Noch mehr merkwürdige Mathematik

Jochen Werner
jochen.christa@t-online.de

Inhaltsverzeichnis

1 Einleitung

Nach *Merkwürdige Mathematik* und *Mehr merkwürdige Mathematik* folgt jetzt *Noch mehr merkwürdige Mathematik*. Angeregt zu diesen Titeln wurde ich durch die Anthologien *Morde*, *Mehr Morde* und *Noch mehr Morde* angelsächsischer Kriminalgeschichten, die Ende der 1950er Jahre von Mary Hottinger herausgegeben wurden. Wieder sollen hier einige (zumindestens für mich) interessante mathematische Probleme und Lösungen dargestellt werden. Das geschieht völlig ungeordnet.

2 John-Löwner Ellipsoide

2.1 Problemstellung, Existenz einer Lösung

In diesem Abschnitt untersuchen wir die Aufgabe, zu einem gegebenen konvexen Körper $K \subset \mathbb{R}^n$ ein umschriebenes bzw. einbeschriebenes Ellipsoid minimalen bzw. maximalen Volumens zu bestimmen. Hierbei heißt eine Menge $K \subset \mathbb{R}^n$ ein *konvexer Körper*, wenn K konvex und kompakt ist sowie das Innere int (K) von K nichtleer ist. Unter einem *Ellipsoid* verstehen wir das Bild der abgeschlossenen (euklidischen) Einheitskugel

$$B[0;1] := \{y \in \mathbb{R}^n : \|y\|_2 \leq 1\}$$

unter einer *nichtsingulären, affin linearen* Abbildung $\Phi(y) := Ay + b$, d. h. die Matrix $A \in \mathbb{R}^{n\times n}$ ist nichtsingulär und $b \in \mathbb{R}^n$. Mit einer nichtsingulären Matrix $A \in \mathbb{R}^{n\times n}$ und

$b \in \mathbb{R}^n$ (dem *Zentrum des Ellipsoids*) ist ein Ellipsoid E (wenn wir die Abhängigkeit von (A, b) betonen wollen, schreiben wir auch $E(A, b)$) also gegeben durch

$$E = b + A(B[0;1]) = \{x \in \mathbb{R}^n : \|A^{-1}(x-b)\|_2 \leq 1\}.$$

Das Volumen von E ist [1]

$$\text{vol}\,(E) = \det(A) \cdot \text{vol}\,(B[0;1]).$$

In der Darstellung eines Ellipsoids $E = b + A(B[0;1])$ als Bild der Einheitskugel unter einer nichtsingulären affin linearen Abbildung $\Phi(y) = Ay + b$ können wir annehmen, dass A symmetrisch und positiv definit ist. Denn bekanntlich besitzt die nichtsinguläre Matrix A eine Singulärwertzerlegung $A = V_1 \Sigma V_2$, wobei $V_1, V_2 \in \mathbb{R}^{n\times n}$ orthogonal sind und $\Sigma \in \mathbb{R}^{n\times n}$ eine Diagonalmatrix mit positiven Einträgen in der Diagonalen ist. Dann ist

$$A = \underbrace{V_1 \Sigma V_1^T}_{=:P} \cdot \underbrace{V_1 V_2}_{=:U} = PU,$$

also A das Produkt einer symmetrischen und positiv definiten Matrix P und einer orthogonalen Matrix U (die sogenannte *Polarzerlegung*). Folglich ist

$$E = b + A(B[0;1]) = b + PU(B[0;1]) = b + P(B[0;1])$$

und daher können wir bei der Darstellung eines Ellipsoids $E(A, b)$ annehmen, dass A symmetrisch und positiv definit ist. Die Aufgabe, zu K ein umschriebenes Ellipsoid minimalen Volumens bzw. ein einbeschriebenes Ellipsoid maximalen Volumens zu bestimmen führt also auf die Aufgabe

(U) Minimiere $\det(A)$ auf $M := \{(A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ K \subset E(A, b)\}$

bzw.

(E) Maximiere $\det(A)$ auf $N := \{(A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ E(A, b) \subset K\}$.

Hierbei ist $\mathcal{S}_+^{n\times n}$ die offene, konvexe Teilmenge der positiv definiten Matrizen im linearen normierten Raum $\mathcal{S}^{n\times n}$ der symmetrischen $n \times n$-Matrizen. Lösungen von (U) bzw. (P) heißen *John-Löwner-Ellipsoide*[2]. Der Sprachgebrauch ist hierbei nicht einheitlich. So erfährt man bei Wikipedia:

- Ein John-Ellipsoid ist in der Mathematik das eindeutig bestimmte Ellipsoid, das in einem konvexen Körper enthalten ist und mit dieser Eigenschaft maximales (voll-dimensionales) Volumen besitzt. Das John-Ellipsoid ist nach dem deutschen Mathematiker Fritz John benannt.

[1] Dass das Volumen der Einheitskugel durch

$$\text{vol}\,(B[0;1]) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$$

gegeben ist, ist für das weitere unerheblich.

[2] Interessante Informationen über Karel Löwner und Fritz John findet man bei M. Henk (2012).

In der Arbeit von F. John (1948) wird eine Multiplikatorenregel für eine Optimierungsaufgabe (mit einer endlichen Zahl von Variablen und möglicherweise unendlich vielen Ungleichungen als Nebenbedingungen) bewiesen, nämlich der Satz von F. John. Diese Multiplikatorenregel wird auf zwei geometrische Probleme angewandt. Dies geschieht in den Abschnitten 3. *Application to minimum sphere containing a set* und 4. *Application to the ellipsoid of least volume containing a set in S in E_m.* Es wird also *nicht* das Problem behandelt, zu einem konvexen Körper das einbeschriebene Ellipsoid mit maximalem Volumen zu bestimmen.

Als Literatur zu Löwner-John-Ellipsoiden sei auf die Arbeiten von K. Ball (1992, 1997) sowie O. Güler, F. Gürtuna (2007), M. Henk (2012) verwiesen.

Wir wollen uns hier schon davon überzeugen, dass (U) bzw. (E) jeweils eine Lösung besitzen.

Satz 2.1 *Bei gegebenem konvexen Körper $K \subset \mathbb{R}^n$ existiert ein umschriebenes Ellipsoid $U(K)$ minimalen Volumens bzw. ein einbeschriebenes Ellipsoid $E(K)$ maximalen Volumens, d. h. die Optimierungsaufgaben (U) und (E) sind lösbar.*

Beweis: Wir betrachten zunächst die Aufgabe (U) und folgen bei dem Beweis der Existenzaussage im wesentlichen der Argumentation von John. Wir wählen uns $(A_0, b_0) \in M$ (da z. B. eine hinreichend große Kugel den konvexen Körper K enthält, ist $M \neq \emptyset$), bilden die Niveaumenge

$$L_0(U) := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : (A,b) \in M,\ \det(A) \leq \det(A_0)\}$$

und zeigen, dass $L_0(U)$ kompakt ist. Da die Zielfunktion von (U) stetig ist, folgt die Existenz einer Lösung von (U).

Da K ein nichtleeres Inneres besitzt, existiert $b_1 \in \text{int}\,(K)$. Folglich enthält K eine (euklidische) Kugel mit einem Radius $r > 0$ und dem Mittelpunkt b_1. Jedes K enthaltende Ellipsoid $E(A,b)$ enthält insbesondere die Kugel $B[b_1; r] \subset K$. Aber auch die geshiftete Kugel $B[b; r]$ ist in $E(A,b)$ enthalten. Um dies einzusehen definieren wir $b_2 := 2b - b_1$ (dann ist b Mittelpunkt von b_1 und b_2) und überlegen uns, dass $B[b_2; r] \subset E(A,b)$. Dies ist richtig, denn für ein beliebiges $z \in B[0;1]$ ist

$$\|A^{-1}(b_2 + rz - b)\|_2 = \|A^{-1}(b + rz - b_1)\|_2 = \|A^{-1}(\underbrace{b_1 - rz}_{\in B[b_1;r]} - b)\|_2 \leq 1,$$

da $B[b_1; r] \subset K \subset E(A,b)$. Für ein beliebiges $x \in B[b; r]$ definieren wir

$$x_1 := b_1 + x - b \in B[b_1; r] \subset E(A,b), \qquad x_2 := b_2 + x - b \in B[b_1; r] \subset E(A,b).$$

Da $E(A,b)$ konvex ist, ist $x = \frac{1}{2}(x_1 + x_2) \in E(A,b)$ und damit $B[b; r] \subset E(A,b)$ für alle $(A,b) \in M$. Wegen $B[b; r] \subset E(A,b)$ ist $\|A^{-1}(b + rz - b)\|_2 = r\,\|A^{-1}z\|_2 \leq 1$ für alle $z \in B[0;1]$. Daher ist $\|A^{-1}\|_2 \leq 1/r$ für alle $(A,b) \in M$. Bezeichnet man mit $\lambda_{\min}(A)$ den kleinsten und mit $\lambda_{\max}(A)$ den größten Eigenwert von $A \in \mathcal{S}_+^{n\times n}$, so ist also $\lambda_{\min}(A) \geq r$ für alle $(A,b) \in M$. Jetzt sind wir bereit, die Kompaktheit, also die Beschränktheit und die Abgeschlossenheit, von $L_0(U)$ zu zeigen. Sei $(A,b) \in L_0(U)$. Dann ist

$$r^{n-1}\lambda_{\max}(A) \leq \det(A) \leq \det(A_0)$$

und daher

$$\|A\|_2 = \lambda_{\max}(A) \leq (1/r)^{n-1} \det(A_0).$$

Mit ein einem beliebigen $x_0 \in K$ ist $\|A^{-1}(x_0 - b)\|_2 \leq 1$ und daher

$$\|x_0 - b\|_2 \leq \|A\|_2 \underbrace{\|A^{-1}(x_0 - b)\|_2}_{\leq 1} \leq \|A\|_2 \leq (1/r)^{n-1} \det(A_0).$$

Damit haben wir die Beschränktheit der Niveaumenge $L_0(U)$ zu bewiesen. Zum Beweis der Abgeschlossenheit von $L_0(U)$ nehmen wir an, $\{(A_k, b_k)\} \subset L_0(U)$ sei eine Folge mit $(A_k, b_k) \to (A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n$. Wegen $\lambda_{\min}(A_k) \geq r$ für alle k ist auch $\lambda_{\min}(A) \geq r$, insbesondere also $A \in \mathcal{S}_+^{n\times n}$. Weiter folgt aus $K \subset E(A_k, b_k) = b_k + A_k(B[0;1])$ für alle $k \in \mathbb{N}$ offenbar, dass auch $K \subset E(A, b)$. Damit ist auch die Abgeschlossenheit von $L_0(U)$ und damit die Existenz einer Lösung von (U) bewiesen.

Jetzt folgen die entsprechenden Untersuchungen für die Aufgabe (E), also zu K ein einbeschriebenes Ellipsoid maximalen Volumens zu bestimmen. Entsprechend dem ersten Teil des Beweises ü wir uns, dass mit vorgegebenem $(A_0, b_0) \in N$ (die Existenz ist gesichert, da int $(K) \neq \emptyset$) die Niveaumenge

$$L_0(E) := \{(A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : (A, b) \in N,\ \det(A) \geq \det(A_0)\}$$

kompakt ist. Für $(A, b) \in N$ ist $b + A(B[0;1]) \subset K$ und hieraus folgt wegen der Beschränktheit von K die Beschränktheit von N und damit die von $L_0(E)$. Ist $\{(A_k, b_k)\} \subset L_0(E)$ eine Folge mit $(A_k, b_k) \to (A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n$, so ist auch $\det(A) \geq \det(A_0) > 0$ und damit $A \in \mathcal{S}_+^{n\times n}$. Also ist auch $(A, b) \in L_0(E)$, damit $L_0(E)$ abgeschlossen. Die Existenz einer Lösung von (E) ist damit bewiesen. □

2.2 Die Funktion $f(A) := -\log\det(A)$

Zum Beweis der Eindeutigkeit eines Löwner-John-Ellipsoids benutzen wir eine auch für sich interessante Hilfsaussage (siehe z. B. J. Werner (2002, S. 26)).

Satz 2.2 *Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n}$ die (konvexe) Teilmenge der positiv definiten Matrizen. Die Funktion $f\colon \mathcal{S}_+^{n\times n} \longrightarrow \mathbb{R}$ sei definiert durch*

$$f(A) := -\log\det(A).$$

Dann gilt:

1. *Für jedes $A \in \mathcal{S}_+^{n\times n}$ existiert die Abbildung $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$, definiert für $P \in \mathcal{S}^{n\times n}$ durch*

   $$f'(A; P) := \lim_{t\to 0} \frac{f(A + tp) - f(A)}{t},$$

   und ist gegeben durch

   $$f'(A; P) = -\mathrm{tr}(A^{-1}P),$$

   wobei $\mathrm{tr}(\cdot)$ einer Matrix ihre Spur zuordnet. Insbesondere ist $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ eine lineare Abbildung.

2. *Für $A, B \in \mathcal{S}_+^{n\times n}$ ist $f'(A; B-A) \leq f(B) - f(A)$. Hier gilt Gleichheit genau dann, wenn $A = B$.*

3. *Die Abbildung $f\colon \mathcal{S}_+^{n\times n} \longrightarrow \mathbb{R}$ ist (auf $\mathcal{S}_+^{n\times n}$) konvex. Für $A, B \in \mathcal{S}_+^{n\times n}$ und $\lambda \in [0,1]$ ist also*

$$\log\det((1-\lambda)A + \lambda B) \geq (1-\lambda)\log\det(A) + \lambda\log\det(B).$$

Gilt hier für ein $\lambda \in (0,1)$ Gleichheit, so ist $A = B$.

Beweis: Für $A \in \mathcal{S}_+^{n\times n}$ und $P \in \mathcal{S}^{n\times n}$ ist $A + tP \in \mathcal{S}_+^{n\times n}$ für alle hinreichend kleinen $|t| > 0$. Für diese t ist

$$\begin{aligned}
\frac{f(A+tP) - f(A)}{t} &= -\frac{\log\det(A+tP) - \log\det(A)}{t} \\
&= -\frac{\log\det(I + tA^{-1/2}PA^{-1/2})}{t} \\
&= -\frac{1}{t}\log\prod_{i=1}^{n}\lambda_i(I + tA^{-1/2}PA^{-1/2}) \\
&= -\frac{1}{t}\sum_{i=1}^{n}\log(1 + t\lambda_i(A^{-1/2}PA^{-1/2})).
\end{aligned}$$

Folglich ist

$$f'(A;P) = -\sum_{i=1}^{n}\lambda_i(A^{-1/2}PA^{-1/2}) = -\mathrm{tr}(A^{-1/2}PA^{-1/2}) = -\mathrm{tr}(A^{-1}P),$$

die Richtungsableitung $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ existiert also für jedes $A \in \mathcal{S}_+^{n\times n}$ und ist linear. Hierbei haben wir mit $\lambda_i(C)$, $i = 1, \ldots, n$, die Eigenwerte einer Matrix $C \in \mathcal{S}^{n\times n}$ bezeichnet und ausgenutzt, dass die Spur einer Matrix gleich der Summe ihrer Eigenwerte ist und daher bei einer Ähnlichkeitsabbildung invariant bleibt.

Jetzt kommen wir zum Beweis des zweiten Teiles von Satz 2.2. Mit $A, B \in \mathcal{S}_+^{n\times n}$ ist

$$\begin{aligned}
f'(A)(B-A) &= -\mathrm{tr}(A^{-1/2}(B-A)A^{-1/2}) \\
&= -\mathrm{tr}(A^{-1/2}BA^{-1/2}) + n \\
&= -n\left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i(A^{-1/2}BA^{-1/2})\right) + n \\
&\leq -n\left(\prod_{i=1}^{n}\lambda_i(A^{-1/2}BA^{-1/2})\right)^{1/n} + n \\
&\qquad \text{(Ungleichung vom geometrisch-arithmetischen Mittel)} \\
&= -n\det(A^{-1/2}BA^{-1/2})^{1/n} + n \\
&= -\frac{n}{\det(A)^{1/n}}[\det(B)^{1/n} - \det(A)^{1/n}] \\
&\leq -n[\log\det(B)^{1/n} - \log\det(A)^{1/n}] \\
&\qquad \text{(Konkavität des Logarithmus auf } \mathbb{R}_{++}) \\
&= f(B) - f(A).
\end{aligned}$$

Gilt Gleichheit, so gilt insbesondere Gleichheit bei der Anwendung der Ungleichung vom geometrisch-arithmetischen Mittel. Dies ist genau dann der Fall, wenn ein $c > 0$ mit $\lambda_i(A^{-1/2}BA^{-1/2}) = c$, $i = 1, \ldots, n$, existiert. Dies wiederum ist genau dann der Fall, wenn $B = cA$. Es muss aber auch Gleichheit in der zweiten der auftretenden Ungleichungen gelten, bei der die Konkavität des Logarithmus auf $\mathbb{R}_{++}$ ausgenutzt wurde. Es ist also

$$\frac{1}{\det(A)^{1/n}}[\det(B)^{1/n} - \det(A)^{1/n}] = \log\det(B)^{1/n} - \log\det(A)^{1/n},$$

wegen $B = cA$ mit $c > 0$ ist also $c - 1 = \log c$. Hieraus folgt aber $c = 1$ bzw. $A = B$.
Zum Beweis des dritten Teiles von Satz 2.2 seien $A, B \in \mathcal{S}_+^{n\times n}$ und $\lambda \in [0, 1]$ vorgegeben. Dann ist $C := (1-\lambda)A+\lambda B \in \mathcal{S}_+^{n\times n}$ wegen der Konvexität von $\mathcal{S}_+^{n\times n}$. Wegen der gerade eben bewiesenen Hilfsaussage in 2. ist

$$f(A) - f(C) \geq f'(C; A - C), \qquad f(B) - f(C) \geq f'(C; B - C).$$

Multipliziert man die erste Ungleichung mit $1 - \lambda$, die zweite mit λ und addiert die entstehenden Ungleichungen anschließend, so erhält man wegen der *Linearität* von $f'(C; \cdot)$, dass

$$(1 - \lambda)f(A) + \lambda f(B) - f((1 - \lambda)A + \lambda B) \geq 0 \quad \text{für alle } A, B \in \mathcal{S}_+^{n\times n} \text{ und } \lambda \in [0, 1].$$

Gilt hier Gleichheit für ein $\lambda \in (0, 1)$, so folgt $A = B$ wegen des zweiten Beweisschrittes. Damit ist Satz 2.2 bewiesen. □

2.3 Einbeschriebenes Ellipsoid: Eindeutigkeit einer Lösung

Den folgenden schönen Eindeutigkeitsbeweis für ein einem konvexen Körper einbeschriebenes Ellipsoid maximalen Volumens habe ich im wesentlichen einer Arbeit von R. Howard (1997) entnommen. Im dritten Teil von Satz 2.9 geben wir einen alternativen Eindeutigkeitsbeweis an, welcher notwendige (und hinreichende) Optimalitätsbedingungen benutzt. Der erste Eindeutigkeitsbeweis für ein einbeschriebenes Ellipsoid maximalen Volumens findet sich wohl bei L. Danzer, D. Laugwitz, H. Lenz (1957, Satz 2). Dagegen ist der zweidimensionale Spezialfall schon von F. Behrend (1938) behandelt worden. Von ihm wird die Existenz genau einer Umellipse und genau einer Inellipse zu einem konvexen Bereich, d. h. einem konvexen Körper in der Ebene, bewiesen.

Satz 2.3 *Zu einem konvexen Körper $K \subset \mathbb{R}^n$ existiert genau ein einbeschriebenes Ellipsoid maximalen Volumens, d. h. die Optimierungsaufgabe*

$$\text{(E)} \qquad \begin{cases} \textit{Maximiere} \quad \det(A) \quad \textit{auf} \\ N := \{(A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ b + A(B[0; 1]) \subset K\} \end{cases}$$

ist eindeutig lösbar.

Beweis: Mit einem Kompaktheitsargument hatten wir uns beim Beweis von Satz 2.1 schon überlegt, dass die *Existenz* einer Lösung von (E) gesichert ist. Zum Beweis der *Eindeutigkeit* einer Lösung zeigen wir zunächst, dass die Menge N der zulässigen Lösungen von (E) konvex ist. Seien hierzu $(A_1, b_1), (A_2, b_2) \in N$ und $\lambda \in [0, 1]$. Dann ist

$$(1-\lambda)(A_1, b_1) + \lambda(A_2, b_2) = ((1-\lambda)A_1 + \lambda A_2, (1-\lambda)b_1 + \lambda b_2) \in \mathcal{S}_+^{n\times n} \times \mathbb{R}.$$

Weiter ist

$$\begin{aligned}(1-\lambda)b_1 + \lambda b_2 + ((1-\lambda)A_1 + \lambda A_2)(B[0;1]) &\subset (1-\lambda)[b_1 + A_1(B[0;1])] \\ &\quad + \lambda[b_2 + A_2(B[0;1])] \\ &\subset (1-\lambda)K + \lambda K \\ &= K,\end{aligned}$$

also $(1-\lambda)(A_1, b_1) + \lambda(A_2, b_2) \in N$. Damit ist die Konvexität von N bewiesen. Seien $(A_1, b_1), (A_2, b_2)$ zwei Lösungen von (E). Dann ist (da der Optimalwert zu (E) eindeutig bestimmt ist) $\det(A_1) = \det(A_2)$. Wir definieren $A_3 := \frac{1}{2}(A_1 + A_2)$ und $b_3 := \frac{1}{2}(b_1 + b_2)$. Wegen der Konvexität von N ist $(A_3, b_3) \in N$. Aus Satz 2.2 erhalten wir, dass

$$\begin{aligned}\log\det(A_3) &= \log\det(\tfrac{1}{2}(A_1 + A_2)) \\ &\geq \frac{1}{2}\log\det(A_1) + \frac{1}{2}\log\det(A_2) \\ &= \log\det(A_1)\end{aligned}$$

und damit $\det(A_3) \geq \det(A_1)$. Da (A_1, b_1) eine Lösung von (E) und der Optimalwert zu (E) eindeutig bestimmt ist, ist $\det(A_3) = \det(A_1)$. Also gilt

$$\det(\tfrac{1}{2}(A_1 + A_2)) = \det(A_1) = \det(A_2).$$

In der Ungleichung

$$\log\det(\tfrac{1}{2}(A_1 + A_2)) \geq \frac{1}{2}\log\det(A_1) + \frac{1}{2}\log\det(A_2)$$

gilt also Gleichheit. Aus Satz 2.2 folgt $A_1 = A_2$. Zu zeigen bleibt $b_1 = b_2$. Angenommen, es wäre $b_1 \neq b_2$. Dann wäre $E_1 := b_1 + A_1(B[0;1])$ eine (echte) Verschiebung von $E_2 := b_2 + A_1(B[0;1])$, d. h. es ist $E_1 = (b_1 - b_2) + E_2$. Sei $K_0 := \text{co}\,(E_1 \cup E_2) \subset K$ die konvexe Hülle der beiden Ellipsoide E_1 und E_2, also die kleinste konvexe Menge, die E_1 und E_2 enthält. Wir veranschaulichen uns die Situation in Abbildung 1. In dieser Abbildung sind die Ellipsen E_1 und E_2 durch Einheitskugeln um b_1 bzw. b_2 gegeben. Dies kann o. B. d. A. angenommen werden, denn nach der Transformation $x \mapsto y := A_1^{-1}x$ wird

$$E_1 = b_1 + A_1(B[0;1]) = \{x \in \mathbb{R}^n : \|A_1^{-1}(x - b_1)\|_2 \leq 1\}$$

zu einer Einheitskugel mit dem Mittelpunkt $A_1^{-1}b_1$. In Abbildung 1 haben wir die Kugeln $E_1 := B[b_1; 1]$, $E_2 := B[b_2; 1]$ sowie (gestrichelt) die Kugel $B[b_3; 1]$ mit $b_3 := \frac{1}{2}(b_1 + b_2)$ eingetragen. Angegeben ist ferner die konvexe Hülle $K_0 := \text{co}\,(E_1 \cup E_2)$ von

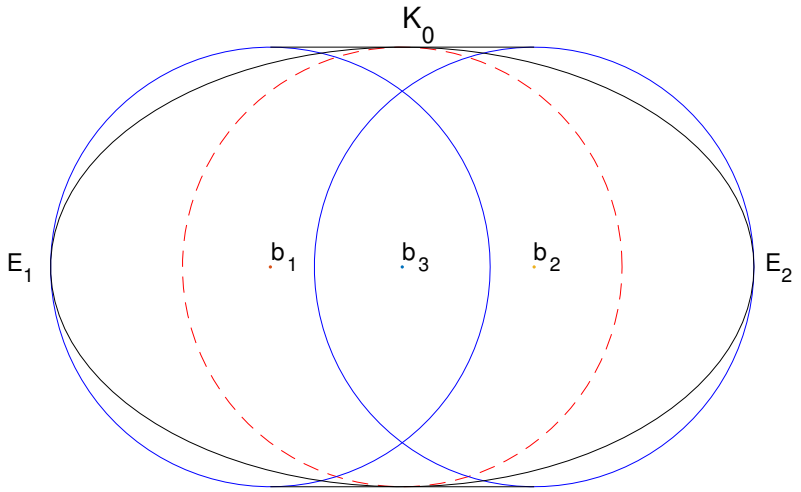


Abbildung 1: Veranschaulichung des Eindeutigkeitsbeweises

E_1 und E_2. Offenbar[3] ist $K_0 = \bigcup_{b\in[b_0,b_1]} B[b;1]$. Wir wollen uns überlegen, dass es in K_0 ein Ellipsoid E mit einem Volumen gibt, welches größer als das von E_1 ist. Dies ergibt dann den gewünschten Widerspruch. Für E machen wir den Ansatz

$$E := b_3 + (I + \epsilon(b_1 - b_2)(b_1 - b_2)^T)(B[0;1])$$

mit einem gewissen $\epsilon > 0$. Dann ist

$$\frac{\text{vol }(E)}{\text{vol }(E_1)} = \det(I + \epsilon(b_1 - b_2)(b_1 - b_2)^T) = 1 + \epsilon\|b_1 - b_2\|_2^2 > 1.$$

Wir haben $\epsilon > 0$ so zu bestimmen, dass $E \subset K_0$. Hierzu überlegen wir uns, dass es bei hinreichend kleinem $\epsilon > 0$ zu jedem $z \in E$ ein $b_z \in [b_1, b_2]$ mit $z \in B[b_z;1]$ gibt. Wegen $\bigcup_{b\in[b_1,b_2]} B[b;1] = K_0$ ist dann $z \in K_0$ bzw. $E \subset K_0$ bewiesen. Zu $z \in E$ gibt es ein $x \in B[0;1]$ mit

$$z = b_3 + (I + \epsilon(b_1 - b_2)(b_1 - b_2)^T)x.$$

Der Punkt x lässt sich eindeutig in der Form

$$x = \alpha(b_1 - b_2) + u \qquad \text{mit } \alpha \in \mathbb{R} \text{ und } u \in \text{span }\{b_1 - b_2\}^\perp$$

[3]Einerseits ist $\bigcup_{b\in[b_1,b_2]} B[b;1]$ konvex (sind nämlich $x \in B[b_x;1], y \in B[b_y;1]$ mit $b_x, b_y \in [b_1, b_2]$ und ist $\lambda \in [0,1]$, so ist $(1-\lambda)x + \lambda y \in B[(1-\lambda)b_x + \lambda b_y;1] \subset \bigcup_{b\in[b_1,b_2]} B[b;1]$) und enthält E_1 und E_2, sodass co $(E_1 \cup E_2) \subset \bigcup_{b\in[b_1,b_2]} B[b;1]$. Ist umgekehrt $x \in B[b;1]$ mit $b = (1-\mu)b_1 + \mu b_2 \in [b_1, b_2]$ (und $\mu \in [0,1]$), so ist $x = (1-\mu)x_1 + \mu x_2$ Konvexkombination zweier Elemente $x_1 := b_1 + x - b \in E_1$ und $x_2 := b_2 + x - b \in E_2$, also ein Element der konvexen Hülle co $(E_1 \cup E_2)$ von E_1 und E_2. Damit ist co $(E_1 \cup E_2) = \bigcup_{b\in[b_1,b_2]} B[b;1]$ bewiesen.

darstellen und es ist

$$\|x\|_2^2 = \alpha^2 \|b_1 - b_2\|_2^2 + \|u\|_2^2 \leq 1,$$

insbesondere ist $|\alpha| \, \|b_1 - b_2\|_2 \leq 1$. Der Punkt z kann dann dargestellt werden als

$$\begin{aligned} z &= b_3 + \epsilon\alpha \|b_1 - b_2\|_2^2 (b_1 - b_2) + x \\ &= \left(\frac{1}{2} + \epsilon\alpha \|b_1 - b_2\|_2^2\right) b_1 + \left(\frac{1}{2} - \epsilon\alpha \|b_1 - b_2\|_2^2\right) b_2 + x. \end{aligned}$$

Nun wähle man $\epsilon > 0$ so klein, dass $\epsilon \, \|b_1 - \|_2 \leq \frac{1}{2}$ und setze anschließend

$$b_z := \left(\frac{1}{2} + \epsilon\alpha \, \|b_1 - b_2\|_2^2\right) b_1 + \left(\frac{1}{2} - \epsilon\alpha \|b_1 - b_2\|_2^2\right) b_2.$$

Wegen

$$\frac{1}{2} + \epsilon\alpha \, \|b_1 - b_2\|_2^2 \leq \frac{1}{2} + \underbrace{\epsilon \, \|b_1 - b_2\|_2}_{\leq \frac{1}{2}} \underbrace{|\alpha| \, \|b_1 - b_2\|_2}_{\leq 1} \leq 1$$

und

$$\frac{1}{2} + \epsilon\alpha \, \|b_1 - b_2\|_2^2 \geq \frac{1}{2} - \underbrace{\epsilon \, \|b_1 - b_2\|_2}_{\leq \frac{1}{2}} \underbrace{|\alpha| \, \|b_1 - b_2\|_2}_{\leq 1} \geq 0$$

ist b_z eine Konvexkombination von b_1 und b_2. Damit ist

$$z \in B[b_z; \|x\|_2] \subset B[b_z; 1] \subset K_0$$

und der Satz ist bewiesen. □

Bemerkung: Im wesentlichen sind wir dem Beweis von R. Howard (1997) gefolgt. Dort wird aber nicht Satz 2.2 benutzt, sondern die folgende Aussage (Proposition 4.2):

- *Seien $A, B \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Dann ist*

$$\det(A + B)^{1/n} \geq \det(A)^{1/n} + \det(B)^{1/n}.$$

Ferner wird von Howard der letzte Teil des Beweises nur angedeutet. □

2.4 Der Satz von F. John

Unser Ziel ist es, den folgenden Satz (Theorem I bei F. John (1948)) zu beweisen. Die Arbeit von F. John findet man auch in dem Sammelband F. John (1985, S. 543–560). In einem Kommentar zu dieser Arbeit, welche man ebenfalls in dem Sammelband findet, schreibt H. W. Kuhn zu Beginn:

> *This paper is a pleasure to read and reread.*

Und am Schluss seines Kommentars schreibt H. W. Kuhn, einer der Pioniere bei der Entwicklung der Optimierung:

The widespread development, dissemination, and use of these results attests to their importance. At present they are the keystones of a broad structure of applied and computational mathematics. However, strictly as a piece of pure mathematics, the present paper stands as a pearl among the many jewels in John's work.

Satz 2.4 (F. John) *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : G(x,y) \leq 0 \text{ für alle } y \in Y\}$.

Hierbei seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *und* $G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *mit einem kompakten metrischen Raum* (Y,d) *stetig*[4], $f(\cdot)$ *und* $G(\cdot,y)$ *bei festem* $y \in Y$ *stetig partiell differenzierbar, ferner seien* $G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *und* $\nabla_x G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *stetig. Sei* $x^* \in M$ *eine lokale Lösung von (P) (die Glattheitsvoraussetzungen an* $f(\cdot)$ *und* $G(\cdot,y)$ *brauchen natürlich nur lokal erfüllt zu sein). Sei*

$$Y^* := \{y \in Y : G(x^*,y) = 0\}.$$

Mit einem $s \in \{0,\ldots,n\}$ *existieren dann* $y_1,\ldots,y_s \in Y^*$ *und sogenannte Multiplikatoren* $\lambda_0,\lambda_1,\ldots,\lambda_s$, *die nicht alle verschwinden*[5], *mit*

$$\lambda_0 \geq 0, \quad \lambda_1 > 0,\ldots,\lambda_s > 0$$

und

$$\lambda_0 \nabla f(x^*) + \sum_{i=1}^{s} \lambda_i \nabla_x G(x^*,y_i) = 0.$$

Beweis: Im folgenden können wir annehmen, dass $Y^* \neq \emptyset$. Denn andernfalls ist die Ungleichungsrestriktion in x^* nicht aktiv und man kann $s := 0$ und $\lambda_0 := 1$ setzen.
Das erste Zwischenergebnis ist entscheidend.

- Es gibt kein $p \in \mathbb{R}^n$ mit

  $$(*) \qquad \nabla f(x^*)^T p < 0, \qquad \max_{y \in Y^*} \nabla_x G(x^*,y)^T p < 0.$$

Denn: Angenommen, es existiert ein $p \in \mathbb{R}^n$, für welches $(*)$ gilt. Wir überlegen uns, dass dann positive Zahlen δ, ϵ existieren mit

$$(**) \quad \nabla f(x)^T p < -\delta, \qquad \max_{y \in Y_\epsilon^*} \nabla_x G(x,y) < -\delta \qquad \text{für alle } (x,y) \in B[x^*;\epsilon] \times Y_\epsilon^*,$$

wobei

$$Y_\epsilon^* := \{y \in Y : d(y,Y^*) \leq \epsilon\}$$

die Menge der Punkte aus Y ist, die von Y^* einen Abstand $\leq \epsilon$ besitzen. Wenn das nicht richtig wäre, so würde es Folgen $\{x_k\} \subset \mathbb{R}^n$, $\{y_k\} \subset Y$ und $\{\eta_k\} \subset Y^*$ geben mit

$$\lim_{k\to\infty} x_k = x^*, \qquad \lim_{k\to\infty} d(y_k,\eta_k) = 0$$

[4]Die Konvergenz in $\mathbb{R}^n \times Y$ ist in naheliegender Weise erklärt.
[5]Der Fall $\lambda_0 = 0$ und $s = 0$ wird hierdurch ausgeschlossen.

und

$$\liminf_{k\to\infty} \nabla f(x_k)^T p \geq 0$$

oder

$$\liminf_{k\to\infty} \nabla_x G(x_k, y_k)^T p \geq 0.$$

Mit Y ist auch Y^* kompakt. Durch Auswahl geeigneter Teilfolgen können wir daher annehmen, dass $\{y_k\}$ und $\{\eta_k\}$ einen (gemeinsamen) Grenzwert $y \in Y^*$ haben. Da $\nabla f(\cdot)$ und $\nabla_x G(\cdot,\cdot)$ stetig sind, folgt $\nabla f(x^*)^T p \geq 0$ oder $\nabla_x G(x^*, y)^T p \geq 0$. Dies ist aber ein Widerspruch zu der Annahme, dass p den Ungleichungen $(*)$ genügt. Also existieren positive Zahlen δ, ϵ, für die $(**)$ gilt. Dann existiert ein $\mu = \mu(\epsilon) > 0$ mit $G(x^*, y) < -\mu$ für alle $y \in Y \setminus Y_\epsilon^*$. Für alle hinreichend kleinen $t > 0$ ist

$$\begin{aligned} f(x^* + tp) &= f(x^*) + t\nabla f(x^* + \theta tp)^T p, \\ G(x^* + tp, y) &= G(x^*, y) + t\nabla_x G(x^* + \theta tp, y)^T p, \end{aligned}$$

wobei θ jeweils eine Zahl zwischen 0 und 1 ist. Wählt man nun $t > 0$ so klein, dass $t \|p\|_2 \leq \epsilon$ (bzw. $x^* + tp \in B[x^*; \epsilon]$) und

$$t \cdot \max_{(x,y)\in B[x^*;\epsilon]\times Y} |\nabla_x G(x, y)^T p| < \mu,$$

so ist

$$f(x^* + tp) = f(x^*) + t\underbrace{\nabla f(x^* + \theta tp)^T}_{<-\delta} < f(x^*) - t\delta < f(x^*)$$

und

$$G(x^* + tp, y) = \underbrace{G(x^*, y)}_{\leq 0} + t\underbrace{\nabla_x G(x^* + \theta tp, y)^T p}_{<-\delta} < 0 \quad \text{für alle } y \in Y_\epsilon^*$$

sowie

$$G(x^* + tp, y) \leq \underbrace{G(x^*, y)}_{<-\mu} + \underbrace{t \cdot \max_{(x,y)\in B[x^*;\epsilon]} |\nabla_x G(x, y)^T p|}_{<\mu} < 0 \quad \text{für alle } y \in Y \setminus Y_\epsilon^*.$$

Damit haben wir einen Widerspruch dazu erhalten, dass x^* eine lokale Lösung von (P) ist.

Nun kommt der zweite Beweisschritt:

- Es ist
$$0 \in \text{co}\,(\{\nabla f(x^*)\} \cup \{\nabla_x G(x^*, y) : y \in Y^*\}).$$

Denn: Zur Abkürzung setzen wir

$$\Sigma := \{\nabla f(x^*)\} \cup \{\nabla_x G(x^*, y) : y \in Y^*\}.$$

Da $Y^* \subset Y$ kompakt ist, ist auch $\Sigma \subset \mathbb{R}^n$ kompakt. Die konvexe Hülle co (Σ) der kompakten Teilmenge Σ des $\mathbb{R}^n$ ist ebenfalls kompakt (siehe z. B. J. Werner (1984,

S. 117)) und insbesondere abgeschlossen. Die Annahme, es sei $0 \notin \text{co}\,(\Sigma)$ liefert wegen des starken Trennungssatzes ein $p \in \mathbb{R}^n$ mit

$$\sup_{q \in \text{co}\,(\Sigma)} q^T p < 0$$

bzw. die Existenz einer Hyperebene im $\mathbb{R}^n$ mit der Eigenschaft, dass co (Σ) in einem davon erzeugten Halbraum und 0 im gegenüberliegenden offenen Halbraum liegt. Hieraus folgt

$$\nabla f(x^*)^T p < 0, \qquad \nabla_x G(x^*, y)^T p < 0 \quad \text{für alle } y \in Y^*.$$

Dies ist ein Widerspruch zum Ergebnis des ersten Beweisschrittes, dass nämlich $(*)$ keine Lösung besitzt.

Zum Schluss des Beweises wenden wir den Satz von Carathéodory an (siehe z. B. J. Werner (1984, S. 43)). Dieser sagt aus, dass sich jeder Punkt (in unserem Falle der Nullpunkt) aus der konvexen Hülle co (Σ) einer Menge $\Sigma \subset \mathbb{R}^n$ als Konvexkombination von höchstens $n+1$ Punkten aus Σ darstellen lässt, wobei einer der Punkte willkürlich gewählt werden kann (in unserem Fall ist dies $\nabla f(x^*)$)[6].

Hieraus erhalten wir offenbar sofort die Behauptung. □

Bemerkung: Die Aussage von Satz 2.4 bleibt offenbar unverändert, wenn in M eine zusätzliche Nebenbedingung $x \in O$ auftritt, wobei $O \subset \mathbb{R}^n$ eine *offene* Menge ist. □

---

[6]Diese geringfügige Verschärfung der üblichen Version des Satzes von Carathéodory formulieren und beweisen wir jetzt.

- Sei $\Sigma := \{x_0\} \cup X$ mit $x_0 \in \mathbb{R}^n$ und $X \subset \mathbb{R}^n$. Ist $x \in \text{co}\,(\Sigma)$, so existieren $\{x_1, \ldots, x_s\} \subset X$ mit $s \in \{0, \ldots, n\}$ sowie $\lambda_0 \geq 0$, $\lambda_1 > 0, \ldots, \lambda_s > 0$ mit $\sum_{i=0}^s \lambda_i = 1$ und $x = \sum_{i=0}^s \lambda_i x_i$.

Denn: Die konvexe Hülle einer Menge besteht aus endlichen Konvexkombinationen ihrer Elemente. Wir nehmen an, x habe eine Darstellung $x = \sum_{i=0}^m \lambda_i x_i$ mit $\{x_1, \ldots, x_m\} \subset X$, $\lambda_0 \geq 0$, $\lambda_1 > 0, \ldots, \lambda_m > 0$ sowie $\sum_{i=0}^m \lambda_i = 1$. Ist $m \leq n$, so sind wir fertig. Daher nehmen wir an, es sei $m > n$. Wir zeigen, dass x eine Konvexkombination von x_0 und $m-1$ der Punkte $\{x_1, \ldots, x_m\} \subset X$ ist. Da die $m > n$ Vektoren $\{x_1 - x_0, \ldots, x_m - x_0\} \subset \mathbb{R}^n$ linear abhängig sind, existieren $r_1, \ldots, r_m$, nicht alle gleich Null, mit $\sum_{i=1}^m r_i(x_i - x_0) = 0$. O. B. d. A. ist $\sum_{i=1}^m r_i \geq 0$ (notfalls ersetze man r_i durch $-r_i$, $i = 1, \ldots, m$). Dann ist $r_0 := -\sum_{i=1}^m r_i \leq 0$. Dann ist $\sum_{i=0}^m r_i = 0$ und daher $\sum_{i=0}^m r_i x_i = 0$. Für jedes $\alpha \in \mathbb{R}$ ist dann $x = \sum_{i=0}^m (\lambda_i - \alpha r_i) x_i$. Nun wähle man $\alpha > 0$ so, dass $\lambda_i - \alpha r_i \geq 0$, $i = 0, \ldots, m$, und $\lambda_j - \alpha r_j = 0$ für wenigstens ein $j \in \{1, \ldots, m\}$. Dies erreicht man, indem man

$$\alpha := \min_{i \in \{1,\ldots,m\}, r_i > 0} \left(\frac{\lambda_i}{r_i} \right) = \frac{\lambda_j}{r_j}$$

setzt. Man beachte hierbei, dass es ein $i \in \{1, \ldots, m\}$ mit $r_i > 0$ gibt, da $\sum_{i=1}^m r_i = 0$ und nicht alle r_i, $i = 1, \ldots, m$, verschwinden. Mit $\mu_i := \lambda_i - \alpha r_i$, $i = 0, \ldots, m$, ist $\mu_0 \geq 0$ (wegen $\lambda_0 \geq 0$ und $r_0 \leq 0$), $\mu_i \geq 0$, $i = 1, \ldots, m$ und $\mu_j = 0$ f, ferner $\sum_{\substack{i=0 \\ i \neq j}}^m \mu_i = 0$. Folglich ist

$$x = \mu_0 x_0 + \sum_{\substack{i=1 \\ i \neq j}}^m \mu_i x_i$$

eine Konvexkombination von x_0 und $m-1$ Punkten aus X. Nach endlich vielen Schritten erhält man das gewünschte Ergebnis.

Im folgenden Satz (Theorem II bei F. John (1948)) wird eine Zusatzbedingung angegeben, die sichert, dass die notwendigen Optimalitätsbedingungen von Satz 2.4 auch *hinreichend* für ein lokales Minimum sind.

Satz 2.5 *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : G(x,y) \le 0 \text{ für alle } y \in Y\}$.

Hierbei seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *und* $G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *mit einem kompakten metrischen Raum* (Y,d) *stetig,* $f(\cdot)$ *und* $G(\cdot,y)$ *bei festem* $y \in Y$ *stetig partiell differenzierbar, ferner seien* $G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *und* $\nabla_x G\colon \mathbb{R}^n \times Y \longrightarrow \mathbb{R}$ *stetig. Sei* $x^* \in M$ *und*

$$Y^* := \{y \in Y : G(x^*,y) = 0\}.$$

Mit einem $s \in \{0,\dots,n\}$ *mögen* $y_1,\dots,y_s \in Y^*$ *und* $\lambda_0,\lambda_1,\dots,\lambda_s$, *die nicht alle verschwinden, existieren mit*

$$\lambda_0 \ge 0, \quad \lambda_1 > 0,\dots,\lambda_s > 0$$

und

$$\lambda_0 \nabla f(x^*) + \sum_{i=1}^{s} \lambda_i \nabla_x G(x^*,y_i) = 0.$$

Zusätzlich habe die Matrix

$$A := \begin{pmatrix} \lambda_0 \nabla f(x^*) & \nabla_x G(x^*,y_1) & \cdots & \nabla_x G(x^*,y_s) \end{pmatrix} \in \mathbb{R}^{n\times(1+s)}$$

den Rang n. *Dann ist* x^* *eine lokale Lösung von (P).*

Beweis: Wir zeigen mehr als behauptet, nämlich dass $x^* \in M$ eine lokale Lösung von

(P_s) Minimiere $f(x)$ auf $M_s := \{x \in \mathbb{R}^n : G(x,y_i) \le 0,\ i = 1,\dots,s\}$

ist. Angenommen, dies wäre nicht der Fall. Dann existiert eine Folge $\{t_k\} \subset \mathbb{R}_{++}$ mit $t_k \to 0$ und eine Folge $\{p_k\} \subset \mathbb{R}^n$ mit $\|p_k\|_2 = 1$, $f(x^*+t_kp_k) < f(x^*)$ und $x^*+t_kp_k \in M_s$ bzw. $G(x^*+t_kp_k,y_i) \le 0$, $i = 1,\dots,s$, $k \in \mathbb{N}$. Mit geeignetem θ zwischen 0 und 1 (von k und i abhängig) ist dann

$$\nabla f(x^* + \theta t_k p_k)^T p_k \le 0$$

und

$$\nabla_x G(x^* + \theta t_k p_k, y_i)^T p_k \le 0, \quad i = 1,\dots,s.$$

Eine geeignete Teilfolge von $\{p_k\}$ konvergiert gegen ein $p \in \mathbb{R}^n$ mit $p \ne 0$ (sogar $\|p\|_2 = 1$), und es ist

$$\nabla f(x^*)^T p \le 0, \qquad \nabla_x G(x^*,y_i)^T p \le 0, \quad i = 1,\dots,s.$$

Dann ist

$$0 = \underbrace{\left(\lambda_0 \nabla f(x^*) + \sum_{i=1}^{s} \lambda_i \nabla_x G(x^*,y_i)\right)}_{=0}{}^T p = \underbrace{\lambda_0 \nabla f(x^*)^T p}_{\le 0} + \sum_{i=1}^{s} \underbrace{\lambda_i}_{>0} \underbrace{\nabla_x G(x^*,y_i)^T p}_{\le 0}$$

und folglich

$$\lambda_0 \nabla f(x^*)^T p = 0, \qquad \nabla_x G(x^*, y_i)^T p = 0, \quad i = 1, \ldots, s.$$

Mit

$$A := \begin{pmatrix} \lambda_0 \nabla f(x^*) & \nabla_x G(x^*, y_1) & \cdots & \nabla_x G(x^*, y_s) \end{pmatrix}$$

ist also $A^T p = 0$. Wegen $p \neq 0$ ist dies ein Widerspruch dazu, dass Rang $(A) = n$ bzw. die Zeilen von A (oder die Spalten von A^T) linear unabhängig sind. Der Satz ist bewiesen. □

Als Anwendung von Satz 2.4 wird von F. John (1948) im dritten Abschnitt seines Aufsatzes die Aufgabe betrachtet, zu einer kompakten Menge $S \subset \mathbb{R}^n$ eine Kugel $B[x; r]$ mit kleinstem positiven Radius r zu finden, welche S enthält. Die Existenz einer solchen Kugel ist klar, wenn die Voraussetzung gemacht wird, dass S mindestens zwei verschiedene Punkte enthält. Wir erhalten damit die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad x_{n+1} \quad \text{auf} \\ M := \{(x, x_{n+1}) \in \mathbb{R}^n \times \mathbb{R} : \frac{1}{2}\|x - y\|_2^2 - x_{n+1} \leq 0 \text{ für alle } y \in S\}. \end{cases}$$

Ist $(x^*, x^*_{n+1}) \in M$ eine Lösung von (P), so ist $B[x^*; \sqrt{2x^*_{n+1}}]$ die gesuchte Umkugel zu S mit minimalem Radius. Wir wenden Satz 2.4 an mit $f\colon \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ definiert durch

$$f(x, x_{n+1}) := x_{n+1},$$

mit $Y := S$ und $G\colon (\mathbb{R}^n \times \mathbb{R}) \times S \longrightarrow \mathbb{R}$ definiert durch

$$G((x, x_{n+1}), y) := \frac{1}{2}\|x - y\|_2^2 - x_{n+1}.$$

Sei $(x^*, x^*_{n+1}) \in M$ eine (lokale) Lösung von (P) und

$$S^* := \{y \in S : G((x^*, x^*_{n+1}), y) = 0\}.$$

Aus Satz 2.4 erhalten wir die Existenz von $s \in \{0, \ldots, n+1\}$ Punkten $y_1^*, \ldots, y_s^* \in S^*$ und Multiplikatoren $\lambda_0, \lambda_1, \ldots, \lambda_s$, die nicht alle verschwinden, mit

$$\lambda_0 \geq 0, \quad \lambda_1 > 0, \ \ldots, \ \lambda_s > 0$$

sowie

$$\lambda_0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sum_{i=1}^{s} \lambda_i \begin{pmatrix} x^* - y_i^* \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Daher ist

$$\lambda_0 = \sum_{i=1}^{s} \lambda_i > 0, \qquad \sum_{i=1}^{s} \lambda_i (x^* - y_i) = 0.$$

Für beliebige $(x, x_{n+1}) \in M$ ist

$$\begin{aligned}
& \sum_{i=1}^{s} \lambda_i \left(x_{n+1} - \frac{1}{2}\|x - y_i^*\|_2^2 \right) \\
= & \sum_{i=1}^{s} \lambda_i \left[x_{n+1} - \left(\frac{1}{2}\|x - x^*\|_2^2 + (x^* - y_i^*)^T (x - x^*) + \frac{1}{2}\|x^* - y_i^*\|_2^2 \right) \right] \\
= & \sum_{i=1}^{s} \lambda_i \left[x_{n+1} - \left(\frac{1}{2}\|x - x^*\|_2^2 + \frac{1}{2}\|x^* - y_i^*\|_2^2 \right) \right] \\
& \qquad (\text{wegen } \textstyle\sum_{i=1}^{s} \lambda_i (x^* - y_i^*) = 0) \\
= & \sum_{i=1}^{s} \lambda_i \left(x_{n+1} - x_{n+1}^* - \frac{1}{2}\|x - x^*\|_2^2 \right) \\
& \qquad (\text{wegen } y_i^* \in S^*,\ i = 1, \ldots, s) \\
= & \lambda_0 \left(x_{n+1} - x_{n+1}^* - \frac{1}{2}\|x - x^*\|_2^2 \right).
\end{aligned}$$

Hieraus schließen wir:

- *Eine Kugel $B[x; r]$ mit $\{y_1, \ldots, y_s\} \subset B[x; r]$ hat einen Radius $r \geq \sqrt{2x_{n+1}^*}$, wobei das Gleichheitszeichen genau dann gilt, wenn $x = x^*$. Daher ist die kleinste S enthaltende Kugel eindeutig bestimmt und stimmt mit der kleinsten $\{y_1, \ldots, y_s\} \subset S^*$ enthaltenden Kugel überein.*

Denn: Wegen $\{y_1^*, \ldots, y_s^*\} \subset B[x; r]$ ist $\|x - y_i^*\|_2 \leq r$, $i = 1, \ldots, s$. Mit $x_{n+1} := \frac{1}{2}r^2$ ist also

$$\frac{1}{2}\|x - y_i^*\|_2^2 - x_{n+1} \leq 0.$$

Aus der obigen Identität erhalten wir

$$0 \leq \sum_{i=1}^{s} \underbrace{\lambda_i}_{>0} \Bigg(\underbrace{x_{n+1} - \frac{1}{2}\|x - y_i^*\|_2^2}_{\geq 0} \Bigg) = \underbrace{\lambda_0}_{>0} \left(x_{n+1} - x_{n+1}^* - \frac{1}{2}\|x - x^*\|_2^2 \right).$$

Daher ist

$$\frac{1}{2}\|x - x^*\|_2^2 \leq x_{n+1} - x_{n+1}^* = \frac{1}{2}r^2 - x_{n+1}^*.$$

Hieraus liest man die Behauptung ab.

