[a(d)+b(d)+c(d)],$$

so ist der Umkugelradius $F(d)$ des Dreiecks $\triangle P_2(d)P_6(d)P_9(d)$ durch

$$F(d) := \frac{a(d)b(d)c(d)}{4\sqrt{s(d)(s(d)-a(d))(s(d)-b(d))(s(d)-c(d))}}$$

gegeben. Mit $d = 0.45$ geben wir die Konstruktion in Abbildung 128 an, wobei der Umkreis zu $\triangle P_2(d)P_6(d)P_9(d)$ rot eingetragen ist. Durch Plotten der Funktion $F(\cdot)$ auf dem Intervall $[\sqrt{2}-1, \frac{1}{2}]$ erkennt man, dass es genau ein $d_V \in [\sqrt{2}-1, \frac{1}{2}]$ mit $F(d_V) = d_V$, also genau einen Fixpunkt d_V von $F(\cdot)$ gibt. Eine genauere Berechnung ergibt, dass $d_V \approx 0.4211897032$. Dann ist $\{P_1(d_V), \ldots, P_{10}(d_V)\}$ eine Konfiguration von zehn Punkten im Einheitsquadrat mit

$$\min_{1 \le i < j \le 10} |P_i(d_V)P_j(d_V)| = d_V,$$

siehe Abbildung 128 rechts. Wegen $d_V > d_S > d_G$ ist die von Valette angegebene Konfiguration besser als die von Schaer (und die von Goldberg). Man beachte, dass es in der Konfiguration von Valette einen "freien" Kreis gibt, in Abbildung 127 links ist dieser rot gezeichnet.

Zum Zeitpunkt des Erscheinens der Arbeit von Valette war die für zehn Punkte optimale Konfiguration schon bekannt. Sie wurde von K. Schlüter (1979) angegeben, ihre Optimalität wurde von ihm in seiner Arbeit allerdings nur vermutet und nicht bewiesen. Der maximale Minimalabstand für zehn Punkte ist $d_{10} \approx 0.4212795440$. Dieses Ergebnis wurde auch von M. Mollard, C. Payan (1990) wiederentdeckt. Angelehnt an deren Darstellung wollen wir nun die Schlütersche Konfiguration schildern.

Wir geben uns ein $d \in [\sqrt{2}-1, 0.5]$ vor und *versuchen*, auf die folgende Weise neun Punkte $P_1(d), \ldots, P_9(d)$ im Einheitsquadrat zu konstruieren, wobei wir hier schon