- *Es gilt die Jung'sche Ungleichung: Sei $S \subset \mathbb{R}^n$ kompakt, r der Umkugelradius, also der Radius der kleinsten S enthaltenden Kugel, und d der Durchmesser von S. Dann ist*

$$d \geq \sqrt{\frac{2(n+1)}{n}}\, r.$$

Denn: Es ist

$$\begin{aligned}\sum_{\substack{i,j=1\\ i\neq j}}^{s}\lambda_i\lambda_j\|y_i-y_j\|_2^2 &= \sum_{i,j=1}^{s}\lambda_i\lambda_j\|y_i-x^*+x^*-y_j\|_2^2\\ &= \sum_{i,j=1}^{s}\lambda_i\lambda_j(\underbrace{\|y_i-x^*\|_2^2}_{=2x^*_{n+1}}+2(y_i-x^*)^T(x^*-y_j)+\underbrace{\|x^*-y_j\|_2^2}_{=2x^*_{n+1}})\\ &= 4\lambda_0^2x^*_{n+1}.\end{aligned}$$

Andererseits ist wegen der Cauchy-Schwarzschen Ungleichung

$$\lambda_0=\sum_{i=1}^{s}\lambda_i=\sum_{i=1}^{s}\lambda_i\cdot 1\le\left(\sum_{i=1}^{s}\lambda_i^2\right)^{1/2}\left(\sum_{i=1}^{s}1\right)^{1/2}=\sqrt{s}\left(\sum_{i=1}^{s}\lambda_i^2\right)^{1/2}$$

und daher

$$\sum_{\substack{i,j=1\\ i\neq j}}^{s}\lambda_i\lambda_j=\left(\sum_{i=1}^{s}\lambda_i\right)^2-\sum_{i=1}^{s}\lambda_i^2=\lambda_0^2-\sum_{i=1}^{s}\lambda_i^2\le\frac{s-1}{s}\lambda_0^2.$$

Folglich ist

$$\begin{aligned}2\lambda_0^2r^2 &= 4\lambda_0^2x^*_{n+1}\\ &= \sum_{\substack{i,j=1\\ i\neq j}}^{s}\lambda_i\lambda_j\|y_i-y_j\|_2^2\\ &\le d^2\sum_{\substack{i,j=1\\ i\neq j}}^{s}\lambda_i\lambda_j\\ &\le \frac{s-1}{s}\lambda_0^2\\ &\le d^2\frac{n}{n+1}\lambda_0^2,\end{aligned}$$

woraus sofort die Jung'sche Ungleichung folgt.

2.5 Umschriebenes Ellipsoid: Charakterisierung und Eindeutigkeit einer Lösung

Wir wollen in diesem Unterabschnitt Satz 2.4 auf die Aufgabe anwenden, zu einem vorgegebenen konvexen Körper $K\subset\mathbb{R}^n$ ein umschriebenes Ellipsoid minimalen Volumens zu bestimmen. Dies führt, wie wir gesehen haben, auf die Optimierungsaufgabe

(U) Minimiere $\det(A)$ auf $M:=\{(A,b)\in\mathcal{S}^{n\times n}\times\mathbb{R}^n : A\in\mathcal{S}_+^{n\times n},\ K\subset E(A,b)\}$,

wobei

$$E(A,b):=b+A(B[0;1])$$

das von (A, b) erzeugte Ellipsoid ist, $\mathcal{S}^{n\times n}$ den linearen Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}^{n\times n}_+ \subset \mathcal{S}^{n\times n}$ die Teilmenge der positiv definiten Matrizen bezeichnet. Wegen

$$E(A, b) = \{x \in \mathbb{R}^n : \|A^{-1}(x - b)\|_2 \leq 1\} = \{x \in \mathbb{R}^n : (x - b)^T A^{-2}(x - b) \leq 1\}$$

ist (U) gleichwertig mit

$$\begin{cases} \text{Minimiere} \quad \log\det(A)^2 \quad \text{auf} \\ M := \{(A, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}^{n\times n}_+,\ (x - b)^T A^{-2}(x - b) \leq 1 \text{ für alle } x \in K\}, \end{cases}$$

wobei wir auch noch ausgenutzt haben, dass das Minimieren von $\det(A)$ auf M gleichwertig mit dem Minimieren von $\log\det(A)^2$ auf M ist. Mit $X := A^{-2}$ ist

$$\log\det(A)^2 = \log\det(A^2) = \log\det(X^{-1}) = \log(1/\det(X)) = -\log\det(X)$$

und daher (U) gleichwertig (wir ändern die Bezeichnung für die Menge der zulässigen Lösungen nicht)

$$\text{(U)} \quad \begin{cases} \text{Minimiere} \quad f(X, b) \quad \text{auf} \\ M := \{(X, b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : X \in \mathcal{S}^{n\times n}_+,\ G((X, b), x) \leq 0 \,\text{für alle } x \in K\}, \end{cases}$$

wobei

$$f(X, b) := -\log\det(X), \qquad G((X, b), x) := (x - b)^T X(x - b) - 1.$$

Ist $(X^*, b^*) \in M$ eine Lösung von (U), so ist also $E(A^*, b^*)$ mit $A^* := (X^*)^{-1/2}$ das gesuchte Ellipsoid. Auf die Optimierungsaufgabe (U) wenden wir Satz 2.4, den Satz von F. John an, und erhalten dadurch notwendige Optimalitätsbedingungen, die sich auch als hinreichend herausstellen werden und auf einfache Weise die Eindeutigkeit einer Lösung implizieren.

Auf $\mathbb{R}^{n\times n}$ und damit auch auf $\mathcal{S}^{n\times n}$ ist ein inneres Produkt durch $\langle X, Y\rangle := \mathrm{tr}(X^T Y)$ definiert[7], auf dem $\mathbb{R}^n$ hat man in gewohnter Weise durch $\langle u, v\rangle := u^T v$ das Skalarprodukt als inneres Produkt definiert. Damit hat man etwa auf $\mathcal{S}^{n\times n} \times \mathbb{R}^n$ das innere Produkt

$$\langle (X, b), (Y, c)\rangle := \langle X, Y\rangle + \langle b, c\rangle = \mathrm{tr}(XY) + b^T c.$$

Die Menge $\mathcal{S}^{n\times n}_+$ ist offen in $\mathcal{S}^{n\times n}$, ist also für die Anwendung von Satz 2.4 irrelevant. Dieser Satz liefert uns die folgenden Informationen:

- Sei $(X^*, b^*) \in M$ eine lokale Lösung von (U) und

  $$K^* := \{x \in K : G((X^*, b^*), x) = 0\}.$$

  Dann existieren $s \in \{0, \ldots, \frac{n(n+3)}{2}\}$ Punkte $x_1, \ldots, x_s \in K^*$ sowie Multiplikatoren $\lambda_0, \lambda_1, \ldots, \lambda_s$, die nicht alle verschwinden, mit

  $$\lambda_0 \geq 0,\ \lambda_1 > 0, \ldots, \lambda_s > 0$$

[7] Die zugehörige Norm ist offenbar die Frobenius-Norm.

und

$$(*) \qquad \lambda_0 \nabla_X f(X^*, b^*) + \sum_{i=1}^{s} \lambda_i \nabla_X G((X^*, b^*), x_i) = 0$$

sowie

$$(**) \qquad \lambda_0 \nabla_b f(X^*, b^*) + \sum_{i=1}^{s} \lambda_i \nabla_b G((X^*, b^*), x_i) = 0.$$

Jetzt gilt es, die jeweiligen Gradienten zu berechnen. Wie wir in Satz 2.2 gesehen haben, ist

$$f'((X^*, b^*); (Y, c)) = -\mathrm{tr}((X^*)^{-1}Y) = \langle \nabla_X f(X^*, b^*), Y \rangle + \langle \nabla_b f(X^*, b^*), c \rangle,$$

und daher

$$\nabla_X f(X^*, b^*) = -(X^*)^{-1}, \qquad \nabla_b f(X^*, b^*) = 0.$$

Weiter ist

$$\begin{aligned}
& \frac{G((X^*, b^*) + t(Y, c), x) - G((X^*, b^*), x)}{t} \\
= \; & \frac{(x - b^* - tc)^T (X^* + tY)(x - b^* - tc) - (x - b^*)^T X^* (x_b^*)}{t} \\
= \; & \{(x - b^*)^T [Y(x - b^*) - X^* c] - c^T X^* (x - b^*)\} + O(t) \\
= \; & (x - b^*)^T Y (x - b^*) - 2(X^*(x - b^*))^T c + O(t) \\
= \; & \mathrm{tr}((x - b^*)(x - b^*)^T Y) - 2(X^*(x - b^*))^T c + O(t)
\end{aligned}$$

und daher

$$\nabla_X G((X^*, b^*), x) = (x - b^*)(x - b^*)^T, \qquad \nabla_b G((X^*, b^*), x) = -2X^*(x - b^*).$$

Die Beziehungen $(*)$ und $(**)$ ergeben

$$-\lambda_0 (X^*)^{-1} + \sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T = 0, \qquad \sum_{i=1}^{s} \lambda_i (x_i - b^*) = 0.$$

Wäre hier $\lambda_0 = 0$, so wäre

$$\sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T = 0$$

und damit

$$\mathrm{tr}\left(\sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T \right) = \sum_{i=1}^{s} \lambda_i \|x_i - b^*\|_2^2 = 0$$

bzw. $s = 0$, ein Widerspruch. O. B. d. A. ist daher $\lambda_0 = 1$. Wir fassen die erhaltenen notwendigen Optimalitätsbedingungen im ersten Teil des folgenden Satzes zusammen und zeigen anschließend, dass diese auch hinreichend für Optimalität sind und die Eindeutigkeit einer Lösung implizieren (siehe z. B. O. Güler, F. Gürtuna (2007, Theorem 2.7)).

Satz 2.6 *Sei $K \subset \mathbb{R}^n$ ein konvexer Körper. Hierzu betrachte man die Optimierungsaufgabe*

$$\text{(U)} \quad \left\{ \begin{array}{c} \textit{Minimiere} \quad f(X,b) \quad \textit{auf} \\ M := \{(X,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : X \in \mathcal{S}_+^{n\times n},\ G((X,b),x) \leq 0 \,\textit{für alle}\, x \in K\}, \end{array} \right.$$

wobei

$$f(X,b) := -\log\det(X), \qquad G((X,b),x) := (x-b)^T X (x-b) - 1.$$

1. *Sei $(X^*,b^*) \in M$ eine (lokale) Lösung von (U) und*

$$K^* := \{x \in K : (x-b^*)^T X^* (x-b^*) = 1\}.$$

Dann existieren $s \in \{0,\ldots,\frac{n(n+3)}{2}\}$ Punkte $x_1,\ldots,x_s \in K^*$ sowie positive Multiplikatoren $\lambda_1,\ldots\lambda_s$ mit

$$(X^*)^{-1} = \sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T, \qquad \sum_{i=1}^{s} \lambda_i (x_i - b^*) = 0.$$

2. *Zu $(X^*,b^*) \in M$ mögen s Punkte $x_1,\ldots,x_s \in K^*$ sowie positive Multiplikatoren $\lambda_1,\ldots\lambda_s$ mit*

$$(X^*)^{-1} = \sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T, \qquad \sum_{i=1}^{s} \lambda_i (x_i - b^*) = 0$$

existieren. Hierbei sei

$$K^* := \{x \in K : (x-b^*)^T X^* (x-b^*) = 1\}.$$

Dann ist (X^*,b^*) eine (globale) Lösung von (U). Die in 1. angegebenen notwendigen Optimalitätsbedingungen sind also auch hinreichend für Optimalität.

3. *Die Aufgabe (U) ist eindeutig lösbar.*

Beweis: Der erste Teil des Satzes ist schon bewiesen. Zum Beweis des zweiten Teiles

gebe man sich $(X, b) \in M$ vor. Dann ist

$$\begin{aligned}
f(X,b) - f(X^*,b^*) &\geq -\mathrm{tr}((X^*)^{-1}(X - X^*)) \\
&\qquad \text{(siehe 1. und 2. von Satz 2.2)} \\
&= -\mathrm{tr}\left(\sum_{i=1}^{s} \lambda_i (x_i - b^*)(x_i - b^*)^T (X - X^*)\right) \\
&= \sum_{i=1}^{s} \lambda_i [\underbrace{(x_i - b^*)^T X^* (x_i - b^*)}_{=1} - (x_i - b^*)^T X (x_i - b^*)] \\
&= \sum_{i=1}^{s} \lambda_i [1 - (x_i - b^*)^T X (x_i - b + b - b^*)] \\
&= \sum_{i=1}^{s} \lambda_i [1 - (x_i - b^*)^T X (x_i - b)] \\
&= \sum_{i=1}^{s} \lambda_i [\underbrace{1 - (x_i - b)^T X (x_i - b)}_{\geq 0} + (b^* - b)^T X (x_i - b)] \\
&\geq (b^* - b)^T X \sum_{i=1}^{s} \lambda_i (x_i - b) \\
&= \left(\sum_{i=1}^{s} \lambda_i\right) (b^* - b)^T X (b^* - b) \\
&\geq 0.
\end{aligned}$$

Damit ist gezeigt, dass (X^*, b^*) eine Lösung von (U) ist.

Die Eindeutigkeit einer Lösung (nur diese ist noch zu zeigen) von (U) liest man aus obiger Gleichungs-Ungleichungskette ab. Sind nämlich $(X, b) \in M$ und $(X^*, b^*) \in M$ zwei Lösungen von (U), so sind die Ungleichungen in obiger Kette in Wahrheit Gleichungen. Gleichheit in der ersten Ungleichung impliziert $X = X^*$ (siehe Teil 2. von Satz 2.2). Gleichheit bei der letzten Ungleichung impliziert $b = b^*$, da $X \in \mathcal{S}_+^{n \times n}$. Damit ist der Satz bewiesen. □

Bemerkung: Die in Satz 2.6 1. und 2. vorkommenden Punkte $x_1, \ldots, x_s \in K^*$ liegen nicht nur auf dem Rand des Ellipsoids $E(A^*, b^*)$ minimalen Volumens, das K enthält, sondern auch auf dem Rand von K. Denn wäre $x_i \in \mathrm{int}\,(K)$ für ein $i \in \{1, \ldots, s\}$, so gäbe es ein $\epsilon > 0$ mit $B[x_i; \epsilon] \subset K$. Da $(X^*, b^*) \in M$, ist $(x - b^*)^T X^* (x - b^*) \leq 1$ für alle $x \in B[x_i; \epsilon]$. Wegen $x_i \in K^*$ wäre $2z^T X^* (x_i - b^*) + \epsilon z^T X^* z \leq 0$ für alle $z \in B[0; 1]$, ein Widerspruch zu $x_i \in K^*$. Die Punkte $x_1, \ldots, x_s$ werden *Kontakt-Punkte* von K und $E(A^*, b^*)$ genannt. □

2.6 Einbeschriebenes Ellipsoid: Charakterisierung und Eindeutigkeit einer Lösung

Die Aufgabe, zu einem konvexen Körper $K \subset \mathbb{R}^n$ das Ellipsoid maximalen Volumens zu bestimmen, das in K enthalten bzw. einbeschrieben ist (wegen Satz 2.1 und Satz

2.3 sind Existenz und Eindeutigkeit eines solchen Ellipsoids gesichert) führt auf die Optimierungsaufgabe

$$\text{(E)} \qquad \begin{cases} \text{Maximiere} \quad \det(A) \quad \text{auf} \\ N := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ E(A,b) \subset K\}, \end{cases}$$

wobei

$$E(A,b) := b + A(B[0;1])$$

das von (A,b) erzeugte Ellipsoid ist. Die Nebenbedingung $E(A,b) \subset K$ ist jetzt ein wenig schwerer zu behandeln als die Bedingung $K \subset E(A,b)$ im Unterabschnitt 2.5. Bei unseren Überlegungen benutzen wir u. a. die Darstellungen bei F. Juhnke (1994) und O. Güler, F. Gürtuna (2007, Abschnitt 3). Hierzu definieren wir:

Definition 2.7 Sei $C \subset \mathbb{R}^n$ nichtleer, konvex und kompakt. Dann heißt die Abbildung $\sigma_C \colon \mathbb{R}^n \longrightarrow \mathbb{R}$, definiert durch

$$\sigma_C(p) := \max_{x\in C} p^T x,$$

das *Stützfunktional zu* C.

Die für uns wichtige Eigenschaft eines Stützfunktionals geben wir im folgenden Lemma an.

Lemma 2.8 *Sind $C, D \subset \mathbb{R}^n$ zwei nichtleere, konvexe und kompakte Teilmengen des $\mathbb{R}^n$, so ist $C \subset D$ genau dann, wenn $\sigma_C(p) \le \sigma_D(p)$ für alle $p \in \mathbb{R}^n$ mit $\|p\|_2 = 1$.*

Beweis: Ist $C \subset D$, so ist offensichtlich $\sigma_C(p) \le \sigma_D(p)$ für alle $p \in \mathbb{R}^n$. Umgekehrt sei $\sigma_C(p) \le \sigma_D(p)$ für alle $p \in \mathbb{R}^n$ mit $\|p\|_2 = 1$. Angenommen, es existiert $x_0 \in C$ mit $x_0 \notin D$. Aus dem starken Trennungssatz folgt, dass sich $\{x_0\}$ und D stark trennen lassen. Dies bedeutet, dass es ein $p \in \mathbb{R}^n$ mit

$$\sigma_D(p) = \max_{x\in D} p^T x < p^T x_0 \le \max_{x\in C} p^T x = \sigma_C(p)$$

gibt, wobei o. B. A. $\|p\|_2 = 1$. Das ist ein Widerspruch und das Lemma ist bewiesen. □

Beispiel: Wir wollen das Stützfunktional zum Ellipsoid

$$E(A,b) := b + A(B[0;1])$$

berechnen, wobei $(A,b) \in \mathcal{S}_+^{n\times n} \times \mathbb{R}^n$. Es ist

$$\begin{aligned} \sigma_{E(A,b)}(p) &= \max_{x\in E(A,b)} p^T x \\ &= p^T b + \max_{\|u\|_2\le 1} p^T A u \\ &= p^T b + \max_{\|u\|_2\le 1} (Ap)^T u \\ &= p^T b + \|Ap\|_2. \end{aligned}$$

Die Nebenbedingung $E(A,b) \subset K$ ist damit gleichwertig mit

$$\|Ap\|_2 + p^T b - s_K(p) \leq 0$$

für alle $p \in S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}$. □

Nun kommen wir zu einer Umformulierung der Aufgabe (E), zu einem konvexen Körper ein hierin enthaltenes Ellipsoid maximalen Volumens zu bestimmen. Zunächst beachte man, dass das Maximieren von $\det(A)$ gleichwertig mit dem Minimieren von $-\log\det(A)$ ist. Daher erhalten wir die Optimierungsaufgabe

$$\text{(E)} \quad \begin{cases} \text{Minimiere} \quad f(A,b) \quad \text{auf} \\ N := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}^{n\times n}_+,\ G((A,b),p) \leq 0 \text{ für alle } p \in S\}, \end{cases}$$

wobei

$$S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}$$

und

$$f(A,b) := -\log\det(A), \qquad G((A,b),p) := \|Ap\|_2 + p^T b - \sigma_K(p).$$

Ist also $(A^*, b^*) \in N$ eine Lösung von (E), so ist $E(A^*, b^*)$ das gesuchte Ellipsoid. Jetzt wenden wir Satz 2.4 auf die Aufgabe (E) an, genau wie wir es im vorigen Abschnitt bei der Aufgabe (U) getan haben. Sei also $(A^*, b^*) \in N$ eine (lokale) Lösung von (E). Zunächst sind $\nabla_A G((A^*, b^*), p)$ und $\nabla_b G((A^*, b^*), p)$ zu berechnen. Für $(B, c) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n$ und $t > 0$ ist

$$\begin{aligned} \frac{G((A^*,b^*) + t(B,c), p) - G((A^*,b^*),p)}{t} &= \frac{p^T A^* B p + p^T B A^* p}{2\|A^* p\|_2} + p^T c + \frac{o(t)}{t} \\ &= \operatorname{tr}\left(\frac{(A^*p)p^T + p(A^*p)^T}{2\|A^*p\|_2} B\right) + p^T c + \frac{o(t)}{t} \end{aligned}$$

und daher

$$\nabla_A G((A^*,b^*),p) = \frac{(A^*p)p^T + p(A^*p)^T}{2\|A^*p\|_2}, \qquad \nabla_b G((A^*,b^*),p) = p.$$

Eine Anwendung von Satz 2.4 liefert mit

$$S^* := \{p \in S : \|A^*p\|_2 + p^T b^* = \sigma_K(p)\}$$

die Existenz von $s \in \{0, \ldots, \frac{n(n+3)}{2}\}$ Punkten $p_1, \ldots, p_s$ sowie von Multiplikatoren $\lambda_0, \lambda_1, \ldots, \lambda_s$, die nicht alle verschwinden, mit

$$\lambda_0 \geq 0,\ \lambda_1 > 0, \ldots, \lambda_s > 0$$

und

$$\lambda_0 (A^*)^{-1} = \sum_{i=1}^{s} \lambda_i \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2\|A^*p_i\|_2}, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0.$$

Wäre $\lambda_0 = 0$, so wäre

$$0 = \operatorname{tr}\left(\sum_{i=1}^{s} \lambda_i \frac{1}{2\|A^* p_i\|_2}[(A^* p_i)p_i^T + p_i(A^* p_i)^T]\right) = \sum_{i=1}^{s} \lambda_i \underbrace{\frac{p_i^T A^* p_i}{\|A^* p_i\|_2}}_{>0}$$

und daher $\lambda_1 = \cdots = \lambda_s = 0$, ein Widerspruch. Daher ist o. B. d. A. $\lambda_0 = 1$. Ganz entsprechend zu Satz 2.6 erhalten wir damit im folgenden Satz eine Charakterisierung einer Lösung von (E) sowie eine Eindeutigkeitsaussage.

Satz 2.9 *Sei $K \subset \mathbb{R}^n$ ein konvexer Körper. Hierzu betrachte man die Optimierungsaufgabe*

$$\text{(E)} \quad \left\{ \begin{array}{c} \textit{Minimiere} \quad f(A,b) \quad \textit{auf} \\ N := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ G((A,b),p) \le 0 \textit{ für alle } p \in S\}, \end{array} \right.$$

wobei

$$f(A,b) := -\log\det(A), \qquad G((A,b),p) := \|Ap\|_2 + p^T b - \sigma_K(p)$$

und

$$S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}, \qquad \sigma_K(p) := \max_{x\in K} p^T x.$$

1. *Sei $(A^*, b^*) \in N$ eine (lokale) Lösung von (E) und*

   $$S^* := \{p \in S : \|A^* p\|_2 + p^T b^* - \sigma_K(p) = 0\}.$$

   Dann existieren $s \in \{0, \ldots, \frac{n(n+3)}{2}\}$ Punkte $p_1, \ldots, p_s \in S^*$ sowie positive Multiplikatoren $\lambda_1, \ldots \lambda_s$ mit

   $$(A^*)^{-1} = \sum_{i=1}^{s} \lambda_i \frac{(A^* p_i)p_i^T + p_i(A^* p_i)^T}{2\|A^* p_i\|_2}, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0.$$

2. *Zu $(A^*, b^*) \in N$ mögen s Punkte $p_1, \ldots, p_s \in S^*$ sowie positive Multiplikatoren $\lambda_1, \ldots \lambda_s$ mit*

   $$(A^*)^{-1} = \sum_{i=1}^{s} \lambda_i \frac{(A^* p_i)p_i^T + p_i(A^* p_i)^T}{2\|A^* p_i\|_2}, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0$$

   existieren. Hierbei sei

   $$S^* := \{p \in S : \|A^* p\|_2 + p^T b^* - \sigma_K(p) = 0\}.$$

   Dann ist (A^*, b^*) eine (globale) Lösung von (E). Die in 1. angegebenen notwendigen Optimalitätsbedingungen sind also auch hinreichend für Optimalität.

3. *Die Aufgabe (E) ist eindeutig lösbar.*

Beweis: Der erste Teil des Satzes ist schon bewiesen. Zum Beweis des zweiten Teiles gebe man sich $(A, b) \in N$ vor. Dann ist

$$\begin{aligned}
f(A,b) - f(A^*,b^*) &\geq -\mathrm{tr}((A^*)^{-1}(A-A^*)) \\
&\qquad \text{(siehe 1. und 2. von Satz 2.2)} \\
&= -\mathrm{tr}\left(\sum_{i=1}^{s} \lambda_i \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2\|A^*p_i\|_2}(A-A^*)\right) \\
&= \sum_{i=1}^{s} \lambda_i \frac{(A^*p_i)^T(A^*-A)p_i}{\|A^*p_i\|_2} \\
&= \sum_{i=1}^{s} \lambda_i \left(\|A^*p_i\|_2 - \frac{(A^*p_i)^T(Ap_i)}{\|A^*p_i\|_2}\right) \\
&\geq \sum_{i=1}^{s} \lambda_i (\|A^*p_i\|_2 - \|Ap_i\|_2) \\
&= \sum_{i=1}^{s} \lambda_i (\sigma_K(p_i) - p_i^T b^* - \|Ap_i\|_2) \\
&\geq \sum_{i=1}^{s} \lambda_i p_i^T (b-b^*) \\
&= \underbrace{\left(\sum_{i=1}^{s} \lambda_i p_i\right)}_{=0}^{T} (b-b^*) \\
&= 0.
\end{aligned}$$

Damit ist gezeigt, dass (A^*, b^*) eine Lösung von (E) ist.

Die Eindeutigkeit einer Lösung (nur diese ist noch zu zeigen) von (E) liest man aus obiger Gleichungs-Ungleichungskette ab. Sind nämlich $(A, b) \in N$ und $(A^*, b^*) \in N$ zwei Lösungen von (E), so sind die Ungleichungen in obiger Kette in Wahrheit Gleichungen. Gleichheit in der ersten Ungleichung impliziert $A = A^*$ (siehe Teil 2. von Satz 2.2). Da die letzte Ungleichung eine Gleichung ist, ist $p_i^T(b-b^*) = 0$, $i = 1, \ldots, s$. Dann ist aber

$$\begin{aligned}
(b-b^*)^T(A^*)^{-1}(b-b^*) &= \sum_{i=1}^{s} \frac{\lambda_i}{2\|A^*p_i\|_2}[(b-b^*)^T(A^*p_i)\underbrace{p_i^T(b-b^*)}_{=0} \\
&\qquad + \underbrace{(b-b^*)^T p_i}_{=0}(A^*p_i)^T(b-b^*)] \\
&= 0
\end{aligned}$$

und folglich $b = b^*$, da $(A^*)^{-1} \in \mathcal{S}_+^{n\times n}$. Damit ist die Eindeutigkeit einer Lösung von (E) (erneut) und damit der ganze Satz bewiesen. □

Bemerkung: Seien $p_1, \ldots, p_s$ und $\lambda_1, \ldots, \lambda_s$ wie im ersten Teil von Satz 2.9 gegeben. Definiert man

$$x_i := b^* + A^*\left(\frac{A^*p_i}{\|A^*p_i\|_2}\right), \qquad i = 1, \ldots, s,$$

so ist

$$(x_i - b^*)^T (A^*)^{-2} (x_i - b^*) = 1, \qquad i = 1, \ldots, s,$$

und daher $x_i \in \partial E(A^*, b^*) \subset K$, $i = 1, \ldots, s$. Wegen

$$p_i^T x_i = p_i^T b^* + \|A^* p_i\|_2 = \sigma_K(p_i), \qquad i = 1, \ldots, s,$$

ist $x_i \in \partial K$, und folglich sind $x_i \in \partial K \cap \partial E(A^*, b^*)$, $i = 1, \ldots, s$, Kontaktpunkte von K und $E(A^*, b^*)$. Mit

$$\mu_i := \lambda_i \, \|A^* p_i\|_2, \qquad i = 1, \ldots, s,$$

ist nach einfacher Rechnung

$$A^* = \sum_{i=1}^{s} \frac{\mu_i}{2} [(x_i - b^*)(x_i - b^*)^T (A^*)^{-1} + (A^*)^{-1} (x_i - b^*)(x_i - b^*)^T]$$

und

$$\sum_{i=1}^{s} \mu_i (x_i - b^*) = 0.$$

Im folgenden Korollar zeigen wir, dass diese Bedingungen auch *hinreichend* für die Optimalität eines Paares (A^*, b^*) ist. □

Korollar 2.10 *Sei $K \subset \mathbb{R}^n$ ein konvexer Körper. Hierzu betrachte man die Optimierungsaufgabe*

$$\text{(E)} \qquad \begin{cases} \text{Minimiere} \quad f(A,b) \quad \text{auf} \\ N := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ G((A,b),p) \le 0 \text{ für alle } p \in S\}, \end{cases}$$

wobei

$$f(A,b) := -\log\det(A), \qquad G((A,b),p) := \|Ap\|_2 + p^T b - \sigma_K(p)$$

und

$$S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}, \qquad \sigma_K(p) := \max_{x \in K} p^T x.$$

Zu $(A^*, b^*) \in N$ mögen s Kontaktpunkte $x_1, \ldots, x_s \in \partial K \cap \partial E(A^*, b^*)$ und positive Multiplikatoren $\mu_1, \ldots, \mu_s$ mit

$$A^* = \sum_{i=1}^{s} \frac{\mu_i}{2} [(x_i - b^*)(x_i - b^*)^T (A^*)^{-1} + (A^*)^{-1} (x_i - b^*)(x_i - b^*)^T]$$

und

$$\sum_{i=1}^{s} \mu_i (x_i - b^*) = 0$$

existieren. Dann ist $(A^*, b^*) \in N$ eine (globale) Lösung von (E).

Beweis: Mit

$$S^* := \{p \in S : \|A^*p\|_2 + p^T b^* - \sigma_K(p) = 0\}$$

definieren wir $p_1, \ldots, p_s \in S^*$ und positive $\lambda_1, \ldots, \lambda_s$ mit der Eigenschaft, dass

$$(A^*)^{-1} = \sum_{i=1}^{s} \lambda_i \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2\|A^*p_i\|_2}, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0.$$

Aus dem zweiten Teil von Satz 2.9 erhalten wir, dass (A^*, b^*) Lösung von (E) ist. Naheliegenderweise definieren wir

$$p_i := \frac{(A^*)^{-2}(x_i - b^*)}{\|(A^*)^{-2}(x_i - b^*)\|_2}, \qquad \lambda_i := \frac{\mu_i}{\|A^*p_i\|_2} \qquad (i = 1, \ldots, s).$$

Offenbar ist $p_i \in S$, $i = 1, \ldots, , s$. Es ist aber sogar $p_i \in S^*$, $i = 1, \ldots, s$. Denn für $i \in \{1, \ldots, s\}$ ist

$$\begin{aligned}
\|A^*p_i\|_2 + p_i^T b^* - \sigma_K(p_i) &= \|A^*p_i\|_2 + p_i^T b^* - p_i^T x_i \\
&\qquad (\text{denn } x_i \in \partial K) \\
&= \frac{\|(A^*)^{-1}(x_i - b^*)\|_2}{\|(A^*)^{-2}(x_i - b^*)\|_2} - \frac{(x_i - b^*)(A^*)^{-2}(x_i - b^*)}{\|(A^*)^{-2}(x_i - b^*)\|_2} \\
&= \frac{1}{\|(A^*)^{-2}(x_i - b^*)\|_2} - \frac{1}{\|(A^*)^{-2}(x_i - b^*)\|_2} \\
&\qquad (\text{denn } x_i \in \partial E(A^*, b^*)) \\
&= 0.
\end{aligned}$$

Folglich ist $p_i \in S^*$, $i = 1, \ldots, s$. Weiter ist

$$\begin{aligned}
&\sum_{i=1}^{s} \lambda_i \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2\|A^*p_i\|_2} \\
= &\sum_{i=1}^{s} \mu_i \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2\|A^*p_i\|_2^2} \\
= &\sum_{i=1}^{s} \mu_i \|(A^*)^{-2}(x_i - b^*)\|_2^2 \frac{(A^*p_i)p_i^T + p_i(A^*p_i)^T}{2} \\
= &\sum_{i=1}^{s} \mu_i \frac{(A^*)^{-1}(x_i - b^*)(x_i - b^*)^T(A^*)^{-2} + (A^*)^{-2}(x_i - b^*)(x_i - b^*)^T(A^*)^{-1}}{2} \\
= &(A^*)^{-1} \underbrace{\left(\sum_{i=1}^{s} \frac{\mu_i}{2} [(x_i - b^*)(x_i - b^*)^T(A^*)^{-1} + (A^*)^{-1}(x_i - b^*)(x_i - b^*)^T] \right)}_{=A^*} (A^*)^{-1} \\
= &(A^*)^{-1}.
\end{aligned}$$

Weiter ist

$$\sum_{i=1}^{s} \lambda_i p_i = \sum_{i=1}^{s} \mu_i \frac{(A^*)^{-2}(x_i - b^*)}{\|A^*p_i\|_2 \, \|(A^*)^{-2}(x_i - b^*)\|_2} = (A^*)^{-2} \underbrace{\sum_{i=1}^{s} \mu_i (x_i - b^*)}_{=0} = 0.$$

Jetzt folgt die Behauptung aus dem zweiten Teil von Satz 2.9. □

2.7 Die Sätze von John-Löwner

Unser Ziel in diesem Unterabschnitt ist es, zwei Sätze zu beweisen, die in einem gewissen Sinne zueinander dual sind. Der erste Satz stammt von F. John (1948, Theorem III), siehe auch O. Güler, F. Gürtuna (2007, Theorem 2.9)).

Satz 2.11 (John) *Sei $K \subset \mathbb{R}^n$ ein konvexer Körper und (X^*, b^*) die Lösung der Aufgabe*

$$\text{(U)} \quad \begin{cases} \text{Minimiere} \quad f(X,b) \quad \text{auf} \\ M := \{(X,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : X \in \mathcal{S}_+^{n\times n},\ G((X,b),x) \le 0 \text{ für alle } x \in K\}, \end{cases}$$

wobei

$$f(X,b) := -\log\det(X), \qquad G((X,b),x) := (x-b)^T X(x-b) - 1,$$

also $E(A^*,b^*) := b^* + A^*(B[0;1])$ mit $A^* := (X^*)^{-1/2}$ das Ellipsoid kleinsten Volumens, das K enthält. Dann ist

$$b^* + \frac{1}{n}A^*(B[0;1]) \subset K \subset b^* + A^*(B[0;1]).$$

Ist K symmetrisch zum Nullpunkt, also $K = -K$, so ist notwendig $b^* = 0$ und der Faktor $1/n$ kann durch $1/\sqrt{n}$ ersetzt werden, es ist also

$$\frac{1}{\sqrt{n}}A^*(B[0;1]) \subset K \subset A^*(B[0;1]).$$

Beweis: Wegen Satz 2.6 existieren $s \in \{0,\dots,\frac{n(n+3)}{2}\}$ Punkte $x_1,\dots,x_s \in K^*$ sowie positive Multiplikatoren $\lambda_1,\dots\lambda_s$ mit

$$(X^*)^{-1} = \sum_{i=1}^s \lambda_i (x_i - b^*)(x_i - b^*)^T, \qquad \sum_{i=1}^s \lambda_i (x_i - b^*) = 0.$$

Hierbei ist

$$K^* := \{x \in K : (x-b^*)^T X^* (x-b^*) = 1\}.$$

Dann ist

$$n = \operatorname{tr}(I) = \operatorname{tr}\left(\sum_{i=1}^s \lambda_i X^*(x_i-b^*)(x_i-b^*)^T\right) = \sum_{i=1}^s \lambda_i \underbrace{(x_i-b^*)^T X^*(x_i-b^*)}_{=1} = \sum_{i=1}^s \lambda_i.$$

Unser Ziel ist es zu zeigen, dass

$$b^* + \frac{1}{n}A^*(B[0;1]) \subset P := \operatorname{co}(\{x_1,\dots,x_s\}),$$

woraus wegen co $(\{x_1, \ldots, x_s\}) \subset K$ die Behauptung folgt. Sei also

$$y \in b^* + \frac{1}{n} A^*(B[0;1]).$$

Dann ist

$$(y-b^*)^T \underbrace{X^*}_{=(A^*)^{-2}} (y-b^*) \leq \frac{1}{n^2} \qquad \text{bzw.} \qquad \| \underbrace{(X^*)^{1/2}}_{=(A^*)^{-1}} (y-b^*)\|_2 \leq \frac{1}{n}.$$

Weiter ist

$$y - b^* = \sum_{i=1}^{s} \lambda_i [(x_i - b^*)^T X^* (y-b^*)](x_i - b^*).$$

Mit

$$\alpha_i := (x_i - b^*)^T X^* (y-b^*), \qquad i = 1, \ldots, s,$$

ist

$$\sum_{i=1}^{s} \lambda_i \alpha_i = \Bigg(\underbrace{\sum_{i=1}^{s} \lambda_i (x_i - b^*)}_{=0} \Bigg)^T X^* (y-b^*) = 0$$

und daher

$$\begin{aligned} y &= b^* + \sum_{i=1}^{s} \lambda_i \alpha_i (x_i - b^*) \\ &= \Bigg(1 - \underbrace{\sum_{i=1}^{s} \lambda_i \alpha_i}_{=0} \Bigg) b^* + \sum_{i=1}^{s} \lambda_i \alpha_i x_i \\ &= \frac{1}{n} \sum_{i=1}^{s} \lambda_i x_i + \sum_{i=1}^{s} \lambda_i \alpha_i x_i \\ &= \sum_{i=1}^{s} \mu_i x_i, \end{aligned}$$

wobei

$$\mu_i := \lambda_i \Big(\frac{1}{n} + \alpha_i \Big), \qquad i = 1, \ldots, s.$$

Dann ist

$$\sum_{i=1}^{s} \mu_i = \Bigg(\underbrace{\sum_{i=1}^{s} \lambda_i}_{=n} \Bigg) \cdot \frac{1}{n} + \underbrace{\sum_{i=1}^{s} \lambda_i \alpha_i}_{=0} = 1.$$

Zu zeigen bleibt also, dass $\mu_i \geq 0$, $i = 1, \ldots, s$. Dies ist einfach, denn wegen der Cauchy-Schwarzschen Ungleichung ist

$$\begin{aligned} |\alpha_i| &= |(x_i - b^*)^T X^* (y - b^*)| \\ &= |(X^*)^{1/2}(x_i - b^*))^T (X^*)^{1/2}(y - b^*)| \\ &\leq \underbrace{\|(X^*)^{1/2}(x_i - b^*)\|_2}_{=1} \underbrace{\|(X^*)^{1/2}(y - b^*)\|_2}_{\leq 1/n} \\ &\leq \frac{1}{n} \end{aligned}$$

und folglich

$$\mu_i = \underbrace{\lambda_i}_{>0} \Big(\underbrace{\frac{1}{n} + \alpha_i}_{\geq 0} \Big) \geq 0, \qquad i = 1, \ldots, s.$$

Damit ist der erste Teil des Satzes bewiesen. Nun nehmen wir an, es sei $K = -K$. Wir zeigen, dass auch $K \subset E(-b^*, A^*)$. Da $E(A^*, b^*)$ und $E(A^*, -b^*)$ dasselbe Volumen besitzen, folgt aus der Eindeutigkeit einer Lösung von (U) (siehe Teil 3 von Satz 2.6), dass $b^* = -b^*$ bzw. $b^* = 0$. Sei also $x \in K$. Dann ist auch $-x \in K \subset E(A^*, b^*)$, folglich $\|(A^*)^{-1}(-x - b^*)\|_2 \leq 1$, dann $\|(A^*)^{-1}(x - (-b^*)\|_2 \leq 1$ und damit $x \in E(A^*, -b^*)$. Also ist $b^* = 0$. Wir zeigen nun, dass

$$\frac{1}{\sqrt{n}} A^*(B[0;1]) \subset D := \text{co}\,(\{\pm x_1, \ldots, \pm x_s\}),$$

woraus wegen $D \subset K$ die Behauptung folgt. Ein direkter Beweis wie im ersten Teil des Satzes ist mir nicht eingefallen. Daher machen wir (wie z. B. O. Güler, F. Gürtuna (2007, Theorem 2.9)) einen Umweg über *polare Mengen*. Ist $C \subset \mathbb{R}^n$ nichtler, konvex und kompakt, so definierten wir das *Stützfunktional* $\sigma_C \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch $\sigma_C(p) := \max_{x \in C} p^T x$. Die zu C *polare Menge* C° ist definiert durch

$$C^\circ := \{p \in \mathbb{R}^n : \sigma_C(p) \leq 1\} = \{p \in \mathbb{R}^n : p^T x \leq 1 \text{ für alle } x \in C\}.$$

Sind $C, D \subset \mathbb{R}^n$ zwei nichtleere, konvexe und kompakte Mengen mit $C \subset D$, so ist $\sigma_C(p) \leq \sigma_D(p)$ für alle $p \in \mathbb{R}^n$ und daher $D^\circ \subset C^\circ$. Nun berechnen wir die polaren Mengen für

$$C := \frac{1}{\sqrt{n}} A^*(B[0;1]), \qquad D := \text{co}\,(\{\pm x_1, \ldots, \pm x_s\}).$$

Es ist

$$\begin{aligned} \sigma_C(p) &= \max_{x \in C} p^T x \\ &= \frac{1}{\sqrt{n}} \max_{\|u\|_2 \leq 1} p^T A^* u \\ &= \frac{1}{\sqrt{n}} \max_{\|u\|_2 \leq 1} (A^* p)^T u \\ &= \frac{1}{\sqrt{n}} \|A^* p\|_2. \end{aligned}$$

Daher ist

$$C^\circ = \left\{ p \in \mathbb{R}^n : \frac{1}{\sqrt{n}} \|A^* p\|_2 \leq 1 \right\}.$$

Weiter ist

$$\begin{aligned} \sigma_D(p) &= \max_{x \in D} p^T x \\ &= \max_{x \in \mathrm{co}\,(\{\pm x_1, \ldots, \pm x_s\})} p^T x \\ &= \max_{i=1,\ldots,s} |p^T x_i| \end{aligned}$$

und daher

$$D^\circ = \{p \in \mathbb{R}^n : |p^T x_i| \leq 1,\ i = 1, \ldots, s\}.$$

Ist $p \in D^\circ$, so ist

$$\begin{aligned} \frac{1}{\sqrt{n}} \|A^* p\|_2 &= \frac{1}{\sqrt{n}} \sqrt{p^T (A^*)^2 p} \\ &= \frac{1}{\sqrt{n}} \sqrt{p^T (X^*)^{-1} p} \\ &= \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^s \lambda_i \underbrace{(p^T x_i)^2}_{\leq 1}} \\ &\leq \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^s \lambda_i} \\ &= 1. \end{aligned}$$

Also ist $D^\circ \subset C^\circ$. Hieraus folgt $C \subset D$. Denn angenommen, es existiert ein $q \in C$ mit $q \notin D$. Die konvexe Hülle endlich vieler Punkte (bzw. von $\{\pm x_1, \ldots, \pm x_s\}$) ist eine nichtleere, konvexe und kompakte Menge. Wegen des starken Trennungssatzes lassen sich $\{q\}$ und D stark trennen. Also existiert $p \in \mathbb{R}^n \setminus \{0\}$ und $\gamma \in \mathbb{R}$ mit $p^T x \leq \gamma < p^T q$ für alle $x \in D$. Dies ist gleichbedeutend mit $|p^T x_i| \leq \gamma < p^T q$, $i = 1, \ldots, s$. Hier ist $\gamma > 0$. Denn andernfalls wäre $p^T x_i = 0$, $i = 1, \ldots, s$, was wegen

$$0 < p^T (X^*)^{-1} p = \sum_{i=1}^s \lambda_i (p^T x_i)^2 = 0$$

zu einem Widerspruch führt. Da man notfalls p durch p/γ ersetzen kann, kann $\gamma = 1$ angenommen werden. Also ist $p \in D^\circ \subset C^\circ$. Wegen $q \in C$ und $p^T q > 1$ ist das ein Widerspruch. Damit ist der Satz schließlich bewiesen. □

Bemerkung: Die Darstellung im symmetrischen Fall (d. h. $K = -K$) ist in der Literatur oft nicht ganz befriedigend. So findet man etwa bei O. Güler, F. Gürtuna (2007, Theorem 2.9) die folgende Aussage:

- Let K be a convex body in $\mathbb{R}^n$ and $E(X, c)$ be its optimal circumscribing ellipsoid. The ellipsoid with the same center c but shrunk by a factor n is contained in K. If K is symmetric ($K = -K$), then the ellipsoid with the same center c but shrunk only by a factor $\sqrt{n}$ is contained in K.

Hierdurch wird nicht deutlich, dass im symmetrischen Fall das Zentrum des optimalen Ellipsoids notwendig der Nullpunkt ist. Weiter sei darauf hingewiesen, dass wir im ersten Teil des Satzes die Aussage

$$b^* + \frac{1}{n} A^*(B[0;1]) \subset K$$

sehr einfach aus den notwendigen Optimalitätsbedingungen direkt beweisen konnten (was in der mir bekannten Literatur nicht getan wird), der Beweis des zweiten Teiles aber etwas aufwändiger ist. □

Bemerkung: In Abbildung 2 geben wir ein gleichseitiges Dreieck und den zugehörigen Umkreis und Inkreis an. Das Verhältnis der Radien von Um- und Inkreis ist 2. Wir

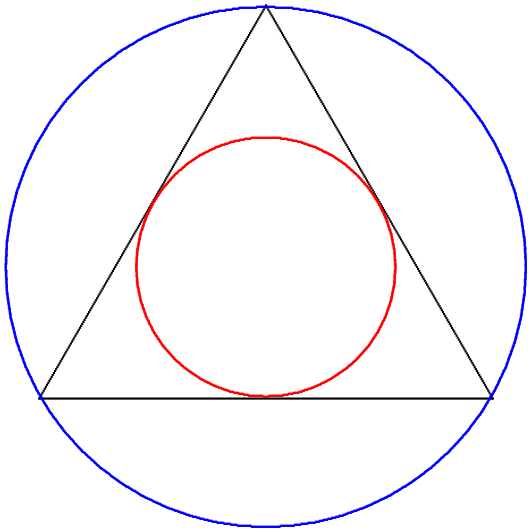


Abbildung 2: Umkeis und Inkreis eines gleichseitigen Dreiecks

sehen, dass im allgemeinen, d. h. nichtsymmetrischen Fall, der Faktor $1/n$ optimal ist. In Abbildung 3 betrachten wir ein (symmetrisches) Quadrat und den zugehörigen Umkreis und Inkreis. Das Verhältnis der beiden Radien ist $\sqrt{2}$. □

Der nächste Satz ist ein Analogon zu Satz 2.11. Ihn findet man z. B. bei O. Güler, F. Gürtuna (2007, Theorem 3.4).

Satz 2.12 *Sei $K \subset \mathbb{R}^n$ ein konvexer Körper und $(A^*, b^*) \in N$ eine Lösung der Optimierungsaufgabe*

$$\text{(E)} \quad \begin{cases} \text{Minimiere} \quad f(A,b) \quad \text{auf} \\ N := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ G((A,b),p) \le 0 \text{ für alle } p \in S\}, \end{cases}$$

wobei

$$f(A,b) := -\log\det(A), \qquad G((A,b),p) := \|Ap\|_2 + p^T b - \sigma_K(p)$$

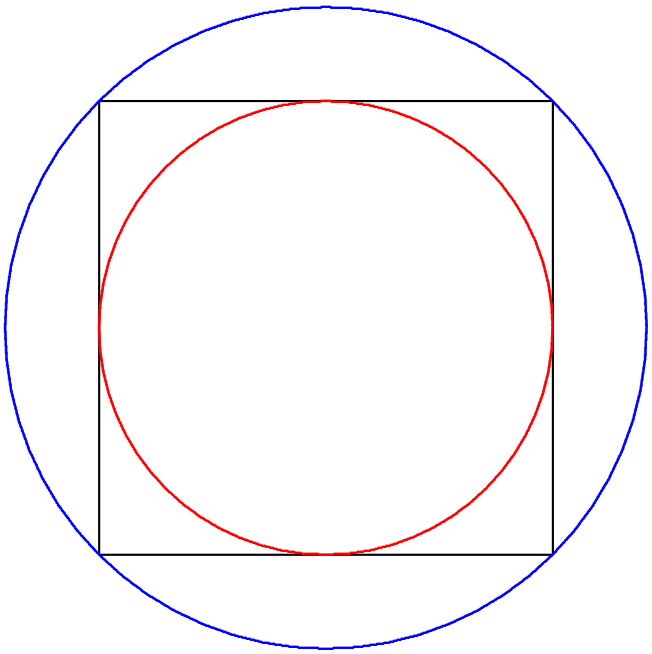


Abbildung 3: Umkeis und Inkreis eines Quadrats

und

$$S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}, \qquad \sigma_K(p) := \max_{x \in K} p^T x,$$

also $E(A^*, b^*) := b^* + A^*(B[0;1])$ *das Ellipsoid größten Volumens, das in* K *enthalten ist. Dann ist*

$$b^* + A^*(B[0;1]) \subset K \subset b^* + nA^*(B[0;1]).$$

Ist K *symmetrisch zum Nullpunkt, also* $K = -K$*, so ist notwendig* $b^* = 0$ *und der Faktor* n *kann durch* $\sqrt{n}$ *ersetzt werden, es ist also*

$$A^*(B[0;1]) \subset K \subset \sqrt{n}A^*(B[0;1]).$$

Beweis: Diesmal folgen wir im wesentlichen dem Beweis bei O. Güler, F. Gürtuna (2007, Theorem 3.4). Wir definieren $K^* := (A^*)^{-1}(K - b^*)$. Dann ist mit K auch K^* ein konvexer Körper im $\mathbb{R}^n$ und $B^* := B[0;1]$ ist das Ellipsoid größten Volumens, das in K^* enthalten ist. Daher ist $(I, 0) \in N^*$ die Lösung der Optimierungsaufgabe

$$(\mathrm{E}^*) \quad \begin{cases} \text{Minimiere} \quad f(A,b) \quad \text{auf} \\ N^* := \{(A,b) \in \mathcal{S}^{n\times n} \times \mathbb{R}^n : A \in \mathcal{S}_+^{n\times n},\ G^*((A,b),p) \le 0 \text{ für alle } p \in S\}, \end{cases}$$

wobei

$$f(A,b) := -\log\det(A), \qquad G^*((A,b),p) := \|Ap\|_2 + p^T b - \sigma_{K^*}(p)$$

und

$$S := \{p \in \mathbb{R}^n : \|p\|_2 = 1\}, \qquad \sigma_{K^*}(p) := \max_{y \in K^*} p^T y.$$

Wegen des ersten Teiles von Satz 2.9, angewandt auf sie Aufgabe (E^*), erhalten wir die Existenz von $s \in \{0, \ldots, \frac{n(n+3)}{2}\}$ Punkten $p_1, \ldots, p_s \in S$ mit $\sigma_{K^*}(p_i) = 1$, $i = 1, \ldots, s$, sowie von positiven Multiplikatoren $\lambda_1, \ldots, \lambda_s$ mit

$$I = \sum_{i=1}^{s} \lambda_i p_i p_i^T, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0.$$

Hieraus folgt

$$n = \operatorname{tr}(I) = \operatorname{tr}\left(\sum_{i=1}^{s} \lambda_i p_i p_i^T\right) = \sum_{i=1}^{s} \lambda_i \underbrace{\|p_i\|_2^2}_{=1} = \sum_{i=1}^{s} \lambda_i.$$

Wir definieren $P := \operatorname{co}(\{p_1, \ldots, p_s\})$ und überlegen uns, dass

$$K^* \subset P^\circ \subset nB[0;1],$$

wobei (siehe den zweiten Teil des Beweises von Satz 2.11) die zu P polare Menge P° durch

$$\begin{aligned} P^\circ &:= \{p \in \mathbb{R}^n : \sigma_P(p) \le 1\} \\ &= \{p \in \mathbb{R}^n : p^T x \le 1 \text{ für alle } x \in P\} \\ &= \{p \in \mathbb{R}^n : p^T p_i \le 1,\ i = 1, \ldots, s\} \end{aligned}$$

gegeben ist. Wegen $\sigma_{K^*}(p_i) = 1$, $i = 1, \ldots, s$, ist $K^* \subset P^\circ$. Ist $p \in P^\circ$, so ist

$$-\|p\|_2 = -\|p\|_2 \, \|p_i\|_2 \le p^T p_i \le 1$$

und daher

$$\begin{aligned} 0 &\le \sum_{i=1}^{s} \underbrace{\lambda_i}_{>0} \underbrace{(1 - p^T p_i)}_{\ge 0} \underbrace{(\|p\|_2 + p^T p_i)}_{\ge 0} \\ &= \underbrace{\left(\sum_{i=1}^{s} \lambda_i\right)}_{=n} \|p\| + p^T \underbrace{\left(\sum_{i=1}^{s} \lambda_i p_i\right)}_{=0} (1 - \|p\|) - \sum_{i=1}^{s} \lambda_i (p^T p_i)^2 \\ &= n\|p\|_2 - \|p\|_2^2, \end{aligned}$$

wobei die letzte Gleichung aus

$$p = \sum_{i=1}^{s} \lambda_i (p_i^T p) p_i, \qquad \|p\|_2^2 = \sum_{i=1}^{s} \lambda_i (p^T p_i)^2$$

folgt. Damit ist auch $P^\circ \subset nB[0;1]$ nachgewiesen. Dann ist also insgesamt

$$K^* = (A^*)^{-1}(K - b^*) \subset P^\circ \subset nB[0;1]$$

und daher, wie behauptet,

$$K \subset b^* + nB[0;1].$$

Damit ist der erste Teil des Satzes bewiesen.

Im zweiten Teil des Beweises nehmen wir an, es sei $K = -K$ und zeigen, dass auch $E(A^*, -b^*) \subset K$ ist, woraus $b^* = 0$ genau wie im entsprechenden Teil des Beweises von Satz 2.11 folgt. Sei also $y \in E(A^*, -b^*)$. Mit einem $x \in B[0;1]$ ist $y = -b^* + A^*x$. Dann ist

$$-y = b^* + A^*(-x) \in E(A^*, b^*) \subset K,$$

also $y \in -K = K$, womit $E(A^*, -b^*) \subset K$ bewiesen ist. Im symmetrischen Fall hat also das K eingeschriebene Ellipsoid maximalen Volumens notwendig den Nullpunkt als Zentrum. Der Rest des Beweises verläuft weitgehend analog zum unsymmetrischen Fall. Wir setzen $K^* := (A^*)^{-1}(K)$ und beachten, dass $B[0;1]$ das Ellipsoid maximalen Volumens ist, welches im konvexen, symmetrischen Körper K^* enthalten ist. Wegen der Symmetrie von K^* ist weiter

$$\sigma_{K^*}(p) := \max_{y \in K^*} p^T y = \max_{y \in K^*} |p^T y|.$$

Wie oben existieren $p_1, \ldots, p_s \in \mathbb{R}^n$ mit $\|p_i\|_2 = 1$ und $\sigma_{K^*}(p_i) = 1$, $i = 1, \ldots, s$, sowie positive Multiplikatoren $\lambda_1, \ldots, s$, mit

$$I = \sum_{i=1}^{s} \lambda_i p_i p_i^T, \qquad \sum_{i=1}^{s} \lambda_i p_i = 0.$$

Diesmal definieren wir

$$P := \text{co}\,(\{\pm p_1, \ldots, \pm p_s\}).$$

Wegen $\sigma_{K^*}(\pm p_i) = 1$ und

$$P^\circ = \{p \in \mathbb{R}^n : |p^T p_i| \leq 1,\ i = 1, \ldots, s\}$$

ist $K^* \subset P^\circ$. Für $p \in P^\circ$ ist

$$\|p\|_2^2 = \sum_{i=1}^{s} \lambda_i \underbrace{(p^T p_i)^2}_{\leq 1} \leq \sum_{i=1}^{n} \lambda_i = n,$$

also $\|p\|_2 \leq \sqrt{n}$. Damit ist

$$K^* = (A^*)^{-1}(K) \subset \sqrt{n} B[0;1]$$

bzw.

$$K \subset \sqrt{n} A^*(B[0;1]).$$

Der Satz ist vollständig bewiesen. □

Die Sätze von John-Löwner sagen also aus: Schrumpft man das Ellipsoid minimalen Volumens, das einen gegebenen konvexen Körper K umschreibt, um den Faktor $1/n$, so erhält man ein in K einbeschriebenes Ellipsoid. Bläst man umgekehrt ein dem Körper K einbeschriebenes Ellipsoid maximalen Volumens um den Faktor n auf, so erhält man ein K umschreibendes Ellipsoid. Für symmetrisches K können diese Faktoren zu $1/\sqrt{n}$ bzw. $\sqrt{n}$ verbessert werden.

2.8 Die Steiner-Inellipse und der Satz von Siebeck-Marden

Wir definieren:

Definition 2.13 Gegeben sei ein Dreieck $K := \triangle z_1 z_2 z_3$ mit den Ecken z_1, z_2, z_3 in der Ebene $\mathbb{R}^2$ bzw. der komplexen Zahlenebene $\mathbb{C}$. Eine K einbeschriebene Ellipse[8] heißt *Steiner-Inellipse zu* K, wenn sie die Seiten des Dreiecks in den Mittelpunkten berührt.

Zunächst zeigen wir:

Satz 2.14 *Zu jedem Dreieck* $K := \triangle z_1 z_2 z_3$ *existiert genau eine Steiner-Inellipse.*

Beweis: Wir können o. B. d. A. annehmen, dass die Punkte z_1 und z_2 auf der x-Achse liegen, also durch $z_1 = (x_1, 0)$, $z_2 = (x_2, 0)$ mit $x_1 < x_2$ gegeben sind. Dies ist notfalls durch eine Verschiebung und eine Drehung erreichbar. Der dritte Eckpunkt sei $z_3 = (x_3, y_3)$ mit $y_3 > 0$. Durch eine lineare, nichtsinguläre Abbildung T, welche z_1 und z_2 fest lässt, transformieren wir das Dreieck $\triangle z_1 z_2 z_3$ in das gleichseitiges Dreieck $\triangle z_1 z_2 z_4$, wobei $z_4 = (0, \frac{\sqrt{3}}{2}(x_2 - x_1))$. Die Abbildung bzw. Matrix T ist durch

$$T := \begin{pmatrix} 1 & (\frac{1}{2}(x_1 + x_2) - x_3)/y_3 \\ 0 & \frac{\sqrt{3}}{2}(x_2 - x_1)/y_3 \end{pmatrix}$$

gegeben. Das gleichseitige Dreieck $\triangle z_1 z_2 z_4$ besitzt offenbar genau eine Steiner-Inellipse, nämlich einen Kreis. Dies veranschaulichen wir in Abbildung 4. Durch die Abbildung

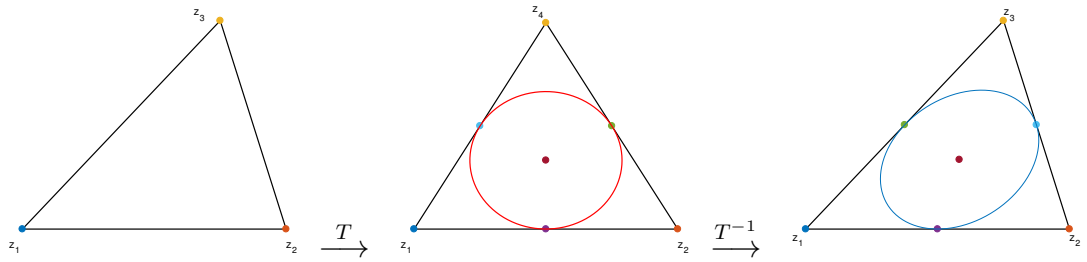


Abbildung 4: Existenz und Eindeutigkeit einer Steiner-Inellipse

T^{-1} wird das Dreieck $\triangle z_1 z_2 z_4$ mit der zugehörigen Steiner-Inellipse (welche ein Inkreis ist) wieder abgebildet auf das Dreieieck $\triangle z_1 z_2 z_3$, wobei der Inkreis in diesem Dreieck in eine Steiner-Inellipse zu $\triangle z_1 z_2 z_3$ übergeht. Seitenmittelpunkte gehen dabei in Seitenmittelpunkte über. Hierdurch ist die Existenz einer Steiner-Inellipse bewiesen. Die Eindeutigkeit ergibt sich durch die Eindeutigkeit im gleichseitigen Fall. □

Einen sozusagen animierten Beweis für die gerade formulierte Existenz- und Eindeutigkeitsaussage findet man hier.

[8] Unter einer Ellipse verstehen wir das zweidimensionale Analogon eines Ellipsoids (und nicht nur dessen Rand).

Im folgenden Satz wird ausgesagt, dass die Steiner-Inellipse zu einem Dreieck sozusagen die John-Ellipse ist.

Satz 2.15 *Die Steiner-Inellipse zu einem Dreieck besitzt unter allen Ellipsen, die diesem Dreieck einbeschrieben sind, maximalen Flächeninhalt.*

Beweis: Einen elementaren Beweis dieser Aussage findet man bei D. Minda, S. Phelps (2008, Corollary 4.2). Wir präsentieren einen Beweis mit Hilfe von Korollar 2.10. Es genügt offenbar zu zeigen (siehe den Existenz- und Eindeutigkeitsbeweis zu Satz 2.14), dass der Steiner-Inkreis zu einem *gleichseitigen* Dreieck unter allen dem Dreieck einbeschriebenen Inellipsen den größten Flächeninhalt besitzt. In Abbildung 5 geben wir ein gleichseitiges Dreieck an, dessen Inkreis $B[0;1]$, also der Kreis um den Nullpunkt mit dem Radius 1, ist. Wir wenden Korollar 2.10 an und zeigen, dass

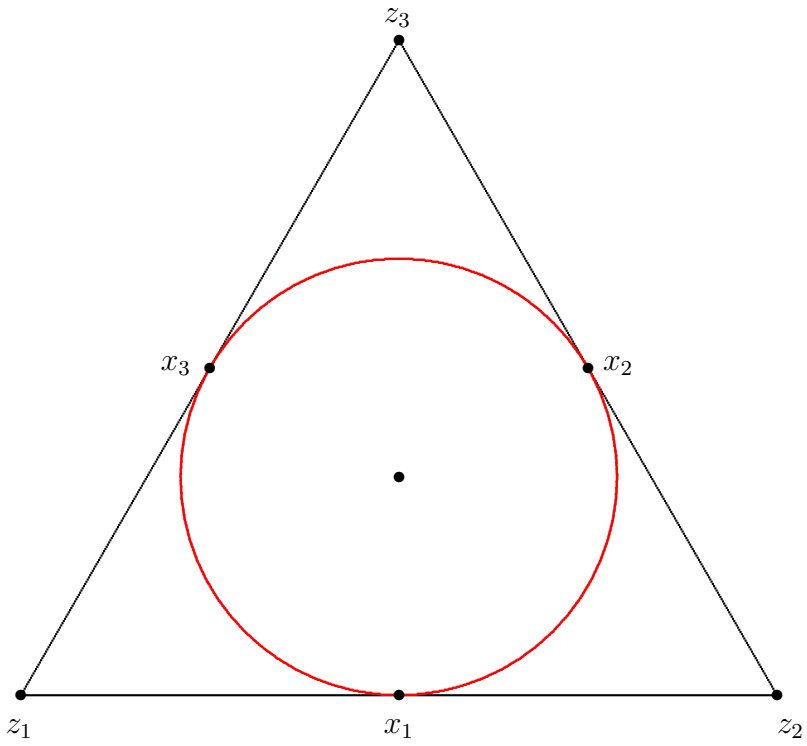


Abbildung 5: Inellipse maximalen Flächeninhalts

$(A^*, b^*) := (I, 0)$ Lösung der Optimierungsaufgabe (E) ist, in dem Dreieck $\triangle z_1 z_2 z_3$ eine einbeschriebene Ellipse maximalen Flächeninhalts zu bestimmen. Wir gehen von einem gleichseitigen Dreieck $\triangle z_1 z_2 z_3$ mit dem Nullpunkt als Schwerpunkt und dem Einheitskreis als Inkreis aus, es sei also etwa

$$z_1 := (-\sqrt{3}, -1), \quad z_2 := (\sqrt{3}, -1), \quad z_3 := (0, 2).$$

Kontaktpunkte von Dreieck und Inkreis sind naheliegenderweise

$$x_1 := (0, -1), \quad x_2 := (\tfrac{1}{2}\sqrt{3}, \tfrac{1}{2}), \quad x_3 := (-\tfrac{1}{2}\sqrt{3}, \tfrac{1}{2}).$$

Für die Multiplikatoren μ_1, μ_2, μ_3 in Korollar 2.10 machen wir den Ansatz $\mu_1 = \mu_2 = \mu_3 = \mu$ mit noch unbekanntem $\mu > 0$. Dann ist

$$\sum_{i=1}^{3} \mu_i x_i x_i^T = \mu\left[\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \frac{3}{4} & \frac{1}{4}\sqrt{3} \\ \frac{1}{4}\sqrt{3} & \frac{1}{4} \end{pmatrix} + \begin{pmatrix} \frac{3}{4} & -\frac{1}{4}\sqrt{3} \\ -\frac{1}{4}\sqrt{3} & \frac{1}{4} \end{pmatrix}\right] = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

wenn man $\mu := \frac{2}{3}$ wählt. Weiter ist

$$\sum_{i=1}^{3} \mu_i x_i = 0,$$

die hinreichenden Optimalitätsbedingungen in Korollar 2.10 sind erfüllt und der Satz ist bewiesen. □

Beim Beweis von Satz 2.14 haben wir ausgenutzt, dass man ein beliebiges Dreieck in der Ebene $\mathbb{R}^2$, eventuell nach einer Verschiebung und Drehung, durch eine lineare, nichtsinguläre Abbildung $T\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ in ein gleichseitiges Dreieck überführen kann. Hierbei werden natürlich Ecken in Ecken, Seiten in Seiten und Seitenmittelpunkte in Seitenmittelpunkte transformiert. Naheliegend ist es, den $\mathbb{R}^2$ mit der Menge $\mathbb{C}$ der komplexen Zahlen zu identifizieren, also einen Punkt $(x, y)^T \in \mathbb{R}^2$ mit der komplexen Zahl $z = x + iy \in \mathbb{C}$. Eine Abbildung $f\colon \mathbb{C} \longrightarrow \mathbb{C}$, definiert durch $f(z) := Az + B\bar{z}$ mit $A, B \in \mathbb{C}$ ist offenbar *reell linear*, d. h. es ist $f(z_1 + z_2) = f(z_1) + f(z_2)$ für alle $z_1, z_2 \in \mathbb{C}$ und $f(\lambda z) = \lambda f(z)$ für alle $\lambda \in \mathbb{R}$ und alle $z \in \mathbb{C}$. Andererseits hat *jede* reell lineare Abbildung $f\colon \mathbb{C} \longrightarrow \mathbb{C}$ diese Form, wie man sich sehr leicht überlegt. Mit $A = a_1 + ia_2$, $B = b_1 + ib_2$ ist die zu f gehörende lineare Abbildung $T\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ durch

$$T = \begin{pmatrix} a_1 + b_1 & -a_2 + b_2 \\ a_2 + b_2 & a_1 - b_1 \end{pmatrix}$$

gegeben. Die Matrix T ist nichtsingulär bzw. die Abbildung f bijektiv genau dann, wenn

$$0 \neq \det T = a_1^2 - b_1^2 - (b_2^2 - a_2^2) = (a_1^2 + a_2^2) - (b_1^2 + b_2^2) = |A|^2 - |B|^2$$

bzw. $|A| \neq |B|$. Eine affin reell lineare Abbildung hat die Form $f(z) = Az + B\bar{z} + C$ mit $A, B, C \in \mathbb{C}$, ist also eine (reell) lineare Abbildung mit anschließender Translation. Eine bijektive affin (reell) lineare Abbildung f transformiert den Einheitskreis

$$B[0; 1] := \{z \in \mathbb{C} : |z| \leq 1\}$$

in der komplexen Zahlenebene in eine Ellipse mit dem Zentrum C. Hierbei wird der Rand des Einheitskreises in den Rand der Ellipse überführt. Als Vorbereitung zum Beweis des schönen Satzes von Siebeck-Marden (von D.Kalman (2008) *"the most marvelous theorem in mathematics"* genannt[9]) beweisen wir die folgende Aussage (siehe D. Minda, S. Phelps (2008, Theorem 3.1)).

[9] Allerdings erscheint der Satz von Siebeck-Marden *nicht* in einer Liste der Top 100 mathematischen Sätze, die man hier finden kann.

Satz 2.16 *Sei $f(z) = Az + B\bar{z} + C$ eine bijektive affin lineare Abbildung. Dann ist das Bild des Einheitkreises $B[0;1]$ eine Ellipse E mit dem Zentrum C und den beiden Brennpunkten $f_{1,2} := C \pm 2\sqrt{AB}$. Die große Halbachse hat die Länge $|A| + |B|$, die kleine Halbachse die Länge $||A| - |B||$.*

Beweis: Die große und die kleine Halbachse a bzw. b werden bestimmt aus

$$a := \max_{t\in[0,2\pi]} |f(e^{it}) - C|, \qquad b := \min_{t\in[0,2\pi]} |f(e^{it}) - C|.$$

Mit

$$A = |A|e^{i\theta}, \qquad B = |B|e^{i\phi}$$

ist offenbar

$$||A| - |B|| \le |f(e^{it}) - C| = ||A|e^{i(\theta+t)} + |B|e^{i(\phi-t)}| \le |A| + |B|.$$

Wir zeigen, dass die obere und die untere Schranke angenommen wird. Nun ist

$$|f(e^{it}) - C| = \begin{cases} |A| + |B|, & \text{für } t = (\phi-\theta)/2,\ t = (\phi-\theta)/2 + \pi, \\ ||A| - |B||, & \text{für } t = (\phi-\theta)/2 + \pi/2,\ t = (\phi-\theta)/2 + 3\pi/2. \end{cases}$$

Daher ist $a := |A|+|B|$ die Länge der großen und $b := ||A|-|B||$ die Länge der kleinen Halbachse. Wegen

$$e^{i\theta/2} = \frac{1}{\sqrt{|A|}}\sqrt{A}, \qquad e^{i\phi/2} = \frac{1}{\sqrt{|B|}}\sqrt{B}$$

ist

$$f(e^{i(\phi-\theta)/2}) = (|A| + |B|)e^{i(\phi+\theta)/2} + C = C + \frac{1}{\sqrt{|AB|}}\sqrt{AB}.$$

In Abbildung 6 haben wir eine Ellipse samt ihrer Halbachsen und Brennpunkte dargestellt. Die Punkte $f_{1,2} := C \pm 2\sqrt{AB}$ sind Brennpunke der Ellipse E, wenn

$$g(t) := |f(e^{it}) - (C + 2\sqrt{AB})| + |f(e^{it} - (C - 2\sqrt{AB})| = 2(|A| + |B|)$$

für alle t ist, wenn also für jeden Punkt des Randes von E die Summe der Abstände zu den beiden Brennpunkten gleich der doppelten Länge der großen Halbachse ist. Nun ist

$$\begin{aligned}
g(t) &= ||A|e^{i(\theta+t)} + |B|e^{i(\phi-t)} - 2\sqrt{|A||B|}e^{i(\theta+\phi)/2}| \\
&\qquad + ||A|e^{i(\theta+t)} + |B|e^{i(\phi-t)} + 2\sqrt{|A||B|}e^{i(\theta+\phi)/2}| \\
&= ||A|e^{is} + |B|e^{-is} - 2\sqrt{|A||B|}| + ||A|e^{is} + |B|e^{-is} + 2\sqrt{|A||B|}| \\
&\qquad \text{(Substitution } t = (\phi-\theta)/2 + s) \\
&= |(\sqrt{|A|}e^{is/2} - \sqrt{|B|}e^{-is/2})^2| + |(\sqrt{|A|}e^{is/2} + \sqrt{|B|}e^{-is/2})^2| \\
&= |\sqrt{|A|}e^{is/2} - \sqrt{|B|}e^{-is/2}|^2 + |\sqrt{|A|}e^{is/2} + \sqrt{|B|}e^{-is/2}|^2 \\
&= (\sqrt{|A|} - \sqrt{|B|})^2\cos^2(s/2) + (\sqrt{|A|} + \sqrt{|B|})^2\sin^2(s/2) \\
&\qquad (\sqrt{|A|} + \sqrt{|B|})^2\cos^2(s/2) + (\sqrt{|A|} - \sqrt{|B|})^2\sin^2(s/2) \\
&= 2(|A| + |B|).
\end{aligned}$$

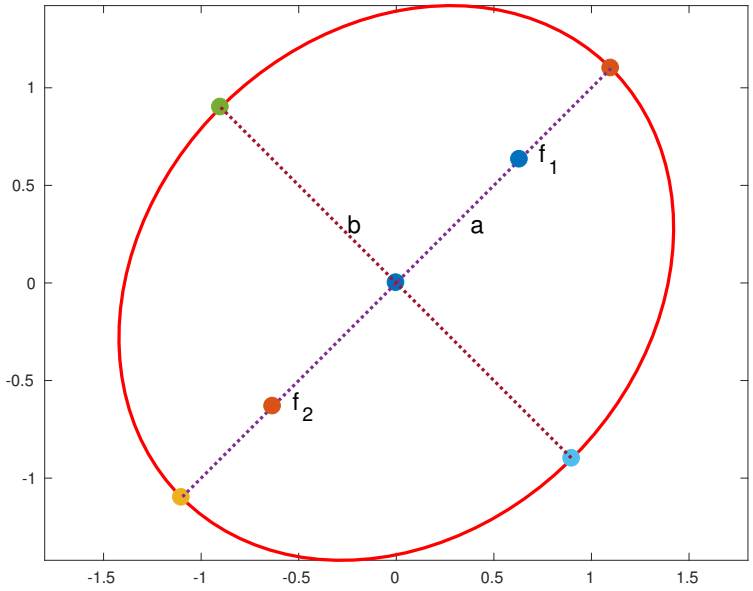


Abbildung 6: Die durch $A = 1 + i$, $B = 0.1(1 + i)$ und $C = 0$ erzeugte Ellipse

Damit ist der Satz bewiesen. □

Der folgende Satz ist eine Präzisierung von Satz 2.14, siehe D. Minda, S. Phelps (2008, Theorem 2.1).

Satz 2.17 *Zu jedem Dreieck $\triangle z_1 z_2 z_3$ mit (nichtkollinearen) $z_1, z_2, z_3 \in \mathbb{C}$ gibt es genau eine Steiner-Inellipse. Deren Brennpunkte sind durch*

$$f_{1,2} := g \pm \sqrt{g^2 - \frac{1}{3}(z_1 z_2 + z_2 z_3 + z_3 z_1)}$$

gegeben, wobei

$$g := \frac{1}{3}(z_1 + z_2 + z_3)$$

das Zentrum von $\triangle z_1 z_2 z_3$ (und der Inellipse) ist.

Beweis: Die Existenz und Eindeutigkeit der Steiner-Inellipse zu dem Dreieck $\triangle z_1 z_2 z_3$ haben wir schon in Satz 2.14 nachgewiesen.

Sei $\omega := e^{2\pi i/3}$ und $\Delta := \triangle \omega^0 \omega^1 \omega^2$ das hiervon erzeugte gleichseitige Dreieck. Dieses besitzt $B[0; \frac{1}{2}] = \frac{1}{2} B[0; 1]$ als Inkreis bzw. als Steiner-Inellipse, siehe Abbildung 7. Wir bestimmen eine affin lineare Abbildung $f(z) = Az + B\bar{z} + C$ mit $f(\omega^0) = z_1$, $f(\omega^1) = z_2$ und $f(\omega^2) = z_3$. Ist uns dies gelungen, so haben wir mit

$$E := f(B[0; \tfrac{1}{2}]) = \frac{1}{2} f(B[0; 1])$$

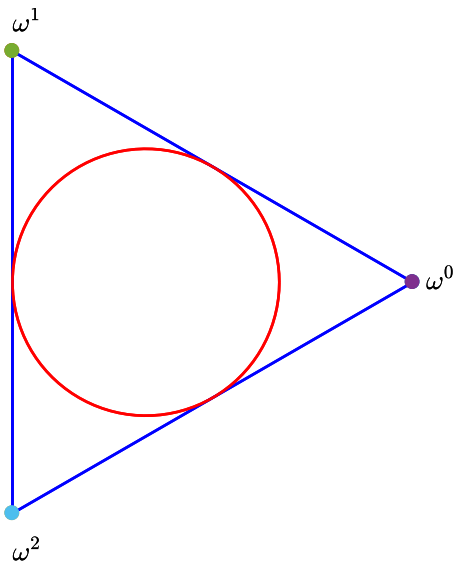


Abbildung 7: Gleichseitiges Dreieck mit Inkreis

die Steiner-Inellipse für $\triangle z_1 z_2 z_3$ gefunden. Nach Satz 2.16 sind deren Brennpunkte durch $f_{1,2} := C \pm \sqrt{AB}$ gegeben. Es kommt jetzt also darauf an, A, B und C zu berechnen. Die Bedingungen

$$f(\omega^0) = z_1, \qquad f(\omega^1) = z_2, \qquad f(\omega^2) = z_3$$

liefern die Gleichungen

$$\begin{pmatrix} 1 & 1 & 1 \\ \omega & \omega^{-1} & 1 \\ \omega^2 & \omega^{-2} & 1 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}.$$

Wegen $1 + \omega + \omega^2 = 0$ erhalten wir hieraus sehr leicht

$$\begin{aligned} A &= \frac{1}{3}(z_1 + \omega^2 z_2 + \omega z_3), \\ B &= \frac{1}{3}(z_1 + \omega z_2 + \omega^2 z_3), \\ C &= \frac{1}{3}(z_1 + z_2 + z_3). \end{aligned}$$

Mit

$$g := C = \frac{1}{3}(z_1 + z_2 + z_3)$$

ist

$$\begin{aligned} AB &= \frac{1}{9}[z_1^2+z_2^2+z_3^2+(\omega+\omega^2)(z_1z_2+z_2z_3+z_3z_1)] \\ &= \frac{1}{9}[z_1^2+z_2^2+z_3^2-(z_1z_2+z_2z_3+z_3z_1)] \\ &= \frac{1}{9}[(z_1+z_2+z_3)^2-3(z_1z_2+z_2z_3+z_3z_1)] \\ &= g^2-\frac{1}{3}(z_1z_2+z_2z_3+z_3z_1). \end{aligned}$$

Daher sind die Brennpunkte der Steiner-Inellipse $E = \frac{1}{2}f(B[0;1])$ zu $\triangle z_1z_2z_3$ durch

$$f_{1,2} := g \pm \sqrt{g^2-\frac{1}{3}(z_1z_2+z_2z_3+z_3z_1)}$$

gegeben. Der Satz ist bewiesen. □

Beispiel: Sei

$$z_1 := 3+14i, \qquad z_2 := 8.5-1.5i, \qquad z_3 := -6-2i,$$

(siehe K. Rohe (2015)). In Abbildung 8 geben wir das Dreieck $\triangle z_1z_2z_3$ und die zu-

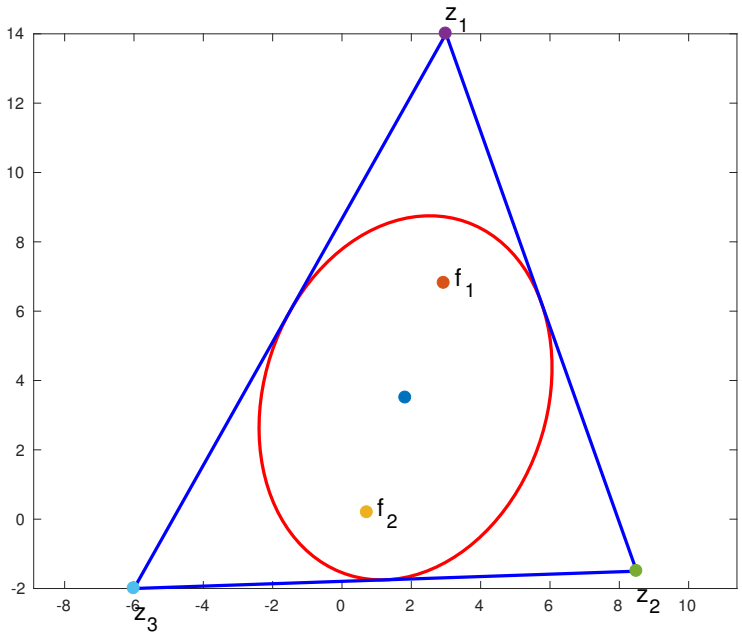


Abbildung 8: Das Dreieck $\triangle z_1z_2z_3$ und die zugehörige Steiner-Inellipse

gehörige Steiner-Inellipse an. Als Ergebnis obiger Formeln zur Berechnung der affin linaren Abbildung $f(z) = Az + B\bar{z} + C$ sowie der Brennpunkte $f_{1,2}$ der zu $\triangle z_1z_2z_3$ gehörenden Inellipse erhalten wir

$$A = 0.7277+1.0642i, \qquad B = 0.4390+9.4358i, \qquad C = 1.8333+3.5000i$$

sowie

$$f_1 = 2.9414 + 6.8091i, \qquad f_2 = 0.7253 + 0.1909i.$$

□

Als leichte Folgerung aus Satz 2.17 erhalten wir den Satz von Siebeck-Marden, siehe J. Siebeck (1864), M. Marden (1945), D. Kalman (2008), B. Bogosel (2017).

Satz 2.18 (Siebeck-Marden) *Seien z_1, z_2, z_3 nichtkollineare Punkte in $\mathbb{C}$ und*

$$p(z) := (z - z_1)(z - z_2)(z - z_3)$$

das hiervon erzeugte kubische Polynom. Dann sind die Nullstellen von p' die Brennpunkte der Steiner-Inellipse zu $\triangle z_1 z_2 z_3$.

Beweis: Es ist

$$\begin{aligned} p(z) &= (z - z_1)(z - z_2)(z - z_3) \\ &= z^3 - (z_1 + z_2 + z_3)z^2 + (z_1 z_2 + z_2 z_3 + z_3 z_1)z - z_1 z_2 z_3 \end{aligned}$$

und daher

$$p'(z) = 3z^2 - 2(z_1 + z_2 + z_3)z + (z_1 z_2 + z_2 z_3 + z_3 z_1).$$

Mit

$$g := \frac{1}{3}(z_1 + z_2 + z_3)$$

sind die Nullstellen von p' gegeben durch

$$f_{1,2} = g \pm \sqrt{g^2 + \frac{1}{3}(z_1 z_2 + z_2 z_3 + z_3 z_1)}$$

und dies sind nach Satz 2.17 genau die Brennpunkte der Steiner-Inellipse zu $\triangle z_1 z_2 z_3$. Der Satz ist bewiesen. □

Bemerkung: Der Satz von Siebeck-Marden macht in Abhängigkeit der Nullstellen z_1, z_2, z_3 eines kubischen Polynoms p eine Aussage über die Nullstellen der Ableitung p'. Eine der bekanntesten Aussagen dieser Art für komplexe Polynome vom Grade n ist der Satz von Gauss-Lucas. Dieser sagt aus (siehe z. B. L. V. Ahlfors (1966, S. 29)):

- *Ist p ein (nichtkonstantes) Polynom mit komplexen Koeffizienten, so gehören alle Nullstellen von p' zur konvexen Hülle der Nullstellen von p.*

Denn: Wir nehmen an, p habe den Grad n und besitze die (nicht notwendig paarweise verschiedenen) Nullstellen $z_1, \ldots, z_n$. Mit $a \in \mathbb{C} \setminus \{0\}$ kann p in der Form

$$p(z) = a \prod_{i=1}^{n} (z - z_i)$$

geschrieben werden. Sei $z \in \mathbb{C}$ mit $p(z) \neq 0$. Für die logarithmische Ableitung erhalten wir

$$\frac{p'(z)}{p(z)} = \sum_{i=1}^{n} \frac{1}{z - z_i}.$$

Ist speziell z eine Nullstelle von p' und $p(z) \neq 0$, so ist

$$\sum_{i=1}^{n} \frac{1}{z - z_i} = 0$$

und damit

$$\sum_{i=1}^{n} \frac{\bar{z} - \bar{z}_i}{|z - z_i|^2} = 0.$$

Mit

$$\Lambda := \sum_{i=1}^{n} \frac{1}{|z - z_i|^2}$$

ist daher

$$\bar{z} = \frac{1}{\Lambda} \sum_{i=1}^{n} \frac{\bar{z}_i}{|z - z_i|^2}$$

bzw.

$$z = \frac{1}{\Lambda} \sum_{i=1}^{n} \frac{z_i}{|z - z_i|^2} \in \text{co}\,(\{z_1, \ldots, z_n\}).$$

Ist dagegen $p(z) = p'(z) = 0$, so ist $z = z_i$ für ein $i \in \{1, \ldots, n\}$ und damit trivalerweise $z \in \text{co}\,(\{z_1, \ldots, z_n\})$. □

3 Der Satz von van der Waerden

3.1 Formulierung des Satzes, einfache Beispiele

Der Satz von van der Waerden (1926) sagt aus[10]:

Satz 3.1 (van der Waerden) *Seien r und k (beliebige) natürliche Zahlen. Dann existiert $W(r,k) \in \mathbb{N}$ mit folgender Eigenschaft: Für alle $N \geq W(r,k)$ und jede Zerlegung*

$$\{1, \ldots, N\} = C_1 \,\dot\cup \cdots \dot\cup\, C_r$$

enthält mindestens eine der Mengen $C_1, \ldots, C_r$ eine arithmetische Progression der Länge k. Oder anders gesagt: Gibt man jeder Zahl aus $\{1, \ldots, N\}$ (genau) eine von r Farben, so existiert eine gleich gefärbte bzw. monochrome arithmetische Progression der Länge k in $\{1, \ldots, N\}$.

Hierbei heißt eine endliche Menge A von k ganzen Zahlen eine *arithmetische Progression der Länge k*, wenn sie sich in der Form

$$A = \{a, a + d, a + 2d, \ldots, a + (k-1)d\}$$

[10]Bei K. Jacobs (1983, S. 102) findet man die folgende Aussage: Der (obige) Satz des jungen van der Waerden stellt eine der bedeutendsten mathematischen Leistungen des 20. Jahrhunderts dar.

mit $a \in \mathbb{Z}$ und $d \in \mathbb{N}$ darstellen lässt, wobei a das *Anfangsglied* und d die *Schrittweite* der arithmetischen Progression genannt wird. Die minimale Zahl $W(r,k)$ mit obiger Eigenschaft wird ebenfalls mit $W(r,k)$ bezeichnet und heißt die *van der Waerden Zahl*.

B. L. VAN DER WAERDEN (1965) beschreibt, dass Ausgangspunkt für seinen Satz eine Vermutung von Baudet war:

- *Teilt man die Gesamtheit der natürlichen Zahlen* $1, 2, \ldots$ *in zwei Klassen ein, so enthält mindestens eine dieser Klassen eine arithmetische Progression von* k *Gliedern, wobei* k *eine beliebig große vorgegebene Zahl ist.*

Eine spannende Darstellung der Geschichte des Satzes von van der Waerden und der Baudetschen Vermutung findet man bei A. SOIFER (2009, S. 309 ff.).

Beispiel: Trivialerweise ist $W(1,k) = k$ für alle $k \in \mathbb{N}$. Ferner ist $W(r,2) = r+1$ für alle $r \in \mathbb{N}$. Denn färbt man $\{1, \ldots, r+1\}$ mit r Farben, so existiert wegen des Schubfachprinzips[11] ein monochromes Paar. □

Beispiel: Sei $r = 2$ und $k = 3$. Dann ist $N = W(2,3)$ die kleinste natürliche Zahl mit der Eigenschaft, dass es unter den beliebig rot oder blau gefärbten Zahlen $\{1, \ldots, N\}$ mindestens eine rote oder eine blaue arithmetische Progression der Länge drei gibt. Es ist $W(2,3) > 8$. Denn färbt man die Zahlen $\{1, \ldots, 8\}$ in den Farben rot und blau z. B. folgendermaßen:

1	2	3	4	5	6	7	8
B	R	R	B	B	R	R	B

so gibt es keine monochrome arithmetische Progresssion der Länge 3. Fügt man aber die Zahl 9 hinzu, so kann man diese rot oder blau färben:

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	R

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	B

Im ersten Fall hat man die rote arithmetische Progression 3 6 9, im zweiten Fall erhält man die blaue arithmetische Progression 1 5 9, jeweils der Länge 3. Wir wollen zeigen, dass $W(2,3) = 9$. Zum Beweis gehen wir systematisch vor und bestimmen *alle* Färbungen der Zahlen $\{1, \ldots, 8\}$ durch die zwei Farben rot und blau, bei denen es *keine* monochrome arithmetische Progression der Länge 3 gibt. Im Anschluss zeigen wir bei diesen Färbungen der Zahlen $\{1, \ldots, 8\}$, dass die Zahl 9 blau oder rot gefärbt werden kann und stets bei den so gefärbten Zahlen $\{1, \ldots, 9\}$ eine monochrome Progression der Länge 3 existiert. O. B. d. A. können wir annehmen, dass 1 blau gefärbt ist.

1 Es ist 2 blau. Dann ist notwendig 3 rot, weil wir andernfalls die blaue Progression 1 2 3 haben. Wir haben also

1	2	3
B	B	R

[11] Wenn $r+1$ Dinge auf r Schubfächer verteilt werden, so muss eines der Fächer mindestens zwei Dinge enthalten.

1.1 Wir nehmen an, es sei 4 blau. Dann müssen 6 und 7 rot sein, weil wir andernfalls die monochromen Progressionen 2 4 6 bzw. 1 4 7 erhalten würden. Dann muss 5 blau sein, weil wir andernfalls die rote Progression 5 6 7 haben:

1	2	3	4	5	6	7
B	B	R	B	B	R	R

Dann kann 8 weder blau noch rot sein. Im ersten Fall hätte man die monochrome Progression 2 5 8, im zweiten Fall die Progression 6 7 8.

Folgerung: Jede Färbung von $\{1, \ldots, 8\}$ durch die Farben blau und rot, die mit 1 2 3 4 beginnt, besitzt eine monochrome arithmetische Progression der Länge 3.

1.2 Nun sei 4 rot. Dann muss 5 blau sein, weil wir andernfalls die monochrome Progression 3 4 5 hätten. Wir haben also:

1	2	3	4	5
B	B	R	R	B

1.2.1 Wir nehmen an, es sei 6 blau. Dann muss 7 rot sein, da wir andernfalls die blaue Progression 5 6 7 haben. 8 muss rot sein, weil wir andernfalls die blaue Progression 2 5 8 hätten. Mit

1	2	3	4	5	6	7	8
B	B	R	R	B	B	R	R

haben wir eine blau-rot-Färbung von $\{1, \ldots, 8\}$ erhalten, zu der es keine monochrome Progression der Länge 3 gibt.

1.2.2 Nun sei 6 rot. Dann muss 8 blau sein, denn andernfalls hätte man die monochrome Progression 4 6 8. Aber dann hat man die blaue Progression 2 5 8.

Folgerung: Jede Färbung von $\{1, \ldots, 8\}$ durch die Farben blau und rot, die mit 1 2 3 4 beginnt, besitzt eine monochrome Progression der Länge 3 bis auf die einzige Ausnahme

1	2	3	4	5	6	7	8
B	B	R	R	B	B	R	R

2 Es ist 2 rot.

2.1 Es ist 3 blau. Dann ist 5 rot und 8 blau, weil man andernfalls die Progressionen 1 3 5 bzw. 3 5 8 hätte. Wir haben also

1	2	3	4	5	6	7	8
B	R	B		R			B

2.1.1 Wir nehmen an, 4 sei blau. Dann ist 7 rot (andernfalls 1 4 7) und anschließend 6 blau (andernfalls 5 6 7). Man hätte also

1	2	3	4	5	6	7	8
B	R	B	B	R	B	R	B

Hier hat man jetzt aber die blaue Progression 4 6 8.

2.1.2 Jetzt ist 4 rot. Dann ist 5 rot (sonst 1 3 5), 6 blau (andernfalls 4 5 6), 8 blau (sonst 2 5 8) und dann schließlich 7 rot (sonst 6 7 8). Man erhält also

1	2	3	4	5	6	7	8
B	R	B	R	R	B	R	B

Diese Färbung der Zahlen $\{1,\dots,8\}$ besitzt offenbar keine monochrome Progression der Länge 3.

Folgerung: Jede Färbung von $\{1,\dots,8\}$ durch die Farben blau und rot, die mit 1 2 3 beginnt, besitzt eine monochrome Progression der Länge 3 bis auf die einzige Ausnahme

1	2	3	4	5	6	7	8
B	R	B	R	R	B	R	B

2.2 Es ist 3 rot. Dann ist notwendig 4 blau und 7 rot. Anschließend erkennt man, dass 5 blau ist (andernfalls 3 5 7) 6 rot und schließlich 8 blau. Man hat also

1	2	3	4	5	6	7	8
B	R	R	B	B	R	R	B

Diese Färbung der Zahlen $\{1,\dots,8\}$ besitzt offenbar keine monochrome Progression der Länge 3.

Wenn man also annimmt, dass 1 blau gefärbt ist (andernfalls kann man sich die Farben als getauscht vorstellen), so erhält man insgesamt drei Färbungen von $\{1,\dots,8\}$ durch die beiden Faben rot und blau, die keine monochrome Progression der Länge 3 besitzen, nämlich

1	2	3	4	5	6	7	8
B	B	R	R	B	B	R	R

1	2	3	4	5	6	7	8
B	R	B	R	R	B	R	B

und

1	2	3	4	5	6	7	8
B	R	R	B	B	R	R	B

Färbt man nun bei diesen drei Färbungen die zusätzliche Zahl 9 blau oder rot, stets erhält man eine monochrome Progression der Länge 3 in $\{1,\dots,9\}$. Damit haben wir schließlich nachgewiesen, dass $W(2,3) = 9$. □

Es ist einleuchtend, dass ein Vorgehen wie im letzten Beispiel für größere r, k, um es vorsichtig zu sagen, nicht sehr erfolgversprechend ist.

Beispiel: Mit einer ausbaufähigen Methode[12] wollen wir zeigen, dass $W(2,3) \leq 325$, dass also für *jede* Färbung

$$c\colon \{1,\dots,325\} \longrightarrow \{\text{blau},\text{rot}\}$$

der Zahlen $\{1,\dots,325\}$ durch die Farben blau und rot eine monochrome arithmetische Progression der Länge 3 existiert. Das Ergebnis ist natürlich nicht so gut wie die optimale Aussage $W(2,3) = 9$, es ist aber eine Existenzaussage wie im Satz von van der Waerden und verwendet keine so einfache Fallunterscheidung wie im letzten Beispiel.