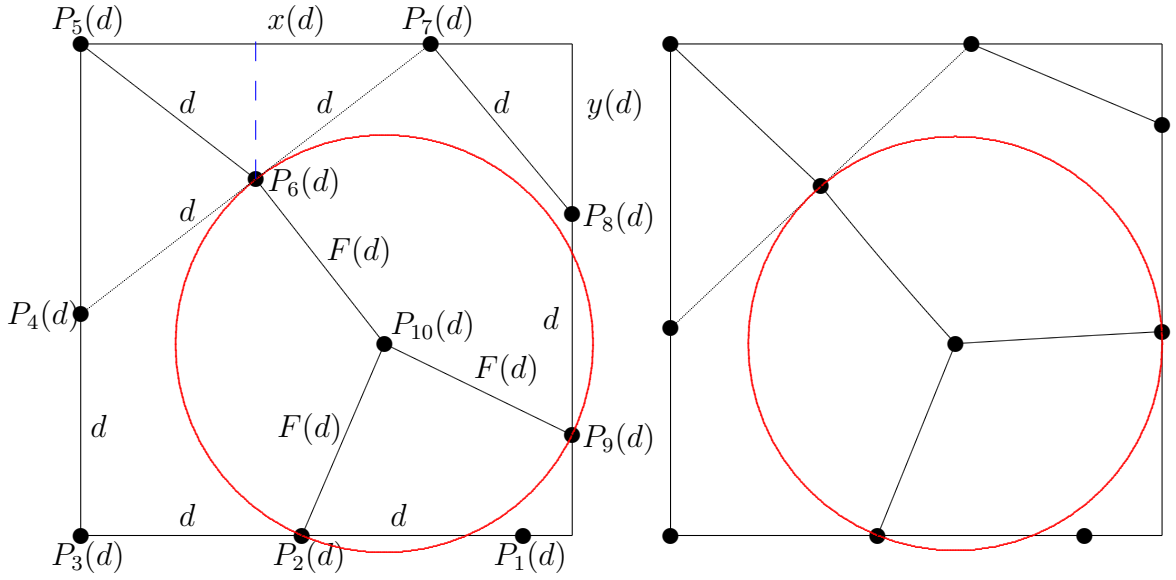


Abbildung 128: Konstruktion und Konfiguration nach Valette

betonen wollen, dass dies nur für gewisse d aus einem kleinen Intervall, z. B. $d \in [0.418, 0.422]$, möglich ist. Weiter geben wir uns ein "kleines" positives ϵ vor. Die Konstruktion veranschaulichen wir uns in Abbildung 129 links, wobei wir $d = 0.42$ und $\epsilon = 0.02$ benutzt haben. Wir wollen die zu konstruierenden Punkte $P_1(d), \ldots, P_9(d)$ so plazieren wie in Abbildung 128 bei der Konstruktion nach Valette. Wir setzen

$$P_3(\epsilon, d) := (0, \epsilon), \qquad P_2(\epsilon, d) := (u(\epsilon, d), 0), \qquad P_1(\epsilon, d) := (u(\epsilon, d) + d, 0),$$

wobei

$$u(\epsilon, d) := \sqrt{d^2 - \epsilon^2}.$$

Anschließend definieren wir

$$P_4(\epsilon, d) := (0, \epsilon + d), \qquad P_5(\epsilon, d) := (0, 1)$$

sowie

$$P_9(1, \epsilon.d) := (1, v(\epsilon, d)), \qquad P_8(\epsilon, d) := (1, v(\epsilon, d) + d)$$

mit

$$v(\epsilon, d) := \sqrt{d^2 - (1 - d - u(\epsilon, d))^2}$$

und

$$P_7(\epsilon, d) := (1 - w(\epsilon, d), 1), \qquad P_6(\epsilon, d) := \Big(\frac{1 - w(\epsilon, d)}{2}, 1 - x(\epsilon, d)\Big),$$

wobei

$$w(\epsilon, d) := \sqrt{d^2 - (1 - d - v(\epsilon, d))^2}, \qquad x(\epsilon, d) := \sqrt{d^2 - \Big(\frac{1 - w(\epsilon, d)}{2}\Big)^2}.$$