[12] Wir folgen ziemlich genau dem Wikipedia-Artikel *Van der Waerden's theorem.*

Wir schreiben $\{1,\dots,325\}$ als Vereinigung von 65 Blöcken der Länge 5:

$$\{1,\dots,325\} = \{1,\dots,5\} \cup \{6,\dots,10\} \cup \dots \cup \{321,\dots,325\} = B_0 \cup B_2 \cup \dots \cup B_{64},$$

wobei

$$B_i := \{5i+1,\dots,5i+5\}, \qquad i = 0,\dots,64.$$

Es gibt $2^5 = 32$ Möglichkeiten einen Block B_i der Länge 5 durch zwei Farben zu färben. Nach dem Schubfachprinzip gibt es unter den 33 ersten Blöcken (mindestens) zwei, etwa B_{i_1} und B_{i_2} mit $i_1 < i_2$, die identisch gefärbt sind. Es gibt also $i_1, i_2 \in \{0, 32\}$ mit

$$c(5i_1 + j) = c(5i_2 + j), \qquad j = 1,\dots,5.$$

Unter den drei natürlichen Zahlen $5i_1+1$, $5i_1+2$ und $5i_1+3$ muss es mindestens zwei mit derselben Farbe geben (etwa wieder wegen des Schubfachprinzips). Dies seien $5i_1+j_1$ und $5i_1+j_2$, wobei $j_1, j_2 \in \{1,2,3\}$ und $j_1 < j_2$. Hierbei nehmen wir o. B. d. A. an, diese beiden Zahlen seien rot gefärbt, andernfalls vertausche man im folgenden rot und blau. Nun sei $j_3 := 2j_2 - j_1$. Dann ist $j_3 = j_2 + (j_2 - j_1) \in \{1,\dots,5\}$. Ist $5i_1 + j_3$ rot, so erhält man die rote arithmetische Progression

$$5i_1 + j_1 \quad 5i_1 + j_2 \quad 5i_1 + j_3.$$

Wir nehmen daher jetzt an $5i_1+j_3$ sei blau. Wegen $j_3 \in \{1,\dots,5\}$ ist $5i_1+j_3 \in B_{i_1}$. Da B_{i_2} genauso gefärbt ist wie B_{i_1}, ist auch $5i_2+j_3$ blau. Ferner hat $5i_2+j_2$ dieselbe Farbe wie $5i_1+j_2$, es ist also $5i_2+j_2$ rot. Nun sei $i_3 := 2i_2-i_1$. Dann ist $i_1 < i_2 < i_3 \le 64$. Wir betrachten jetzt die Zahl $5i_3 + j_3 \in \{1,\dots,325\}$ und machen eine Fallunterscheidung. Ist $5i_3 + j_3$ rot, so hat man die rote arithmetische Progression

$$5i_1 + j_1 \quad 5i_2 + j_2 \quad 5i_3 + j_3.$$

Ist dagegen $5i_3 + j_3$ blau, so hat man die blaue arithmetische Progression

$$5i_1 + j_3 \quad 5i_2 + j_3 \quad 5i_3 + j_3.$$

Damit ist $W(2,3) \le 325$ bewiesen. □

3.2 Beweis des Satzes von van der Waerden nach Graham-Rothschild

Wir bringen eine ausführliche Version des Aufsatzes von R. L. Graham, B. L. Rothschild (1974), siehe auch R. L. Graham (1981, S. 12–13) und R. L. Graham, B. L. Rothschild, J. H. Spencer (1990, S. 33–34). Hingewiesen sei auch auf den Wikipedia-Beitrag *Van der Waerden's theorem*, siehe auch C. J. Moreno, S. S. Wagstaff Jr. (2005, S. 263 ff.).

Sind $a, b \in \mathbb{Z}$ zwei ganze Zahlen mit $a \le b$, so bezeichnen wir mit $[a, b]$ die Menge der ganzen Zahlen z mit $a \le z \le b$. Nun definieren wir:

Definition 3.2 Seien $k, m \in \mathbb{N}$. Dann heißen die beiden m-Tupel

$$x = (x_1, \ldots, x_m),\ x' = (x'_1, \ldots, x'_m) \in [0,k]^m$$

k-äquivalent, wenn ein $j \in [0,m]$ existiert mit $x_i = x'_i$ für $i \in [1,j]$ und $x_i, x'_i \neq k$ für $i \in [j+1,m]$. Unter einer *k-Äquivalenzklasse* von $[0,k]^m$ verstehen wir die Menge aller m-Tupel aus $[0,k]^m$, die zueinander k-äquivalent sind.

Z. B. sind zwei m-Tupel x und x' aus $[0,k]^m$, die k nicht enthalten bzw. aus $[0,k-1]^m$ sind, offensichtlich (setze in der Definition $j := 0$) k-äquivalent. Für $m = 1$ liegen daher $0, 1, \ldots, k-1$ in einer k-Äquivalenzklasse von $[0,k]$. Ferner sind z. B. die 5-Tupel $(3,4,4,0,1)$ und $(3,4,4,1,0)$ 4-äquivalent, während es $(2,4,4,0,1)$ und $(3,4,4,1,0)$ nicht sind.

In Abhängigkeit von $k, m \in \mathbb{N}$ betrachten wir die Aussage

A(k,m) Für jedes $r \in \mathbb{N}$ existiert $W(r,k,m) \in \mathbb{N}$ derart, dass es für jede Funktion

$$C\colon [1, W(r,k,m)] \longrightarrow [1,r]$$

Zahlen $a, d_1, \ldots, d_m \in \mathbb{N}$ gibt, für die $a + \sum_{i=1}^{m} kd_i \in [1, W(r,k,m)]$ und $C(a + \sum_{i=1}^{m} x_i d_i)$ konstant auf jeder k-Äquivalenzklasse von $[0,k]^m$ ist.

Beispiel: Die Aussage A$(k,1)$ ist: Für jedes $r \in \mathbb{N}$ (bzw. jede Anzahl r von Farben) existiert $W(r,k) = W(r,k,1) \in \mathbb{N}$ derart, dass für jede Funktion

$$C\colon [1, W(r,k)] \longrightarrow [1,r]$$

(bzw. jede Färbung der Zahlen $\{1, \ldots, W(r,k)\}$ durch eine der r Farben) es Zahlen $a, d \in \mathbb{N}$ (bzw. eine arithmetische Progression $(a, a+d, \ldots, a+(k-1)d)$ der Länge k) mit $C(a) = C(a+id)$, $i \in [0,k-1]$, gibt (bzw. eine monochrome arithmetische Progression der Länge k). Die Aussage A$(k,1)$ ist daher genau dann richtig, wenn der Satz von van der Waerden gültig ist. □

Ist daher der folgende Satz bewiesen, so ist auch der Satz von van der Waerden bewiesen.

Satz 3.3 *Für alle $k, m \in \mathbb{N}$ gilt die Aussage A(k,m).*

Beweis: Sei $k \in \mathbb{N}$ vorgegeben. Zunächst gilt:

(a) Die Aussage A$(1,1)$ ist richtig.

Denn: Die 1-Äquivalenzklassen von $[0,1]$ sind $\{0\}$ und $\{1\}$. Für $r \in \mathbb{N}$ setze man $W(r,1,1) := 2$. Ist dann $C\colon [1,2] \longrightarrow [1,r]$ eine Abbildung, die $[1,2]$ eine von r Farben zuordnet, so setze man $a := 1$, $d_1 := 1$ und $C(a + xd)$ ist trivialerweise konstant auf den einpunktigen Mengen $\{0\}$ und $\{1\}$.

(b) Gelten A$(k,1)$ und A(k,m), so gilt A$(k,m+1)$.

Denn: Sei $r \in \mathbb{N}$ vorgegeben. Wir definieren $M := W(r,k,m)$, $M' := W(r^M,k,1)$, was jeweils wohldefiniert ist, da $\mathrm{A}(k,m)$ und $\mathrm{A}(k,1)$ nach Annahme richtig sind. Wir geben uns eine Funktion

$$C\colon [1,MM'] \longrightarrow [1,r]$$

vor. Hierdurch ist also eine Färbung der Zahlen $[1,MM']$ durch r Farben gegeben. Diese induziert eine Färbung

$$C'\colon [1,M'] \longrightarrow [1,r^M]$$

der Zahlen $[1,M']$ durch r^M Farben. Denn man stelle sich die Zahlen $[1,MM']$ aufgeteilt in M' Blöcke

$$B_i := \{(i-1)M+j : j \in [1,M]\} = [(i-1)M+1,M], \qquad i \in [1,M'],$$

der Länge M vor. Jeder der M' Blöcke der Länge M kann auf r^M Weisen mit r Farben gefärbt werden. Diese r^M Möglichkeiten, einen Block der Länge M mit r Farben zu färben, denke man sich durchnummeriert und definiere $C'\colon [1,M'] \longrightarrow [1,r^M]$ so, dass $C'(i) \in [1,r^M]$ für $i \in [1,M']$ die "Farbe" des Blockes B_i ist. Man stelle sich etwa vor, dass

$$C'(i) = \begin{pmatrix} C((i-1)M+1) \\ \vdots \\ C(iM) \end{pmatrix}$$

die "Farbe" des Blocks B_i ist[13]. Da die Aussage $\mathrm{A}(k,1)$ nach Voraussetzung richtig ist, gibt es $a',d' \in \mathbb{N}$ derart, dass $a'+kd' \leq M'$ und $C'(a'+xd')$ konstant für $x \in [0,k-1]$ ist. In der durch C' gefärbten Menge der M' Blöcke B_i der Länge M gibt es also eine monochrome arithmetische Progression

$$A' = \{a', a'+d', \ldots, a'+(k-1)d'\}$$

[13]Wir wollen hierzu ein Beispiel angeben. Nehmen wir an, es sei $r=2$, die Farben also z. B. blau und rot. Sei $M=3$ und $M'=4$. Die Abbildung C gibt den Zahlen $[1,MM'] = [1,12]$ jeweils eine Farbe. Sei C etwa gegeben durch

C:

1	2	3	4	5	6	7	8	9	10	11	12
B	B	R	B	R	B	B	R	R	R	B	B

Dann hat man die Blöcke

$$B_1 = \{1,2,3\}, \quad B_2 = \{4,5,6\}, \quad B_3 = \{7,8,9\}, \quad B_4 = \{10,11,12\}.$$

Die Abbildung

$$C'\colon \{1,\ldots,4\} \longrightarrow \{\mathrm{BBB},\mathrm{BBR},\mathrm{BRB},\mathrm{BRR},\mathrm{RRR},\mathrm{RBR},\mathrm{RBB},\mathrm{RRB}\} = \{1,2,3,4,5,6,7,8\}$$

ist dann gegeben durch

C':

1	2	3	4
BBR	BRB	BRR	RBB

$=$

1	2	3	4
2	3	4	7

.

der Länge k. D. h. die k gleichabständigen (mit einem Abstand $d'M$) Blöcke

$$B_{a'}, B_{a'+d'}, \ldots, B_{a'+(k-1)d'}$$

sind exakt gleich gefärbt, was wiederum

$$C((a'-1)M+j) = C((a'+xd')M+j), \quad x \in [1,k-1], \quad j \in [1,M]$$

bedeutet. Wir definieren die Menge natürlicher Zahlen

$$I := [(a'-1)M+1, a'M].$$

In I sind also genau die M natürlichen Zahlen zusammengefasst, die im Block $B_{a'}$ stehen. Da die Aussage $\mathrm{A}(k,m)$ gilt und auf I angewandt (eine Translation einer arithmetischen Progression ist wieder eine arithmetische Progression) werden kann, existieren $a, d_1, \ldots, d_m \in \mathbb{N}$ derart, dass $a + \sum_{i=1}^{m} x_i d_i \in I$ falls $x_i \in [0,k]$, $i \in [1,m]$, und $C(a + \sum_{i=1}^{m} x_i d_i)$ konstant auf k-Äquivalenzklassen von $[0,k]^m$ ist. Wir setzen

$$d_i' := \begin{cases} d_i, & i \in [1,m], \\ d'M, & i = m+1, \end{cases}$$

und überlegen uns, dass $a + \sum_{i=1}^{m+1} x_i d_i' \in [1, MM']$ für $x_i \in [0,k]$, $i \in [1,m+1]$, und $C(a + \sum_{i=1}^{m+1} x_i d_i')$ konstant auf jeder k-Äquivalenzklasse von $[0,k]^{m+1}$ ist. Ist dies gelungen, so ist nachgewiesen, dass $W(r,k,m+1) := MM'$ gewählt werden kann und daher auch $\mathrm{A}(k,m+1)$ gilt. Für $x_i \in [0,k]$, $i \in [1,m+1]$, ist

$$\begin{aligned} a + \sum_{i=1}^{m+1} x_i d_i' &= \underbrace{a + \sum_{i=1}^{m} x_i d_i}_{\in [(a'-1)M+1, a'M]} + x_{m+1} d'M \\ &\leq a'M + kd'M \\ &\leq a'M + MM' - a'M \\ &= MM'. \end{aligned}$$

Die Elemente der k-Äquivalenzklassen von $[0,k]^{m+1}$ bestehen aus $(m+1)$-Tupeln $(x_1, \ldots, x_{m+1})$ mit Komponenten x_i aus $[0,k]$. Die erste Äquivalenzklasse ist durch $[0,k-1]^{m+1}$ gegeben, die Komponenten x_i aller Elemente dieser Äquivalenzklasse liegen also in $[0,k-1]$. Nach Wahl von $a, d_1, \ldots, d_m$ ist $C(a + \sum_{i=1}^{m} x_i d_i)$ unabhängig von dem Vertreter $(x_1, \ldots, x_m)$ der Äquivalenzklasse $[0,k-1]^m$. Da ferner

$$a + \sum_{i=1}^{m} x_i d_i \in B_{a'}, \qquad a + \sum_{i=1}^{m} x_i d_i + x_{m+1} d'M \in B_{a'+x_{m+1}d'}$$

an ein und derselben Stelle im Block $B_{a'}$ bzw. $B_{a'+x_{m+1}d'}$ stehen und diese Blöcke identisch gefärbt sind, haben alle Elemente von $[0,k-1]^{m+1}$ dieselbe Farbe. Ganz genauso kann man für k-Äquivalenzklassen argumentieren, bei denen mit einem $j \in$

$[1,m]$ Vertreter in der j-ten Komponente mit k besetzt sind (und die Komponenten danach aus $[0,k-1]$ sind). Es bleiben nur noch k-Äquivalenzklassen von $[0,k]^{m+1}$ zu betrachten, deren Vertreter in der $(m+1)$-ten Komponente mit k besetzt. Nach Definition der k-Äquivalenz sind diese Äquivalenzklassen aber notwendig einpunktig und daher trivialerweise konstant gefärbt.

(c) Gilt $\mathrm{A}(k,m)$ für alle $m \in \mathbb{N}$, so gilt $\mathrm{A}(k+1,1)$.

Denn: Für ein festes $r \in \mathbb{N}$ sei die Abbildung

$$C\colon [1,2W(r,k,r)] \longrightarrow [1,r]$$

(also eine Färbung der Zahlen $[1,2W(r,k,r)]$ durch r Farben) gegeben. Da insbesondere $\mathrm{A}(k,r)$ gilt, existieren $a,d_1,\ldots,d_r \in \mathbb{N}$ mit $a+\sum_{i=1}^{r} kd_i \leq W(r,k,r)$ und der Eigenschaft, dass $C(a+\sum_{i=1}^{r} x_i d_i)$ auf jeder k-Äquivalenzklasse von $[0,k]^r$ konstant ist. Durch

$$(0,0,0,\ldots,0),\quad (k,0,0,\ldots,0),\quad (k,k,0,\ldots,0),\ldots,(k,k,k,\ldots,k)$$

als Vertreter hat man $r+1$ verschiedene k-Äquivalenzklassen von $[0,k]^r$ gegeben. Da andererseits zur Färbung nur r Farben zur Verfügung stehen, existieren wegen des Schubfachprinzips ganze Zahlen $0 \leq u < v \leq r$ mit

$$C\Big(a+\sum_{i=1}^{u} kd_i\Big) = C\Big(a+\sum_{i=1}^{v} kd_i\Big).$$

Wir setzen

$$a^* := a+\sum_{i=1}^{u} kd_i, \qquad d^* := \sum_{i=u+1}^{v} d_i$$

und zeigen, dass hiermit und $W(r,k+1,1) := 2W(r,k,r)$ die Aussage $\mathrm{A}(k+1,1)$ gilt. Hierzu beachten wir, dass

$$\begin{aligned} a^*+(k+1)d^* &= a+\sum_{i=1}^{u} kd_i + (k+1)\sum_{i=u+1}^{v} d_i \\ &= a+\sum_{i=1}^{v} kd_i + \sum_{i=u+1}^{v} d_i \\ &\leq \underbrace{a+\sum_{i=1}^{r} kd_i}_{\leq W(r,k,r)} + \sum_{i=1}^{r} d_i \\ &\leq 2W(r,k,r), \end{aligned}$$

sodass zu zeigen bleibt, dass $C(a^*+xd^*)$ konstant auf jeder $(k+1)$-Äquivalenzklasse von $[0,k+1]$ ist. Eine der $(k+1)$-Äquivalenzklassen besteht aus $\{0,1,\ldots,k\}$, die andere

aus $\{k+1\}$. Daher ist zu zeigen, dass $C(a^* + xd^*)$ konstant für $x \in [0,k]$ ist. Wir unterscheiden zwei Fälle. Im ersten ist $x \in [0,k-1]$. Dann ist das r-Tupel

$$(\underbrace{k,\ldots,k}_{u},\underbrace{x,\ldots,x}_{v-u},\underbrace{0,\ldots,0}_{r-v}) \in [0,k]^r$$

offenbar k-äquivalent zu

$$(\underbrace{k,\ldots,k}_{u},\underbrace{0,\ldots,0}_{r-u}) \in [0,k]^r.$$

Für $x \in [0,k-1]$ haben daher

$$a^* + xd^* = a + \sum_{i=1}^{u} kd_i + \sum_{i=u+1}^{v} xd_i$$

und

$$a^* = a + \sum_{i=1}^{u} kd_i$$

dieselbe Farbe. Im zweiten Fall ist $x = k$. Dann hat aber (nach Wahl von u und v)

$$a^* + xd^* = a^* + kd^* = a + \sum_{i=1}^{v} kd_i$$

dieselbe Farbe wie a^* und die Behauptung ist bewiesen.

(d) Für alle $k, m \in \mathbb{N}$ gilt die Aussage $\mathrm{A}(k,m)$.

Denn: Der Beweis hierfür durch (doppelte) vollständige Induktion ist wegen der vorangegangenen Schritte offensichtlich. Damit ist der Satz bewiesen. □

Völlig zu Recht steht am Schluss der Arbeit von R. L. Graham, B. L. Rothschild (1974) die Aussage: The authors point out that while previous proofs follow essentially the argument above, the one given is hopefully clearer.

Bemerkung: Mit $W(r,k,m)$ haben wir die in der Aussage $\mathrm{A}(k,m)$ auftretende natürliche Zahl bezeichnet, wobei wir nicht auf einer Minimalitätseigenschaft bestanden haben. So ist also z. B. $W(r,k,1)$ eine obere Schranke für die van der Waerden Zahl $W(r,k)$. In den Beweisschritten (a), (b) bzw. (c) von Satz 3.3 haben wir nachgewiesen:

(a) $W(r,1,1) = 2$,

(b) $W(r,k,m+1) \leq W(r,k,m) \cdot W(r^{W(r,k,m)},k,1)$,

(c) $W(r,k+1,1) \leq 2W(r,k,r)$.

Wir wollen einmal ausrechnen, welche obere Abschätzung für $W(2,3)$ wir auf diese Weise erhalten. Um die Abschätzungen zu vereinfachen, benutzen wir hierbei statt (a),

dass $W(r,2,1) = r+1$ (siehe das Beispiel im Anschluss an die Formulierung des Satzes von van der Waerden). Es ist

$$\begin{aligned} W(2,3) &\leq W(2,3,1) \\ &\leq 2W(2,2,2) \\ &\qquad \text{(wegen (c))} \\ &\leq 2W(2,2,1) \cdot W(2^{W(2,2,1)},2,1) \\ &\qquad \text{(wegen (b))} \\ &\leq 2 \cdot 3 \cdot W(2^3,2,1) \\ &\qquad \text{(wegen } W(r,2,1) = r+1) \\ &\leq 2 \cdot 3 \cdot (2^3+1) \\ &= 54. \end{aligned}$$

Diese Abschätzung ist also etwas besser als die Abschätzung $W(2,3) \leq 325$, die wir im letzten Unterabschnitt hergeleitet haben. □

3.3 Eine Verschärfung des Satzes von van der Waerden

In diesem Unterabschnitt werden wir die folgende Aussage beweisen.

Satz 3.4 *Seien $r,k,s \in \mathbb{N}$ gegeben. Dann existiert $N(r,k,s) \in \mathbb{N}$ mit folgender Eigenschaft: Färbt man $[1,N]$ mit $N \geq N(r,k,s)$ durch r Farben, so existieren $a,d \in \mathbb{N}$ derart, dass die arithmetische Progression $\{a, a+d, \ldots, (k-1)d\}$ der Länge k und $\{sd\}$, also das s-fache der Schrittweite d, in $[1,N]$ enthalten sind und dieselbe Farbe haben.*

Beweis: Wir beweisen den Satz durch vollständige Induktion nach r, der Anzahl der Farben. Für $r = 1$ (nur eine Farbe) müssen wir $a,d \in \mathbb{N}$ nur so wählen, dass

$$a + (k-1)d \leq N(1,k,s), \qquad sd \leq N(1,k,s).$$

Dies erreichen wir, indem wir $a := 1$, $d := 1$ und $N(1,k,s) := \max(k,s)$ setzen. Bei der Induktionsannahme wird angenommen, es sei $r \geq 2$ und $N(r-1,k,s)$ existiere für alle $k,s \in \mathbb{N}$. Wir setzen

$$N^* := sW(r, kN(r-1,k,s)),$$

wobei W die van der Waerden Zahl zum entsprechenden Argument bedeutet. Die Zahlen $[1,N^*]$ seien durch r Farben gefärbt. Wegen des Satzes von van der Waerden können wir eine monochrome arithmetische Progression der Länge $kN(r-1,k,s)$ in $[1, W(r,kN(r-1,k,s))]$ finden, ihre Farbe sei etwa rot. Das Anfangsglied dieser arithmetischen Progression werde mit a, die Schrittweite mit d' bezeichnet. Nun gibt es zwei Möglichkeiten:

- Für ein $j \in [0, N(r-1,k,s)-1]$ hat $sd'(j+1)$ die Farbe rot. Man beachte hierzu, dass
$$sd'(j+1) \leq sd'N(r-1,k,s) \leq sW(r,kN(r-1,k,s)) = N^*,$$
sodass $sd'(j+1)$ wirklich gefärbt ist.

Man setze $d := (j+1)d'$. Dann sind $a, a+d, \ldots, a+(k-1)d$ Teile der arithmetischen Progression $\{a, a+d', \ldots, (kN(r-1,k,s)-1)d'\}$ und damit sämtlich rot gefärbt, da

$$id = i(j+1)d' \leq (k-1)N(r-1,k,s) \leq kN(r-1,k,s), \quad i \in [0,k-1].$$

Da auch $sd = sd'(j+1)$ rot gefärbt ist, haben wir eine monochrome arithmetische Progression $\{a, a+d, \ldots, a+(k-1)\}$ der Länge k gefunden, die zudem noch dieselbe Farbe wie sd hat. Daher existiert $N(r,k,s)$ (und ist $\leq sW(r, kN(r-1,k,s))$).

- Für alle $j \in [0, N(r-1,k,s)-1]$ ist $sd'(j+1)$ *nicht* rot gefärbt.

Dann sind die Punkte $sd'(j+1)$, $j \in [0, N(r-1,k,s)-1]$, durch $r-1$ Farben gefärbt. Hierdurch wird eine Färbung von $[1, N(r-1,k,s)]$ durch $r-1$ Farben induziert, indem man $j \in [1, N(r-1,k,s)]$ die Farbe von $sd'j$ gibt. Nach Definition von $N(r-1,k,s)$ existieren $A, D \in \mathbb{N}$ derart, dass $\{A, A+D, \ldots, (k-1)D\}$ sowie sD dieselbe Farbe haben, etwa blau, und in $[1, N(r-1,k,s)]$ liegen. Daher sind auch die arithmetische Progression $\{sd'A, sd'(A+D), \ldots, sd'(A+(k-1)D)\}$ sowie $\{sd'(sD)\}$ sämtlich blau gefärbt. Daher existiert $N(r,k,s)$ auch in diesem Falle. Der Satz ist bewiesen. □

4 Ramsey-Theorie

Ausgangspunkt der Ramsey-Theorie ist die Arbeit F. P. Ramsey (1930), welche man auch bei I. Gessel, G.-C. Rota (1987) nachlesen kann. Wir wollen einen kleinen Eindruck von dieser Theorie geben, wobei wir eine Bestätigung der oft genannten Aussage

- Every large system contains a large well-organized subsystem

erhalten werden. Standardliteratur für die Ramsey-Theorie ist wohl nach wie vor R. L. Graham, B. L. Rothschild, J. H. Spencer (1990). In einem Handbook of Combinatorics findet man einen Übersichtsartikel von J. Nešetřil (1995). Aber auch in vielen Büchern über Graphentheorie, etwa bei B. Bollobás (1998, S. 181 ff.) findet man einen Überblick.

4.1 Der Satz von Ramsey für vollständige Graphen

Ein *Graph* $G = (V, E)$ besteht bekanntlich aus einer endlichen Menge V von *Ecken* und einer Menge E von *Kanten*, bestehend aus Paaren von Ecken. Ein *vollständiger Graph* ist ein Graph, in dem jede Ecke mit jeder anderen Ecke durch eine Kante verbunden ist. Ein vollständiger Graph mit N Ecken wird mit K_N bezeichnet. Die Anzahl der Kanten von K_N ist offenbar $N(N-1)/2$.

Der Satz von Ramsey für vollständige Graphen (und zwei Farben) sagt aus:

Satz 4.1 (Ramsey) *Zu je zwei natürlichen Zahlen m, n gibt es eine (von m und n abhängige) natürliche Zahl R mit der Eigenschaft, dass jeder vollständige Graph K_N mit $N \geq R$ Ecken, dessen Kanten entweder rot oder blau gefärbt sind, einen*

roten vollständigen Teilgraphen K_m oder einen blauen vollständigen Teilgraphen K_n enthält[14].

Die kleinste natürliche Zahl, die als ein solches R bei gegebenen m, n gewählt werden kann, heißt die zu m und n gehörige *Ramsey-Zahl* und wird mit $R(m, n)$ bezeichnet.

Beispiel: In Abbildung 9 geben wir eine Färbung des K_5 durch die Farben rot und blau an, zu der es kein rotes oder blaues Dreieck bzw. einen monochromen K_3 als Teilgraphen gibt. Also ist $R(3, 3) > 5$ bzw. $R(3, 3) \geq 6$. Andererseits kann man den K_6 beliebig

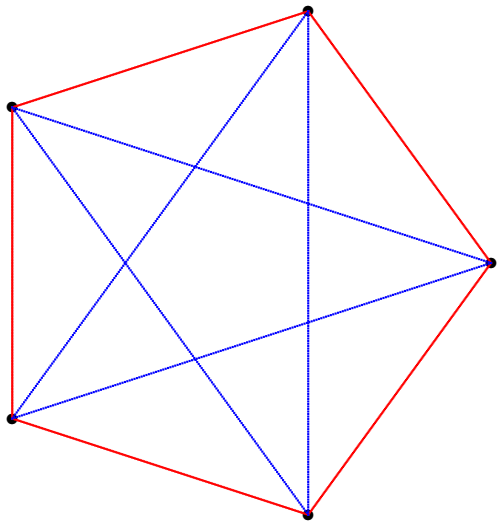


Abbildung 9: Eine 2-Färbung von K_5 ohne monochromes K_3

mit den Farben rot und blau färben und stets ein rotes oder blaues Dreieck finden. Denn man greife sich eine beliebige Ecke v heraus. Dann gibt es fünf Kanten, die mit dieser Ecke inzidieren. Unter diesen fünf Kanten sind mindestens drei gleich gefärbt, o. B. d. A. mit der Farbe rot. Ist eine der drei Kanten, die die Eckpunkte der drei rot gefärbten Kanten verbinden, ebenfalls rot, so hat man ein rotes Dreieck bzw. einen rot gefärbten K_3. Andernfalls sind die verbindenden Kanten alle blau gefärbt und man hat ein blaues Dreieck. Diesen Sachverhalt veranschaulichen wir uns in Abbildung 10 links bzw. rechts. Insgesamt ist daher $R(3, 3) = 6$. Eine Interpretation dieses Sachverhalts ist die folgende:

- Bei jeder Party mit sechs Teilnehmern gibt es drei Personen, die sich vorher schon kannten, oder drei Personen, die sich bei der Party das erste Mal treffen.

Natürlich werden die Teilnehmer als Ecken in dem vollständigen Graphen K_6 interpretiert und die entsprechende Kante rot gefärbt, wenn die Teilnehmer sich schon kannten, im anderen Falle blau. Da $R(3, 3) = 6$, gibt es einen roten oder einen blauen Teilgraphen K_3, also drei Personen, die sich vorher kannten oder nicht kannten.

Im allgemeinen Fall stellt sich bei gegebenen $m, n \in \mathbb{N}$ die folgende Frage:

[14] Monochrome vollständige Teilgraphen werden auch eine *Clique* genannt.

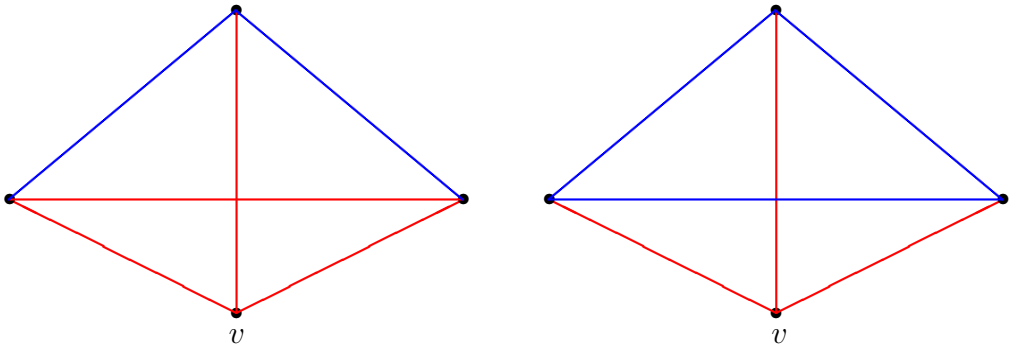


Abbildung 10: Ein rotes bzw. blaues Dreieck

- Wie viele Personen müssen zu einer Party mindestens eingeladen werden um sicher zu sein, dass sich mindestens m Teilnehmer vorher schon kannten oder mindestens n Personen vorher sich noch nie getroffen haben?

Der Satz von Ramsey macht die Aussage, dass es eine solche Mindestzahl gibt. Allerdings sind nur für eher kleine m, n die Ramsey-Zahl $R(m, n)$ genau bekannt, meistens gibt es nur relativ grobe Abschätzungen. □

Durch den folgenden Satz erhalten wir insbesondere die *Existenz* der Ramsey-Zahlen und damit einen *Beweis* des Satzes von Ramsey, siehe z. B. M. Aigner, G. M. Ziegler (2018, S. 346).

Satz 4.2 *Seien $m, n \in \mathbb{N}$. Dann gilt:*

1. $R(m, 1) = 1 = R(1, m)$.
2. $R(m, 2) = m = R(2, m)$.
3. *Seien $m, n \geq 2$. Wenn $R(m-1, n)$ und $R(m, n-1)$ beide existieren, dann existiert auch $R(m, n)$ und es ist*
$$R(m, n) \leq R(m-1, n) + R(m, m-1).$$
4. *Es ist*
$$R(m, n) \leq \binom{m+n-2}{m-1}.$$
Insbesondere existiert $R(m, n)$ für alle $m, n \in \mathbb{N}$, es gilt also der Satz von Ramsey.
5. *Sind $R(m-1, n)$ und $R(m, n-1)$ beide gerade, so ist*
$$R(m, n) \leq R(m-1, n) + R(m, n-1) - 1.$$

Beweis:

1. Es ist $R(m,1) = 1 = R(1,n)$, da K_1 einfarbig ist.

2. Weiter ist $R(m,2) = m$, da alle Kanten von K_m rot gefärbt sind oder eine blaue Kante existiert. Entsprechend ist $R(2,m) = m$.

3. Wir nehmen an, dass $R(m-1,n)$ und $R(m,n-1)$ beide existieren und setzen

$$N := R(m-1,n) + R(m,n-1).$$

Die Kanten des vollständigen Graphen K_N seien beliebig mit den Farben rot und blau gefärbt. Wir zeigen, dass ein roter K_m oder ein blauer K_n in K_N als Teilgraph enthalten ist. Sei v eine beliebige Ecke von K_N. Wir bezeichnen mit R_v bzw. B_v die Menge der Ecken von K_N, die mit v durch eine rote bzw. blaue Kante verbunden sind. Da R_v und B_v disjunkt sind und v nicht enthalten, ist $|R_v| + |B_v| = N - 1$ und folglich entweder $|R_v| \geq R(m-1,n)$ oder $|B_v| \geq R(m,n-1)$. Wir nehmen an, es sei $|R_v| \geq R(m-1,n)$. Nach Definition von $R(m-1,n)$ bedeutet dies, dass in dem (blau-rot gefärbten) vollständigen Graphen mit Ecken aus R_v ein roter K_{m-1} oder ein blauer K_n enthalten ist. Ist ersteres der Fall, so erhält man aus diesem roten K_{m-1} zusammen mit der Ecke v einen vollständigen K_m. Falls $|R_v| \geq R(m-1,n)$, so ist also in K_N ein roter K_m oder ein blauer K_n enthalten. Da man für $|B_v| \geq R(m,n-1)$ analog argumentieren kann, folgt aus der Existenz von $R(m-1,n)$ und $R(m,n-1)$ die Existenz von $R(m,n)$ sowie

$$R(m,n) \leq R(m-1,n) + R(m,n-1).$$

4. Nun zeigen wir, dass

$$R(m,n) \leq \binom{m+n-2}{m-1}$$

für alle $m,n \in \mathbb{N}$. Dies ist für $m = 1,2$ für alle $n \in \mathbb{N}$ und für $n = 1,2$ für alle $m \in \mathbb{N}$ richtig, da

$$R(1,n) = 1 = \binom{n-1}{0}, \qquad R(2,n) = n = \binom{n}{1}$$

und

$$R(m,1) = 1 = \binom{m-1}{m-1}, \qquad R(m,2) = m = \binom{m}{m-1}$$

richtig ist. Den Induktionsbeweis machen wir uns mit Hilfe von Abbildung 11 klar. Die Aussage ist für Punkte auf den ersten beiden waagerechten bzw. senkrechten Linien richtig. Dann kann man offenbar zeilenweise vorgehen. Die Gültigkeit der Aussage für Punkte auf der dritten Linie erhält man sukzessive aus

$$(2,3),(3,2) \to (3,3),(4,2) \to (4,3),(5,2) \to (5,3),(6,2) \to (6,3),(7,2) \to \ldots.$$

Wir zeigen durch vollständige Induktion nach n:

- Für $n \in \mathbb{N}$ ist $R(m,n) \leq \binom{m+n-2}{m-1}$ für alle $m \in \mathbb{N}$.

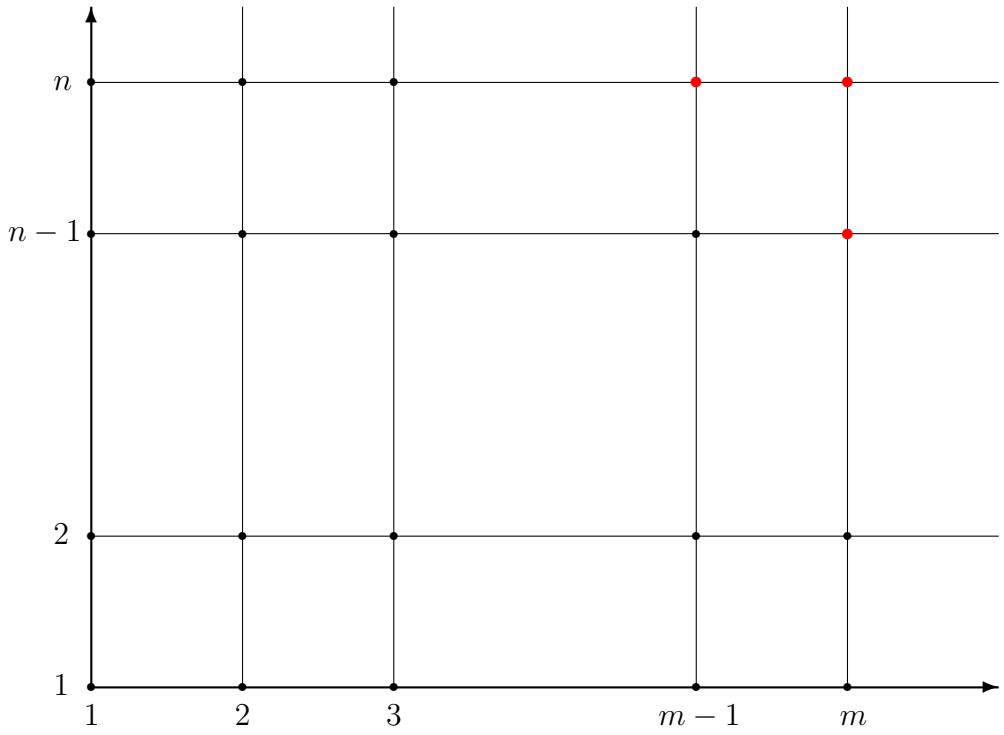


Abbildung 11: Veranschaulichung des Induktionsbeweises

Die Aussage ist für $n = 1, 2$ richtig. Angenommen, sie ist für $n-1$ richtig, es sei also

$$R(m, n-1) \leq \binom{m+n-3}{m-1} \quad \text{für alle } m \in \mathbb{N}.$$

Beim Induktionschluss ist zu zeigen, dass

$$R(m, n) \leq \binom{m+n-2}{m-1} \quad \text{für alle } m \in \mathbb{N}.$$

Dies wiederum wird durch Induktion nach m bewiesen. Die Aussage ist für $m = 1, 2$ richtig. Angenommen, sie sei für $m-1$ richtig. Dann ist

$$\begin{aligned} R(m, n) &\leq R(m-1, n) + R(m, n-1) \\ &\leq \binom{m+n-3}{m-2} + \binom{m+n-3}{m-1} \\ &= \binom{m+n-2}{m-1}, \end{aligned}$$

wobei wir am Schluss die Pascalsche Identität benutzt haben. Damit ist der Induktionsbeweis abgeschlossen und der Satz bewiesen.

5. Seien $R(m-1, n)$ und $R(m, n-1)$ beide gerade. Wieder setzen wir

$$N := R(m-1, n) + R(m, n-1)$$

und zeigen, dass im beliebig rot-blau gefärbten K_{N-1} ein roter K_m oder ein blauer K_n als Teilgraph enthalten ist, woraus dann die Behauptung folgt. Für eine Ecke v von K_{N-1} (diese seien mit $1, \ldots, N-1$ bezeichnet) sei wieder R_v bzw. B_v die Menge der Ecken von K_{N-1}, die mit v durch eine rote bzw. blaue Kante verbunden sind. Dann ist $\sum_{v=1}^{N-1} |R_v|$ offenbar eine gerade Zahl, nämlich genau das Doppelte der Anzahl roter Kanten. Als Summe von zwei geraden Zahlen ist N ebenfalls gerade und daher $N-1$ ungerade. Da $\sum_{v=1}^{N-1} |R_v|$ gerade ist, ist $|R_v|$ für mindestens ein $v \in \{1, \ldots, N-1\}$ gerade. Sei etwa $|R_w|$ für ein gewisses $w \in \{1, \ldots, N-1\}$ gerade. Wegen $|R_w| + |B_w| = N-2$ ist auch $|B_w|$ gerade. Weiter ist $|R_w| \geq R(m-1,n) - 1$ oder $|B_w| \geq R(m, n-1)$. Nun ist $R(m-1,n)-1$ ungerade und $|R_w|$ gerade. Daher ist $|R_w| \geq R(m-1,n)$ oder $|B_w| \geq R(m,n-1)$. Wenn $|R_w| \geq R(m-1,n)$, so können wir genau wie oben argumentieren: Nach Definition von $R(m-1,n)$ bedeutet dies, dass in dem (blau-rot gefärbten) vollständigen Graphen mit Ecken aus R_w ein roter K_{m-1} oder ein blauer K_n enthalten ist. Ist ersteres der Fall, so erhält man aus diesem roten K_{m-1} zusammen mit der Ecke w einen vollständigen K_m. Falls $|R_w| \geq R(m-1,n)$, so ist also in K_{N-1} ein roter K_m oder ein blauer K_n enthalten. Da man für $|B_w| \geq R(m,n-1)$ analog argumentieren kann, ist die Behauptung bewiesen.

Damit ist der Satz bewiesen. □

Beispiel: Nur wenige Ramsey-Zahlen sind bekannt, für die meisten gibt es nur relativ grobe Abschätzungen nach oben und evtl. nach unten. Im letzten Beispiel haben wir $R(3,3) = 6$ gezeigt. Hierzu zeigten wir zunächst, dass $R(3,3) > 5$, da die Kanten des K_5 so rot-blau gefärbt können, dass keine monochromen Dreiecke auftreten. Anschließend haben wir $R(3,3) \leq 6$ gezeigt bzw. dass in einem beliebig rot-blau-gefärbten K_6 stets ein monochromes Dreieck auftritt. Ähnlich wollen wir nun nachweisen, dass $R(3,4) = 9$. Da $R(2,4) = 4$ und $R(3,3) = 6$ beide gerade sind, folgt aus Teil 5. von Satz 4.2, dass

$$R(3,4) \leq R(2,4) + R(3,3) - 1 = 9.$$

Um nachzuweisen, dass $R(3,4) > 8$ ist, müssen wir noch eine rot-blau-Färbung des K_8 angeben, bei der es keinen roten K_3 und keinen blauen K_4 als Teilgraphen gibt. Eine solche Färbung des K_8 geben wir in Abbildung 12 an. Der Übersichtlichkeit halber haben wir die roten und die blauen Kanten getrennt gezeichnet. Ähnlich kann auch $R(3,5) = 14$ gezeigt werden. Wegen Teil 3. von Satz 4.2 ist nämlich

$$R(3,5) \leq R(2,5) + R(3,4) = 5 + 9 = 14.$$

Daher genügt es, eine rot-blaue Färbung des K_{13} anzugeben, in welcher es keinen roten K_3 und keinen blauen K_5 gibt. Näheres hierzu kann man `http://www.cut-the-knot.org/arithmetic/combinatorics/Ramsey53.shtml` (dort sind die Farben vertauscht) finden. Weiter kann man zeigen, dass $R(3,6) = 18$. Durch Teil 5. von Satz 4.2 erhalten wir lediglich

$$R(3,6) \leq \underbrace{R(2,6)}_{=6} + \underbrace{R(3,5)}_{=14} - 1 = 19.$$

Hier muss also noch etwas subtiler argumentiert werden. □

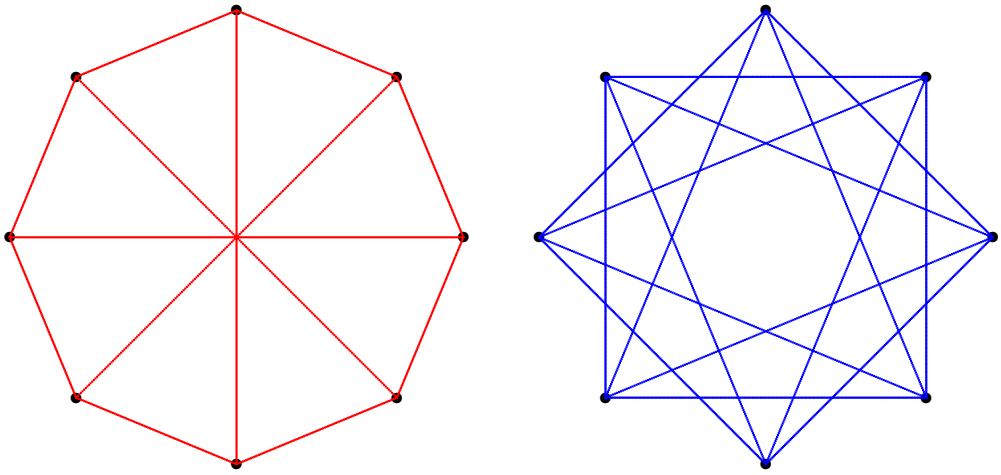


Abbildung 12: Es ist $R(3,4) > 8$

Bemerkung: Bisher haben wir Färbungen eines vollständigen Graphen durch *zwei* Farben betrachtet. Dies kann natürlich verallgemeinert werden. Es gilt nämlich der Satz von Ramsey für mehrere Farben:

- *Gegeben seien $r \in \mathbb{N}$ (Zahl der Farben) sowie $n_1, \ldots, n_r \in \mathbb{N}$. Dann existiert eine (von $n_1, \ldots, n_r$ abhängende) natürliche Zahl R mit der Eigenschaft, dass jeder vollständige Graph K_N mit $N \geq R$ Ecken, dessen Kanten mit einer der r Farben gefärbt ist, für ein $i \in \{1, \ldots, r\}$ einen monochromen vollständigen Teilgraphen K_{n_i} besitzt.*

Die kleinste Zahl R mit dieser Eigenschaft wird mit $R(n_1, \ldots, n_r)$ bezeichnet.

Der *Beweis* der obigen Aussage erfolgt durch vollständige Induktion nach r, der Anzahl der Farben. Die Aussage ist für $r = 1$ trivial (es ist $R(n_1) = n_1$) und für $r = 2$ oben bewiesen. Sei daher jetzt $r > 2$. Wir nehmen an, die Aussage sei für $r - 1$ richtig. Jetzt zeigen wir, dass

$$(*) \qquad R(n_1, \ldots, n_r) \leq R(n_1, \ldots, n_{r-2}, R(n_{r-1}, n_r)).$$

Die rechte Seite dieser Ungleichung ist nach Induktionsannahme wohldefiniert, weil in ihr nur die Ramsey-Zahl $R(n_{r-1}, n_r)$ für eine Färbung mit zwei Farben sowie eine für $r - 1$ Farben vorkommt. Daher kommt es nur noch darauf an, die Ungleichung $(*)$ zu beweisen. Wir definieren

$$N := R(n_1, \ldots, n_{r-2}, R(n_{r-1}, n_r))$$

und nehmen an, die Kanten des vollständigen Graphen K_N seien mit r Farben gefärbt. Jetzt stelle man sich vor, man könne die Farben $r - 1$ und r nicht unterscheiden (bzw. diese seien durch eine Mischfarbe ersetzt), sodass der Graph K_N nur durch

$r-1$ Farben gefärbt ist, nämlich durch $r-2$ "reguläre" Farben und eine Mischfarbe. Nach Induktionsannahme existiert ein $i \in \{1, \ldots, r-2\}$ mit der Eigenschaft, dass K_{n_i} monochrom mit der Farbe i gefärbt ist, oder $K_{R(n_{r-1},n_r)}$ ist mit der Mischfarbe gefärbt. Nach Definition von $R(n_{r-1}, n_r)$ existiert in $K_{R(n_{r-1},n_r)}$ ein mit der Farbe $r-1$ gefärbter monochromer vollständiger Teilgraph $K_{n_{r-1}}$ oder ein durch r gefärbter monochromer vollständiger Teilgraph K_{n_r}. Jedenfalls ist damit die Ungleichung $(*)$ und damit auch die Existenz der Ramsey-Zahlen für mehrere Farben bewiesen.

Nur sehr wenige Ramsey-Zahlen für mehr als zwei Farben sind bekannt. Zu den wenigen gehört $R(3,3,3) = 17$. Wegen obiger Aussage $(*)$ ist

$$R(3,3,3) \leq R(3, R(3,3)) = R(3,6) = 17.$$

Hierbei haben wir $R(3,6) = 17$ ausgenutzt. Um $R(3,3,3) > 16$ zu zeigen, muss man eine Färbung des K_{16} durch drei Farben angeben, bei der keine monochromen Dreiecke auftreten. Hierauf wollen wir aber nicht mehr eingehen. □

4.2 Der Satz von Ramsey für Mengen bzw. uniforme Hypergraphen

Im nächsten Unterabschnitt benötigen wir zum Beweis von Satz 4.4 einen Satz von Ramsey für Mengen, von dem wir die folgende Version angeben. Für eine Menge X und $i \in \mathbb{N}$ bezeichnen wir hierbei mit $\binom{X}{i}$ die Menge der i-elementigen Teilmengen von X, d. h. es ist

$$\binom{X}{i} := \{Y \subset X : |Y| = i\}.$$

Satz 4.3 (Ramsey) *Seien $i, k, l \in \mathbb{N}$ mit $k \geq i$, $l \geq i$ gegeben. Dann gibt es eine (von i, k, l abhängende) Zahl $R \in \mathbb{N}$ mit folgender Eigenschaft: Ist $N \geq R$ eine natürliche Zahl und X eine N-elementige Menge, ist ferner*

$$\binom{X}{i} = C_1 \dot{\cup} C_2$$

eine beliebige disjunkte Zerlegung[15] *der i-elementigen Teilmengen von X, so existiert*[16]

- *$A_1 \in \binom{X}{k}$ und der Eigenschaft, dass $\binom{A_1}{i} \subset C_1$*

oder es existiert

- *$A_2 \in \binom{X}{l}$ und der Eigenschaft, dass $\binom{A_2}{i} \subset C_2$.*

[15] Z. B. kann man sich vorstellen, dass die Elemente von C_1 rot und die von C_2 blau gefärbt sind.

[16] Etwas anders ausgedrückt: Es existiert eine k-elementige Teilmenge A_1 von X, deren i-elementige Teilmengen alle rot gefärbt sind oder es existiert eine l-elementige Teilmenge A_2 von X, deren i-elementige Teilmengen sämtlich blau gefärbt sind.

Angenommen, die Aussage von Satz 4.3 sei für gegebene $i, k, l \in \mathbb{N}$ mit $k \geq i$, $l \geq i$ bewiesen. Die kleinstmögliche Zahl R mit der im Satz angegebenen Eigenschaft wird mit $R_i(k, l)$ bezeichnet und ebenfalls *Ramsey-Zahl* genannt.

Beweis von Satz 4.3: Beim Beweisaufbau folgen wir im wesentlichen P. ERDÖS, G. SZEKERES (1935). Wir überlegen uns zunächst:

(a) Die Aussage des Satzes ist richtig für $i = 1$ und beliebige $k, l \in \mathbb{N}$.