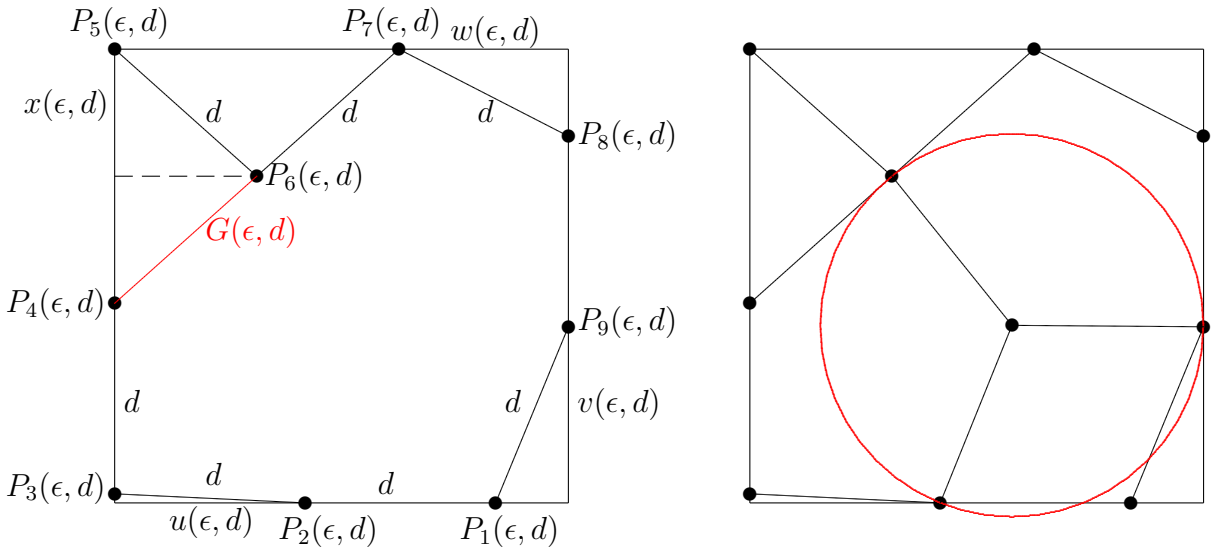


Abbildung 129: Konstruktion und Konfiguration nach Schlüter

Hierbei sind naturlich $P_3(\epsilon,d)$ von d und $P_5(\epsilon,d)$ von ϵ und d unabhängig. Schließlich setzen wir

$$G(\epsilon,d) := |P_6(\epsilon,d)P_4(\epsilon,d)| = \sqrt{\Big(\frac{1-w(\epsilon,d)}{2}\Big)^2 + (1-\epsilon-d-x(\epsilon,d))^2}$$

und bestimmen in Abhängigkeit von d ein positives $\epsilon^*(d)$ mit $G(\epsilon^*(d),d) = d$. Wir werden uns später überlegen, dass die Gleichung $G(\epsilon,d) = d$ für jedes $d \in [0.418, 0.422]$ genau eine Lösung $\epsilon^*(d)$ in $[0, 0.05]$ besitzt. Anschließend setzen wir

$$P_i(d) := P_i(\epsilon^*(d),d), \qquad i = 1,\ldots,9.$$

Danach bestimmen wir zum Dreieck $\triangle P_2(d)P_6(d)P_9(d)$ den Umkreismittelpunkt $P_{10}(d)$ und den Umkreisradius $F(d)$. Dies stellen wir in Abbildung 129 rechts dar. Die Schlütersche Konfiguration erhalten wir, indem wir einen Fixpunkt d^* von $F(\cdot)$ in $[0.418, 0.422]$ bestimmen. Um Approximationen an d^* zu erhalten, haben wir dass *Sekantenverfahren* auf

$$f(d) := F(d) - d = 0$$

angewandt, d. h. d_{k+1} wird aus d_k und d_{k-1} mit Hilfe der Vorschrift

$$d_{k+1} := d_k - \frac{(d_k - d_{k-1})f(d_k)}{f(d_k) - f(d_{k-1})}, \qquad k = 1,2,\ldots,$$

berechnet. In Tabelle 12 geben wir die mit $d_0 := 0.418$, $d_1 := 0.42$ erhaltenen Ergebnisse an. Der erhaltene Wert $d^* \approx 0.4212795440$ stimmt außerordentlich gut mit dem in der Literatur angegebenen Wert, siehe z. B. K. Schlüter (1979), überein. Bei `http://www.packomania.com` ist der Näherungswert

$$d^* \approx 0.42127954398390343276882176065 1$$

k	d_k	$\epsilon^*(d_k)$	$F(d_k)$	$F(d_k) - d_k$
0	0.418000000000000	0.034692493343576	0.423621022321586	0.005621022321586
1	0.420000000000000	0.020325208261048	0.422138406842236	0.002138406842236
2	0.421228046481110	0.011829131348824	0.421312881646208	0.000084835165098
3	0.421278778348486	0.011482968035470	0.421280038900702	0.000001260552215
4	0.421279543535014	0.011477749689059	0.421279544274065	0.000000000739051
5	0.421279543983899	0.011477746627815	0.421279543983906	0.000000000000006