Denn: Wir wollen uns überlegen, dass die Aussage des Satzes mit $R := k + l - 1$ richtig ist. Sei $N \geq R$. Angenommen, die Aussage sei nicht richtig. Dann wäre $|C_1| < k$ und $|C_2| < l$ und daher

$$k + l - 1 = R \leq N = |C_1| + |C_2| \leq (k - 1) + (l - 1) = k + l - 2,$$

ein Widerspruch.

Sei daher jetzt $i > 1$.

(b) Die Aussage des Satzes ist richtig, falls $k = i$ oder $l = i$.

Denn: Sei etwa $k = i$. Wir wollen uns überlegen, dass die Aussage des Satzes mit $R := l$ richtig ist. Sei $N \geq R$. Ist $A_1 \in \binom{X}{k} = \binom{X}{i}$, so ist $\binom{A_1}{i} = \{A_1\}$. Gilt also die erste Alternative nicht[17], so ist notwendigerweise $C_1 = \emptyset$ und folglich $\binom{X}{i} = C_2$. Wegen $N \geq R = l$ ist daher $\binom{A_2}{i} \subset C_2$ für ein beliebiges $A_2 \in \binom{X}{l}$, d. h. die zweite Alternative ist erfüllt.

Die Argumentation für $l = i$ verläuft entsprechend.

(c) Sei $k > i$. Angenommen, die Aussage des Satzes gelte für $i - 1$ und alle k, l, ferner für i, $k - 1$, l und i, k, $l - 1$. Dann gilt die Aussage des Satzes auch für i, k, l, sodass der Satz wegen (a) und (b) bewiesen sein wird.

Denn: Nach Annahme existieren $R_{i-1}(k, l)$ für alle k, l sowie $k' := R_i(k - 1, l)$ und $l' := R_i(k - 1, l)$, daher ist

$$R := R_{i-1}(k', l') + 1$$

definiert. Sei $N \geq R$, X eine N-elementige Menge und

$$\binom{X}{i} = C_1 \dot{\cup} C_2$$

eine beliebige disjunkte Zerlegung der i-elementigen Teilmengen von X (bzw. eine beliebige rot-blau-Färbung der i-elementigen Teilmengen von X). Man wähle $x \in X$ beliebig und setze $Y := X \setminus \{x\}$. Dann ist

$$|Y| = N - 1 \geq R_{i-1}(k', l').$$

[17] Dann ist keine $k = i$-elementige Teilmenge von X rot gefärbt, d. h. alle i-elementigen Teilmengen von X sind blau gefärbt.

Eine durch C_1, C_2 induzierte disjunkte Zerlegung

$$\binom{Y}{i-1} = D_1 \dot\cup D_2$$

(bzw. rot-blau-Färbung der $(i-1)$-elementigen Teilmengen von Y) erhalten wir durch

$$D_1 := \left\{ B \in \binom{Y}{i-1} : B \cup \{x\} \in C_1 \right\}, \quad D_2 := \left\{ B \in \binom{Y}{i-1} : B \cup \{x\} \in C_2 \right\}.$$

Eine $(i-1)$-elementige Teilmenge B von $X \setminus \{x\}$ erhält also die Farbe von $B \cup \{x\}$. Nach Induktionsannahme existiert

- $B_1 \in \binom{Y}{k'}$ und der Eigenschaft, dass $\binom{B_1}{i-1} \subset D_1$

oder es existiert

- $B_2 \in \binom{Y}{l'}$ und der Eigenschaft, dass $\binom{B_2}{i-1} \subset D_2$.

Wir nehmen an, ersteres sei der Fall. Der Beweis für den zweiten Fall verläuft völlig entsprechend. Es existiert also $B_1 \in \binom{X\setminus\{x\}}{R_i(k-1,l)}$ und der Eigenschaft, dass $\binom{B_1}{i-1} \subset D_1$, also alle $(i-1)$-elementigen Teilmengen von B_1 rot gefärbt sind. Dann existiert aber

- $A \in \binom{B_1}{k-1}$ und der Eigenschaft, dass $\binom{A}{i} \subset C_1$

oder es existiert

- $A' \subset \binom{B_1}{l}$ und der Eigenschaft, dass $\binom{A'}{i} \subset C_2$.

Falls der zweite Fall vorliegt, so sind wir fertig. Man setze nämlich $A_2 := A'$ und hat damit eine Menge $A_1 \in \binom{B_1}{l} \subset \binom{X}{l}$ mit der Eigenschaft, dass $\binom{A_2}{i} \subset C_2$. Liegt der erste Fall vor, so setze man $A_1 := A \cup \{x\}$. Da A genau $k-1$ Elemente enthält, aber x nicht, ist $A_1 \in \binom{X}{k}$. Zu zeigen bleibt $\binom{A_1}{i} \subset C_1$, dass also alle i-elementigen Teilmengen von A_1 rot gefärbt sind. Wenn eine i-elementige Teilmenge von A_1 den Punkt x enthält, so hat sie wegen der induzierten Färbung dieselbe Farbe wie ohne x, also rot. Enthält dagegen eine i-elementige Teilmenge von A_1 den Punkt x nicht, so ist sie eine Teilmenge von A, also rot gefärbt bzw. in C_1 enthalten. Damit ist Satz 4.3 bewiesen. Ferner ist

$$R_i(k,l) \leq R_{i-1}(R_i(k-1,l), R_i(k,l-1)) + 1$$

nachgewiesen. □

Bemerkung: Ein Paar (V, E) mit einer endlichen Menge V und einer Menge E von (nicht notwendig allen) i-elementigen Teilmengen von V, den sogenannten *Hyperkanten*, wird ein *i-uniformer Hypergraph* genannt. Daher wird Satz 4.3 auch *Satz von Ramsey für uniforme Hypergraphen* genannt. Offenbar ist ein 2-uniformer Hypergraph ein Graph. □

Bemerkung: Im Anschluss an den Satz von Ramsey für vollständige Graphen und zwei Farben, nämlich Satz 4.1, hatten wir eine Version für mehrere Farben angegeben. Dies kann man entsprechend auf den Satz von Ramsey für Mengen bzw. uniforme Hypergraphen übertragen. Genauer gilt:

- *Sei $r \in \mathbb{N}$ (Anzahl der Farben) mit $r \geq 2$ gegeben, ferner seien $i, n_1, \ldots, n_r \in \mathbb{N}$ mit $n_j \geq i$, $j = 1, \ldots, r$, gegeben. Dann gibt es eine (von $i, n_1, \ldots, n_r$ abhängende) Zahl $r \in \mathbb{N}$ mit folgender Eigenschaft: Ist $N \geq R$ eine natürliche Zahl und X eine N-elementige Menge, ist ferner*

$$\binom{X}{i} = C_1 \dot{\cup} \cdots \dot{\cup} C_r$$

eine beliebige disjunkte Zerlegung der i-elementigen Teilmengen von X, so existiert ein $j \in \{1, \ldots, r\}$ und $A_j \in \binom{X}{n_j}$ mit $\binom{A_j}{i} \subset C_j$.

Die kleinstmögliche Zahl R mit der in obiger Aussage angegebenen Eigenschaft wird mit $R_i(n_1, \ldots, n_r)$ bezeichnet.

Der Beweis der obigen Aussage erfolgt durch Induktion nach r, der Anzahl der Farben, mit denen die Elemente von $\binom{X}{i}$ gefärbt werden. Der Induktionsanfang liegt bei $r = 2$, was wegen Satz 4.3 gerechtfertigt ist. Wir nehmen an, die Aussage sei für $r - 1$ richtig und setzen (ähnlich wie bei der Färbung eines vollständigen Graphen durch mehr als zwei Farben)

$$R := R_i(n_1, \ldots, n_{r-2}, R_i(n_{r-1}, n_r)).$$

Wegen unserer Induktionsannahme ist R wohldefiniert und es gilt:

- Es gibt ein $j \in \{1, \ldots, r-2\}$ und eine Menge $A_j \in \binom{X}{n_j}$ mit der Eigenschaft, dass $\binom{A_j}{i} \subset C_j$

oder

- Es gibt eine Menge $A \in \binom{X}{R_i(n_{r-1}, n_r)}$ mit $\binom{A}{i} \subset C_1 \cup C_2$, d. h. die i-elementigen Teilmengen von A haben sozusagen eine "Mischfarbe". Dann gibt es n(nach Definition von $R_i(n_{r-1}, n_r)$) eine Menge $A_{r-1} \in \binom{A}{n_{r-1}}$ und der Eigenschaft, dass $\binom{A_{r-1}}{i} \subset C_{r-1}$ oder eine Menge $A_r \in \binom{A}{n_r}$ und der Eigenschaft, dass $\binom{A_r}{i} \subset C_r$.

Insgesamt ist der Satz von Ramsey für uniforme Hypergraphen und mehrere Farben bewiesen. Ferner ist

$$R_i(n_1, \ldots, n_r) \leq R_i(n_1, \ldots, n_{r-2}, R_i(n_{r-1}, n_r))$$

nachgewiesen worden. □

4.3 Der Satz von Erdös-Szekeres

Ziel dieses Unterabschnittes ist es, den folgenden Satz von P. Erdös, G. Szekeres (1935) zu beweisen.

Satz 4.4 (Erdös-Szekeres) *Sei $k \in \mathbb{N}$ mit $k \geq 4$ gegeben. Dann gibt es ein (von k abhängendes) $N \in \mathbb{N}$ mit der Eigenschaft, dass unter N Punkten in der Ebene in allgemeiner Lage es k Punkte gibt, die Eckpunkte eines konvexen k-gons bilden.*

Einen Überblick zum Satz von Erdös-Szekeres findet man bei W. Morris, V. Soltan (2000)[18]. Hierbei heißt eine Menge von Punkten der Ebene *in allgemeiner Lage*, wenn je drei von ihnen nicht kollinear sind, also nicht auf einer Geraden liegen. Ferner heißt ein Polygon mit k Ecken auch ein k-gon. Für $k = 3$ bzw. $k = 4$ nennen wir ein k-gon auch ein Dreieck bzw. Viereck.

Im folgenden Lemma betrachten wir den Spezialfall $k = 4$ und werden zeigen, dass es zu fünf Punkten in der Ebene eine Teilmenge von vier Punkten gibt, welche ein konvexes Viereck bilden. Diese Aussage wurde von P. Erdös *Happy Ending Theorem* genannt, da sie von Esther Klein, der Freundin und späteren Frau von G. Szekeres, gestellt (und für $k = 4$ gelöst) wurde. In Abbildung 13 geben wir links ein konvexes Viereck bzw. 4-gon, in der Mitte ein nicht konvexes 4-gon und rechts fünf Punkte an, zu denen wir vier Punkte finden können, die ein konvexes Viereck bilden.

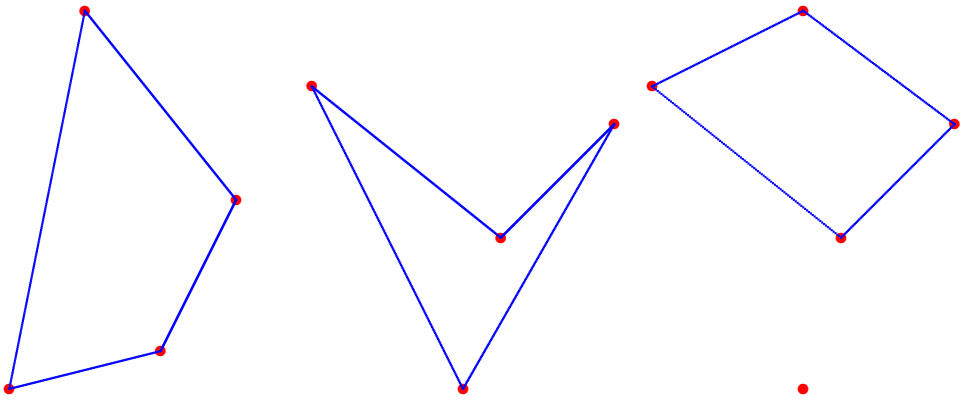


Abbildung 13: Illustration zum Happy Ending Theorem

Bei gegebenem $k \in \mathbb{N}$ mit $k \geq 4$ heißt die kleinste natürliche Zahl N mit der Eigenschaft, dass unter N Punkten in der Ebene in allgemeiner Lage es k Punkte gibt, die Eckpunkte eines konvexen k-gons bilden, die *Erdös-Szekeres-Zahl* und wird mit $ES(k)$ bezeichnet.

Satz 4.5 (Happy Ending Theorem) *Zu fünf Punkten in der Ebene in allgemeiner Lage gibt es eine Teilmenge von vier Punkten, welche die Ecken eines konvexen Vierecks sind.*

Beweis: Gegeben sind fünf Punkte in der Ebene in allgemeiner Lage. Man bilde die konvexe Hülle dieser fünf Punkte. Diese sei ein (konvexes) m-gon. Ist $m = 5$ oder $m = 4$, so sind wir fertig, siehe Abbildung 14. Sei daher $m = 3$, da die gegebenen Punkte in allgemeiner Lage sind, ist dies der einzig mögliche verbleibende Fall. Dann ist die konvexe Hülle der gegebenen 5 Punkte A, B, C, D und E ein Dreieck, etwa das Dreieck $\triangle ABC$, welches die restlichen beiden Punkte D und E im Innern (auf dem Rand können sie nicht liegen, da man sonst drei kollineare Punkte hätte) enthält. Die Gerade durch D und E teilt das Dreieck $\triangle ABC$ in zwei Teile. Eines dieser Teile enthält zwei Ecken des Dreiecks, etwa A und B. Dann ist $\square ABDE$ das gesuchte Viereck, siehe Abbildung 15. Damit ist das Happy Ending Theorem bewiesen. $\square$

[18]Diese Autoren schreiben zu Beginn ihres Artikels: The following problem attracts the attention of many mathematicians by its beauty and elementary character.

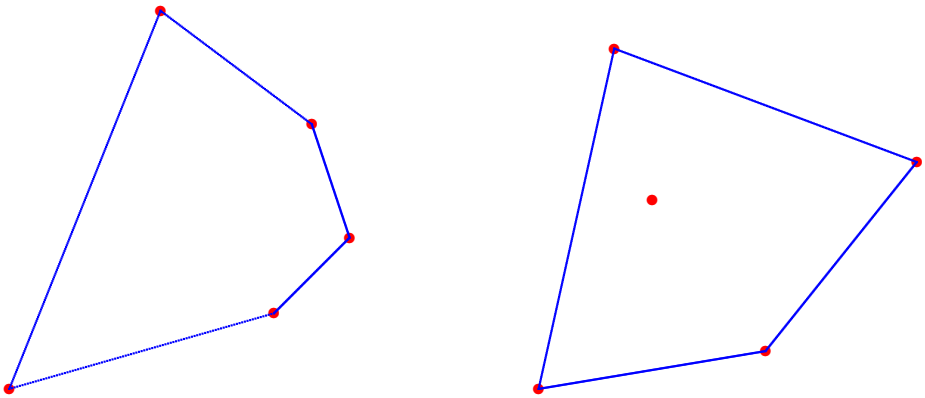


Abbildung 14: Konvexe Hülle von 5 Punkten

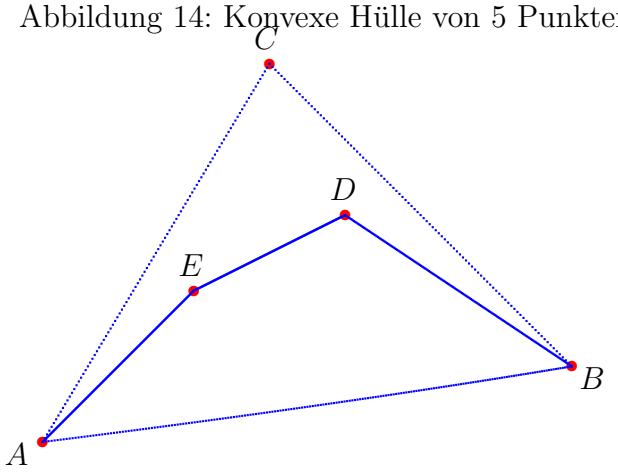


Abbildung 15: Ein Dreieck als konvexe Hülle von 5 Punkten

Beispiel: Das Happy Ending Theorem zeigt, dass $ES(4) \leq 5$. Da es natürlich ein nichtkonvexes Viereck gibt, ist $ES(4) > 4$ und damit insgesamt $ES(4) = 5$. In der Arbeit P. Erdös, G. Szekeres (1935) wird bemerkt, dass E. Makai $ES(5) = 9$ bewiesen hat. Klar ist, dass $ES(5) > 8$. Denn in der Konfiguration in Abbildung 16 haben wir 8 Punkte, zu denen wir kein konvexes Fünfeck finden können. Platziert man einen weiteren Punkt, so kann man sich in dieser speziellen Situation überlegen, dass dann ein konvexes Fünfeck erzeugt wird. Das ist aber natürlich kein Beweis für $ES(5) \leq 9$. Weiter ist bekannt, dass $ES(6) = 17$, siehe G. Szekeres, L. Peters (2006). Dies führt zu der (nach wie vor unbewiesenen und nicht widerlegten) Erdös-Szekeres Vermutung aus P. Erdös, G. Szekeres (1935), dass $ES(k) = 2^{k-2} + 1$ für alle $k \in \mathbb{N}$ mit $k \geq 4$. Von P. Erdös, G. Szekeres (1960) ist bewiesen worden, dass $2^{k-2} + 1 \leq ES(k)$ für alle $k \in \mathbb{N}$ mit $k \geq 4$. Dies geschieht dadurch, dass 2^{k-2} Punkte in der Ebene konstruiert werden, die kein konvexes k-gon enthalten. Dieses Ergebnis findet man auch bei W. Morris, V. Soltan (2000, Theorem 2.6). □

In der Arbeit von P. Erdös, G. Szekeres (1935) findet man, nachdem obiges Happy Ending Theorem bewiesen wurde und bevor in einen ersten Beweis (mit Hilfe des Ramsey Theorems) eingestiegen wird, die folgende Aussage:

- *Now it can be easily proved by induction that n points determine a convex polygon if and only if any 4 points of them form a convex quadrilateral.*

Abbildung 16: Es ist $ES(5) > 8$

Diese Aussage (bzw. eine Richtung davon) formulieren wir nun als Lemma:

Lemma 4.6 *In der Ebene seien n (paarweise verschiedene) Punkte in allgemeiner Lage gegeben. Für je 4 Punkte dieser n Punkte sei das hierdurch gebildete Viereck konvex. Dann sind die n Punkte Ecken eines konvexen n-gons.*

Beweis: Angenommen, die gegebenen n Punkte seien nicht Eckpunkte eines konvexen n-gons. Dann ist einer der n Punkte im Inneren (auf dem Rand kann er ja nicht liegen, da die gegebenen Punkte sich in allgemeiner Lage befinden) der konvexen Hülle $\mathcal{H}$ der n Punkte enthalten, dies sei etwa der Punkt P. Sei Q eine der Ecken von $\mathcal{H}$. Man verbinde Q mit allen anderen Ecken von H. Dadurch erhält man $\mathcal{H}$ als eine Vereinigung von Dreiecken. Der Punkt P liegt im Inneren eines dieser Dreiecke, dessen Ecken zu den gegebenen n Punkten gehören. Zusammen mit P erhält man vier Punkte, die ein nichtkonvexes Viereck bilden. Das ist ein Widerspruch zur Voraussetzung. □

Nun kommen zum **Beweis von Satz 4.4**. Wie angegeben sei $k \in \mathbb{N}$ mit $k \geq 4$ gegeben. Sei X eine Menge von N Punkten der Ebene in allgemeiner Lage, wobei $N \geq R_4(k, 5)$. Mit

$$\binom{X}{4} := \{Y \subset X : |Y| = 4\}$$

bezeichnen wir (wie schon bei entsprechenden Situationen im letzten Unterabschnitt) die Menge aller 4-elementigen Teilmengen von X. Die Menge $\binom{X}{4}$ ist disjunkte Vereinigung von Mengen C_1 und C_2, wobei

$$C_1 := \left\{Y \in \binom{X}{4} : \text{Das durch } Y \text{ bestimmte Viereck ist konvex}\right\},$$
$$C_2 := \left\{Y \in \binom{X}{4} : \text{Das durch } Y \text{ bestimmte Viereck ist nicht konvex}\right\}.$$

Wegen des Satzes 4.3 von Ramsey für Mengen (setze dort $i := 4$ und $l := 5$) existiert

- $A_1 \in \binom{X}{k}$ und der Eigenschaft, dass $\binom{A_1}{4} \subset C_1$ (bzw. für jede 4-elementige Teilmenge von A_1 das hierdurch bestimmte Viereck konvex ist)

oder es existiert

- $A_2 \in \binom{X}{5}$ und der Eigenschaft, dass $\binom{A_2}{4} \subset C_2$ (bzw. für jede 4-elementige Teilmenge von A_2 das hierdurch erzeugte Viereck nicht konvex ist).

Wegen des Happy Ending Theorems 4.4 ist die zweite Alternative nicht möglich. Daher ist die Existenz einer k-elementigen Teilmenge A_1 der vorgegebenen Menge X von N Punkten in allgemeiner Lage gesichert, welche die Eigenschaft hat, dass jede 4-elementige Teilmenge von A_1 ein konvexes Viereck bestimmt. Wegen Lemma 4.6 sind die k Elemente von A_1 Ecken eines konvexen k-gons und der Satz ist bewiesen. □

4.4 Eine untere Schranke für die Ramsey-Zahl $R(k,k)$

In diesem kurzen Unterabschnitt wollen wir ein bemerkenswertes Ergebnis von P. Erdös (1947, Theorem I) angeben, siehe auch M. Aigner, G. M. Ziegler (2018, S. 347). Das wesentliche neue Ergebnis ist eine *untere Schranke* für die Ramsey-Zahl $R(k,k)$.

Satz 4.7 *Für $k \geq 3$ ist*

$$2^{k/2} < R(k,k) \leq \binom{2k-2}{k-1} < 4^{k-1}.$$

Beweis: Die Abschätzung

$$R(k,k) \leq \binom{2k-2}{k-1}$$

ist schon in Teil 4. von Satz 4.2 bewiesen worden. Die rechte Ungleichung zeigen wir durch vollständige Induktion nach k. Für $k = 3$ ist sie richtig. Für den Induktionsschluss von k nach $k+1$ beachten wir, dass

$$\binom{2k}{k} = \underbrace{\binom{2k-2}{k-1}}_{<4^{k-1}} \cdot \underbrace{\frac{(2k-1)(2k)}{k^2}}_{<4} < 4^k.$$

Es bleibt, $2^{k/2} < R(k,k)$ für alle $k \geq 2$ nachzuweisen. Wegen $R(3,3) = 6$ ist dies für $k = 3$ richtig. Sei daher jetzt $k \geq 4$ und $N \in \mathbb{N}$ beliebig mit $N \leq 2^{k/2}$. Unser Ziel ist es, zu zeigen:

- K_N besitzt eine rot-blaue Kantenfärbung mit der Eigenschaft, dass es keinen monochromen (roten oder blauen) vollständigen Teilgraphen K_k enthält.

Um dieses Ziel zu erreichen, zählen wir die Anzahl der Färbungen des K_N, welche einen monochromen (roten oder blauen) vollständigen Teilgraphen K_k besitzen. Die Ecken von K_N seien $\{1,\ldots,N\}$ und $S \subset \{1,\ldots,N\}$ beliebig mit $|S| = k$, mit K_S bezeichnen wir den durch S induzierten vollständigen Teilgraphen von K_N. Sei A_S die Menge der rot-blau-Färbungen des K_N mit der Eigenschaft, dass K_S ein monochromer, also rot

oder blau gefärbter, Teilgraph ist. Es gibt 2 Färbungs-Möglichkeiten, nämlich rot oder blau, für die Kanten eines monochromen K_S. Die verbleibenden

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

Kanten des K_N, die nicht zu K_S gehören, können jeweils auf zweierlei Weise gefärbt werden. Daher ist

$$|A_S| = 2 \cdot 2^{\binom{N}{2}-\binom{k}{2}}.$$

Wegen

$$\left| \bigcup_{\substack{S \subset \{1,\ldots,N\} \\ |S|=k}} A_S \right| \leq \sum_{\substack{S \subset \{1,\ldots,N\} \\ |S|=k}} |A_S| = \binom{N}{k} \cdot 2 \cdot 2^{\binom{N}{2}-\binom{k}{2}}$$

ist die Zahl der rot-blau-Färbungen des K_N, welche einen monochromen (roten oder blauen) K_k enthalten, höchstens gleich $\binom{N}{k} \cdot 2 \cdot 2^{\binom{N}{2}-\binom{k}{2}}$. Die Gesamtzahl der rot-blau-Färbungen des K_N ist $2^{\binom{N}{2}}$. Ist daher

$$\binom{N}{k} \cdot 2 \cdot 2^{\binom{N}{2}-\binom{k}{2}} < 2^{\binom{N}{2}}$$

bzw.

$$(*) \qquad \binom{N}{k} \cdot 2^{-\binom{k}{2}} < \frac{1}{2},$$

so gibt es eine rot-blau-Färbung des K_N, welche keinen monochromen Teilgraphen K_k enthält. Nun ist aber

$$\begin{aligned}
\binom{N}{k} \cdot 2^{-\binom{k}{2}} &= \frac{N(N-1)\cdots(N-k+1)}{k!} \cdot 2^{-\binom{k}{2}} \\
&\leq \frac{N^k}{k!} \cdot 2^{-\binom{k}{2}} \\
&\leq \frac{N^k}{2^{k-1}} \cdot 2^{-\binom{k}{2}} \\
&< 2^{k^2/2-\binom{k}{2}-k+1} \\
&\qquad \text{(wegen } N < 2^{k/2}\text{)} \\
&= 2^{1-k/2} \\
&\leq \frac{1}{2} \\
&\qquad \text{(wegen } k \geq 4\text{)}.
\end{aligned}$$

Damit ist der Satz schließlich bewiesen. □

Bemerkung: Die untere Schranke für die Ramsey-Zahl $R(k,k)$ kann offenbar noch ein wenig verbessert werden. Wie der Beweis von Satz 4.7 nämlich zeigt, gilt sogar:

- *Ist $N \in \mathbb{N}$ und*

$$\binom{N}{k} \cdot 2^{-\binom{k}{2}} < \frac{1}{2},$$

so ist $R(k,k) > N$.

□

Beispiel: Die Abschätzung von Satz 4.7 liefert $6 \leq R(5,5) \leq 70$. Wegen $\binom{11}{5} \cdot 2^{-\binom{5}{2}} \approx 0.4512$ liefert die anschließende Bemerkung die untere Schranke $R(5,5) > 11$ bzw. $R(5,5) \geq 12$. Der genaue Wert von $R(5,5)$ ist bisher nicht bekannt, die beste bekannte Abschätzung ist $43 \leq R(5,5) \leq 48$, siehe V. Angeltveit, B. D. McKay (2017). Hier findet man auch die Aussage:

- *The actual value of $R(5,5)$ is widely believed to be 43, because a lot of computer resources have been expended in an unsuccessful attempt to construct a Ramsey(5,5)-graph of order 43.*

Eine oft zitierte Aussage von P. Erdös ist die folgende:

- *Suppose aliens invade the earth and threaten to obliterate it in a year's time unless human beings can find the Ramsey number for red five and blue five. We could marshal the world's best minds and fastest computers, and within a year we could probably calculate the value. If the aliens demanded the Ramsey number for red six and blue six, however, we would have no choice but to launch a preemptive attack.*

Weitere Aussprüche von P. Erdös findet man bei `http://www.azquotes.com/author/4538-Paul_Erdos`. Über den ganz ungewöhnlichen Mathematiker Paul Erdös findet man im Internet sehr viel Interessantes. Wir geben nur die Internetseite `http://www.math.ucsd.edu/~erdosproblems/About.html` an. □

4.5 Ein Satz von Schur

Wie zu Beginn des Unterabschnitts 3.2 benutzen wir für $a, b \in \mathbb{Z}$ die Bezeichnung $[a,b]$ für die Menge ganzer Zahlen z mit $a \leq z \leq b$. In diesem sehr kurzen Unterabschnitt[19] formulieren und beweisen wir einen Satz von I. Schur (1917), der als erster Satz der Ramsey-Theorie gilt.

Satz 4.8 (Schur) *Sei $r \in \mathbb{N}$ (Anzahl der Farben) mit $r \geq 2$ gegeben. Dann gibt es $s \in \mathbb{N}$ derart, dass es zu einer beliebigen Färbung $c\colon [1,s] \longrightarrow [1,r]$ der Zahlen $[1,s]$ durch r Farben ein monochromes Tripel (x,y,z) natürlicher Zahlen aus $[1,s]$ gibt mit $x+y=z$ (sowie $c(x)=c(y)=c(z)$).*

Beweis: Sei

$$N := R(\underbrace{3,\ldots,3}_{r}) =: R(3;r), \qquad s := N-1.$$

[19] Eine ausführlichere Darstellung findet man z. B. bei B. M. Landman, A. Robertson (2014, Chapter 8).

Ist dann $c\colon [1,s] \longrightarrow [1,r]$ eine Färbung der Zahlen $[1,s]$ durch r Farben, so erhält man eine r-Färbung des K_N, indem man der Kante (i,j) die Farbe $c(|i-j|) \in [1,r]$ gibt. Wegen des Satzes von Ramsey für mehrere Farben existiert ein monochromes Dreieck $\triangle(i,j,k)$, es sei $1 \leq i < j < k \leq s$. Wir setzen

$$x := j - i, \quad y := k - j, \quad z := k - i.$$

Dann ist $x + y = z$ und $c(x) = c(y) = c(z)$, wir haben also das gesuchte monochrome Tripel gefunden. □

Bemerkung: Bei gegebenem $r \in \mathbb{N}$ nennt man die kleinste natürliche Zahl $s = s(r)$ mit der Eigenschaft, dass es zu einer beliebigen r-Färbung von $[1:s]$ ein monochromes Schur-Tripel (x,y,z) aus $[1,s]$ mit $x + y = z$ gibt, *Schur-Zahl*. Wir haben im Beweis des Satzes von Schur nachgewiesen, dass $s(r) \leq R(3;r) - 1$. □

Beispiel: Die einzigen bekannten Schur-Zahlen sind $s(1) = 2$, $s(2) = 5$, $s(3) = 14$ und $s(4) = 45$. Wir wollen uns hier davon überzeugen, dass $s(2) = 5$. Zunächst ist $s(2) \geq 5$. Denn bei einer Färbung der Zahlen $[1,4]$ durch zwei Farben, etwa rot und blau, gemäß

1	2	3	4
R	B	B	R

gibt es kein Schur-Tripel. Nun wollen wir zu einer beliebigen rot-blau-Färbung von $[1,5]$ die Existenz eines Schur-Tripels zeigen. O. B. d. A. sei 1 rot gefärbt. Wir nehmen im Gegensatz zu der Behauptung an, in $[1,5]$ gäbe es kein Schur-Tripel. Wegen $1+1=2$ muss 2 blau gefärbt sein, da $2+2=4$ ist 4 rot gefärbt. Schließlich muss 5 wegen $1+4=5$ blau gefärbt sein. Daher muss nur noch 3 gefärbt werden. Ist 3 rot, so ist $(1,3,4)$ ein rotes Schur-Tripel. Ist dagegen 3 blau gefärbt, so ist $(2,3,5)$ ein blaues Schur-Tripel. Damit ist $s(2) = 5$ nachgewiesen. □

4.6 Der Satz von Rado

Durch den Satz von Rado (siehe R. Rado (1933)) wird die folgende Frage beantwortet:

- Sei $A \in \mathbb{Z}^{m\times n}$, also A eine $m \times n$-Matrix mit ganzzahligen Einträgen. Unter welchen Voraussetzungen an A besitzt die Gleichung $Ax = 0$ für beliebiges $r \in \mathbb{N}$ und eine beliebige Abbildung $c\colon \mathbb{N} \longrightarrow [1,r]$ eine Lösung $x = (x_j) \in \mathbb{N}^n$ mit $c(x_1) = \cdots = c(x_n)$? Mit anderen Worten: Die natürlichen Zahlen seien mit beliebigem $r \in \mathbb{N}$ durch r Farben gefärbt. Unter welchen Voraussetzungen besitzt das homogene Gleichungssystem $Ax = 0$ eine monochrome Lösung $x \in \mathbb{N}^n$?

Ist z. B. $A := \begin{pmatrix} 1 & 1 & -1 \end{pmatrix}$, so fragen wir, ob es bei einer Färbung von $\mathbb{N}$ durch endlich viele Farben ein Schurtripel gibt, also ein monochromes Tripel $\{x_1,x_2,x_3\}$ mit $x_1 + x_2 = x_3$. Durch den Satz 4.8 von Schur wird hierauf eine positive Antwort gegeben.

4.6.1 Partitionsreguläre Matrizen, Spaltenbedingung

In der folgenden Definition werden wir zunächst der gewünschten Eigenschaft der Matrix $A \in \mathbb{Z}^{m\times n}$ den Namen *partitionsregulär* gegeben, danach wird die sogenannte *Spaltenbedingung* formuliert.

Definition 4.9 1. Eine Matrix $A \in \mathbb{Z}^{m\times n}$ heißt *partitionsregulär*, wenn es zu jeder endlichen Zerlegung der natürlichen Zahlen, etwa durch

$$\mathbb{N} = N_1 \dot\cup \cdots \dot\cup N_r,$$

ein $x = (x_j) \in \mathbb{N}^n$ und ein $k \in [1, r]$ mit $Ax = 0$ und $x_1, \ldots, x_n \in N_k$ gibt. Anders gesagt ist eine Matrix $A \in \mathbb{Z}^{m\times n}$ partitionsregulär, wenn es zu jedem $r \in \mathbb{N}$ und jeder (Färbungs-) Funktion $c\colon \mathbb{N} \longrightarrow [1, r]$ eine monochrome Lösung $x \in \mathbb{N}^n$ von $Ax = 0$ gibt, also eine Lösung $x = (x_j)$ mit $c(x_1) = \cdots = c(x_n)$.

2. Eine Matrix $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{Z}^{m\times n}$ (mit $a_j \in \mathbb{Z}^m$ wird also die j-te Spalte von A bezeichnet) genügt der *Spaltenbedingung*, wenn es eine Zerlegung

$$[1, n] = D_1 \dot\cup \cdots \dot\cup D_s$$

der Spaltenindizes in disjunkte, nichtleere Mengen $D_1, \ldots, D_s$ gibt derart, dass

$$\sum_{j\in D_1} a_j = 0$$

und

$$\sum_{j\in D_k} a_j \in \operatorname{span}_Q\{a_j : j \in D_1 \cup \cdots \cup D_{k-1}\}, \qquad k = 2, \ldots, s.$$

Hierbei verstehen wir unter $\operatorname{span}_{\mathbb{Q}}\{a_j : j \in D\}$ die Menge der rationalen Linearkombinationen von $\{a_j\}_{j\in D}$.

Bemerkung: Eine Matrix $A \in \mathbb{Z}^{m\times n}$ genügt offenbar genau dann der Spaltenbedingung, wenn sie nach einer eventuellen Permutation der Spalten die Form

$$A = \begin{pmatrix} A^{(1)} & A^{(2)} & \cdots & A^{(r)} \end{pmatrix}$$

besitzt, wobei die Summe der Spalten von $A^{(1)}$ der Nullvekor ist und für $k = 2, \ldots, r$ die Summe der Spalten von $A^{(k)}$ eine rationale Linearkombination der Spalten von

$$\begin{pmatrix} A^{(1)} & \cdots & A^{(k-1} \end{pmatrix}$$

ist. □

Beispiele: 1. Die Matrix $A := \begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \in \mathbb{Z}^{1\times 3}$ genügt der Spaltenbedingung. Denn mit der Zerlegung

$$[1, 3] = D_1 \dot\cup D_2$$

mit

$$D_1 := \{1, 3\}, \qquad D_2 := \{2\}$$

sind die entsprechenden Bedingungen erfüllt. Die Matrix A ist wegen des Satzes 4.8 von Schur offenbar partitionsregulär.

2. Die Matrix

$$A := \begin{pmatrix} 1 & 0 & 2 & -3 & 2 \\ 0 & -1 & 2 & -2 & 1 \\ 4 & 2 & -5 & 1 & 6 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \end{pmatrix}$$

genügt der Spaltenbedingung. Man setze nämlich

$$D_1 := \{1, 3, 4\}, \qquad D_2 := \{2, 5\}.$$

Dann ist

$$\sum_{j \in D_1} a_j = 0, \qquad \sum_{j \in D_2} a_j = \begin{pmatrix} 2 \\ 0 \\ 8 \end{pmatrix} = 2 \cdot a_1.$$

Dagegen genügt die Matrix

$$A := \begin{pmatrix} 1 & 3 & 0 & -3 & 1 \\ -2 & -1 & 2 & 7 & 0 \\ 5 & 3 & -5 & 2 & -3 \\ 1 & -2 & 3 & 4 & -5 \end{pmatrix}$$

nicht der Spaltenbedingung. Denn es gibt keine Menge von Spalten, die summiert den Nullvektor ergeben.

3. Die Matrix

$$A := \begin{pmatrix} -2 & 1 & 1 & 3 & 0 & 1 \\ 1 & -2 & 1 & 0 & -3 & 1 \\ 1 & 1 & -2 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{pmatrix}$$

genügt der Spaltenbedingung. Denn setzt man

$$D_1 := \{1, 2, 3\}, \quad D_2 := \{4, 5\}, \quad D_3 := \{6\},$$

so ist

$$\sum_{j \in D_1} a_j = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \sum_{j \in D_2} a_j = \begin{pmatrix} 3 \\ -3 \\ 0 \end{pmatrix} = 2a_2 + a_3$$

und

$$\sum_{j \in D_3} a_j = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{3}(a_4 - a_5).$$

4. Für einzeilige Matrizen kann die Spaltenbedingung einfach charakterisiert werden:

- *Eine Matrix* $A = \begin{pmatrix} a_2 & \cdots & a_n \end{pmatrix} \in \mathbb{Z}^{1 \times n}$ *genügt genau dann der Spaltenbedingung, wenn* $A = 0$ *oder es eine nichtleere Indexmenge* $J \subset [1, n]$ *mit* $a_j \neq 0$, $j \in J$, *und* $\sum_{j \in J} a_j = 0$ *gibt.*

Denn sei die Spaltenbedingung wie in der Definition erfüllt. Sind in D_1 Indizes von Null verschiedener a_j enthalten, so kann man $J := \{j \in D_1 : a_j \neq 0\}$ setzen. Ist dies nicht der Fall, ist also $a_j = 0$ für alle $j \in D_1$, so ist $\sum_{j \in D_2} a_j = 0$. Wenn $\{j \in D_2 : a_j \neq 0\} \neq \emptyset$, so sind wir fertig, andernfalls kann man entsprechend fortfahren und erhält, dass $A = 0$ oder eine nichtleere Indexmenge $J \subset [1, n]$ mit $a_j \neq 0$, $j \in J$, und $\sum_{j \in J} a_j = 0$ existiert. Ist umgekehrt $A \neq 0$ (andernfalls genügt A trivialerweise

der Spaltenbedingung) und existiert eine nichtleere Indexmenge $J \subset [1,n]$ mit $a_j \neq 0$, $j \in J$, und $\sum_{j\in J} a_j = 0$, so setze man $D_1 := J$ und $D_2 := [1,n] \setminus J$. Mit einem beliebigen $k \in J = D_1$ ist dann $\sum_{j\in D_1} a_j = 0$ und

$$\sum_{j\in D_2} a_j = \frac{\sum_{j\in D_2} a_j}{a_k} a_k \in \operatorname{span}_{\mathbb{Q}}\{a_j : j \in D_1\}.$$

Also ist die Spaltenbedingung erfüllt. Daher erfüllt $A := \begin{pmatrix} 1 & 1 & -1 \end{pmatrix}$ die Spaltenbedingung, während $A := \begin{pmatrix} 1 & -2 \end{pmatrix}$ der Spaltenbedingung nicht genügt. Wir wollen uns davon überzeugen, dass diese Matrix A auch nicht partitionsregulär ist. Hierzu färben wir die Menge $\mathbb{N}$ der natürlichen Farben durch zwei Farben, etwa rot und blau. Dies machen wir so, dass für jedes $n \in \mathbb{N}$ die Zahlen n und $2n$ unterschiedliche Farben haben und damit keine monochrome Lösung zu

$$Ax = \begin{pmatrix} 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 - 2x_2 = 0$$

existieren kann. Dies erreichen wir, indem wir die ungeraden Zahlen rot färben, während wir eine gerade Zahl $2n$ blau färben, wenn n rot gefärbt ist, andernfalls blau. Damit haben n und $2n$ verschiedene Farben. Wir erhalten die folgende Färbung von $\mathbb{N}$:

$$\mathbb{N} = \{1,2,3,4,5,6,7,8,9,\ldots\}.$$

□

Ziel in diesem Unterabschnitt ist es, den folgenden Satz zu beweisen.

Satz 4.10 (Rado) *Eine Matrix $A \in \mathbb{Z}^{m\times n}$ ist genau dann partitionsregulär, wenn sie der Spaltenbedingung genügt.*

Beispiel: Mit vorgegebenem $s \in \mathbb{N}$ sei

$$A := \begin{pmatrix} -1 & 2 & -1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 & 0 \\ -s & s & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 \end{pmatrix} \in \mathbb{Z}^{(n-2)\times n}.$$

Dann genügt A der Spaltenbedingung. Denn bezeichnet man mit $a_1,\ldots,a_n \in \mathbb{Z}^{n-2}$ die Spalten von A, so ist mit

$$D_1 := \{1,2,\ldots,n-1\}, \qquad D_2 := \{n\}$$

einerseits

$$[1,n] = D_1 \,\dot\cup\, D_2$$

und andererseits

$$\sum_{j\in D_1} a_j = 0$$

sowie

$$\sum_{j\in D_2} a_j = a_n = \frac{1}{s}a_1 - \sum_{j=3}^{n-1}\frac{j-2}{s}a_j \in \operatorname{span}_{\mathbb{Q}}\{a_j : j \in D_1\}.$$

Wegen des (noch unbewiesenen) Satzes von Rado ist A partitionsregulär. Bei einer beliebigen Färbung von $\mathbb{N}$ durch endlich viele Farben gibt es also eine monochrome Lösung $x = (x_j) \in \mathbb{N}^n$ von $Ax = 0$. Dies bedeutet, dass

$$-x_j + 2x_{j+1} - x_{j+2} = 0, \qquad j \in [1, n-3],$$

und

$$-sx_1 + sx_2 - x_n = 0.$$

Also ist $x_n = s(x_2 - x_1)$ und folglich $d := x_2 - x_1 \in \mathbb{N}$ und $x_n = sd$. Aus

$$x_{j+2} - x_{j+1} = x_{j+1} - x_j, \qquad j \in [1, n-3],$$

erhalten wir

$$x_j = x_1 + (j-1)(x_2 - x_1) = x_1 + (j-1)d, \qquad j \in [1, n-1].$$

Wir haben also aus dem (noch unbewiesenen!) Satz von Rado erhalten, dass es eine monochrome arithmetische Progression $\{x_1, x_1 + d, \ldots, x_1 + (n-2)d\}$ gibt und sd dieselbe Farbe hat. Dies ist genau die Aussage des verschärften Satzes von van der Waerden, siehe Satz 3.4. □

4.6.2 Der Satz von Rado für eine Gleichung

Im folgenden Satz (wir folgen im wesentlichen I. Leader (2013, S. 15)) zeigen wir für den Spezialfall $m = 1$ eine Richtung des Satzes von Rado, dass nämlich die Spaltenbedingung eine *notwendige* Bedingung dafür ist, dass eine Matrix $A \in \mathbb{Z}^{1\times n}$ partitionsregulär ist. Hierbei können wir o. B. d. A. annehmen, dass die Einträge in der einen Zeile von A sämtlich von 0 verschieden sind, da man das Problem andernfalls auf eine niederdimensionale Aufgabe zurückführen kann.

Satz 4.11 *Seien $a_1, \ldots, a_n \in \mathbb{Z} \setminus \{0\}$ und sei $A := \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{Z}^{1\times n}$ partitionsregulär. Dann existiert eine nichtleere Menge $J \subset [1, n]$ derart, dass $\sum_{j\in J} a_j = 0$, d. h. (siehe obiges Beispiel) A genügt der Spaltenbedingung.*

Beweis: Man wähle eine Primzahl p mit $p > \sum_{j=1}^{n} |a_j|$. Die Abbildung

$$d\colon \mathbb{N} \longrightarrow [1, p-1]$$

sei auf die folgende Weise definiert:

- Sei $x \in \mathbb{N}$. Man bilde die p-adische Entwicklung von x, stelle x also dar als

$$x = d_r p^r + d_{r-1} p^{r-1} + \cdots + d_1 p + d_0 \quad \text{mit} \quad d_i \in [0, p-1], \quad i = 1, \ldots, r.$$

  Man beachte, dass nicht alle d_i gleich Null sein können, da $x \in \mathbb{N}$ und damit $x \neq 0$.

- Sei $L(x) := \min\{i \in [0, r] : d_i \neq 0\}$ und $d(x) := d_{L(x)}$, also $d(x)$ der niedrigste von Null verschiedene Koeffizient in der p-adischen Entwicklung von x.

Jetzt denken wir uns $\mathbb{N}$ durch die Abbildung $d\colon \mathbb{N} \longrightarrow [1, p-1]$ gefärbt. Da A als partitionsregulär vorausgesetzt wurde, existiert ein $x = (x_j) \in \mathbb{N}^n$ mit $Ax = 0$ bzw. $\sum_{j=1}^n a_j x_j = 0$, welches bezüglich dieser Färbung monochrom ist, es gelte also $d(x_j) = d$, $j \in [1, n]$, mit $d \in [1, p-1]$. Dann ist $J \subset [1, n]$ nichtleer. Wir wollen uns überlegen, dass $\sum_{j \in J} a_j = 0$, womit die Behauptung bewiesen sein wird.
Da $\sum_{j=1}^n a_j x_j = 0$, teilt p^{L+1} trivialerweise $\sum_{j=1}^n a_j x_j$, wir schreiben hierfür

$$p^{L+1} \,\Big|\, \sum_{j=1}^{n} a_j x_j.$$

Bei der p-adischen Entwicklung von x_j können wir annehmen, dass der höchste Koeffizientenindex r für alle j derselbe ist, da man notfalls mit verschwindenden Koeffizienten auffüllen kann. Wegen $L \leq L(x_j)$, $j \in [1, n]$, können wir entsprechend annehmen, dass der niedrigste Koeffizientenindex für alle j derselbe, nämlich L, ist. Daher können wir annehmen, dass x_j die p-adische Entwicklung

$$x_j = d_{r,j} p^r + d_{r-1,j} p^{r-1} + \cdots + d_{L+1,j} p^{L+1} + d_{L,j} p^L, \qquad j \in [1, n]$$

besitzt. Für $j \notin J$ ist der niedrigste Koeffizientenindex $\geq L+1$, so dass $p^{L+1} | \sum_{j \notin J} a_j x_j$, damit $p^{L+1} | \sum_{j \in J} a_j x_j$ und folglich

$$p \,\Big|\, \sum_{j \in J} a_j d_{L,j}.$$

Wegen $d_{L,j} = d$, $j \in J$, gilt

$$p \,\Big|\, \sum_{j \in J} a_j d.$$

Nun ist $d \in [1, p-1]$ und folglich gilt

$$p \,\Big|\, \sum_{j \in J} a_j.$$

Da wir die Primzahl p so gewählt haben, dass $p > \sum_{j=1}^n |a_j| \geq \sum_{j \in J} |a_j|$, folgt $\sum_{j \in J} a_j = 0$. Der Satz ist damit bewiesen. □
Das folgende Lemma benötigen wir zum Beweis der Aussage, dass eine die Spaltenbedingung erfüllende Matrix $A \in \mathbb{Z}^{1 \times n}$ partitionsregulär ist.

Lemma 4.12 *Sei $\lambda \in \mathbb{Q}$ gegeben. Dann existiert zu jedem $r \in \mathbb{N}$ ein $n \in \mathbb{N}$ mit der Eigenschaft, dass es zu jeder Färbungsfunktion $c\colon [1,n] \longrightarrow [1,r]$ ein monochromes Tripel $\{x,y,z\} \subset [1,n]$ mit $x + \lambda y = z$ gibt.*

Beweis: O. B. A. ist $\lambda > 0$. Denn für $\lambda = 0$ ist die Aussage trivial, während wir für $\lambda < 0$ die Bestimmungsgleichung $x + \lambda y = z$ auch als $z - \lambda y = x$ schreiben können (man vertausche also x und z und beachte, dass $-\lambda > 0$). Sei

$$\lambda = \frac{p}{q} \quad \text{mit } p, q \in \mathbb{N}.$$

Wir beweisen das Lemma durch vollständige Induktion nach r, der Anzahl der Farben. Für $r = 1$ ist die Aussage richtig, denn mit $n := \max(p,q) + 1$ sowie

$$x := 1, \quad y := q, \quad z := p + 1$$

ist $\{x,y,z\} \subset [1,n]$ ein monochromes (es gibt nur eine Farbe!) Tripel mit

$$x + \lambda y = 1 + \frac{p}{q} q = 1 + p = z.$$

In der Induktionsannahme gehen wir davon aus, dass die Aussage für $r-1$ richtig ist. Es existiert also ein $n_0 \in \mathbb{N}$ mit der Eigenschaft, dass bei beliebiger Färbung von $[1,n_0]$ durch $r-1$ Farben es ein monochromes Tripel $\{x_0,y_0,z_0\} \subset [1,n_0]$ mit $x_0 + \lambda y = z$ gibt. Für den Induktionsschluss setzen wir $t := \max(p,q)$ und $n := W(r, n_0 t + 1)$, wobei $W(r,k)$ die nach Satz 3.1 von van der Waerden existierende van der Waerden Zahl ist. Sei $c\colon [1,n] \longrightarrow [1,r]$ eine beliebige Färbungsfunktion. Nach Definition von n bzw. wegen des Satzes von van der Waerden existiert in $[1,n]$ eine monochrome arithmetische Progression der Länge $n_0 t + 1$, sie sei mit

$$\{a, a+d, \ldots, a + n_0 t d\}$$

bezeichnet. Sei $i \in [1,r]$ die Farbe dieser arithmetischen Progression. Für die in $[1,n]$ enthaltene Progression

$$\{qd, 2qd, \ldots, n_0 qd\}$$

der Länge n_0 gibt es zwei Möglichkeiten:

(a) Es existiert $j \in [1,n_0]$ mit $c(jqd) = i$.

Dann hat man ein gesuchtes Tripel gefunden, man setze nämlich

$$x := a, \quad y := jqd, \quad z := a + jpd.$$

Dann ist $\{x,y,z\} \subset [1,n]$, da $x = a \in [1:n]$ und y, z natürliche Zahlen mit

$$y = jqd \leq n_0 qd \leq n_0 td \leq a + n_0 td \leq n$$

und

$$z = a + jpd \leq a + n_0 pd \leq a + n_0 td \leq n$$

sind. Als Elemente der Progression $\{a, a+d, \ldots, n_0 td\}$ haben x und z dieselbe Farbe i, die auch y hat. Daher ist $\{x,y,z\}$ ein monochromes Tripel. Ferner ist

$$x + \lambda y = a + \frac{p}{q} jqd = a + pjd = z.$$

(b) Es existiert kein $j \in [1, n_0]$ mit $c(jqd) = i$.

Dann ist die arithmetische Progession $\{qd, 2qd, \ldots, n_0qd\} \subset [1, n]$ durch maximal $r-1$ Farben mit Hilfe von $c\colon [1, n] \longrightarrow [1, r]$ gefärbt. Wir definieren

$$c_0\colon [1, n_0] \longrightarrow [1, r-1]$$

durch

$$c_0(j) := c(jqd), \quad j \in [1, n_0].$$

Nach Induktionsannahme existiert ein Tripel $\{x_0, y_0, z_0\} \subset [1, n_0]$ mit

$$c_0(x_0) = c_0(y_0) = c_0(z_0), \qquad x_0 + \lambda y_0 = z_0.$$

Dann ist durch

$$x := x_0qd, \quad y := y_0qd, \quad z := z_0qd$$

das gesuchte (bezüglich c) monochrome Tripel $\{x, y, z\} \subset [1, n]$ mit $x + \lambda y = z$ gefunden. Das Lemma ist bewiesen. □

Durch den folgenden Satz wird der Satz von Rado für den Fall einer einzigen Gleichung bewiesen sein.

Satz 4.13 *Seien $a_1, \ldots, a_n \in \mathbb{Z} \setminus \{0\}$. Dann ist die Matrix $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{Z}^{1\times n}$ genau dann partitionsregulär, wenn sie der Spaltenbedingung genügt bzw. eine nichtleere Menge $J \subset [1, n]$ mit $\sum_{j\in J} a_j = 0$ existiert.*

Beweis: Durch Satz 4.11 haben wir schon bewiesen, dass eine partitionsreguläre Matrix $A \in \mathbb{Z}^{1\times n}$ der Spaltenbedingung genügt. Daher nehmen wir jetzt an, es existiere eine nichtleere Menge $J \subset [1, n]$ mit $\sum_{j\in J} a_j = 0$ und zeigen, dass A partitionsregulär ist. Wir können annehmen, dass J eine *echte* Teilmenge von $[1, n]$ ist, da andernfalls A wegen $Ae = 0$, wobei e der Vektor in $\mathbb{N}^n$ ist, dessen Komponenten alle gleich 1 sind, trivialerweise partitionsregulär ist. Die Menge $\mathbb{N}$ der natürlichen Zahlen sei mit endlich vielen Farben gefärbt. Wir wählen ein beliebiges $j_0 \in J$ und definieren

$$\lambda := \frac{\sum_{j\notin J} a_j}{a_{j_0}}.$$

Wegen Lemma 4.12 existiert ein monochromes Tripel $\{x, y, z\} \subset \mathbb{N}$ mit $x+\lambda y = z$. Nun definieren wir $x \in \mathbb{N}^n$ (die zweierlei Bedeutungen von x, einmal aus $\mathbb{N}$, zum anderen aus $\mathbb{N}^n$, sollten zu keinen größeren Verwirrungen führen) durch

$$x_j := \begin{cases} x, & j = j_0, \\ y, & j \notin J, \\ z, & j \in J \setminus \{j_0\}. \end{cases}$$

Offensichtlich ist $x \in \mathbb{N}^n$ monochrom. Weiter ist

$$\begin{aligned}
Ax &= \sum_{j=1}^{n} a_j x_j \\
&= a_{j_0} x_{j_0} + \sum_{j \notin J} a_j x_j + \sum_{j \in J \setminus \{j:0\}} a_j x_j \\
&= a_{j_0} x + \left(\sum_{j \notin J} a_j \right) y + \left(\sum_{j \in J \setminus \{j_0\}} a_j \right) z \\
&= a_{j_0} \Bigg(\underbrace{x + \frac{\sum_{j \notin J} a_j}{a_{j_0}} y}_{=z} + \frac{\sum_{j \in J \setminus \{j_0\}} a_j}{a_{j_0}} z \Bigg) \\
&= \left(\sum_{j \in J} a_j \right) z \\
&= 0.
\end{aligned}$$

Damit ist der Satz bewiesen. □

4.6.3 Der allgemeine Fall

Der folgende Satz ist das Analogon von Satz 4.11 für ein System von Gleichungen, durch ihn wird eine Richtung des Satzes von Rado bewiesen sein. Als Hilfsmittel für den Beweis formulieren wir aber vorher noch das folgende Lemma. Wir benutzen hier und im folgenden auch die Arbeit von C. Liu (2016).

Lemma 4.14 *Seien $v_1, \ldots, v_n, v \in \mathbb{Z}^m$. Ist $v \notin \operatorname{span}_{\mathbb{Q}}\{v_1, \ldots, v_n\}$, lässt sich v also nicht als rationale Linearkombination von $v_1, \ldots, v_n$ darstellen, so existiert $u \in \mathbb{Z}^m$ mit $u^T v_j = 0$, $j = 1, \ldots, n$, und $u^T v \neq 0$.*

Beweis: Sei $S := \operatorname{span}_{\mathbb{Q}}\{v_1, \ldots, v_n\}$. Wir können $w_1, \ldots, w_k \in \mathbb{Q}^m$ mit $k \leq n$ bestimmen derart, dass $S = \operatorname{span}_{\mathbb{Q}}\{w_1, \ldots, w_k\}$ und $\{w_1, \ldots, w_k\}$ linear unabhängig sind. Durch das Verfahren von Gram-Schmidt können wir aus $\{w_1, \ldots, w_k\}$ eine orthonormale Basis $\{e_1, \ldots, e_k\} \subset \mathbb{Q}^m$ von S berechnen. Da $v \notin S$, ist $v \notin \operatorname{span}_{\mathbb{Q}}\{e_1, \ldots, e_k\}$. Das Orthonormalsystem $\{e_1, \ldots, e_k\}$ kann zu einem Orthonormalsystem $\{e_1, \ldots, e_{k'}\} \subset \mathbb{Q}^m$ mit $v \in \operatorname{span}_{\mathbb{Q}}\{e_1, \ldots, e_{k'}\}$ erweitert werden. Bezüglich dieses Orthonormalsystems hat v eine eindeutige Darstellung

$$v = \sum_{i=1}^{k'} \alpha_i e_i.$$

Es existiert $i \in \{k+1, \ldots, k'\}$ mit $\alpha_i \neq 0$, da andernfalls $v \in S$. Setze $u := \alpha e_i$, wobei wir $\alpha \in \mathbb{Z} \setminus \{0\}$ so wählen, dass $u \in \mathbb{Z}^m$. Dann steht u auf S senkrecht, es gilt also $u^T v_j = 0$, $j \in [1, n]$, ferner ist $u^T v = \alpha_i \alpha \neq 0$. Das Lemma ist bewiesen. □

Satz 4.15 *Die Matrix $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{Z}^{m \times n}$ sei partitionsregulär. Dann genügt A der Spaltenbedingung.*

Beweis: Sei p eine beliebige Primzahl und $d(x)$ in der p-adischen Entwicklung von $x \in \mathbb{N}$, also einer Darstellung

$$x = d_r p^r + d_{r-1} p^{r-1} + \cdots + d_1 p + d_0, \quad d_i \in [0, p-1],$$

der letzte von Null verschiedene Koeffizient und $L(x)$ sein Index, genau wie im Beweis von Satz 4.11. Insbesondere ist also $d(x) \in [1, p-1]$ für alle $x \in \mathbb{N}$. Die Färbungsfunktion $c\colon \mathbb{N} \longrightarrow [1, p-1]$ sei durch $c(x) := d(x)$, $x \in \mathbb{N}$, gegeben. Da wir vorausgesetzt haben, dass A partitionsregulär ist, existiert eine monochrome Lösung $x = (x_j) \in \mathbb{N}^n$ von $Ax = 0$, d. h. es ist $\sum_{j=1}^n a_j x_j = 0$ und $d(x_1) = \cdots = d(x_n) = d$ mit $d \in [1, p-1]$. Also haben $x_1, \ldots, x_n$ in einer p-adischen Entwicklung dieselbe letzte von Null verschiedene Ziffer d. Indizes $j \in [1, n]$, bei denen diese Ziffer an derselben Position stehen, werden zu einer Indexmenge D_k zusammengefasst. Genauer setzen wir, fast genau wie im Beweis von Satz 4.11

$$L_1 := \min_{j \in [1,n]} L(x_j), \qquad D_1 := \{j \in [1, n] : L(x_j) = L_1\}.$$

In D_1 sind also die Indizes $j \in [1, n]$ zusammengefasst, für die in der p-adischen Entwicklung von x_j der Koeffizient d möglichst weit rechts steht bzw. der zugehörige Index kleinstmöglich ist. Entsprechend seien in D_2 diejenigen $j \in [1, n] \setminus D_1$ enthalten, für die in der p-adischen Entwicklung von x_j der Koeffizient d möglichst weit rechts steht bzw. der zugehörige Index kleinstmöglich ist. In dieser Weise kann man fortfahren und erhält zu jeder Primzahl p eine Zerlegung

$$(*) \qquad [1, n] = D_1 \dot\cup \cdots \dot\cup D_s.$$

Ist z. B.

$$\begin{array}{lcccccccccc}
x_1: & \cdot & \cdot & \cdot & \cdot & \cdot & d & 0 & 0 & 0 \\
x_2: & \cdot & \cdot & \cdot & \cdot & \cdot & d & 0 & 0 & 0 \\
x_3: & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & d & 0 & 0 \\
x_4: & \cdot & \cdot & \cdot & \cdot & d & 0 & 0 & 0 & 0 \\
x_5: & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & d & 0 & 0
\end{array}$$

so ist

$$D_1 = \{3, 5\}, \qquad D_2 = \{1, 2\}, \qquad D_3 = \{4\}.$$

Die Partitionen hängen natürlich von der gewählten Primzahl ab. Die Anzahl der Partitionen von $[1, n]$ ist endlich, während es unendlich viele Primzahlen gibt. Daher gibt es eine Menge P unendlich vieler Primzahlen, zu der es ein und dieselbe Zerlegung $(*)$ von $[1, n]$ gibt. Diese Zerlegung genügt der Spaltenbedingung. Denn:

- Es ist
$$\sum_{j \in D_1} a_j = 0.$$

Denn: Wir können wörtlich so vorgehen wie im Beweis von Satz 4.11, wobei lediglich D_1 die Rolle von J spielt und die Teilbarkeitseigenschaft *komponentenweise* zu verstehen ist. Wir erhalten jedenfalls, dass

$$p \,\Big|\, \sum_{j \in D_1} a_j \quad \text{für alle } p \in P.$$

Hieraus folgt

$$\sum_{j\in D_1} a_j = 0.$$

Nun zeigen wir:

- Es ist
$$\sum_{j\in D_k} a_j \in \operatorname{span}_Q\{a_j : j \in D_1 \cup \cdots \cup D_{k-1}\}, \qquad k = 2, \ldots, s.$$

Denn: Für $k \in [2, s]$ sei $L_k := L(x_j)$ für $j \in B_k$ der zu d gehörende Index in der p-adischen Entwicklung von x_j. Offenbar ist $L_1 < \cdots < L_s$. Dann hat x_j in der p-adischen Entwicklung eine Darstellung

$$x_j = d_{r,j}p^r + d_{r-1,j}p^{r-1} + \cdots + d_{L_k+1,j}p^{L_k+1} + dp^{L_k}, \qquad j \in D_k, \quad k \in [1, s].$$

Nun sei $k \in [2, s]$ fest gewählt. Es ist

$$0 = \sum_{j=1}^{n} a_j x_j = \sum_{j\in D_k} a_j x_j + \sum_{j\in D_1\cup\cdots\cup D_{k-1}} a_j x_j + \sum_{j\in D_{k+1}\cup\cdots\cup D_s} a_j x_j.$$

Die linke Seite ist trivialerweise durch p^{L_k+1} teilbar, daher ist es auch die rechte Seite. Auf der rechten Seite sind einige Terme offensichtlich durch p^{L_k+1} teilbar, z. B. alle Summanden im dritten Term und in der ersten Summe bei der p-adischen Entwicklung alle Terme bis auf den letzten. Daher müssen auch die Terme, von denen es zunächst nicht offensichtlich ist, durch p^{L_k+1} teilbar sein. Mit anderen Worten gilt

$$p^{L_k+1} \;\Big|\; \left(p^{L_k} d \sum_{j\in D_k} a_j + \sum_{j\in D_1\cup\cdots\cup D_{k-1}} a_j x_j\right)$$

bzw.

$$(*) \qquad p^{L_k} d \sum_{j\in D_k} a_j + \sum_{j\in D_1\cup\cdots\cup D_{k-1}} a_j x_j = p^{L_k+1} b_k \quad \text{mit } b_k \in \mathbb{Z}^m.$$

Angenommen, die Behauptung wäre nicht richtig, für ein $k \in [2, s]$ wäre also

$$\sum_{j\in D_k} a_j \notin \operatorname{span}_Q\{a_j : j \in D_1 \cup \cdots \cup D_{k-1}\}.$$

Wegen Lemma 4.14 existiert $u \in \mathbb{Z}^m$ mit

$$u^T a_j = 0, \quad j \in D_1 \cup \cdots \cup D_{k-1}$$

und

$$u^T \sum_{j\in D_k} a_j \neq 0.$$

Aus $(*)$ erhält man daher

$$p^{L_k} d u^T \sum_{j \in D_k} a_j = p^{L_k+1} u^T b_k$$

bzw.

$$d u^T \sum_{j \in D_k} a_j = p u^T b_k.$$

Da $p \not| d$ gilt daher

$$p \,\Big|\, u^T \sum_{j \in D_k} a_j \quad \text{für alle } P \in P.$$

Daher ist

$$u^T \sum_{j \in D_k} a_j = 0,$$

ein Widerspruch. Der Satz ist bewiesen. □

Es bleibt zu zeigen, dass eine Matrix $A \in \mathbb{Z}^{m \times n}$, die der Spaltenbedingung genügt, partitionsregulär ist. Das wird ein wenig schwieriger sein als die bisherigen Betrachtungen. Für $p \in \mathbb{N}$ schreiben wir $[-p, p]$ statt $\{-p, -p+1, \ldots, p-1, p\}$. Die folgende Definition findet sich bei W. Deuber (1973).

Definition 4.16 Seien $m, p, c \in \mathbb{N}$ gegeben. Eine Menge $S \subset \mathbb{N}$ heißt eine (m, p, c)-*Menge*, wenn $y = (y_i) \in \mathbb{N}^m$ existiert derart, dass

$$S = \left\{ \sum_{i=1}^{m} \lambda_i y_i : \lambda_i \begin{cases} = & 0, & i < j, \\ = & c, & i = j, \\ \in & [-p, p], & i > j, \end{cases} \quad j \in [1:m] \right\}.$$

Wir sagen, dass $y \in \mathbb{N}^m$ die (m, p, c)-Menge S *erzeugt*.

Daher enthält eine durch $y = (y_i) \in \mathbb{N}^m$ erzeugte (m, p, c)-Menge S *alle* Ausdrücke der Form

$$\begin{aligned} c y_1 + \lambda_2 y_2 + \lambda_3 y_3 + \cdots + \lambda_{m-1} y_{m-1} + \lambda_m y_m, &\qquad (j = 1) \\ c y_2 + \lambda_3 y_3 + \cdots + \lambda_{m-1} y_{m-1} + \lambda_m y_m, &\qquad (j = 2) \\ \vdots &\qquad (j = 3, \ldots, m-2) \\ c y_{m-1} + \lambda_m y_m, &\qquad (j = m-1) \\ c y_m, &\qquad (j = m) \end{aligned}$$

wobei $\lambda_i \in [-p, p]$, $i = 2, \ldots, m$. In naheliegender Weise sprechen wir von den m *Zeilen* einer (m, p, c)-Menge S.

Beispiele: 1. Eine Menge $S \subset \mathbb{N}$ ist eine $(2, p, 1)$-Menge, wenn $(y_1, y_2) \in \mathbb{N} \times \mathbb{N}$ mit

$$S = \{y_1 - p y_2, y_1 - (p-1) y_2, \ldots, y_1, y_1 + y_2, \ldots, y_1 + p y_2\} \cup \{y_2\}$$

existieren. Also ist S eine arithmetische Progression der Länge $2p+1$ mitsamt ihrer Schrittweite.

2. Eine Menge $S \subset \mathbb{N}$ ist eine $(2,2,3)$-Menge, wenn $(y_1, y_2) \in \mathbb{N} \times \mathbb{N}$ mit

$$S = \{3y_1 - 2y_2, 3y_1 - y_2, 3y_1, 3y_1 + y_2, 3y_1 + 2y_2\} \cup \{3y_2\}$$

existieren. Daher ist S eine arithmetische Progression der Länge 5, wobei der mittlere Term durch 3 teilbar ist, mitsamt des Dreifachen der Schrittweite. □

Mit Hilfe der folgenden beiden Sätze wird der noch fehlende Teil zum Beweis des Satzes von Rado (siehe Satz 4.10) erbracht sein. Der folgende Satz stammt von W. Deuber (1973, Satz 2.1).

Satz 4.17 *Die ganzzahlige Matrix* $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$ *genüge der Spaltenbedingung. Dann existieren* $m, p, c \in \mathbb{N}$ *derart*[20]*, dass jede* (m,p,c)*-Menge* S *eine Lösung von* $Ax = 0$ *enthält. Genauer: Es existieren* $x_j \in S$, $j \in [1,n]$, *mit* $\sum_{j=1}^{n} x_j a_j = 0$.

Beweis: Da A der Spaltenbedingung genügt, existiert eine Zerlegung

$$[1,n] = D_1 \dot\cup \cdots \dot\cup D_s$$

der Spaltenindizes in disjunkte, nichtleere Mengen $D_1, \ldots, D_s$ derart, dass

$$\sum_{j \in D_1} a_j = 0$$

und

$$\sum_{j \in D_k} a_j \in \operatorname{span}_Q\{a_j : j \in D_1 \cup \cdots \cup D_{k-1}\}, \qquad k = 2, \ldots, s.$$

Sei etwa

$$\sum_{j \in D_k} a_j = \sum_{j \in D_1 \cup \cdots \cup D_{k-1}} q_{jk} a_j, \qquad k = 2, \ldots, s,$$

mit $q_{jk} \in \mathbb{Q}$ für $j \in D_1 \cup \cdots \cup D_{k-1}$ und $k \in [2,s]$. Diese Beziehung ist auch für $k = 1$ richtig, da eine leere Summe 0 ergibt. Wir definieren die Matrix $G = (g_{jk}) \in \mathbb{Q}^{n \times s}$ durch

$$g_{jk} := \begin{cases} 0, & j \notin D_1 \cup \cdots \cup D_k, \\ 1, & j \in D_k, \\ -q_{jk}, & j \in D_1 \cup \cdots \cup D_{k-1}, \end{cases} \qquad j \in [1,n],\ k \in [1,s].$$

Dann ist

$$\sum_{j=1}^{n} g_{jk} a_j = \sum_{j \in D_k} a_j - \sum_{j \in D_1 \cup \cdots \cup D_{k-1}} q_{jk} a_j = 0, \qquad k \in [1,s].$$

Nun setzen wir $m := s$, nehmen an, die rationalen Zahlen q_{jk} hätten eine Darstellung $q_{jk} = a_{jk}/b_{jk}$ mit $a_{jk} \in \mathbb{Z}$, $b_{jk} \in \mathbb{N}$ und definieren $c \in \mathbb{N}$ als das kleinste gemeinsame Vielfache der Nenner b_{jk}, wobei $k \in [2,s]$ und $j \in D_1 \cup \cdots \cup D_{k-1}$, sowie anschließend

$$p := c \max_{j,k} |a_{jk}| \geq c.$$

[20]Man beachte: Wir haben vermieden, der Zeilenanzahl von A einen Namen, etwa m, zu geben, um die Bedeutung von m noch in der Hand zu haben.

Hierbei ist das Maximum ebenfalls über alle (j,k) mit $k \in [2,s]$, $j \in D_1 \cup \cdots \cup D_{k-1}$, zu nehmen. Jetzt zeigen wir, dass jede (m,p,c)-Menge S eine Lösung von $Ax = 0$ enthält. Sei $y = (y_k) \in \mathbb{N}^s$ ein Erzeugendensystem von $S \subset \mathbb{N}$. Man setze $x := c \cdot Gy$ bzw.

$$x_j := c \sum_{k=1}^{s} g_{jk} y_k, \qquad j \in [1,n].$$

In dieser Darstellung von x_j sind alle Koeffizienten cg_{jk} ganzzahlig (da c als das kleinste gemeinsame Vielfache der Nenner der q_{jk} gewählt wurde), für $j \in D_k$ ist $cg_{jk} = c$ und alle Koeffizienten sind aus $[-p,p]$. Daher ist $x_j \in S$, $j \in [1,n]$. Ferner ist $Ax = 0$ bzw. $\sum_{j=1}^{n} x_j a_j = 0$, da

$$\sum_{j=1}^{n} x_j a_j = c \sum_{j=1}^{n} \Big(\sum_{k=1}^{s} g_{jk} y_k \Big) a_j = c \sum_{k=1}^{s} y_k \underbrace{\Big(\sum_{j=1}^{n} g_{jk} a_j \Big)}_{=0} = 0.$$

Damit ist der Satz bewiesen. □

Beispiel: Sei

$$A = \begin{pmatrix} a_1 & \cdots & a_6 \end{pmatrix} := \begin{pmatrix} 1 & 3 & -1 & 0 & 1 & 0 \\ 2 & 2 & -2 & 4 & 0 & 1 \\ 3 & 1 & -3 & 8 & 1 & 0 \end{pmatrix}.$$

Dann genügt A der Spaltenbedingung. Denn mit

$$D_1 := \{1,3\}, \qquad D_2 := \{2,4\}, \qquad D_3 := \{5,6\}$$

hat man die disjunkte Zerlegung

$$[1:6] = D_1 \dot\cup D_2 \dot\cup D_2 \dot\cup D_3$$

mit

$$\sum_{j \in D_1} a_j = a_1 + a_3 = 0, \quad \sum_{j \in D_2} a_j = a_2 + a_4 = 3a_1, \quad \sum_{j \in D_3} a_j = a_5 + a_6 = \frac{1}{4} a_1 + \frac{1}{4} a_2.$$

Daher ist

$$\begin{pmatrix} \sum_{j \in D_1} a_j \\ \sum_{j \in D_2} a_j \\ \sum_{j \in D_3} a_j \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{pmatrix} \begin{pmatrix} 0 & 3 & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Im Beweis des letzten Satzes wird $(m,p,c) := (3,12,4)$ gesetzt, die Matrix $G \in \mathbb{Q}^{6\times 3}$ ist durch

$$G = \begin{pmatrix} 1 & -3 & \frac{1}{4} \\ 0 & 1 & -\frac{1}{4} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

gegeben. Mit einem Erzeugendensystem $y \in \mathbb{N}^3$ der $(3, 12, 4)$-Menge erhalten wir durch $x = 4 \cdot Gy$ eine Lösung x von $Ax = 0$ sozusagen in parametrischer Form:

$$\begin{array}{rcrcrcr} x_1 & = & 4y_1 & - & 12y_2 & - & y_3 \\ x_2 & = & & & 4y_2 & - & y_3 \\ x_3 & = & 4y_1 & & & & \\ x_4 & = & & & 4y_2 & & \\ x_5 & = & & & & & 4y_3 \\ x_6 & = & & & & & 4y_3 \end{array}$$

Natürlich ist $x \in \mathbb{N}^6$ nur für gewisse $y \in \mathbb{N}^3$, klar ist aber, dass es solche gibt. □

Mit Hilfe des folgenden Satzes wird der Beweis des Satzes von Rado komplettiert.

Satz 4.18 *Die Menge $\mathbb{N}$ der natürlichen Zahlen sei durch $r \in \mathbb{N}$ Farben gefärbt, diese Färbungsfunktion sei etwa durch $f\colon \mathbb{N} \longrightarrow [1, r]$ gegeben. Für alle $m, p, c \in \mathbb{N}$ gibt es eine monochrome (m, p, c)-Menge S, d. h. alle Elemente von S haben dieselbe Farbe.*

Beweis: Seien $m, p, c \in \mathbb{N}$ gegeben. Wir definieren $M := r(m-1)+1$. Unser Ziel ist es zunächst zu zeigen:

(a) Es existiert eine (M, p, c)-Menge T, die zeilenweise monochrom ist, d. h. alle M Zeilen von T sind monochrom.

Denn: Sei $n_1 \in \mathbb{N}$ gegeben und so groß gewählt, dass n_1 durch $(M-1)pc$ teilbar ist. Mit $f\colon \mathbb{N} \longrightarrow [1, r]$ haben wir die gegebene Färbungsfunktion bezeichnet. Diese induziert durch $g(i) := f(ic)$ die Färbungsfunktion $g\colon \mathbb{N} \longrightarrow [1, r]$. Wegen des Satzes von van der Waerden existiert für $n \geq W(r, 2n_1 + 1)$ eine bezüglich der Färbungsfunktion g monochrome arithmetische Progression

$$\{x_1 - n_1 d_1', \ldots, x_1 - d_1', x_1, x_1 + d_1', \ldots, x_1 + n_1 d_1'\} \subset [1, n]$$

der Länge $2n_1 + 1$. Also existiert $r_1 \in [1, r]$ mit

$$g(x_1 - id_1') = f(cx_1 - icd_1') = f(cx_1 - id_1) = r_1, \qquad i \in [-n_1, n_1],$$

wobei wir $d_1 := cd_1'$ gesetzt haben. Daher ist

$$R_1 := \{cx_1 - n_1 d_1, \ldots, cx_1, \ldots, cx_1 + n_1 d_1\}$$

eine in

$$A_1 := \{c, 2c, \ldots, nc\}$$

enthaltene arithmetische Progression der Länge $2n_1 + 1$, deren Elemente sämtlich mit der Farbe r_1 gefärbt sind. Nun werde

$$B_1 := \left\{ d_1, 2d_1, \ldots, \frac{n_1}{(M-1)p} d_1 \right\}$$

gesetzt. Sind $x_2, \ldots, x_M \in B_1$ und $\lambda_2, \ldots, \lambda_M \in [-p, p]$, so ist

$$cx_1 + \sum_{i=2}^{M} \lambda_i x_i \leq cx_1 + (M-1)p \cdot \frac{n_1}{(M-1)p} d_1 = cx_1 + n_1 d_1$$

und entsprechend

$$cx_1 + \sum_{i=2}^{M} \lambda_i x_i \geq cx_1 - n_1 d_1.$$

Da andererseits $cx_1 + \sum_{i=2}^{M} \lambda_i x_i = cx_1 + \alpha d_1$ mit einem gewissen $\alpha \in \mathbb{Z}$, ist gezeigt:

- Sind $x_2, \ldots, x_M \in B_1$ und $\lambda_2, \ldots, \lambda_M \in [-p, p]$ beliebig, so ist

  $$cx_1 + \sum_{i=2}^{M} \lambda_i x_i \in R_1$$

  und ist daher mit der Farbe r_1 gefärbt.

Für $k = 2, \ldots, M$ mache man nun die folgenden Schritte:

(1) Definiere

$$A_k := \left\{ cd_{k-1}, 2cd_{k-1}, \ldots, \frac{n_{k-1}}{(M-k+1)pc} cd_{k-1} \right\} \subset B_{k-1}$$

und wähle $n_k \in \mathbb{N}$ so, dass n_k durch $(M-k)pc$ teilbar ist. Indem wir $n_{k-1} \in \mathbb{N}$ notfalls vergrößern, etwa können wir

$$\frac{n_{k-1}}{(M-k+1)pc} \geq W(r, 2n_k + 1)$$

wählen, können wir annehmen, dass in A_k wegen des Satzes von van der Waerden eine monochrome arithmetische Progression der Länge $2n_k + 1$, nämlich

$$R_k := \{x'_k cd_{k-1} - n_k d_k, \ldots, x'_k cd_{k-1}, \ldots, x'_k cd_{k-1} + n_k d_k\},$$

enthalten ist, siehe den entsprechenden Beweisschritt am Anfang. Diese habe etwa die Farbe $r_k \in [1, r]$. Weiter ist

$$x_k := x'_k d_{k-1} \leq \frac{n_{k-1}}{(M-k+1)p}$$

und folglich $x_k \in B_{k-1}$.

(2) Nun definieren wir

$$B_k := \left\{ d_k, 2d_k, \ldots, \frac{n_k}{(M-k)p} d_k \right\}.$$

Nun gilt:

– Sind $x_{k+1}, \ldots, x_M \in B_k$ und $\lambda_{k+1}, \ldots, \lambda_M \in [-p, p]$ beliebig, so ist

$$cx_k + \sum_{i=k+1}^{M} \lambda_i x_i \in R_k$$

und ist daher mit der Farbe r_k gefärbt.

Denn: Einerseits existiert $\alpha \in \mathbb{Z}$ mit

$$c_k + \sum_{i=k+1}^{M} \lambda_i x_i = c_k + \alpha d_k,$$

andererseits ist

$$cx_k + \sum_{i=k+1}^{M} \lambda_i x_i \leq cx_k + (M-k)p \cdot \frac{n_k}{(M-k)p} d_k = cx_k + n_k d_k$$

und entsprechend

$$cx_k + \sum_{i=k+1}^{M} \lambda_i x_i \geq cx_k - n_k d_k.$$

Weiter gilt:

– Es ist $B_k \subset B_{k-1}$.

Denn: Wegen

$$\begin{aligned} R_k &= \{cx_k - n_k d_k, \ldots, cx_k - d_k, cx_k, cx_k + d_k, \ldots, cx_k + n_k d_k\} \\ &\subset \left\{cd_{k-1}, 2cd_{k-1}, \ldots, \frac{n_{k-1}}{(M-k+1)pc} cd_{k-1}\right\} \\ &= A_k \\ &\subset B_{k-1} \\ &= \left\{d_{k-1}, 2d_{k-1}, \ldots, \frac{n_{k-1}}{(M-k+1)p} d_{k-1}\right\} \end{aligned}$$

existiert für $j = 0, \ldots, n_k$ ein

$$\alpha_j \in \left[1, \frac{n_{k-1}}{(M-k+1)p}\right]$$

mit

$$cx_k + jd_k = \alpha_j d_{k-1}.$$

Für $j \in [1, n_k]$ ist also $jd_k = (\alpha_j - \alpha_0) d_{k-1} \in B_{k-1}$. Daher ist

$$\begin{aligned} B_k &= \left\{d_k, 2d_k, \ldots, \frac{n_k}{(M-k)p} d_k\right\} \\ &\subset \{d_k, 2d_k, \ldots, n_k d_k\} \\ &\subset B_{k-1}. \end{aligned}$$

Damit ist nachgewiesen, dass die durch $x_1, \ldots, x_M$ erzeugte (M, p, c)-Menge T zeilenweise monochrom ist. Zu beachten ist, dass im k-ten Schritt eventuell $n_{k-1}, \ldots, n_1$ vergrößert werden müssen. Da die Anzahl der Schritte aber endlich ist, ist dies für den Beweis kein Problem.

Der Schluss des Beweises ist eine einfache Anwendung des Schubfachprinzips:

(b) Es gibt mindestens m Zeilen von T, die durch ein und dieselbe Farbe gefärbt sind.

Denn seien genau m_i Zeilen durch die Farbe i gefärbt, $i \in [1, r]$. Wäre $m_i < m$ bzw. $m_i \leq m - 1$, $i = 1, \ldots, r$, so erhielten wir aus

$$r(m-1) + 1 = M = \sum_{i=1}^{r} m_i \leq r(m-1)$$

einen Widerspruch. Also existiert ein $i \in [1, r]$ mit $m_i \geq m$, also eine Farbe i, mit der mindestens m Zeilen von T gefärbt sind.

Seien $i_1, \ldots, i_m \in [1, M]$ gleich gefärbte Zeilen der durch $x_1, \ldots, x_M$ erzeugten (M, p, c)-Menge. Dann ist die durch $x_{i_1}, \ldots, x_{i_m}$ erzeugte (m, p, c)-Menge monochrom und damit ist die Aussage des Satzes bewiesen. □

Jetzt ist der **Beweis des Satzes 4.10** von Rado einfach. Dieser schöne Satz sagt aus, dass eine Matrix A mit ganzzahligen Einträgen genau dann partitionsregulär ist, wenn sie der Spaltenbedingung genügt. Durch Satz 4.15 wurde bewiesen, dass eine partitionsreguläre Matrix der Spaltenbedingung genügt. Ist umgekehrt für die ganzzahlige Matrix A die Spaltenbedingung erfüllt, so existieren wegen Satz 4.17 Zahlen $m, p, c \in \mathbb{N}$ mit der Eigenschaft, dass jede (m, p, c)-Menge eine Lösung von $Ax = 0$ enthält. In Satz 4.18 ist nachgewiesen, dass es zu jedem und insbesondere diesem Tripel (m, p, c) eine monochrome (m, p, c)-Menge gibt. Daher besitzt $Ax = 0$ eine monochrome Lösung. Also ist A partitionsregulär. □

5 Die Cayley-Formel

5.1 Definitionen, Formulierung der Cayley-Formel, Beispiele

Einen mathematischen Begriff kann man *fundamental* nennen, wenn es für ihn mehrere äquivalente Definitionen gibt. Man denke z. B. an den Begriff der Kompaktheit. Einen mathematischen Satz kann man als besonders *schön* empfinden, wenn er einfach zu formulieren ist und es für ihn unterschiedliche oder besonders überraschende Beweise gibt. Genannt seien hier nur der Fundamentalsatz der Algebra und der Brouwersche Fixpunktsatz. In diesem Sinne ist die Cayley-Formel bzw. der Satz von Cayley ein schöner Satz. Das ist auch ein Grund, weshalb M. Aigner, G. M. Ziegler (2018) die Cayley-Formel mit vier verschiedenen Beweisen in ihre Sammlung aufgenommen haben.

Ein Graph $G = (V, E)$ besteht bekanntlich aus einer endlichen Menge $V = V(G)$, der Menge der *Ecken* (**V**ertices), und einer Teilmenge $E = E(G)$ von (ungeordneten)

Paaren aus V, der Menge der *Kanten* (**E**dges). Der *Grad* $\deg(v)$ einer Ecke $v \in V$ ist gegeben durch

$$\deg(v) := |\{w \in V : (v, w) \in E\}|.$$

Also ist $\deg(v)$ die Anzahl der Ecken, die von v aus durch eine Kante erreicht werden können bzw. die Anzahl der *Nachbarn* von v. Unter einem *Weg* in einem Graphen (V, E) versteht man eine Folge von paarweise verschiedenen Ecken $v_1, \ldots, v_k \in V$, wobei $(v_i, v_{i+1}) \in E$, $i = 1, \ldots, k-1$. Dieser heißt ein geschlossener Weg oder ein *Kreis*, wenn darüberhinaus $(v_k, v_1) \in E$. Ein Graph heißt *zusammenhängend*, wenn je zwei Ecken durch einen Weg verbunden können. Ein zusammenhängender Graph, der keine Kreise enthält (oder *kreisfrei* ist), heißt ein *Baum*. Wie man leicht nachweisen kann, ist die Summe der Grade aller Ecken eines Baumes (V, E) gegeben durch

$$\sum_{v \in V} \deg(v) = 2(|V| - 1).$$

Denn wegen des "handshaking lemma" ist

$$\sum_{v \in V} \deg(v) = 2\,|E|$$

für einen beliebigen Graphen (V, E) und in einem Baum (V, E) gilt $|E| = |V| - 1$. Ein zusammenhängender Graph $G = (V, E)$ ist sogar genau dann ein Baum, wenn $|E| = |V| - 1$. Eine Ecke in einem Baum mit dem Grad 1 heißt ein *Blatt*. In einem Baum gibt es mindestens zwei Ecken mit dem Grad 1, denn gäbe es im Baum höchstens eine Ecke vom Grad 1, so wäre

$$\sum_{v \in V} \deg(v) \geq 1 + (2|V| - 1) > 2(|V| - 1),$$

ein Widerspruch. Wir sprechen von einem *bezeichneten Baum* (oder auch *beschrifteten Baum*, engl.: labeled tree), wenn die Ecken eines Baumes eine Bezeichnung oder einen Namen tragen. Die Art der Bezeichnung ist unwichtig, wichtig ist im folgenden aber, dass die Ecken durch ihre Bezeichnung der Größe nach geordnet werden können. Bei n Ecken benutzen wir daher die Bezeichnungen $1, 2, \ldots, n$ oder allgemeiner $x_1 < \cdots < x_n$. Als bezeichnete Graphen sind z. B. die drei in Abbildung 17 angegebenen Graphen als

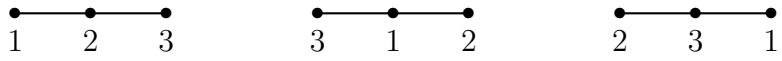


Abbildung 17: Drei verschiedene bezeichnete Bäume mit drei Ecken

(paarweise) verschieden anzusehen. Und dies sind offenbar alle bezeichneten Bäume mit 3 Ecken. Sie unterscheiden sich nur darin, welche der drei Ecken den Grad 2 hat. Dagegen sind natürlich die beiden bezeichneten Bäume in Abbildung 18 gleich. Wenn man auf Bezeichnungen verzichtet, so gibt es nur zwei verschiedene Bäume mit vier

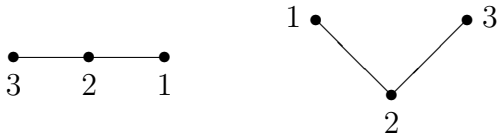


Abbildung 18: Zwei gleiche bezeichnete Bäume

Abbildung 19: Zwei Bäume mit vier Ecken

Ecken, diese beiden sind in Abbildung 19 angegeben. In Abbildung 20 geben wir die 16 verschiedenen bezeichneten Bäume mit 4 Ecken an. Das sind zunächst vier bezeichnete Bäume mit der Reihe nach $1, 2, 3, 4$ als Ecken mit dem Grade 3. Danach jeweils zwei bezeichnete Bäume mit $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$ und $\{3, 4\}$ als Ecken vom Grade 2. Die drei in Abbildung 21 angegebenen bezeichneten Bäume sind gleich. Für einen Baum mit fünf Ecken (die Summe der Grade in den fünf Ecken ist gleich 8) gibt es nur drei (im Sinne der Graphentheorie) unterschiedliche Bäume, nämlich die in Abbildung 22 angegebenen. Links sieht man einen Baum mit einer Ecke vom Grad 4 (und vier Ecken vom Grad 1), in der Mitte einen Baum mit einer Ecke vom Grad 3 und einer Ecke vom Grad 2 (und drei Ecken vom Grad 1), rechts einen Baum mit drei Ecken vom Grad 2 (und zwei Ecken vom Gad 1). Zum ersten Baum gibt es offenbar 5 verschiedene bezeichnete Bäume, zu den beiden weiteren jeweils 60 bezeichnete Bäume, das sind also insgesamt 125 bezeichnete Bäume mit 5 Ecken.

Bezeichnet man mit t_n die Anzahl der bezeichneten Bäume mit n Ecken, so ist also $t_3 = 3 = 3^{3-2}$, $t_4 = 16 = 4^{4-2}$ und $t_5 = 125 = 5^{5-2}$. In dem folgenden Satz (siehe A. Cayley (1889)) wird die Cayley-Formel angegeben.

Satz 5.1 (Cayley) *Es gibt n^{n-2} verschiedene bezeichnete Bäume mit n Ecken.*

Es ist unser Ziel, einige besonders schöne Beweise dieses Satzes anzugeben. Dies geschieht auch in einigen Lehrbüchern über Kombinatorik, Graphentheorie oder allgemein

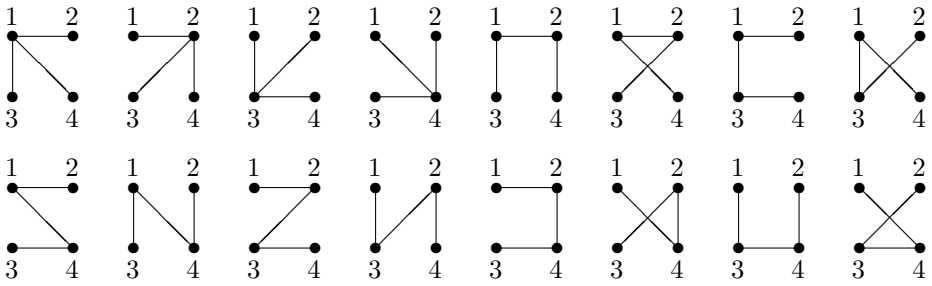


Abbildung 20: Die sechzehn verschiedenen bezeichneten Bäume mit vier Ecken

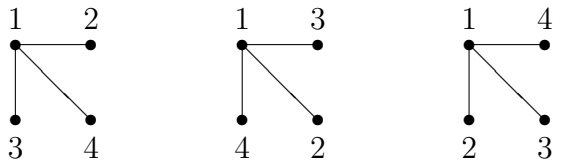


Abbildung 21: Drei gleiche bezeichnete Bäume

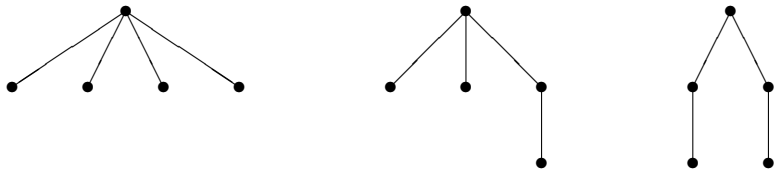


Abbildung 22: Drei unterschiedliche Bäume mit fünf Knoten

Diskrete Mathematik wie z.B. J. H. van Lint, R. M. Wilson (1992, S. 11 ff.)), B. Bollobás (1998, S. 277) oder J. Matoušek, J. Nešetřil (2002, S. 249–270).

5.2 Beweis durch Bijektion (Prüfer)

Von H. Prüfer (1918) stammt ein Beweis der Cayley-Formel, der darauf beruht, eine bijektive Abbildung zwischen der Menge der bezeichneten Bäume mit n Ecken und einer Menge, die ganz offensichtlich aus n^{n-2} Elementen besteht, herzustellen. Zur Einführung in das Problem schreibt H. Prüfer:

- Ich bediene mich hierbei einer anschaulichen geometrischen Einkleidung, in der Herr Prof. Schur den Satz im Mathematischen Proseminar der Universität Berlin als Aufgabe gestellt hat:

  Man denke sich ein Land mit n Städten. Diese n Städte sollen durch ein Eisenbahnnetz von $n-1$ Einzelstrecken (der kleinsten in Betracht kommenden Zahl) so verbunden werden, daß man von jeder Stadt nach jeder anderen gelangen kann. Es gibt n^{n-2} verschiedene Eisenbahnnetze dieser Art..

  Unter einer Einzelstrecke ist dabei natürlich eine Eisenbahnstrecke zu verstehen, die nur 2 Städte mit einander verbindet.

  Der Satz kann dadurch bewiesen werden, daß man jedem Eisenbahnnetz ein gewisses Symbol $\{a_1, a_2, \ldots, a_{n-2}\}$, dessen $n-2$ Elemente unabhängig von einander die Werte $1, 2, \ldots, n$ annehmen können, umkehrbar eindeutig zuordnet. Da es n^{n-2} solche Symbole gibt, ist mit der Angabe einer eineindeutigen Zuordnung der Netze und Symbole der Beweis erbracht.

Sei $\mathcal{T}_n$ die Menge der bezeichneten Bäume mit n Ecken $\{x_1, \ldots, x_n\}$, die durch $x_1 < \cdots < x_n$ geordnet seien. Sei weiter $\mathcal{S}_n$ die Menge aller $(n-2)$-Tupel $[p_1, \ldots, p_{n-2}]$, wobei $p_k \in \{x_1, \ldots, x_n\}$, $k = 1, \ldots, n-2$. Elemente von $\mathcal{S}_n$ heißen *Prüfer-Codes*. Offensichtlich ist $|\mathcal{S}_n| = n^{n-2}$. Wir zeigen die Existenz einer Bijektion $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$. Dann ist natürlich auch

$$t_n = |\mathcal{T}_n| = n^{n-2}.$$

Die Abbildung $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$, welche einem $T \in \mathcal{T}_n$ einen Prüfer-Code $\phi_n(T) \in \mathcal{S}_n$ zuordnet, wird induktiv definiert.

- Für $n = 2$ gibt es einen eindeutigen bezeichneten Baum $T \in \mathcal{T}_2$, man setze $\phi_2(T) := [\,]$.
- Angenommen, ϕ_{n-1} sei für bezeichnete Bäume mit $n-1$ Ecken definiert. Man bestimme in $T \in \mathcal{T}_n$ das Blatt, also die Ecke vom Grad 1, mit der kleinsten Bezeichnung x_i. Sei x_j die Bezeichnung des (eindeutigen) Nachbarn dieser Ecke. Man setze[21] $\phi_n(T) := [x_j, \phi_{n-1}(T \setminus \{x_i\})]$. Hierbei ist $T \setminus \{x_i\}$ der Baum, der aus T dadurch hervorgeht, dass die Ecke x_i und die Kante (x_i, x_j) aus T entfernt werden. Man beachte, dass die Eckenmenge $\{x_1, \ldots, x_n\} \setminus \{x_i\}$ von $T \setminus \{x_i\}$ natürlich auch der Größe nach geordnet werden kann und damit $T \setminus \{x_i\} \in \mathcal{T}_{n-1}$ gilt.

Auf diese Weise wird jedem bezeichneten Baum $T \in \mathcal{T}_n$ auf eindeutige Weise ein Prüfer-Code $\phi_n(T) \in \mathcal{S}_n$ zugeordnet. Im folgenden Satz zeigen wir, dass $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$ eine Bijektion ist. Zunächst geben wir aber einen Algorithmus bzw. eine nichtrekursive Fassung dieser Definition an und veranschaulichen diese an einem Beispiel.

- Sei ein Baum $T \in \mathcal{T}_n$ mit n Ecken $\{x_1, \ldots, x_n\}$ und $n \geq 3$ gegeben. Setze $p := [\,]$.
- Für $k = 1, \ldots, n-2$:
  - Bestimme in T das Blatt (Ecke vom Grad 1 bezüglich T) mit der kleinsten Bezeichnung x_i. Sei x_j die Bezeichnung des (eindeutigen) Nachbarn dieser Ecke. Setze $T := T \setminus \{x_i\}$, man entferne also aus T die Ecke mit der Bezeichnung x_i und die hiervon ausgehende Kante, und $p := [p, x_j]$.
- Setze $\phi_n(T) := p$.

Beispiel: Gegeben sei der in Abbildung 23 angegebene bezeichnete Baum mit 7 Ecken. Wir wollen den zugehörigen Prüfer-Code bestimmen. In Abbildung 24 geben wir die

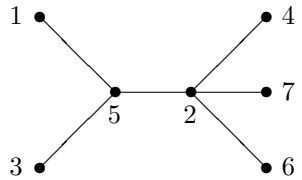


Abbildung 23: Ein bezeichneter Baum mit sieben Ecken

ersten drei Schritte des obigen Algorithmus an. Die zu entfernende Ecke haben wir jeweils eingekreist. In Abbildung 25 folgen die letzten beiden Schritte des Algorithmus. Der Prüfer-Code zu dem in Abbildung 23 angegebenen bezeichneten Baum $T \in \mathcal{T}_7$ ist daher durch $\phi_7(T) = [5, 5, 2, 2, 2]$ gegeben. □

[21] Hier nutzen wir aus, dass wir Ecken mit ihrer Bezeichnung x_i bzw. deren Index i identifizieren.