Tabelle 12: Anwendung des Sekantenverfahrens auf $F(d) - d = 0$

angegeben. Außerdem ist $d^* > d_V$, die Schlütersche Konfiguration also besser als die von Valette. Wir geben sie auch in Abbildung 127 rechts an und man erkennt, dass sie sich nur geringfügig (den Unterschied erkennt man mit bloßem Auge kaum) von der von Valette unterscheidet. Statt des Punktes $(0, 0)$ hat man nur einen Punkt $(0, \epsilon^*)$ mit $\epsilon^* \approx 0.0115$, sonst ist die Konstruktion der Konfiguration im wesentlichen dieselbe wie bei Valette.

Bemerkung: Mit den oben eingeführten Bezeichnungen gilt:

- Für jedes $d \in [0.418, 0.422]$ besitzt die Gleichung $G(\epsilon, d) = d$ in $[0, 0.05]$ genau eine Lösung $\epsilon^*(d)$.
- Die Gleichung $F(d) = d$ besitzt genau eine Lösung d^* in $[0.418, 0.422]$.

Denn: Statt einer exakten Begründung verweisen wir lediglich auf Abbildung 130. Links haben wir $g_d(\cdot) := G(\cdot, d) - d$ für $d = 0.418$, $d = 0.42$ und $d = 0.422$ auf $[0, 0.05]$ sowie die entsprechenden Nullstellen von $g_d(\cdot)$ angegeben, rechts ist $f(d) := F(d) - d$ auf $[0.418, 0.422]$ skizziert sowie die Nullstelle von $f(\cdot)$ bzw. der Fixpunkt von $F(\cdot)$ angegeben. □

Bemerkung: Die Optimalität der Schlüterschen Konfiguration ist von C. de Groot, R. Peikert, D. Würtz (1990) mittels eines Computerbeweises nachgewiesen worden, siehe auch R. Peikert (1994). Hierauf wollen wir nicht eingehen. □

Literatur

[1] Aigner, M. und G. M. Ziegler (2002) *Das BUCH der Beweise*. 2. Auflage. Springer-Verlag, Berlin-Heidelberg-New York.

[2] Aldous, J. M. and R. J. Wilson (2000) *Graphs and Applications. An Introductory Approach*. Springer-Verlag, London-Berlin-Heidelberg.

[3] Almkvist, G. and B. Berndt (1988) Gauss, Landen, Ramanujan, the arithmetic-geometric mean, ellipses, pi and the Ladies Diary. Amer. Math. Monthly 95, pp. 585-608.