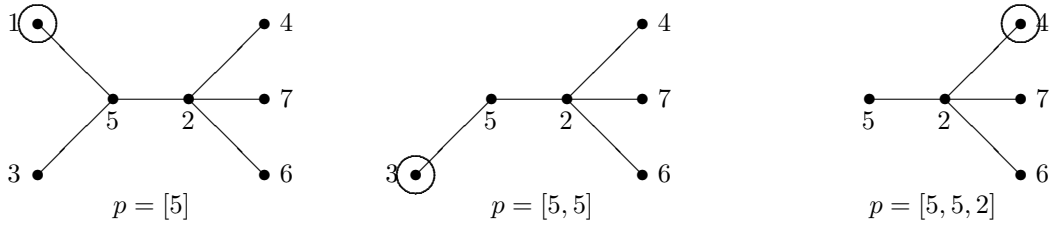


Abbildung 24: Die ersten drei Schritte des obigen Algorithmus

Abbildung 25: Die letzten beiden Schritte des Algorithmus

Satz 5.2 *Die oben definierte Abbildung* $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$ *ist eine Bijektion. Folglich ist* $|\mathcal{T}_n| = |\mathcal{S}_n| = n^{n-2}$.

Beweis: Entscheidend für den Beweis ist die folgende Aussage:

- Sei $T \in \mathcal{T}_n$ ein bezeichneter Baum mit Eckenmenge $\{x_1, \ldots, x_n\}$. Für $k = 1, \ldots, n$ ist $\deg(x_k) = 1 + c_k$, wobei c_k angibt, wie oft x_k in $\phi_n(T)$ vorkommt. Insbesondere ist die Ecke x_k in T (bzw. die Ecke mit der Bezeichnung x_k) genau dann ein Blatt bzw. $\deg(x_k) = 1$, wenn x_k im Prüfer-Code $\phi_n(T)$ *nicht* vorkommt.

Denn: Wir überlegen uns die Richtigkeit dieser Beobachtung durch vollständige Induktion nach n. Für $n = 2$ hat $T \in \mathcal{T}_2$ zwei Ecken, die beide den Grad 1 besitzen. Da $\phi_2(T) = [\,]$, ist die Behauptung in diesem Falle richtig. Die Behauptung sei für bezeichnete Bäume mit $\leq n-1$ Ecken richtig. Sei $T \in \mathcal{T}_n$ und x_i das Blatt (es ist also $\deg(x_i) = 1$) mit der niedrigsten Bezeichnung und x_j der eindeutige Nachbar. Dann ist i in $\phi_n(T) = [\phi_{n-1}(T \setminus \{x_i\}), j]$ nicht enthalten. Denn x_i hat den Grad 1 und kann außer x_j keinen weiteren Nachbarn besitzen, die obige Aussage ist also für $k = i$ richtig. Sei nun $k \neq i, j$. Dann stimmt der Grad $\deg(x_k)$ von x_k bezüglich T mit dem Grad $\deg_{T\setminus\{x_i\}}(x_k)$ von x_k bezüglich $T \setminus \{x_i\}$ überein. Daher kommt x_k nach Induktionsannahme $(\deg(x_k) - 1)$-mal in $\phi_n(T)$ vor. Die obige Aussage ist also zumindest für alle $k \in \{1, \ldots, n\} \setminus \{j\}$ richtig. Sie ist aber auch für $k = j$ richtig. In $T \setminus \{x_i\}$ kommt die Kante (x_i, x_j) nicht vor. Daher ist $\deg_{T\setminus\{i\}}(x_j) = \deg(x_j) - 1$. Nach Induktionsannahme kommt x_j in $\phi(T \setminus \{x_i\})$ genau $(\deg_{T\setminus\{x_i\}}(x_j) - 1)$-mal vor. Berücksichtigt man also den letzten Eintrag x_j in $\phi_n(T)$, so kommt x_j in $\phi_n(T)$ also $(\deg(x_j) - 1)$-mal vor. Damit ist obige Aussage bewiesen.

Nun zeigen wir durch Induktion nach n, dass $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$ injektiv ist. Für $n = 2$ ist dies offensichtlich richtig. Seien $T_1, T_2 \in \mathcal{T}_n$ verschiedene durch $x_1 < \cdots < x_n$ bezeichnete Bäume. Wir haben $\phi_n(T_1) \neq \phi_n(T_2)$ zu zeigen. Hierzu unterscheiden wir zwei Fälle.

(a) Es ist $(\deg_{T_1}(x_1), \ldots, \deg_{T_1}(x_n)) \neq (\deg_{T_2}(x_1), \ldots, \deg_{T_2}(x_n))$.

Dann existiert $k \in \{1, \ldots, n\}$ mit $\deg_{T_1}(x_k) \neq \deg_{T_2}(x_k)$. Wegen der obigen Aussage kommt x_k in $\phi_n(T_1)$ und $\phi_n(T_2)$ unterschiedlich oft vor. Daher ist in diesem Falle $\phi_n(T_1) \neq \phi_n(T_2)$.

(b) Es ist $(\deg_{T_1}(x_1), \ldots, \deg_{T_1}(x_n)) = (\deg_{T_2}(x_1), \ldots, \deg_{T_2}(x_n))$.

Insbesondere stimmen die Blätter von T_1 und T_2 überein, also die Ecken, deren Grad gleich 1 ist. Daher gibt es ein kleinstes i mit $\deg_{T_1}(x_i) = \deg_{T_2}(x_i) = 1$. Der (eindeutige) Nachbar von x_i in T_1 sei x_{j_1}, der (eindeutige) Nachbar von x_i in T_2 sei x_{j_2}. Ist $j_1 \neq j_2$, so ist offenbar $\phi_n(T_1) \neq \phi_n(T_2)$, da bei der Berechnung von $\phi_n(T_1)$ bzw. $\phi_n(T_2)$ der erste Eintrag x_{j_1} bzw. x_{j_2} ist und diese voneinander verschieden sind. Ist dagegen $j_1 = j_2$, so ist $T_1 \setminus \{x_i\} \neq T_2 \setminus \{x_i\}$ (andernfalls wäre $T_1 = T_2$) und daher nach Induktionsannahme $\phi_{n-1}(T_1 \setminus \{x_i\}) \neq \phi_{n-1}(T_2 \setminus \{x_i\})$ und folglich $\phi_n(T_1) \neq \phi_n(T_2)$. Damit ist schließlich gezeigt, dass $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$ injektiv ist.

Zu zeigen bleibt, dass $\phi_n \colon \mathcal{T}_n \longrightarrow \mathcal{S}_n$ surjektiv ist. Dies geschieht ebenfalls durch Induktion nach n, wobei der Fall $n = 2$ evident ist. Sei $p = [p_1, \ldots, p_{n-2}] \in \mathcal{S}_n$, also $p_k \in \{x_1, \ldots, x_n\}$, $k = 1, \ldots, n$. Zu zeigen ist die Existenz eines Baumes $T \in \mathcal{T}_n$ mit $\phi_n(T) = p$. Sei i minimal unter der Nebenbedingung, dass x_i in p als Eintrag nicht vorkommt. Da $n - 2 < n$ muss es ein solches i geben. In dem gesuchten Baum T ist x_i ein Blatt, hat also den Grad 1. Sei x_j der erste Eintrag in p. Man gewinne $p' \in \mathcal{S}_{n-1}$ aus p dadurch, dass man x_j weglässt. Nach Induktionsannahme gibt es zur Eckenmenge $\{x_1, \ldots, x_n\} \setminus \{x_i\}$ einen Baum $T' \in \mathcal{T}_{n-1}$ mit $\phi_{n-1}(T') = p'$. Man gewinne den Baum T aus T' dadurch, dass man die Ecke x_i zu T' hinzufügt und dies mit x_j verbindet. Dann ist $p = [x_j, p']$ der Prüfer-Code von T bzw. $\phi_n(T) = p$. Damit ist nachgewiesen, dass ϕ_n surjektiv ist. Der ganze Satz ist bewiesen. □

Bemerkung: Der Beweis, dass es zu jedem Prüfer-Code $p \in \mathcal{S}_n$ einen bezeichneten Baum $T \in \mathcal{T}_n$ gibt, erfolgte oben rekursiv. Eine nichtrekursive Fassung könnte folgendermaßen aussehen.

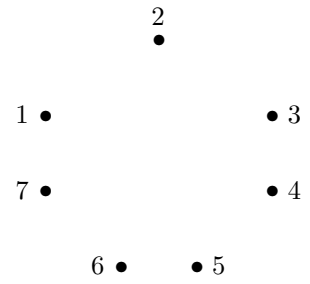
- Gegeben seien n durch $\{1, \ldots, n\}$ bezeichnete Ecken, aber noch keine verbindenden Kanten.
- Gegeben sei $p = [p_1, \ldots, p_{n-2}] \in \mathcal{S}_n$ mit $p_k \in \{1, \ldots, n\}$, $k = 1, \ldots, n-2$. Sei $b := [\,]$.
- Für $k = 1, \ldots, n-2$:
  - Bestimme das kleinste i, das weder in p noch in b als Eintrag enthalten ist. Setze $b := [b, i]$ und verbinde die Ecke i mit der Ecke j, wobei j der erste Eintrag von p ist[22].
  - Entferne den ersten Eintrag von p und nenne das Ergebnis wieder p.

[22] In dem gesuchten Baum $T \in \mathcal{T}_n$ ist i das Blatt mit der kleinsten Bezeichnung und j der eindeutige Nachbar.

- Jetzt sind $n-2$ Kanten bestimmt und b hat $n-2$ Einträge. Man erhält die $(n-1)$-te Kante dadurch, dass man die beiden Ecken miteinander verbindet, deren Bezeichnungen nicht in b als Einträge enthalten sind.

□

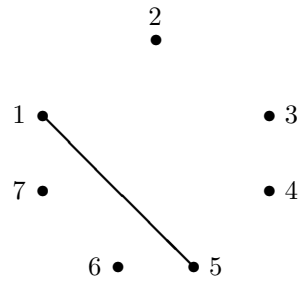
Beispiel: Wir kommen auf das letzte Beispiel zurück und wollen zum Prüfer-Code $p = [5, 5, 2, 2, 2]$ den zugehörigen bezeichneten Baum bestimmen. Die Ausgangssituation geben wir in Abbildung 26 an. Die ersten drei Schritte des obigen Algorithmus



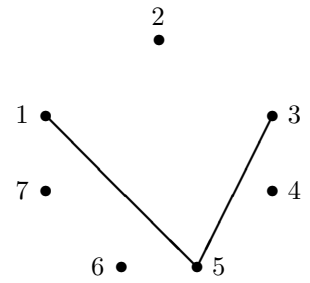
$p = [5, 5, 2, 2, 2]$, $b = [\,]$

Abbildung 26: Bestimmung von $T \in \mathcal{T}_7$ zum Prüfer-Code $p = [5, 5, 2, 2, 2]$

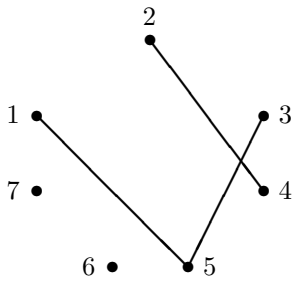
findet man in Abbildung 27. In Abbildung 28 geben wir die restlichen drei Schritte



$p = [5, 2, 2, 2]$, $b = [1]$



$p = [2, 2, 2]$, $b = [1, 3]$



$p = [2, 2]$, $b = [1, 3, 4]$

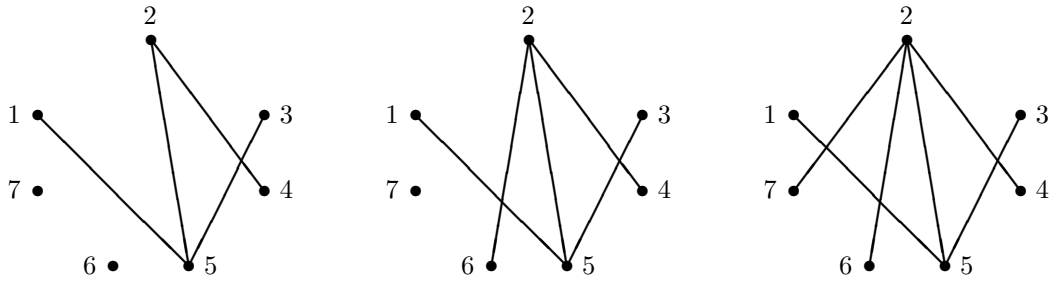
Abbildung 27: Die ersten drei Schritte

des Algorithmus an. Der letzte Baum in Abbildung 28 ist der gesuchte Baum zum gegebenen Prüfer-Code $p = [5, 5, 2, 2, 2]$. Er sieht auf den ersten Blick anders aus als der in Abbildung 23, es ist aber derselbe! □

5.3 Beweis durch Bijektion (Joyal)

Wir wollen einen weiteren Beweis der Cayley-Formel durch ein Bijektionsargument kennenlernen. Dieses Argument geht auf A. Joyal (1981) zurück, siehe auch G. E. Lee, D. Zeilenberger (2012) und M. Aigner, G. M. Ziegler (2018, S. 262).

Wieder sei t_n die Anzahl bezeichneter Bäume mit n Ecken. Um $t_n = n^{n-2}$ zu zeigen, zeigen wir $n^2 t_n = n^n$. Die rechte Seite n^n ist die Anzahl der Abbildungen



$p = [2],\ b = [1, 3, 4, 5]$ $\qquad$ $p = [\,],\ b = [1, 3, 4, 5, 6]$

Abbildung 28: Die letzten drei Schritte

$f\colon \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$, während links die Anzahl $n^2 t_n$ der "doppelt verwurzelten" bezeichneten Bäume mit n Ecken steht. Damit meinen wir genauer: Wird in der Menge aller bezeichneten Bäume eine beliebige der n Ecken als Wurzel ausgezeichnet, so hat diese Menge der *Wurzelbäume* (diese spielen in Unterabschnitt 5.6 eine Rolle) nt_n Elemente. Werden zwei der Ecken ausgezeichnet, wobei diese zusammenfallen dürfen, so hat die entsprechende Menge der "doppelt verwurzelten" bezeichneten Bäume $n^2 t_n$ Elemente. Wenn wir also eine bijektive Abbildung zwischen der Menge aller Abbildungen $f\colon \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ und der Menge doppelt verwurzelter bezeichneter Bäume mit n Ecken angeben können, so ist erneut die Cayley-Formel bewiesen.

Sei $f\colon \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ eine beliebige Abbildung. Wir stellen diese in der Form

$$\begin{pmatrix} 1 & \cdots & n \\ f(1) & \cdots & f(n) \end{pmatrix}$$

dar. Dieser Abbildung ordnen wir einen gerichteten Graphen $\vec{G}_f$ zu, indem wir von i nach $f(i)$, $i = 1, \ldots, n$, eine gerichtete Kante bilden. Da $i = f(i)$ nicht ausgeschlossen ist, kann $\vec{G}_f$ *Schleifen* (oder *Schlingen*) enthalten.

Ist z. B.

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 5 & 4 & 6 & 2 & 1 & 7 \end{pmatrix},$$

so erhält man den in Abbildung 29 angegebenen gerichteten Graphen:

Jede der n Ecken von $\vec{G}_f$ ist Ausgangsecke genau einer gerichteten Kante. Daher gibt es in $\vec{G}_f$ mindestens einen gerichteten Kreis. Sei $M \subset \{1, \ldots, n\}$ Vereinigung der Ecken aller dieser Kreise, sei etwa $M = \{i_1, \ldots, i_r\}$ mit $i_1 < \cdots < i_r$. Die Einschränkung

$$f|_M = \begin{pmatrix} i_1 & \cdots & i_r \\ f(i_1) & \cdots & f(i_r) \end{pmatrix}$$

von f auf M ist eine bijektive Abbildung von M nach M. Genauer ist M eine maximale Teilmenge von $\{1, \ldots, n\}$ mit dieser Eigenschaft. Hieraus erhalten wir einen doppelt verwurzelten bezeichneten Baum, der aus dem Weg von $f(i_1)$ nach $f(i_r)$ mit den Wurzeln $f(i_1)$ und $f(i_r)$ besteht und die restlichen Knoten von $\vec{G}_f$ und die zugehörigen (ungerichteten) Kanten ergänzt wird.

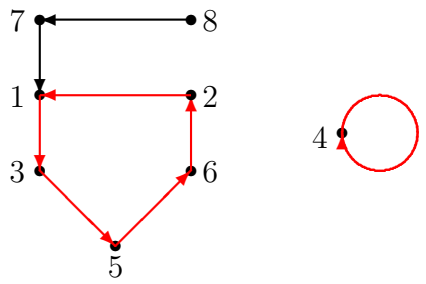


Abbildung 29: Gerichteter Graph $\vec{G}_f$ Graph zu obiger Abbildung f

In Abbildung 29 haben wir die beiden Kreise rot gezeichnet. Wir erhalten $M = \{1, 2, 3, 4, 5, 6\}$ und

$$f|_M = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 4 & 6 & 2 \end{pmatrix}$$

und damit den in Abbildung 30 angegebenen doppelt verwurzelten bezeichneten Baum.

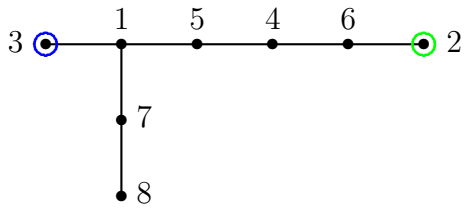


Abbildung 30: Doppelt verwurzelter Baum zu obiger Abbildung f

Umgekehrt sei ein doppelt verwurzelter bezeichneter Baum mit n Ecken gegeben. Dies gibt uns einen eindeutigen Weg P von der einen zur anderen Wurzel und damit die Menge M und die Abbildung $f|_M$. Die restlichen Zuweisungen $i \to f(i)$ werden mit Hilfe der eindeutigen Wege von i nach P ermittelt.

In unserem Beispiel erhalten wir aus dem doppelt verwurzelten Baum in Abbildung 30 zunächst $M = \{3, 1, 5, 4, 6, 2\}$ (Ecken des Weges von der einen zur anderen Wurzel), anschließend

$$f|_M = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 4 & 6 & 2 \end{pmatrix},$$

wobei in der ersten Zeile die Elemente aus M, der Größe nach geordnet, auftreten. Die noch fehlenden Ecken 7 und 8 haben 1 bzw. 7 als Nachbarn. Daher ist

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 5 & 4 & 6 & 2 & 1 & 7 \end{pmatrix}.$$

Beispiel: Sei

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 1 & 5 & 7 & 6 & 3 & 8 & 7 & 4 \end{pmatrix}.$$

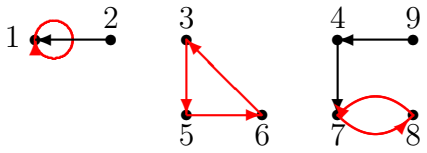


Abbildung 31: Der gerichtete Graph $\vec{G}_f$ zur Abbildung f

Den zugehörigen gerichteten Graphen $\vec{G}_g$ geben wir in Abbildung 31 an. Rot haben wir die Kreise in den jeweiligen Komponenten angegeben. Dann ist

$$M = \{1,3,5,6,7,8\}, \qquad g|_M = \begin{pmatrix} 1 & 3 & 5 & 6 & 7 & 8 \\ 1 & 5 & 6 & 3 & 8 & 7 \end{pmatrix}.$$

In Abbildung 32 geben wir den zu g gehörenden doppelt verwurzelten Baum an. Um

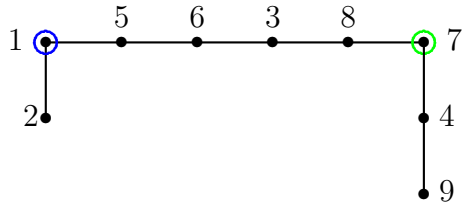


Abbildung 32: Doppelt verwurzelter Baum zu obiger Abbildung g

umgekehrt aus dem doppelt verwurzelten Baum aus 32 die zugrunde liegende Abbildung g zu rekonstruieren, gehen wir vor wie oben. Zunächst erhalten wir

$$M = \{1,3,5,6,7,8\}, \qquad g|_M = \begin{pmatrix} 1 & 3 & 5 & 6 & 7 & 8 \\ 1 & 5 & 6 & 3 & 8 & 7 \end{pmatrix},$$

anschließend

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 1 & 5 & 7 & 6 & 3 & 8 & 7 & 4 \end{pmatrix}.$$

□

5.4 Beweis durch Rekursion

Eine klassische Beweismethode besteht darin, das gegebene Problem mit Hilfe eines Parameters in eine allgemeinere Problemklasse einzubetten, diese mit Hilfe eines Induktionsbeweises bzw. einer Rekursion zu lösen und dadurch eine Lösung des eigentlichen Problems zu erhalten. Diese Vorgehensweise wollen wir hier demonstrieren.

Ein kreisfreier, aber nicht notwendig zusammenhängender Graph heißt ein *Wald*. Die (endlich vielen) Zusammenhangskomponenten eines Waldes sind also Bäume. Die Ecken des gegebenen Graphen seien mit $V := \{1,\ldots,n\}$ bezeichnet. Mit $k \in \{1,\ldots,n\}$ sei $t_{n,k}$ die Anzahl der (bezeichneten) Wälder mit k Komponenten, für die die Ecken $1,\ldots,k$ in verschiedenen Komponenten liegen. Die acht verschiedenen Wälder mit den vier Ecken

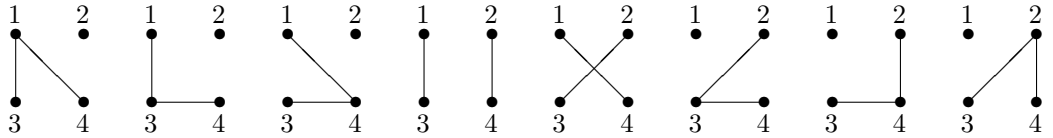


Abbildung 33: Es ist $t_{4,2} = 8$

$\{1, 2, 3, 4\}$, zwei Komponenten und der Eigenschaft, dass 1 und 2 in verschiedenen Zusammenhangskomponenten liegen, geben wir in Abbildung 33 an. Unser Ziel besteht darin, $t_{n,k} = kn^{n-k-1}$ nachzuweisen. Da $t_n = t_{n,1}$ die Anzahl bezeichneter Bäume mit n Knoten ist, ist dann auch die Cayley-Formel bewiesen. Wir wollen also den folgenden Satz beweisen, wobei wir uns im wesentlichen an L. Takács (1990), siehe aber auch M. Aigner, G. M. Ziegler (2018, S. 264 ff.) und C. Casarotto (2006), halten werden.

Satz 5.3 *Sei $n \in \mathbb{N}$ und $k \in \{1, \ldots, n\}$. Die Anzahl $t_{n,k}$ bezeichneter Wälder mit der Eckenmenge $\{1, \ldots, n\}$, k Zusammenhangskomponenten und der Eigenschaft, dass die Ecken $1, \ldots, k$ in verschiedenen Zusammenhangskomponenten liegen, ist durch*

$$t_{n,k} = kn^{n-k-1}$$

gegeben. Insbesondere ist die Anzahl t_n bezeichneter Bäume mit n Ecken durch

$$t_n = t_{n,1} = n^{n-2}$$

gegeben.

Beweis: Der Beweis basiert auf der Rekursionsformel

$$(*) \qquad t_{n,k} = \sum_{j=0}^{n-k} \binom{n-k}{j} t_{n-1,k-1+j},$$

wobei wir $t_{0,0} := 1$ und $t_{n,0} := 0$ für $n > 0$ setzen. Um $(*)$ zu beweisen, betrachten wir einen Wald zur Eckenmenge $\{1, \ldots, n\}$ mit k Zusammenhangskomponenten und der Eigenschaft, dass $1, \ldots, k$ in verschiedenen Zusammenhangskomponenten liegen. Die Ecke 1 kann keine der Ecken $2, \ldots, k$ als Nachbarn besitzen, daher hat sie $j \in \{0, \ldots, n-k\}$ (dies erklärt $\sum_{j=0}^{n-k}$ in $(*)$) Nachbarn in $\{k+1, \ldots, n\}$. Diese j Ecken nennen wir Ecken zweiter Art. Unter den $n-k$ Ecken $\{k+1, \ldots, n\}$ können wir auf $\binom{n-k}{j}$ unterschiedliche Weise j Ecken zweiter Art auswählen. Entfernt man die Ecke 1 und die j von 1 ausgehenden Kanten aus dem Wald, so erhält man einen Wald mit der Eckenmenge $\{2, \ldots, n\}$, der aus $k + j - 1$ Zusammenhangskomponenten (also Bäumen) besteht derart, dass die $n-1$ Ecken $\{2, \ldots, n\}$ und die j Ecken zweiter Art zu verschiedenen Bäumen gehören. Die Anzahl solcher Wälder ist $t_{n-1,k+j-1}$. Addieren wir $\binom{n-k}{j} t_{n-1,k+j-1}$ für alle möglichen Werte von j, nämlich $j = 0, \ldots, n-k$, so erhalten wir $t_{n,k}$. Damit ist $(*)$ bewiesen.

Wir zeigen nun $t_{n,k} = kn^{n-k-1}$ für $n \in \mathbb{N}$ und $k \in \{1, \ldots, n\}$ durch vollständige Induktion nach n. Die Aussage ist offensichtlich für $n = 1$ richtig. Wir nehmen an, es sei $t_{n-1,i} = i(n-1)^{n-i-2}$ für $i \in \{1, \ldots, n-1\}$ und $n > 1$. Wegen $(*)$ ist dann

$$\begin{aligned}
t_{n,k} &= \sum_{j=0}^{n-k} \binom{n-k}{j} t_{n-1,k-1+j} \\
&= \sum_{j=0}^{n-k} \binom{n-k}{j} (k-1+j)(n-1)^{n-k-j-1} \\
&= \sum_{i=0}^{n-k} \binom{n-k}{n-k-i} (n-1-i)(n-1)^{i-1} \\
&\qquad \text{(setze } i := n-k-j\text{)} \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1-i)(n-1)^{i-1} \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^i - \sum_{i=0}^{n-k} \binom{n-k}{i} i(n-1)^{i-1} \\
&= (1+n-1)^{n-k} - (n-k) \sum_{i=1}^{n-k} \binom{n-1-k}{i-1} (n-1)^{i-1} \\
&\qquad \text{(binomischer Lehrsatz)} \\
&= n^{n-k} - (n-k) \sum_{i=0}^{n-1-k} \binom{n-1-k}{i} (n-1)^i \\
&= n^{n-k} - (n-k) n^{n-1-k} \\
&\qquad \text{(binomischer Lehrsatz)} \\
&= kn^{n-1-k}.
\end{aligned}$$

Damit ist der Satz bewiesen. □

5.5 Beweis mit Hilfe linearer Algebra: Die Anzahl der einen zusammenhängenden Graphen aufspannenden Bäume

Wir halten uns in diesem Unterabschnitt vor allem an M. Aigner, G. M. Ziegler (2018, S. 263 ff.) und J. Matoušek, J. Nešetřil (2002, S. 259 ff.). Grundlegend in diesem Unterabschnitt ist der folgende Begriff. Ist $G = (V, E)$ ein zusammenhängender Graph, so heißt ein Untergraph $G' = (V, E')$ (dieselbe Eckenmenge wie G und eine Kantenmenge $E' \subset E$), der ein Baum ist, ein (den Graphen G) *aufspannender Baum* (engl.: *spanning tree*). Offenbar besitzt jeder zusammenhängende Graph G (mindestens) einen aufspannenden Baum. Entweder ist G bereits ein Baum, dann sind wir fertig, oder G besitzt einen Kreis C. Entfernen wir aus C eine beliebige Kante e, so ist $G_1 = (V, E \setminus \{e\})$ nach wie vor zusammenhängend. Auf G_1 kann dasselbe Verfahren angewandt werden und nach endlich vielen Schritten erhält man einen aufspannenden

Baum. Mit $t(G)$ bezeichnen wir die Anzahl der den Graphen G aufspannenden Bäume. Offenbar ist $t_n = t(K_n)$, also ist die Anzahl der (bezeichnete) Bäume mit n Ecken gleich der Anzahl der den vollständigen Graphen K_n aufspannenden Bäume. In Abbildung 20 haben wir die 16 den K_4 aufspannenden Bäume angegeben. Unser Ziel besteht darin, eine Formel für $t(G)$ für einen beliebigen zusammenhängenden Graphen anzugeben, aus der dann leicht für den Spezialfall $G := K_n$ die Cayley-Formel folgt.

Sei $G = (V, E)$ ein zusammenhängender Graph mit der Eckenmenge $V = \{1, \ldots, n\}$ und der Kantenmenge E, einer Menge von Paaren (i, j) mit $i, j \in \{1, \ldots, n\}$ und $i \neq j$. Wir gehen von *ungerichteten* Graphen aus, d. h. mit $(i, j) \in E$ ist auch $(j, i) \in E$. Daher stimmt eine Kante (i, j) mit (j, i) überein, sodass wir eigentlich besser $\{i, j\}$ statt (i, j) schreiben sollten. Mit diesem Hinweis wollen wir es aber bewenden lassen. Die *Laplace-Matrix* $L = (l_{ij}) \in \mathbb{Z}^{n \times n}$ zum Graphen $G = (V, E)$ ist definiert durch

$$l_{ij} := \begin{cases} \deg(i), & \text{falls } i = j, \\ -1, & \text{falls } i \neq j \text{ und } (i, j) \in E, \\ 0, & \text{falls } i \neq j \text{ und } (i, j) \notin E. \end{cases}$$

Mit der zu G gehörenden *Gradmatrix* $D := \operatorname{diag}(\deg(1), \ldots, \deg(n))$ sowie der *Adjazenzmatrix* $A = (a_{ij})$ mit

$$a_{ij} := \begin{cases} 1, & \text{falls } i \neq j \text{ und } (i, j) \in E, \\ 0, & \text{sonst,} \end{cases}$$

ist

$$L = D - A.$$

Die Laplace-Matrix eines Graphen steht in enger Verbindung mit der *Inzidenzmatrix*. Die Eckenmenge sei nach wie vor $V = \{1, \ldots, n\}$, sei $m := |E|$, also etwa $E = \{e_1, \ldots, e_m\}$. Die Kanten seien also auf eine bestimmte Weise durchnummeriert. Jetzt nehmen wir aber an, jede Kante sei *gerichtet*, hätte also einen Anfangs- und eine Endecke. Diese (beliebig) gerichteteten Kanten bezeichnen wir mit $\vec{e}_1, \ldots, \vec{e}_m$. Die zu diesem so entstandenen *gerichteten Graphen* $\vec{G}$ gehörige Inzidenzmatrix $B = (b_{ik}) \in \mathbb{Z}^{n \times m}$ ist definiert durch

$$b_{ik} := \begin{cases} 1, & i \text{ ist Anfangsecke von } \vec{e}_k, \\ -1, & i \text{ ist Endecke von } \vec{e}_k, \\ 0, & \text{sonst.} \end{cases}$$

In jeder Spalte der Inzidenzmatrix steht also genau eine 1 und eine -1. Der Zusammenhang zwischen der Laplace-Matrix L zu einem Graphen (diese hängt nur von der Nummerierung der Ecken von G ab) und einer Inzidenzmatrix B (die Nummerierung der Ecken sei dieselbe wie bei der Laplace-Matrix) ist dann gegeben durch:

- Es ist $L = BB^T$.

Denn: Es ist

$$(BB^T)_{ij} = \sum_{k=1}^{m} b_{ik} b_{jk}.$$

Für $i = j$ ist

$$b_{ik} b_{jk} = b_{ik}^2 = \begin{cases} 1, & i \text{ ist Anfangs- oder Endecke der gerichteten Kante } \vec{e}_k, \\ 0, & \text{sonst.} \end{cases}$$

Daher ist $\sum_{k=1}^{m} b_{ik}^2$ die Anzahl der Kanten, die i als Anfangs- oder Endecke besitzen, und das ist genau die Anzahl der i benachbarten Ecken bzw. der Grad $\deg(i)$ von i. Für $i \neq j$ ist

$$b_{ik} b_{jk} = \begin{cases} -1, & \text{falls } \vec{e}_k = (i,j) \text{ oder } \vec{e}_k = (j,i), \\ 0, & \text{sonst,} \end{cases}$$

und daher

$$\sum_{k=1}^{m} b_{ik} b_{jk} = \begin{cases} -1, & \text{falls } i \neq j \text{ und } (i,j) \in E, \\ 0, & \text{falls } i \neq j \text{ und } (i,j) \notin E. \end{cases}$$

Damit ist $L = BB^T$ bewiesen.

Beispiel: Gegeben sei der in Abbildung 34 angegebene Graph G (siehe den Wikipedia-Eintrag zur Laplace-Matrix). Die Laplace-Matrix zu diesem Graphen ist offenbar

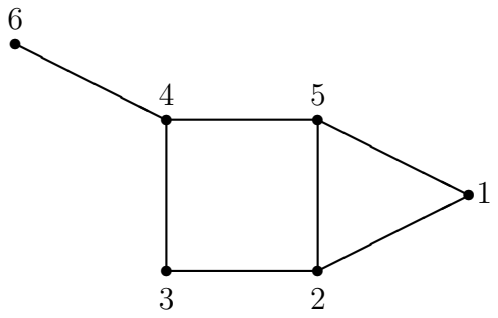


Abbildung 34: Was ist die Laplace-Matrix zu diesem Graphen?

$$L = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

Jetzt führen wir den Graphen G aus Abbildung 34 in einen gerichteten Graphen $\vec{G}$ über und nummerieren die gerichteten Kanten, siehe Abbildung 35. Die Inzidenzmatrix zu

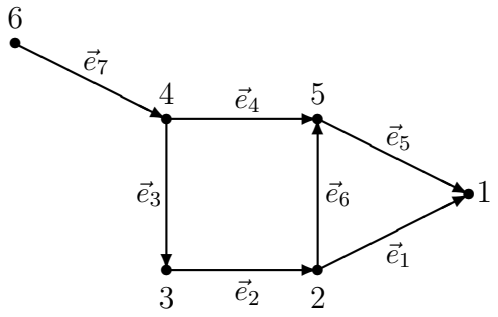


Abbildung 35: Was ist die Inzidenzmatrix zu diesem Graphen?

dem in Abbildung 35 angegebenen gerichteten Graphen ist offenbar

$$B = \begin{pmatrix} -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Die Beziehung $L = BB^T$ haben wir allgemein nachgewiesen. Wenn man unbedingt will, so kann man sie in diesem speziellen Beispiel überprüfen. □

Unser Ziel besteht darin, den folgenden Satz (häufig Matrix-Baum-Satz von Kirchhoff genannt) zu beweisen.

Satz 5.4 *Sei G ein zusammenhängender Graph mit n Ecken und L die zugehörige Laplace-Matrix. Dann ist die Anzahl $t(G)$ den Graphen G aufspannender Bäume gleich* $\det(L_{ii})$, $i = 1, \ldots, n$, *wobei L_{ii} durch Streichen der i-ten Zeile und der i-ten Spalte aus L hervorgeht.*

Beispiel: Für den Graphen G, in Abbildung 34 erhält man $\det(L_{11}) = 11$, nach dem (noch nicht bewiesenen) Satz 5.4 gibt es also genau elf verschiedene G aufspannende (bezeichnete) Bäume. Diese geben wir in Abbildung 36 an. □

Beispiel: Wir wollen uns überlegen, dass aus Satz 5.4 die Gültigkeit der Cayley-Formel folgt. Die Anzahl bezeichneter Bäume mit n Ecken ist gleich der Anzahl $t(K_n)$ der den vollständigen Graphen K_n aufspannenden Bäume. Die Laplace-Matrix zum K_n ist

$$L = \begin{pmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \cdots & n-1 \end{pmatrix} \in \mathbb{Z}^{n \times n}.$$

Streicht man hier die i-te Zeile und die i-te Spalte, so hat die resultierende Matrix

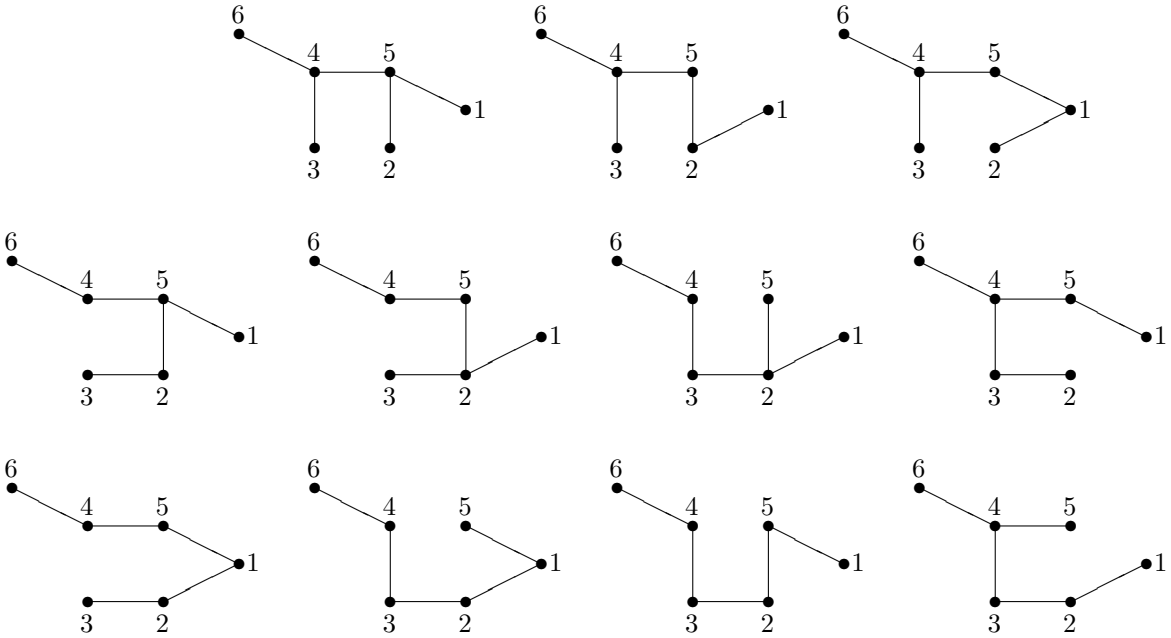


Abbildung 36: Elf aufspannende Bäume

dieselbe Form wie L, nur natürlich eine Zeile und eine Spalte weniger. D. h. es ist

$$L_{ii} = \begin{pmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \cdots & n-1 \end{pmatrix} = nI - ee^T \in \mathbb{Z}^{(n-1)\times(n-1)},$$

wobei I die $(n-1) \times (n-1)$-Einheitsmatrix und $e = (1, \ldots, 1)^T \in \mathbb{Z}^{n-1}$ ist. Mit Hilfe der Sherman-Morrison-Formel (siehe z. B. J. Werner (1992a, S. 29)) kann man leicht die Determinante von L_{ii} berechnen. Es gilt nämlich:

- *Sei $A \in \mathbb{R}^{m\times m}$ nichtsingulär und $u, v \in \mathbb{R}^m$. Dann gilt:*

  1. *Die Matrix $A + uv^T$ ist genau dann nichtsingulär, wenn $1 + v^T A^{-1} u \neq 0$.*
  2. *Ist $1 + v^T A^{-1} u \neq 0$, so gilt die Sherman-Morrison-Formel*

  $$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

  3. *Es ist* $\det(A + uv^T) = (1 + v^T A^{-1} u) \det(A)$.

Setzt man $m := n-1$, $A := nI$, $u := e$ und $v := -e$, so erhält man

$$\begin{aligned}
\det(L_{ii}) &= \det(A + uv^T) \\
&= (1 + v^T A^{-1} u) \det(A) \\
&= \left(1 - \frac{1}{n} e^T e\right) \det(nI) \\
&= \left(1 - \frac{n-1}{n}\right) n^{n-1} \\
&= n^{n-2}.
\end{aligned}$$

Damit ist nachgewiesen, dass die Cayley-Formel aus Satz 5.4 folgt. Etwas einfacher könnte man auch argumentieren, dass L_{ii} den $(n-2)$-fachen Eigenwert n mit Eigenvektoren aus dem $(n-2)$-dimensionalen linearen Raum span $\{e\}^\perp$ und den Eigenwert 1 zum Eigenvektor e besitzt. Da die Determinante einer Matrix das Produkt ihrer Eigenwerte ist, erhält man auch hier die Cayley-Formel. □

Wichtiges Hilfsmittel für den Beweis von Satz 5.4 ist der folgende Satz.

Satz 5.5 (Binet-Cauchy) *Sei*

$$A = (a_{ik}) = \begin{pmatrix} a_1 & \cdots & a_m \end{pmatrix} \in \mathbb{R}^{n \times m}, \qquad B = (b_{kj}) = \begin{pmatrix} b_1^T \\ \vdots \\ b_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}$$

mit $n \leq m$ gegeben. Dann ist

$$\det(A \cdot B) = \sum_{1 \leq k_1 < \cdots < k_n \leq m} \det(A^{(k_1,\ldots,k_n)}) \det(B_{(k_1,\ldots,k_n)}).$$

Hierbei ist

$$A^{(k_1,\ldots,k_n)} := \begin{pmatrix} a_{k_1} & \cdots & a_{k_n} \end{pmatrix} \in \mathbb{R}^{n \times n}, \qquad B_{(k_1,\ldots,k_n)} := \begin{pmatrix} b_{k_1}^T \\ \vdots \\ b_{k_n}^T \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Bemerkung: Ist $m = n$ in Satz 5.5, sind also A und B quadratische Matrizen derselben Dimension, so sagt der Satz aus, dass $\det(A \cdot B) = \det(A) \det(B)$. □

Beispiel: Seien

$$A := \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}, \qquad B := \begin{pmatrix} -8 & 7 \\ 6 & -5 \\ -4 & 3 \\ 2 & -1 \end{pmatrix}.$$

Wegen des (noch unbewiesenen) Satzes von Binet-Cauchy ist

$$\begin{aligned}
\det(A\cdot B) &= \det(A^{(1,2)})\cdot\det(B_{(1,2)}) + \det(A^{(1,3)})\cdot\det(B_{(1,3)}) \\
&\quad + \det(A^{(1,4)})\cdot\det(B_{(1,4)}) + \det(A^{(2,3)})\cdot\det(B_{(2,3)}) \\
&\quad + \det(A^{(2,4)})\cdot\det(B_{(2,4)}) + \det(A^{(3,4)})\cdot\det(B_{(3,4)}) \\
&= \det\begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix}\cdot\det\begin{pmatrix} -8 & 7 \\ 6 & -5 \end{pmatrix} + \det\begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix}\cdot\det\begin{pmatrix} -8 & 7 \\ -4 & 3 \end{pmatrix} \\
&\quad + \det\begin{pmatrix} 1 & 4 \\ 5 & 8 \end{pmatrix}\cdot\det(\begin{pmatrix} -8 & 7 \\ 2 & -1 \end{pmatrix} \\
&\quad + \det\begin{pmatrix} 2 & 3 \\ 6 & 7 \end{pmatrix}\cdot\det\begin{pmatrix} 6 & -5 \\ -4 & 3 \end{pmatrix} \\
&\quad + \det\begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix}\cdot\det\begin{pmatrix} 6 & -5 \\ 2 & -1 \end{pmatrix} \\
&\quad + \det\begin{pmatrix} 3 & 4 \\ 7 & 8 \end{pmatrix}\cdot\det\begin{pmatrix} -4 & 3 \\ 2 & -1 \end{pmatrix} \\
&= (-4)\cdot(-2) + (-8)\cdot 4 + (-12)\cdot(-6) \\
&\quad + (-4)\cdot(-2) + (-8)\cdot 4 + (-4)\cdot(-2) \\
&= 32.
\end{aligned}$$

Andererseits ist

$$A\cdot B = \begin{pmatrix} 0 & 2 \\ -16 & 18 \end{pmatrix}$$

und daher $\det(A\cdot B) = 32$. Die obige Rechnung zeigt sehr deutlich, dass der Satz von Binet-Cauchy *keine* effiziente Methode zur Berechnung von $\det(A\cdot B)$ beinhaltet! □

Beweis von Satz 5.5, dem Satz von Binet-Cauchy: Die *Determinante* einer quadratischen Matrix $C = (c_{ij}) \in \mathbb{R}^{n\times n}$ ist gegeben durch

$$\det(C) := \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) c_{1\sigma(1)}\cdots c_{n\sigma(n)}.$$

Hierbei ist S_n die Menge der Permutationen von $\{1,\ldots,n\}$, also der bijektiven Abbildungen $\sigma\colon \{1,\ldots,n\} \longrightarrow \{1,\ldots,n\}$, und

$$\operatorname{sgn}(\sigma) := (-1)^{|\operatorname{inv}(\sigma)|}$$

mit

$$\operatorname{inv}(\sigma) := \{(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\} : i<j, \sigma(i) > \sigma(j)\}$$

das *Vorzeichen* oder *Signum* der Permutation $\sigma \in S_n$. Dann ist

$$(A\cdot B)_{ij} = \sum_{k=1}^{m} a_{ik}b_{ki}, \qquad (i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$$

und daher

$$\begin{aligned}
\det(A\cdot B) &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\underbrace{\Big(\sum_{k_1=1}^{m}a_{1k_1}b_{k_1\sigma(1)}\Big)}_{=(A\cdot B)_{1\sigma(1)}}\cdots\underbrace{\Big(\sum_{k_n=1}^{m}a_{nk_n}b_{k_n\sigma(n)}\Big)}_{=(A\cdot B)_{n\sigma(n)}}\\
&= \sum_{k_1,\ldots,k_n=1}^{m}a_{1k_1}\cdots a_{nk_n}\underbrace{\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)b_{k_1\sigma(1)}\cdots b_{k_n\sigma(n)}}_{=\det(B_{(k_1,\ldots,k_n)})}\\
&= \sum_{(k_1,\ldots,k_n)\in[1:m]^n}a_{1k_1}\cdots a_{nk_n}\det(B_{(k_1,\ldots,k_n)})\\
&= \sum_{(k_1,\ldots,k_n)\in K}a_{1k_1}\cdots a_{nk_n}\det(B_{(k_1,\ldots,k_n)}),
\end{aligned}$$

wobei $[1:m]:=\{1,\ldots,m\}$ und

$$K:=\{(k_1,\ldots,k_n)\in[1:m]^n : k_i\neq k_j \text{ für } i\neq j\}.$$

Die letzte Gleichung gilt offensichtlich, denn stimmen zwei Indizes k_i und k_j überein, so sind in der $n\times n$-Matrix $B^{(k_1,\cdots,k_n)}$ zwei Zeilen gleich und daher verschwindet ihre Determinante, der entsprechende Summand kann also weggelassen werden. Nun definieren wir

$$K_<:=\{(k_1,\ldots,k_n)\in[1:m]^n : k_1<\cdots<k_n\}\subset K.$$

Dann gilt:

- Ist $(k_1,\ldots,k_n)\in K$, so existiert $\sigma\in S_n$ mit
$$(k'_1,\ldots,k'_n):=(k_{\sigma(1)},\ldots,k_{\sigma(n)})\in K_<.$$
- Ist $(k'_1,\ldots,k'_n)\in K_<$ und $\sigma\in S_n$, so ist
$$(k_1,\ldots,k_n):=(k'_{\sigma(1)},\ldots,k'_{\sigma(n)})\in K.$$

Daher ist

$$\begin{aligned}
\det(A\cdot B) &= \sum_{(k_1,\ldots,k_n)\in K}a_{1k_1}\cdots a_{nk_n}\det(B_{(k_1,\ldots,k_n)})\\
&= \sum_{(k'_1,\ldots,k'_n)\in K_<}\sum_{\sigma\in S_n}a_{1k'_{\sigma(1)}}\cdots a_{nk'_{\sigma(n)}}\det(B_{(k'_{\sigma(1)},\ldots,k'_{\sigma(n)})})\\
&= \sum_{(k'_1,\ldots,k'_n)\in K_<}\sum_{\sigma\in S_n}a_{1k'_{\sigma(1)}}\cdots a_{nk'_{\sigma(n)}}\operatorname{sgn}(\sigma)\det(B_{(k'_1,\ldots,k'_n)}\\
&= \sum_{(k'_1,\ldots,k'_n)\in K_<}\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{1k'_{\sigma(1)}}\cdots a_{nk'_{\sigma(n)}}\Big)\det(B_{(k'_1,\ldots,k'_n)}\\
&= \sum_{(k'_1,\ldots,k'_n)\in K_<}\det(A^{(k'_1,\ldots,k'_n)})\det(B_{(k'_1,\ldots,k'_n)})\\
&= \sum_{1\le k_1<\cdots<k_n\le m}\det(A^{(k_1,\ldots,k_n)})\det(B_{(k_1,\ldots,k_n)})
\end{aligned}$$

und das war zu zeigen. Der Satz von Binet-Cauchy ist bewiesen. □

Beweis von Satz 5.4: Sei $G = (V, E)$ ein zusammenhängender Graph mit den n Ecken $V = \{1, \ldots, n\}$ und $m := |E|$ Kanten. Wir können annehmen, es sei $m \geq n$. Denn wäre $m = n-1$, so wäre G selbst schon ein Baum und daher natürlich $t(G) = 1$. Sei $L \in \mathbb{Z}^{n\times n}$ die zugehörige Laplace-Matrix und $B \in \mathbb{Z}^{n\times m}$ eine (von der Orientierung der Kanten abhängende) Inzidenzmatrix, wobei die n Zeilen den Ecken und die m Spalten den (nummerierten) Spalten von G entsprechen. Greifen wir uns auf (beliebige) Art $n-1$ Spalten von B bzw. Kanten von G heraus, so hat man zusammen mit den n Ecken einen Untergraphen T von G, der offenbar genau dann ein (G aufspannender) Baum ist, wenn er zusammenhängend ist. Überraschenderweise kann dies mit Hilfe der Laplace- und der Inzidenzmatrix sowie des Satzes von Binet-Cauchy geklärt werden.