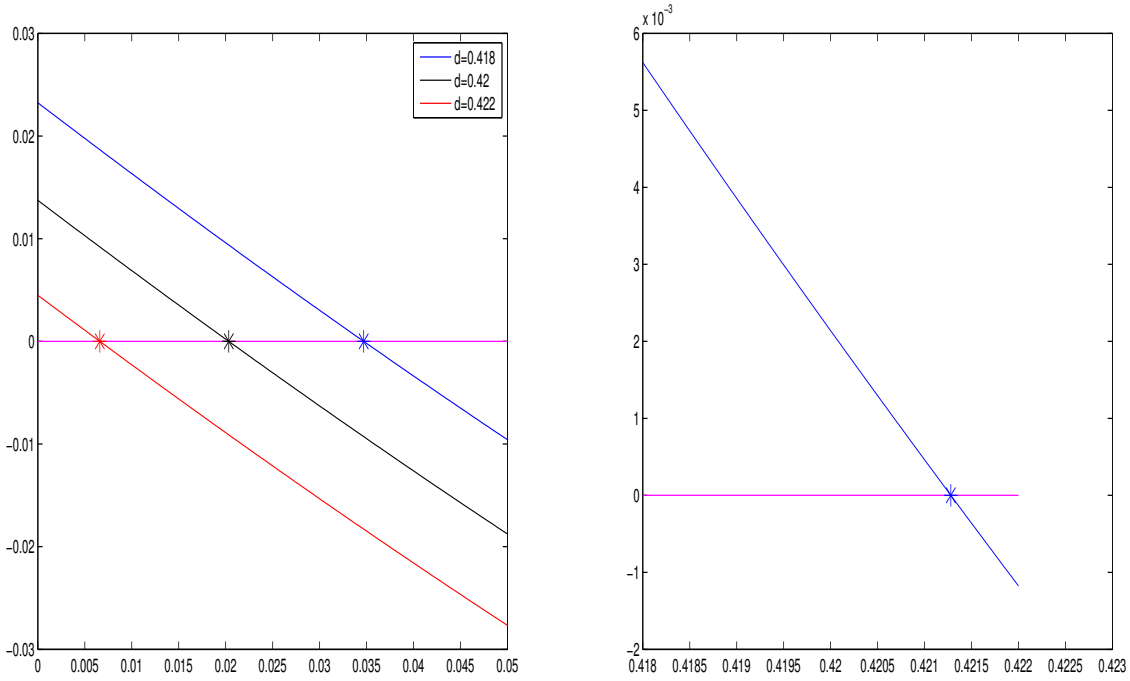


Abbildung 130: Veranschaulichung obiger Aussagen

[4] Arndt, J. und C. Haenel (2000) *Pi: Algorithmen, Computer, Arithmetik*. 2., überarbeitete und erweiterte Auflage. Springer-Verlag, Berlin-Heidelberg-New York.

[5] Artin, E. (1931) *Einführung in die Theorie der Gammafunktion*. Hamburger Mathematische Einzelschriften. B. G. Teubner, Leipzig.

[6] Babu, P., K. Pelckmans, P. Stoica and J. Li (2010) Linear Systems, Sparse Solutions, and Sudoku. IEEE Signal Processing Letters, Vol. 17, 40–42.

[7] Bammel, S. F. and J. Rothstein (1975) The number of 9×9 Latin Squares. Discrete Mathematics 11, 93–95.

[8] Barankin, E. and R. Dorfman (1958) On quadratic programming. University of California Publications in Statistics 2, 258–318.

[9] Bartlett, A. C., T. P. Chartier, A. N. Langville and T. P. Rankin (2008) An Integer Programming Model for the Sudoku Problem. Siehe `http://langvillea.people.cofc.edu/sudoku5.pdf` oder auch `http://www.maa.org/joma/volume8/bartlett/solvingpuzzle.html`. Hier findet man auch einen Hinweis auf ein MATLAB-script-file `sudoku.m`, in welchem dieser Ansatz umgesetzt wurde, siehe `http://www.math.washington.edu/~greenbau/Math_498/sudoku.m`.

[10] Basieux, P. (2007) *Die Top Ten der schönsten mathematischen Sätze*. Rowohlt Taschenbuch Verlag, Reinbek.

[11] Berggren, L., J. Borwein and P. Borwein (1997) *Pi: A Source Book*. Springer-Verlag, New York-Berlin-Heidelberg.

[12] Beutelsbacher, A. und B. Petri (2000) *Der goldene Schnitt.* 2., überarbeitete und erweiterte Auflage. Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford.

[13] Brent, R.P. (1975) Multiple-precision zero finding methods and the complexity of elementary function evaluation, in: *Analytic Computational Complexity* (edited by J. F. Traub), Academic Press, New York.

[14] Bryan, K. and T. Leise (2006) The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. SIAM Review 48, 569–581.

[15] Bourbaki, N. (1961) *Éléments de Mathématique. Livre IV. Fonctions d'une Variable Réelle.* Hermann, Paris.

[16] Cantor, M. (1880) *Vorlesungen über Geschichte der Mathematik., 1. Band*. B. G. Teubner, Leipzig.

[17] Collatz, L. und W. Wetterling (1971) *Optimierungsaufgaben*. Springer-Verlag, Berlin-Heidelberg-New York.

[18] A. Connes (1998) A new proof of Morley's theorem. Inst. Hautes Études Sci. In: Les relations entre les mathématiques et la physique théoretique: Festschrift for the 40th anniversary of the IHÉS, 43–46.

[19] Cooley, J. W. and J. W. Tukey (1965) An algorithm for the machine calculation of complex Fourier series. Math. Comp. 19, 297–301.

[20] Courant, R. und H. Robbins (1967) *Was ist Mathematik?* Zweite Auflage. Springer-Verlag, Berlin-Heidelberg-New York.

[21] Coxeter, H. S. M. (1948) A problem of collinear points. Amer. Math. Monthly 55, 26–28.

[22] Coxeter, H. S. M. (1969) *Introduction to Geometry*. Second Edition. John Wiley & Sons, New York.

[23] Dantzig, G. B. (1966) *Lineare Programmierung und Erweiterungen*. Springer-Verlag, Berlin-Heidelberg-New York.

[24] Debreu, G. and I. N. Herstein (1953) Nonnegative Square Matrices. Econometrica 21, 597–607.

[25] Deuflhard, P. und A. Hohmann (2008) *Numerische Mathematik 1. Eine algorithmisch orientierte Einführung*. 4. Auflage. De Gruyter, Berlin-New York.

[26] Erdős, P., A. Rényi and V. Sós (1966) On a problem of graph theory. Studia Sci. Math. 1, 215–235.

[27] Felgenhauer, B. and F. Jarvis (2006) Mathematics of Sudoku I. http://afjarvis.staff.shef.ac.uk/sudoku.

[28] Fra Luca Pacioli (1889) *De Divina Proportione. Die Lehre vom goldenen Schnitt.* Nach der venezianischen Ausgabe von 1509. Neu herausgegeben, übersetzt und erläutert von Constantin Winterberg. Quellenschriften für Kunstgeschichte und Kunsttechnik des Mittelalters und der Neuzeit. Verlag Carl Graeser, Wien.

[29] Franklin, J. (1980) *Methods of Mathematical Economics.* Springer, New York-Heidelberg-Berlin.

[30] Frobenius, G. (1912) Über Matrizen aus nicht negativen Elementen. Sitzungsber. Königl. Preuss. Akad. Wiss.: 456–477.

[31] Geiges, H. (2001) Beweis des Satzes von Morley nach A. Connes. Elem. Math. 56, 137–142.

[32] Goldberg, M (1970) On the densest packing of equal circles in a square. Math. Mag. 43, 24–30.

[33] Govan, A., C. D. Meyer and R. Albright (2008) Generalizing Google's Pagerank to rank National Football League Teams. North Carolina State University, SAS Institute, Inc., Paper 151-2008.

[34] de Groot, C., R. Peikert and D. Würtz (1990) The optimal packing of ten equal circles in a square. IPS Research Report No. 90-12, ETH Zürich.

[35] Halmos, P. R. and H. E. Vaughan (1950) The marriage problem. Amer. J. Math. 72, 214–215.

[36] Howard, R. (2004) The Milnor-Rogers proof of the Brouwer fixed point theorem. Mimeo, University of South Carolina, siehe http://www.math.sc.edu/~howard/.

[37] Huneke, C. (2002) The Friendship Theorem. The American Mathematical Monthly, Vol. 109, No. 2, 192-194.

[38] Isaacson, E. and H. B. Keller (1966) *Analysis of numerical methods*. John Wiley & Sons, Inc., New York-London-Sydney.

[39] Jacobs, K. und D. Jungnickel (2004) *Einführung in die Kombinatorik, 2. Auflage.* de Gruyter, Berlin-New York.

[40] Jung, H. W. E. (1901) Über die kleinste Kugel, die eine räumliche Figur enthält. Journal Reine Angew. Math. 123, 241–257.

[41] Keener, J. P. (1993) The Perron-Frobenius Theorem and the Ranking of Football Teams. SIAM Review, Vol. 35, S. 80-93.

[42] Knuth, D. E. (1968) *The Art of Computer Programming. Volume 1: Fundamental Algorithms*. Addison-Wesley, Reading.

[43] Kuhn, H. W. and A. W. Tucker (1951) Nonlinear Programming. Proc. Second Berkeley Symp. on Math. Statist. and Prob. (Univ. of Calif. Press), 481–492.

[44] Langville, A. N. and C. D. Meyer (2006) *Google's PageRank and Beyond. The Science of Search Engine Rankings*. Princeton University Press, Princeton.

[45] Langville, A. N. and C. D. Meyer (2012) *Who's # 1: The Science of Rating and Ranking*. Princeton University Press, Princeton.

[46] van Lint, J. H. and R. M. Wilson (1992) *A Course in Combinatorics*. Cambridge University Press, Cambridge.

[47] Löbbing, M. und I. Wegener (1996) Knight moves–was macht der Springer allein auf dem Schachbrett? In: *Highlights aus der Informatik* (I. Wegener, Hrsg.). Springer-Verlag, Berlin-Heidelberg-New York.

[48] Maak, W. (1950, S. 234) *Fastperiodische Funktionen*. Springer-Verlag, Berlin-Göttingen-Heidelberg.

[49] Mangasarian, O. L. (1969) *Nonlinear Programming*. McGraw-Hill Book Company, New York.

[50] McGuire, G. et al. (2012) There is no 16-Clue Sudoku: Solving the Minimum Number of Clues Problem, siehe `http://arxiv.org/abs/1201.0749`.

[51] Melissen, J. B. M. (1994) Densest packing of six equal circles. Elem. Mat. 49, 27–31.

[52] Meyer, C. (2000) *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia.

[53] Miel, G. (1983) Of Calculations Past and Present: The Archimedean Algorithm. The American Mathematical Monthly 90, 17–35.

[54] Milnor, J. (1978) Analytic proofs of the hairy ball theorem and the Brouwer fixed-point theorem. Amer. Math. Monthly 85, 521–524.

[55] Mollard, M. and C. Payan (1990) Some progress in the packing of equal circles in a square. Discrete Mathematics 84, 303–307.

[56] Morris, S. (2007) *Rätsel für Denker und Tüftler*. Zweite Auflage. DuMont, Köln.

[57] Moser, L. (1960) Problem 24 (corrected). Canad. Math. Bull. 3, 78.

[58] Motzkin, T. The lines and planes connecting the points of a finite set. Transactions of the Amer. Math. Society 70, 451–464.

[59] Naas, J. und W. Tutschke (2009) *Große Sätze und schöne Beweise in der Mathematik*. 3. Auflage. Verlag Harri Deutsch, Frankfurt am Main.

[60] von Neumann, J. (1928) Zur Theorie der Gesellschaftsspiele. Math. Annalen 100, 295–320.

[61] Newman, D. J. (1996) The Morley Miracle. Mathematical Intelligencer Volume 18, Number 1, 31–32.

[62] Niven, I. (1947) A simple proof that π is irrational. Bulletin Amer. Math. Soc. 53, 509.

[63] von Randow, G. (2010) *Das Ziegenproblem. Denken in Wahrscheinlichkeiten.* Rowohlt Taschenbuch Verlag, Reinbek.

[64] Peikert, R. (1994) Dichteste Packungen von gleichen Kreisen in einem Quadrat. Elem. Math. 49, 16–26.

[65] Peikert, R., D. Würtz, M. Monagan and C. de Groot (1992) Packing Circles in a Square: A Review and New Results. In: *System Modelling and Optimization 1991*, P.Kall, ed., Springer Lecture Notes Control Inf. Sci. 180, 45–54.

[66] Perron, O. (1907) Zur Theorie der Matrizen. Math. Ann. 64, 248–263.

[67] Peterson, E. L. and J. G. Ecker (1969) Geometric programming: Duality in quadratic programming and l_p-approximation I. In: *Proceedings of the Princeton Symposion on Mathematical Programming* (H. W. Kuhn, Ed.), 445–480. Princeton University Press, Princeton.

[68] Peterson, E. L. and J. G. Ecker (1969) Geometric programming: Duality in quadratic programming and l_p-approximation II (canonical programs). SIAM J. Appl. Math. 17, 317–340.

[69] Peterson, E. L. and J. G. Ecker (1970) Geometric programming: Duality in quadratic programming and l_p-approximation III (degenerate programs).' J. Math. Anal. Appl. 29, 365–383.

[70] Renteln, P. and A. Dundes (2005) Foolproof: A Sampling of Mathematical Folk Humor. Notices of the AMS Vol. 52 Issue 1, 24–34. Siehe auch `http://www.ams.org/notices/200501/index.html`.

[71] Rogers, C. A. (1980) A less strange version of Milnor's proof of Brouwer's fixed-point theorem. Amer. Math. Monthly 87, 525–527. Siehe `http://www.jstor.org/pss/2321416`

[72] Rudio, F. (1892) *Archimedes, Huygens, Lambert, Legendre. Vier Abhandlungen über die Kreismessung.* Teubner, Leipzig.

[73] Russell, E. and F. Jarvis (2006) Mathematics of Sudoku II. `http://www.afjarvis.staff.shef.ac.uk/sudoku/russell_jarvis_spec2.pdf`.

[74] Salamin, E. (1976) Computation of π using the arithmetic-geometric mean. Math. Comp. 30, 565–570.

[75] Schaer, J. (1965) The densest packing of 9 circles in a square. Canad. Math. Bull. 8, 273–277.

[76] Schaer, J. (1971) On the packing of ten equal circles in a square. Math. Mag. 44, 139–140.

[77] Schaer, J. and A. Meir (1965) On a geometric extremum problem. Canad. Math. Bull. 8, 21–27.

[78] Schlüter, K. (1979) Kreispackung in Quadraten. Elem. Math. 34, 12–14.

[79] Sierpinski, W. (1988) *Elementary theory of numbers*. (North Holland Mathematical Library Vol. 31). Elsevier, Amsterdam.

[80] Sperner, E. (1928) Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes. Abh. Math. Sem. Hamburg 6, 265–272.

[81] Steinhagen, P. (1921) Über die größte Kugel in einer konvexen Punktmenge. Abh. Math. Sem. Univ. Hamburg 1, 15–26.

[82] Sylvester, J. J. (1857) A Question in the Geometry of Situation. Quarterly J. Pure and Appl. Math. 1, 79.

[83] Taalman, L. (2007) Taking Sudoku Seriously. Math Horizons September 2007, 5–9.

[84] Tikhomirov, V. M. (1990) *Stories about Maxima and Minima*. American Mathematical Society.

[85] Valette, G. (1989) A Better Packing of Ten Equal Circles in a Square. Discrete Mathematics 76, 57–59.

[86] Van Loan, C. (1992) *Computational Frameworks for the Fast Fourier Transform*. SIAM, Philadelphia.

[87] Volkmann, L. (1991) *Graphen und Digraphen. Eine Einführung in die Graphentheorie*. Springer-Verlag, Wien-New York.

[88] Wagon, S. (1985) The Collatz Problem. Mathematical Intelligencer 7, 72–76.

[89] Walter, W. (1985) *Analysis I*. Springer-Verlag, Berlin-Heidelberg-New York-Tokyo.

[90] Walter, W. (1990) *Analysis II*. Springer-Verlag, Berlin-Heidelberg- New York-Tokyo.

[91] Wanner, G. (2004) (2004) Elementare Beweise des Satzes von Morley. Elem. Math. 59, 144–150.

[92] Werner, J. (1992) *Numerische Mathematik 1*. Vieweg-Verlag, Braunschweig-Wiesbaden.

[93] Wille, F. (1982) *Humor in der Mathematik*. Vandenhoek & Ruprecht, Göttingen.

[94] Wirsching, G. J. (2001) Das Collatz Problem. In: *Lexikon der Mathematik*. SPEKTRUM Akademischer Verlag, Heidelberg.