Es ist $L = BB^T$ wie wir uns oben überlegt haben. Wir haben zu zeigen, dass $\det(L_{ii})$ die Anzahl der G aufspannenden Bäume ist, wobei $L_{ii} \in \mathbb{Z}^{(n-1)\times(n-1)}$ aus L durch Streichen der i-ten Zeile und der i-ten Spalte entsteht. Wie hängt aber L_{ii} von der Inzidenzmatrix B ab? Ist

$$B = \begin{pmatrix} b_1^T \\ \vdots \\ b_n^T \end{pmatrix},$$

so ist

$$L = BB^T = \begin{pmatrix} b_1^T \\ \vdots \\ b_n^T \end{pmatrix} \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix} = \begin{pmatrix} b_1^T b_1 & \cdots & b_1^T b_n \\ \vdots & \ddots & \vdots \\ b_n^T b_1 & \cdots & b_n^T b_n \end{pmatrix}.$$

Daher ist $L_{ii} = B_i B_i^T$, wobei $B_i \in \mathbb{Z}^{(n-1)\times m}$ aus B dadurch entsteht, dass die i-te Zeile gestrichen wird. Wegen des Satzes 5.5 von Binet-Cauchy ist

$$\det(L_{ii}) = \sum_{1\leq k_1<\cdots<k_{n-1}\leq m} \det(B_i^{(k_1,\ldots,k_{n-1})})^2,$$

wobei $B_i^{(k_1,\ldots,k_{n-1})} \in \mathbb{Z}^{(n-1)\times(n-1)}$ aus B_i dadurch entsteht, dass nur die $(n-1)$ Spalten mit den Indizes $k_1, \ldots, k_{n-1}$ aufgenommen werden. Den $n-1$ Spalten von $B_i^{(k_1,\ldots,k_{n-1})}$ kann man einen Untergraphen T von G mit den n Ecken $V = \{1, \ldots, n\}$ und den $n-1$ Kanten $\{e_{k_1}, \ldots, e_{k_{n-1}}\}$ zuordnen. Mit der folgenden Hilfsaussage (siehe J. Matoušek, J. Nešetřil (2002, S. 261)) wird Satz 5.4 bewiesen sein.

- Sei $T = (V, E)$ ein Graph mit der Eckenmenge $V = \{1, \ldots, n\}$ mit $n \geq 2$ und der Kantenanzahl $|E| = n-1$. Weiter sei $\vec{T}$ ein zugehöriger (beliebig) gerichteter Graph und $B \in \mathbb{Z}^{n\times(n-1)}$ die entsprechende Inzidenzmatrix. Die Matrix $B_i \in \mathbb{Z}^{(n-1)\times(n-1)}$ entstehe aus B durch Streichen der i-ten Zeile, wobei $i \in \{1, \ldots, n\}$. Dann ist $\det(B_i) \in \{0, 1, -1\}$ und es ist $\det(B_i) \neq 0$ genau dann, wenn T ein Baum ist.

Denn: Wir führen vollständige Induktion über n. Der Induktionsanfang liegt bei $n = 2$. Dann hat T genau eine Kante, ist also ein aufspannender Baum. Je nach der Orientierung der Kante hat man die Situation links oder rechts in Abbildung 37. Man erhält

Abbildung 37: Graph mit zwei Ecken und einer Kante

die Inzidenzmatrizen $B = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ bzw. $B = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, nach dem Streichen einer Zeile hat man für die 1×1-Matrix B_i den Eintrag 1 oder -1.

Sei nun $n > 2$. Wir nehmen an, die Aussage sei für einen Graphen mit $n-1$ Ecken und $n-2$ Kanten richtig. Es werden zwei Fälle unterschieden.

Im ersten Fall nehmen wir an, eine der Ecken aus $\{1, \ldots, n\} \setminus \{i\}$ habe den Grad 1. Sei dies etwa die Ecke j. Diese ist also Ausgangs- oder Endecke genau einer (gerichteten) Kante $\vec{e}_k$. Für die Inzidenzmatrix B bedeutet dies, dass in der j-ten Zeile lauter Nullen stehen, außer in der k-ten Spalte, wo in der Position (j,k) als Eintrag 1 oder -1 steht. Entsprechendes gilt für die Matrix B_i, die aus B durch Streichen der i-ten Zeile entsteht. Entwickelt man $\det(B_i)$ nach der j-ten (bzw. $(j-1)$-ten) Zeile[23] , so erhält man

$$|\det(B_i)| = |\det(B_i^{(j,k)})|,$$

wobei $B_i^{(j,k)} \in \mathbb{Z}^{(n-2)\times(n-2)}$ aus B_i dadurch entsteht, dass die Zeile zur Ecke j und die k-te Spalte gestrichen werden. Nun betrachte man den gerichteten Graphen $\vec{T}^{(j,k)}$, der aus T durch Streichen der Ecke j und der einzigen von j ausgehenden oder dort endenden Kante $\vec{e}_k$ entsteht. Der Graph $\vec{T}^{(j,k)}$ besitzt $n-1$ Ecken, $n-2$ Kanten und hat die Matrix $B^{(j,k)}$, die aus B durch Streichen der j-ten Zeile sowie der k-ten Spalte entsteht, als Inzidenzmatrix. Die Induktionsannahme liefert, dass $|\det(B_i^{(j,k)})|$ gleich 1 oder 0 ist, je nachdem ob $T^{(j,k)}$ (die ungerichtete Version von $\vec{T}^{(j,k)}$) ein Baum ist oder nicht. Da T genau dann ein Baum ist, wenn es $T^{(j,k)}$ ist, ist die Aussage im ersten Fall bewiesen.

Im zweiten Fall hat keine der Ecken $\{1, \ldots, n\} \setminus \{i\}$ den Grad 1. Dann muss es in T eine isolierte Ecke, also eine Ecke mit dem Grad 0 geben. Denn andernfalls haben die $n-1$ Ecken aus $\{1, \ldots, n\} \setminus \{i\}$ mindestens den Grad 2 und die Ecke i mindestens den Grad 1, es wäre also

$$\sum_{j=1}^{n} \deg(j) \geq 2(n-1) + 1 > 2(n-1),$$

ein Widerspruch zur Aussage des "handshaking lemma", nach welchem $\sum_{v \in V} \deg(v) = 2|E|$ in einem beliebigen Graphen $G = (V, E)$ gilt. Also ist T kein Baum. Wir haben zu zeigen, dass $\det(B_i) = 0$. Existiert eine isolierte Ecke unter den Ecken $\{1, \ldots, n\} \setminus \{i\}$, so enthält B_i in der entsprechenden Zeile eine Nullzeile und insbesondere ist $\det(B_i) = 0$. Ist dagegen die Ecke i isoliert, so ist die Summe der Zeilen von B_i der Nullvektor, da

[23] Genauer: Ist $j < i$, so ist es die j-te Zeile, ist $j > i$, so ist es die $(j-1)$-te Zeile.

die Zeilensumme der Inzidenzmatrix B der Nullvektor ist und die weggelassene i-te Zeile eine Nullzeile ist. Auch in diesem Fall ist $\det(B_i) = 0$.

Der Satz 5.4 ist vollständig bewiesen. □

Beispiel: Bei N. HARTSFIELD, G. RINGEL (1990, S. 100 ff.) kann man nachlesen, dass die Anzahl den vollständigen bipartiten Gaphen $K_{m,n}$ aufspannender Bäume durch $t(K_{m,n}) = m^{n-1}n^{m-1}$ gegeben ist und der Beweis hierfür schwierig sei. Die Fälle $m = 2$ und $m = 3$ werden gesondert betrachtet. Wir wollen hier $t(K_{2,n}) = n2^{n-1}$ mit Hilfe von Satz 5.4 nachweisen. In Abbildung 38 stellen wir den vollständigen bipartiten Graphen $K_{2,n}$ dar. Die zugehörige Laplace-Matrix $L \in \mathbb{Z}^{(n+2)\times(n+2)}$ ist offenbar durch

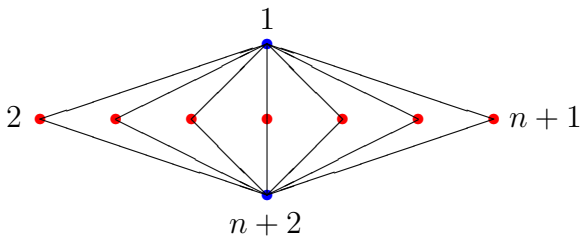


Abbildung 38: Der vollständige bipartite Graph $K_{2,n}$

$$L = D - A$$

mit

$$D := \operatorname{diag}(n, 2, \ldots, 2, n), \qquad A := \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 & 1 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 1 & 1 & \cdots & 1 & 1 & 0 \end{pmatrix}$$

gegeben. Gewinnt man $L_{11} \in \mathbb{Z}^{(n+1)\times(n+1)}$ aus L durch Streichen der ersten Zeile und der ersten Spalte, so erhält man

$$L_{11} = \begin{pmatrix} 2 & 0 & \cdots & 0 & -1 \\ 0 & 2 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2 & -1 \\ -1 & -1 & \cdots & -1 & n \end{pmatrix} = 2I + ne_{n+1}e_{n+1}^T - (ee_{n+1}^T + e_{n+1}e^T).$$

Hierbei ist natürlich I die $(n+1) \times (n+1)$-Einheitsmatrix, e der $(n+1)$-Vektor, dessen Komponenten sämtlich gleich 1 sind, und e_{n+1} der $(n+1)$-te Einheitsvektor. Die Matrix L_{11} hat 2 als $(n-1)$-fachen Eigenwert mit zugehörigen Eigenvektoren aus dem $(n-1)$-dimensionalen linearen Raum $\operatorname{span}\{e, e_{n+1}\}^\perp$. Wegen

$$\begin{aligned} L_{11}(\alpha e + \beta e_{n+1}) &= \alpha(e - e_{n+1}) + \beta((n+1)e_{n+1} - e) \\ &= (\alpha - \beta)e + (\beta(n+1) - \alpha)e_{n+1} \end{aligned}$$

erhält man die beiden restlichen Eigenwerte $\lambda_{1,2}$ von L_{11} als Eigenwerte der 2×2-Matrix $\begin{pmatrix} 1 & -1 \\ -1 & n+1 \end{pmatrix}$, also ist

$$\lambda_1\lambda_2 = \det \begin{pmatrix} 1 & -1 \\ -1 & n+1 \end{pmatrix} = n.$$

Insgesamt haben wir damit nachgewiesen, dass $\det(L_{11}) = n2^{n-1}$ die Anzahl der den vollständigen bipartiten Graphen $K_{2,n}$ aufspannenden Bäume ist. Wir haben hier einen Beweis mit Hilfe von Satt 5.4 angegeben. Allerdings ist ein direkter Beweis einfach. Wir stellen uns vor, im $K_{2,n}$ seien zwei Ecken a und b blau gefärbt, n Ecken seien rot. In jedem den $K_{2,n}$ aufspannenden Baum hat der eindeutige Weg von a nach b die Länge 2 und verläuft über eine rote Ecke x. Es gibt n Möglichkeiten, x zu wählen und für jeden der verbleibenden $n-1$ Ecken gibt es zwei Möglichkeiten, sie können nämlich zu a oder b benachbart sein. Dies ergibt insgesamt $n2^{n-1}$ aufspannende Bäume. □

5.6 Beweis durch doppeltes Zählen

Der folgende Beweis der Cayley-Formel stammt von J. Pitman (1999) und wird von J. Matoušek, J. Nešetřil (2002, S. 267 ff.) als der zurzeit wohl einfachste Beweis bezeichnet, was natürlich eine subjektive Bewertung ist. Wir werden auch die Darstellung bei M. Aigner, G.M. Ziegler (2018, S. 265 ff.) (zum Teil wörtlich) benutzen.

Wir nennen einen Graphen mit den Ecken $V = \{1, \ldots, n\}$ einen *Wurzelwald* (rooted forrest) auf $\{1, \ldots, \}$, wenn er ein Wald ist, also jede der Zusammenhangskomponenten ein Baum ist, und in jedem Komponentenbaum eine Ecke als Wurzel ausgezeichnet ist. Mit $\mathcal{F}_{n,k}$ bezeichnen wir die Menge aller Wurzelbäume mit n Ecken und k Zusammenhangsbäumen. Insbesondere ist $\mathcal{F}_{n,1}$ die Menge aller Wurzelbäume und $|\mathcal{F}_{n,1}| = nt_n$, wobei t_n die Anzahl bezeichneter Bäume mit n Ecken ist. Denn in jedem Baum haben wir n Möglichkeiten, die Wurzel auszuwählen.

Ein Wurzelwald $F_{n,k} \in \mathcal{F}_{n,k}$ kann als ein gerichteter Graph aufgefasst werden, in dem alle Kanten von den Wurzeln wegzeigen. Denn in jeder Komponente gibt es von der Wurzel zu einer Ecke dieser Komponente einen eindeutigen Weg und damit eine Folge gerichteter Kanten, die von der Wurzel wegzeigen. Die Richtung jeder Kante ist dann eindeutig bestimmt. Wir sagen, ein Wurzelwald F auf $\{1, \ldots, n\}$ *enthält* einen anderen Wurzelwald F' auf $\{1, \ldots, n\}$, in Zeichen $F \supset F'$, wenn im Sinne von gerichteten Graphen F' in F enthalten ist, also jede (gerichtete) Kante von F' auch eine (gerichtete) Kante in F ist. In Abbildung 39 geben wir ein Beispiel an. Wir definieren nun:

Definition 5.6 Wir nennen $F_1, \ldots, F_k$ eine *verfeinernde Kette*, wenn $F_i \in \mathcal{F}_{n,i}$, $i = 1, \ldots, k$, und $F_1 \supset F_2 \supset \cdots \supset F_k$.

Ist also $F_1, \ldots, F_k$ eine verfeinernde Kette, so ist F_1 ein Wurzelbaum mit n Ecken und die weiteren Wurzelwälder erhält man, indem sukzessive einzelne Kanten weggelassen werden. Wir zeigen nun den folgenden Satz, siehe J. Pitman (1999, Lemma 1).

Satz 5.7 *Sei $F_k \in \mathcal{F}_{n,k}$ ein (fester) Wurzelwald und $N(F_k)$ die Anzahl der Wurzelbäume mit n Ecken, die F_k enthalten. Dann ist $N(F_k) = n^{k-1}$.*

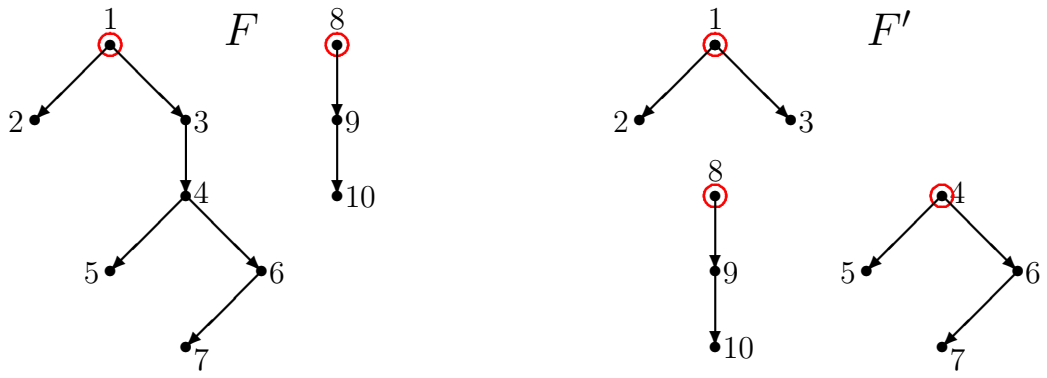


Abbildung 39: Der linke Wurzelwald F enthält den rechten F'

Beweis: Die tolle Idee beim Beweis besteht darin, zunächst $N^*(F_k)$ als die Anzahl verfeinernder Ketten $F_1, \ldots, F_k$ zu definieren und anschließend $N^*(F_k)$ auf zweierlei Weise zu berechnen, nämlich indem wir einmal bei F_1 und zum anderen bei F_k beginnen. Ist F_1 ein F_k enthaltender Wurzelbaum, so erhält F_1 genau $k-1$ Kanten, die nicht in F_k enthalten sind. Diese können in einer beliebigen Reihenfolge entfernt werden, um eine verfeinernde Kette $F_1, \ldots, F_k$ zu erhalten. Daher ist

$$(*) \qquad N^*(F_k) = N(F_k)(k-1)!.$$

Nun beginnen wir bei F_k und fragen uns, wie viele in F_k enthaltene Wurzelbäume $F_{k-1} \in \mathcal{F}_{n,k-1}$ es gibt. Hierzu wähle man beliebig eine der n Ecken und verbinde sie durch eine gerichtete Kante mit einer der $k-1$ Wurzeln derjenigen Teilbäume, die die ausgewählte Ecke nicht enthalten. Dies sind insgesamt $n(k-1)$ Möglichkeiten. Entsprechend gibt es $n(k-2)$ Wurzelbäume $F_{k-2} \in \mathcal{F}_{n,k-2}$, die in einem $F_{k-1} \in \mathcal{F}_{n,k-1}$ enthalten sind. Daher ist die Anzahl $N^*(F_k)$ verfeinernder Ketten $F_1, \ldots, F_k$ gegeben durch

$$(**) \qquad N^*(F_k) = n^{k-1}(k-1)!.$$

Aus $(*)$ und $(**)$ folgt

$$N(F_k) = n^{k-1} \qquad \text{für } F_k \in \mathcal{F}_{n,k}.$$

Der Satz ist damit bewiesen. □

Insbesondere ist $N(F_n) = n^{n-1}$. Da $F_n \in \mathcal{F}_{n,n}$ aus n isolierten Ecken besteht und eine der n Ecken als Wurzel gewählt werden kann, ist $N(F_n) = nt_n$, wobei t_n die Anzahl bezeichneter Bäume mit n Ecken ist. Daher erhalten wir als Folgerung aus Satz 5.7, dass $n^{n-1} = N(F_n) = nt_n$ bzw. $t_n = n^{n-2}$, also erneut die Gültigkeit der Cayley-Formel.

6 Projektive Ebenen

6.1 Definition, Beispiele

Zunächst werden projektive Ebenen sozusagen axiomatisch eingeführt.

Definition 6.1 Eine projektive Ebene ist durch ein Tripel $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ gegeben. Hierbei ist $\mathcal{P}$ eine Menge von *Punkten*, $\mathcal{G}$ eine Menge von *Geraden* und $\mathcal{I} \subset \mathcal{P} \times \mathcal{G}$ eine *Inzidenzstruktur*, wobei die folgenden Axiome erfüllt seien:

(a) Zu je zwei Punkten $x, y \in \mathcal{P}$ mit $x \neq y$ gibt es genau eine Gerade $L \in \mathcal{G}$, die mit x und y inzidiert, für die also $(x, L) \in \mathcal{I}$ und $(y, L) \in \mathcal{I}$. Diese Gerade werden wir häufig mit xy bezeichnen.

(b) Zu je zwei Geraden $L_1, L_2 \in \mathcal{G}$ mit $L_1 \neq L_2$ gibt es genau einen Punkt $x \in \mathcal{P}$, der mit L_1 und L_2 inzidiert, für den also $(x, L_1) \in \mathcal{I}$ und $(x, L_2) \in \mathcal{I}$. Diesen Punkt werden wir häufig mit $L_1 \cap L_2$ bezeichnen.

(c) Es existieren vier Punkte in $\mathcal{P}$ mit der Eigenschaft, dass es keine Gerade $L \in \mathcal{G}$ gibt, welche mit mehr als zwei dieser Punkte inzidiert.

Sind $\mathcal{P}$ und $\mathcal{G}$ endliche Mengen, so spricht man von einer *endlichen* projektiven Ebene.

Man beachte, dass die euklidische Ebene *keine* projektive Ebene ist, da (b) nicht erfüllt ist. Prominentestes Beispiel einer projektiven Ebene ist die *projektive Ebene über einem Körper*.

Definition 6.2 Sei $\mathbb{K}$ ein Körper, $\mathcal{P}$ die Menge der 1-dimensionalen Unterräume von $\mathbb{K}^3$, $\mathcal{G}$ die Menge der 2-dimensionalen Unterräume von $\mathbb{K}^3$ und

$$\mathcal{I} := \{(x, L) \in \mathcal{P} \times \mathcal{G} : x \subset L\}.$$

Dann nennt man das Tripel $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ eine *projektive Ebene über dem Körper* $\mathbb{K}$ und bezeichnen diese mit $\mathbb{KP}^2$.

Wir müssen uns noch davon überzeugen, dass $\mathbb{KP}^2$ den in Definition 6.1 geforderten Bedingungen genügt.

Satz 6.3 *Bei gegebenem Körper $\mathbb{K}$ ist $\mathbb{KP}^2$ eine projektive Ebene, genügt also den in Definition 6.1 formulierten Bedingungen.*

Beweis: (a) Seien x, y Punkte in $\mathbb{KP}^2$ bzw. 1-dimensionale Unterräume von $\mathbb{K}^3$ mit $x \neq y$. Dann existieren $\xi, \eta \in \mathbb{K}^3 \setminus \{0\}$ mit

$$x = \{\lambda\xi : \lambda \in \mathbb{K}\}, \qquad y = \{\mu\eta : \mu \in \mathbb{K}\}.$$

Wegen $x \neq y$ sind ξ und η linear unabhängig. Dann ist

$$L := \{\lambda\xi + \mu\eta : \lambda, \mu \in \mathbb{K}\}$$

eine x und y enthaltende Gerade in der projektiven Ebene $\mathbb{KP}^2$, und zwar offensichtlich die einzige Gerade mit dieser Eigenschaft.
(b) Sind L_1, L_2 Geraden in $\mathbb{KP}^2$ bzw. 2-dimensionale Unterräume von $\mathbb{K}^3$, so existieren linear unabhängige $\xi^{(1)}, \eta^{(1)} \in \mathbb{K}^3 \setminus \{0\}$ und $\xi^{(2)}, \eta^{(2)} \in \mathbb{K}^3 \setminus \{0\}$ mit

$$L_1 = \{\lambda_1\xi^{(1)} + \mu_1\eta^{(1)} : \lambda_1, \mu_1 \in \mathbb{K}\}, \qquad L_2 = \{\lambda_2\xi^{(2)} + \mu_2\eta^{(2)} : \lambda_2, \mu_2 \in \mathbb{K}\}.$$

Es ist $\xi^{(2)} \notin L_1$ oder $\eta^{(2)} \notin L_1$, denn andernfalls wäre $L_1 = L_2$. Wir nehmen o. B. d. A. an, es sei $\xi^{(2)} \notin L_1$. Dann sind $\xi^{(1)}, \eta^{(1)}, \xi^{(2)} \in \mathbb{K}^3 \setminus \{0\}$ linear unabhängig. Denn ist

$$\lambda_1 \xi^{(1)} + \mu_1 \eta^{(1)} + \lambda_2 \xi^{(2)} = 0,$$

so ist notwendig $\lambda_2 = 0$, denn andernfalls wäre $\xi^{(2)} \in L_1$, was wir gerade ausgeschlossen haben. Wegen der linearen Unabhängigkeit von $\xi^{(1)}, \eta^{(1)}$ ist auch $\lambda_1 = \mu_1 = 0$, insgesamt sind also $\xi^{(1)}, \eta^{(1)}, \xi^{(2)}$ linear unabhängig. Insbesondere ist $L_1 + L_2 = \mathbb{K}^3$. Wegen der Dimensionsformel für Unterräume ist

$$\dim(L_1 \cap L_2) = \dim(L_1) + \dim(L_2) - \dim(\underbrace{L_1 + L_2}_{=\mathbb{K}^3}) = 2 + 2 - 3 = 1.$$

Also ist $L_1 \cap L_2$ ein 1-dimensionaler Unterraum von $\mathbb{K}^3$ bzw. ein Punkt in $\mathbb{KP}^2$.
(c) Man setze

$$\xi^{(1)} := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \xi^{(2)} := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \xi^{(3)} := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \xi^{(4)} := \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

und anschließend

$$x_i := \{\lambda \xi^{(i)} : \lambda \in \mathbb{K}\}, \qquad i = 1, 2, 3, 4.$$

Da je drei der $\xi^{(i)} \in \mathbb{K}^3 \setminus \{0\}$, $i = 1, 2, 3, 4$, linear unabhängig sind, gibt es keine Gerade in $\mathbb{KP}^2$, die mehr als zwei der Punkte x_1, x_2, x_3, x_4 enthält. □

Beigegebenem Körper $\mathbb{K}$ sei $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$. Definiert man in $\mathbb{K}^3 \setminus \{0\}$ eine Äquivalenzrelation $\sim$ durch

$$\xi \sim \eta \Longleftrightarrow \text{Es existiert } \lambda \in \mathbb{K}^* \text{ mit } \xi = \lambda\eta,$$

so bestehen Punkte aus $\mathbb{KP}^2$ aus Äquivalenzklassen $[\xi]$, da alle zu einem $\xi \in \mathbb{K}^3 \setminus \{0\}$ äquivalenten Elemente aus $\mathbb{K}^3 \setminus \{0\}$ denselben 1-dimensionalen Unterraum erzeugen.

6.2 Endliche projektive Ebenen

Der folgende Satz ist entscheidend für die Untersuchung endlicher projektiver Ebenen. In ihm wird u. a. gezeigt, dass in einer endlichem projektiven Ebene die Anzahl der Punkte gleich der Anzahl der Geraden ist.

Satz 6.4 *Sei $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ eine endliche projektive Ebene. Man definiere*

$$x(L) := \{x \in \mathcal{P} : (p, L) \in \mathcal{I}\}, \quad L \in \mathcal{G},$$

(Anzahl der Punkte auf einer Geraden) sowie

$$L(x) := \{L \in \mathcal{G} : (x, L) \in \mathcal{I}\}, \quad x \in \mathcal{P},$$

(Anzahl der Geraden durch einen Punkt). Dann existiert $q \in \mathbb{N}$ mit

(a) Es ist $|x(L)| = q + 1$ für alle $L \in \mathcal{G}$,

(b) Es ist $|L(x)| = q + 1$ für alle $x \in \mathcal{P}$,

(c) Es ist $|\mathcal{P}| = q^2 + q + 1$,

(d) Es ist $|\mathcal{G}| = q^2 + q + 1$.

Beweis: (a) Seien $L_1, L_2 \in \mathcal{G}$ mit $L_1 \neq L_2$ gegeben. Wir zeigen, dass $|x(L_1)| = |x(L_2)|$, dass also auf jeder Geraden dieselbe Anzahl von Punkten liegt. Sei hierzu x der eindeutige Punkt, der auf L_1 und L_2 liegt und y ein Punkt, der weder auf L_1 noch L_2 liegt. Seien $x_1, \ldots, x_q$ die von x verschiedenen Punkte auf der Geraden L_1. Die zu y und x_i nach Definition 6.1 (a) eindeutig existierende Verbindungsgerade yx_i hat wegen Definition 6.1 (b) einen eindeutigen Schnittpunkt y_i mit L_2, $i = 1, \ldots, q$. Dies veranschaulichen wir uns in Abbildung 40. Die Punkte $y_1, \ldots, y_q$ auf L_2 sind

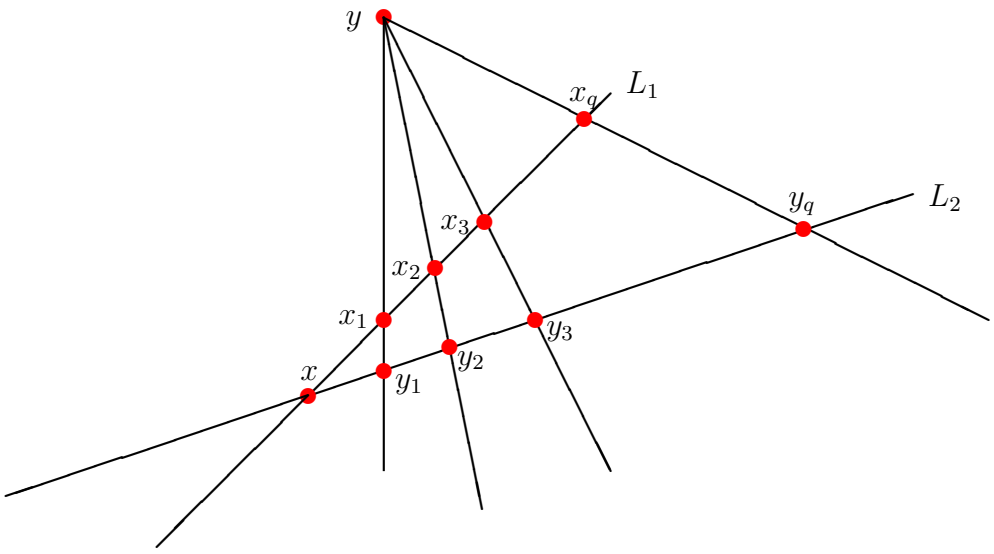


Abbildung 40: Zwei Geraden haben dieselbe Anzahl von Punkten

paarweise voneinander verschieden. Denn wäre $y_i = y_j$, so hätte die y und $y_i = y_j$ verbindende Gerade zwei voneinander verschiedene Schnittpunkte mit der Geraden L_1, ein Widerspruch. Daher ist die Anzahl der Punkte auf L_2 mindestens so groß wie auf L_1 bzw. $|x(L_2)| \geq |x(L_1)|$. Eine Vertauschung der Rollen von L_1 und L_2 ergibt die andere Ungleichung. Daher existiert ein $q \in \mathbb{N}$ mit $|x(L)| = q + 1$ für alle $L \in \mathcal{G}$.

(b) Hier überlegen wir uns zunächst:

- Jede Gerade $L \in \mathcal{G}$ enthält mindestens drei Punkte.

Denn: Sei $L \in \mathcal{G}$ eine Gerade und $x_1, x_2, x_3, x_4 \in \mathcal{P}$ die vier Punkte aus Definition 6.1 (c) mit der Eigenschaft, dass keine Gerade mehr als zwei dieser Punkte enthält bzw. jede Gerade höchstens zwei dieser Punkte enthält. Wir können daher o. B. d. A. annehmen, dass die Gerade L die Punkte x_1 und x_2 nicht enthält. Die Verbindungsgeraden x_1x_2, x_1x_3 und x_1x_4 sind paarweise verschieden wegen Definition 6.1 (c). Denn wäre z. B. $x_1x_2 = x_1x_3$, so wäre dies eine Gerade, die die drei Punkte x_1, x_2, x_3 enthält, was gerade ausgeschlossen ist. Dann haben aber die drei Geraden x_1x_2, x_1x_3 und x_1x_4 paarweise

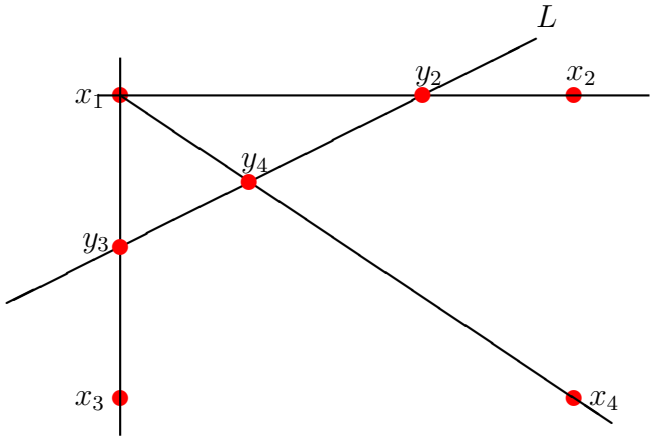


Abbildung 41: Jede Gerade enthält mindestens drei Punkte

verschiedene Schnittpunkte y_2, y_3, y_4 mit der Geraden L. Die Situation stellen wir in Abbildung 41 dar. Denn wäre z. B. $y_2 = y_3$, so würde dieser von x_1 verschiedene Punkt sowohl auf x_1x_2 als auch auf x_1x_3 liegen. Diese Geraden müssten also übereinstimmen, was nicht der Fall ist. Mit y_2, y_3, y_4 hat man damit paarweise verschiedene Punkte auf der Geraden L gefunden.

Als zweite Hilfsaussage formulieren und beweisen wir:

- Zu jedem Punkt $x \in \mathcal{P}$ existiert eine Gerade $L \in \mathcal{G}$, welche x nicht enthält.

Denn: Sei x ein Punkt und $L_1, L_2 \in \mathcal{G}$ zwei Geraden mit $L_1 \neq L_2$. Wir können annehmen, dass L_1 und L_2 den Punkt x enthalten, dieser also Schnittpunkt von L_1 und L_2 sind, weil man andernfalls fertig wäre. Wegen der gerade eben bewiesenen Hilfsaussage (diese wird nicht voll ausgenutzt) gibt es auf L_1 einen von x verschiedenen Punkt x_1 und auf L_2 einen von x verschiedenen Punkt x_2, siehe Abbildung 42. Die x_1 und x_2

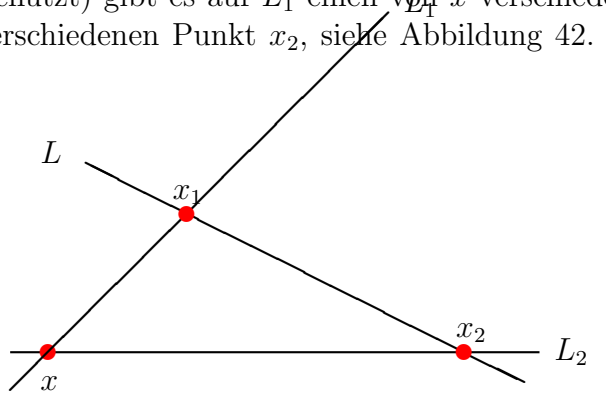


Abbildung 42: Zu jedem Punkt gibt es eine diesen Punkt nicht enthaltende Gerade

enthaltende Gerade L enthält den Punkt x nicht, da andernfalls L_1 gleich L_2 wäre.

Jetzt kommen wir zum Beweis dafür, dass jeder Punkt in $q + 1$ Geraden enthalten ist. Sei hierzu $x \in \mathcal{P}$ beliebig und $L \in \mathcal{G}$ eine Gerade, die x nicht enthält. Von deren Existenz haben wir uns in der zweiten Hilfsbehauptung überzeugt. Seien $x_1, \ldots, x_{q+1}$ die $q + 1$ Punkte, die in L nach (a) enthalten sind. Nach Definition 6.1 (a) gibt es für

$i = 1, \ldots, q+1$ eindeutige Geraden L_i, die x und x_i enthalten. Für $i \neq j$ ist $L_i \neq L_j$. Denn wäre $L_i = L_j$, so wären in L_i die Punkte x_i und x_j enthalten, wegen Definition 6.1 (a) wäre also $L_i = L_j = L$, ein Widerspruch dazu, dass x in L_i, aber nicht in L enthalten ist. Also gibt es mindestens $q+1$ Geraden, die x enthalten. Diesen einfachen Sachverhalt stellen wir in Abbildung 43 dar. Gäbe es eine weitere x enthaltende Gerade M, so gäbe

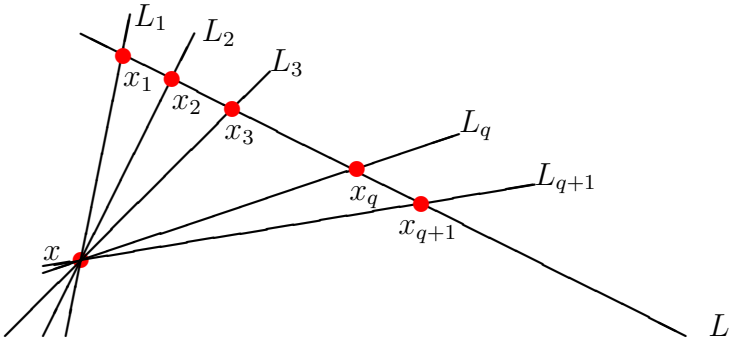


Abbildung 43: Jeder Punkt ist in $q+1$ Geraden enthalten

es wegen Definition 6.1 (b) auf L einen weiteren, von $x_1, \ldots, x_{q+1}$ verschiedenen Punkt, nämlich den Schnittpunkt von L und M. Ein Widerspruch zu (a).

(c) Wegen (a) und (b) wissen wir, dass jede Gerade $q+1$ Punkte enthält und es zu jedem Punkt $q+1$ Geraden gibt, die diesen Punkt enthalten. Sei x ein beliebiger Punkt der projektiven Ebene. Jede der $q+1$ Geraden, die x enthalten, enthält q weitere Punkte. Diese müssen sämtlich paarweise verschieden sein. Daher gibt es $(q+1)q+1 = q^2+q+1$ Punkte in der projektiven Ebene.

(d) Sei $L \in \mathcal{G}$ eine beliebige Gerade. Diese enthält wegen (a) $q+1$ Punkte $x_1, \ldots, x_{q+1}$. Zu jedem dieser $q+1$ Punkte x_i gibt es q weitere Geraden, die x_i enthalten, $i = 1, \ldots, q+1$. Diese sind paarweise verschieden. Daher gibt es $(q+1)q+1 = q^2+q+1$ Geraden in der projektiven Ebene.

Der Satz ist bewiesen. □

Jetzt spezialisieren wir uns auf den Fall, dass $\mathbb{K} := \mathbb{F}_q$ ein (endlicher) Körper mit q Elementen ist. Uns ist bekannt, dass dann notwendigerweise q eine Primzahlpotenz ist und andererseits es in diesem Falle im wesentlichen, d. h. bis auf Isomorphie, genau einen Körper mit q Elementen gibt, siehe z. B. J. Werner (2017, Abschnitt 4). Klar ist, dass die projektive Ebene $\mathbb{F}_q\mathbb{P}^2$ über dem endlichen Körper $\mathbb{F}_q$ eine endliche projektive Ebene ist. Wir wollen die Anzahl der in einer Geraden L enthaltenen Punkte berechnen. Seien $\xi, \eta \in \mathbb{F}_q^3 \setminus \{0\}$ linear unabhängig und

$$L := \{\lambda\xi + \mu\eta : \lambda, \mu \in \mathbb{F}_q\}$$

der hiervon erzeugte 2-dimensionale Unterraum bzw. Gerade in $\mathbb{F}_q\mathbb{P}^2$. Dann ist $|L| = q^2$ bzw. $|L \setminus \{0\}| = q^2 - 1$. Von 0 verschiedene Vielfache aus F_q (deren Anzahl ist $q-1$) von Elementen aus L erzeugen denselben 1-dimensionalen Unterraum von L. Daher ist

$$\frac{q^2-1}{q-1} = q+1$$

die Anzahl der 1-dimensionalen, in L enthaltenen Unterräume bzw. die Anzahl der in der Geraden L enthaltenen Punkte. Wegen Satz 6.4 ist $q+1$ auch die Anzahl der Geraden durch einen Punkt in $\mathbb{F}_q\mathbb{P}^2$ sowie q^2+q+1 jeweils die Anzahl der Punkte bzw. Geraden in $\mathbb{F}_q\mathbb{P}^2$. Die Anzahl der Punkte in $\mathbb{F}_q\mathbb{P}^2$ kann man natürlich auch direkt berechnen. Wir geben hierzu zwei Möglichkeiten an. Punkte in $\mathbb{F}_q\mathbb{P}^2$ sind 1-dimensionale Unterräume von $\mathbb{F}_q^3$ bzw. die Menge skalarer Vielfacher von Elementen aus $\mathbb{F}_q^3 \setminus \{0\}$. Offenbar gibt es q^3-1 Elemente in $\mathbb{F}_q^3 \setminus \{0\}$. Bei festem $\xi \in \mathbb{F}_q^3 \setminus \{0\}$ gibt es $q-1$ Elemente $\lambda\xi$, $\lambda \in \mathbb{F}_q^*$. Diese erzeugen alle denselben 1-dimensionalen Unterraum von $\mathbb{F}_q^3$ bzw. denselben Punkt in $\mathbb{F}_q\mathbb{P}^2$. Daher gibt es

$$\frac{q^3-1}{q-1} = 1+q+q^2$$

Punkte in $\mathbb{F}_q\mathbb{P}^2$. Dasselbe Ergebnis erhält man durch eine etwas andere Argumentation. Die Menge der Äquivalenzklassen $\mathbb{F}_q^3\setminus\{0\}/\sim$ kann man nämlich darstellen als disjunkte Zerlegung von Äquivalenzklassen, für deren Repräsentanten $\xi = (\xi_1,\xi_2,\xi_3)^T \in \mathbb{F}_q^3 \setminus \{0\}$ gilt: (a) $\xi_3 \neq 0$ und dann o. B. d. A. $\xi_3 = 1$, (b) $\xi_3 = 0$, $\xi_2 \neq 0$ und dann o. B. d. A. $\xi_2 = 1$und (c) $\xi_2 = \xi_3 = 0$ und dann o. B. d. A. $\xi_1 = 1$. Daher ist

$$\begin{aligned} |\mathbb{F}_q^3 \setminus \{0\}/\sim| &= \underbrace{\left|\left\{\left[\begin{pmatrix}\xi_1\\ \xi_2\\ 1\end{pmatrix}\right] : \xi_1,\xi_2 \in \mathbb{F}_q\right\}\right|}_{=q^2} + \underbrace{\left|\left\{\left[\begin{pmatrix}\xi_1\\ 1\\ 0\end{pmatrix}\right] : \xi_1 \in \mathbb{F}_q\right\}\right|}_{=q} \\ &\quad + \underbrace{\left|\left\{\left[\begin{pmatrix}1\\ 0\\ 0\end{pmatrix}\right]\right\}\right|}_{=1} \\ &= q^2+q+1. \end{aligned}$$

Beispiel: Sei $\mathbb{F}_3 := \{0,1,2\}$, wobei die Addition $+$ und die Multiplikation $\cdot$ mudulo 3 zu verstehen sind, also durch die folgenden Additions- bzw. Muliplikationstabellen gegeben sind:

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Wie wir uns gerade überlegt haben, gibt es $3^2+3+1 = 13$ Punkte in $\mathbb{F}_3\mathbb{P}^2$. Repräsentanten der entsprechenden Äquivalenzklassen sind

$$\begin{pmatrix}0\\0\\1\end{pmatrix}, \begin{pmatrix}0\\1\\1\end{pmatrix}, \begin{pmatrix}0\\2\\1\end{pmatrix}, \begin{pmatrix}1\\0\\1\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1\\2\\1\end{pmatrix}, \begin{pmatrix}2\\0\\1\end{pmatrix}, \begin{pmatrix}2\\1\\1\end{pmatrix}, \begin{pmatrix}2\\2\\1\end{pmatrix}$$

sowie

$$\begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}2\\1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}.$$

□

Beispiel: Der Körper $\mathbb{F}_2$ besteht aus den Elementen $\{0, 1\}$, die Additions- bzw. Multiplikationstabelle ist durch

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

gegeben. Die projektive Ebene $\mathbb{F}_2\mathbb{P}^2$ besteht aus 7 Punkten und 7 Geraden. Jede Gerade enthält 3 Punkte und durch jeden Punkt gehen 3 Geraden. Die entsprechende projektive Ebene heißt *Fano-Ebene* und ist in Abbildung 53 angegeben. Die die Punkte x_2, x_4, x_6

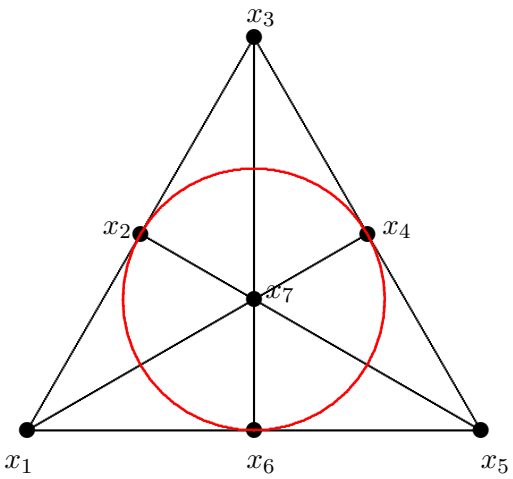


Abbildung 44: Die Fano-Ebene

enthaltende "Gerade" ist hierbei der rot gezeichnete Kreis. □

Bemerkung: Da es für jede Primzahlpotenz $q = p^m$ den (im wesentlichen eindeutigen) endlichen Körper $\mathbb{F}_q$ gibt, existiert für jede Primzahlpotenz q die projektive Ebene $\mathbb{F}_q\mathbb{P}^2$ über dem Körper $\mathbb{F}_q$. Es wird vermutet, dass eine projektive Ebene der Ordnung q *nur* für Primzahlpotenzen q existiert, siehe z. B. E. W. Weisstein (2018). Insbesondere ist durch einen Computerbeweis gezeigt worden, dass es keine projektive Ebene der Ordnung 10 gibt, siehe C. W. H. Lam (1996). □

Zum Schluss dieses Unterabschnitts über projektive Ebenen wollen wir noch den Begriff der *Inzidenzmatrix* einer endlichen projektiven Ebene einführen. Gegeben sei also eine projektive Ebene der Ordnung q. Mit $Q := q^2 + q + 1$ seien (in einer willkürlichen Nummerierung) $\{x_1, \ldots, x_Q\}$ die Punkte und $\{L_1, \ldots, L_Q\}$ die Geraden der projektiven Ebene. Die zugehörige Inzidenzmatrix $A = (a_{ij}) \in \{0, 1\}^{Q \times Q}$ ist definiert durch

$$a_{ij} := \begin{cases} 1, & \text{falls die Gerade } L_i \text{ den Punkt } x_j \text{ den Punkt } x_j \text{ enthält,} \\ 0, & \text{sonst.} \end{cases}$$

Beispiel: Wir kehren zur Fano-Ebene zurück, siehe Abbildung 53. Um die zugehörige Inzidenzmatrix aufzustellen, müssen wir auch die Geraden nummerieren. Wir geben

die entsprechenden Geraden durch drei Punkte an:

$$\begin{array}{rcl} L_1 & := & \{x_1, x_2, x_3\}, \\ L_2 & := & \{x_3, x_4, x_5\}, \\ L_3 & := & \{x_5, x_6, x_1\}, \\ L_4 & := & \{x_1, x_7, x_4\}, \\ L_5 & := & \{x_3, x_7, x_6\}, \\ L_6 & := & \{x_5, x_7, x_2\}, \\ L_7 & := & \{x_2, x_4, x_6\}. \end{array}$$

Als Inzidenzmatrix erhalten wir

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

Man erkennt, dass in jeder Zeile und jeder Spalte der Inzidenzmatrix A genau drei Einsen und sonst nur Nullen als Einträge enthalten sind. □

Einfache Eigenschaften einer Inzidenzmatrix einer endlichen projektiven Ebene formulieren wir in dem folgenden Satz.

Satz 6.5 *Mit $Q := q^2 + q + 1$ sei $A \in \{0,1\}^{Q \times Q}$ Inzidenzmatrix einer projektiven Ebene der Ordnung q. Dann gilt:*

1. *Es ist $AA^T = qI + E$, wobei I die $Q \times Q$-Einheitsmatrix und E die $Q \times Q$-Matrix ist, deren Einträge sämtlich gleich 1 sind.*
2. *Es ist $|\det A| = (q+1)q^{(q^2+q)/2}$.*

Beweis: Sei

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_Q^T \end{pmatrix},$$

also

$$a_i^T = \begin{pmatrix} a_{i1} & \cdots & a_{iQ} \end{pmatrix}$$

die i-te Zeile von A, $i = 1, \ldots, Q$. Berücksichtigt man, dass auf jeder Geraden L_i es genau $q+1$ Punkte gibt, für $i \neq j$ es genau einen Punkt x_k gibt, der auf L_i und L_j liegt, so erhalten wir

$$(AA^T)_{ij} = a_i^T a_j = \begin{cases} q+1, & \text{falls } i = j, \\ 1, & \text{falls } i \neq j, \end{cases}$$

womit der erste Teil des Satzes bewiesen ist. Für den zweiten Teil des Satzes wenden wir die Sherman-Morrison-Formel an, siehe z. B. J. Werner (1992a, S. 29). Mit dem Q-Vektor e, dessen Komponenten sämtlich gleich 1 sind, ist daher

$$\begin{aligned}
\det(A)^2 &= \det(AA^T) \\
&= \det(qI + ee^T) \\
&= \left(1 + e^T \frac{1}{q} Ie\right) \det(qI) \\
&= \left(1 + \frac{Q}{q}\right) q^Q \\
&= (q+Q)q^{Q-1} \\
&= (q+1)^2 q^{q^2+q}
\end{aligned}$$

und folglich, wie behauptet,

$$|\det(A)| = (q+1)q^{(q^2+q)/2}.$$

Damit ist der Satz bewiesen. □

6.3 Der Satz von Bruck-Ryser

6.3.1 Formulierung des Satzes von Bruck-Ryser

Wir wissen, dass es für jede Primzahlpotenz q eine projektive Ebene der Ordnung q gibt, nämlich die projektive Ebene $\mathbb{F}_q\mathbb{P}^2$ über dem endlichen Körper $\mathbb{F}_q$. Durch den Satz von Bruck-Ryser, siehe R. H. Bruck, H. J. Ryser (1949) kann für gewisse $q \in \mathbb{N}$ nachgewiesen werden, dass es *keine* projektive Ebene der Ordnung q gibt. Einen Beweis findet man auch in vielen Lehrbüchern, z. B. bei D. R. Hughes, F. C. Piper (1973, S. 87–89), D. R. Hughes, F. C. Piper (1985, S. 55 ff.), J. H. van Lint, R. M. Wilson (1992, S. 202 ff.), P. J. Cameron (1994, Section 9.8) oder auch bei S. Ball, Z. Weiner (2011). Eine Version des Satzes von Bruck-Ryser ist die folgende.

Satz 6.6 (Bruck-Ryser) *Existiert eine projektive Ebene $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ der Ordnung q und ist $q \equiv 1 \pmod 4$ oder $q \equiv 2 \pmod 4$, so ist q die Summe der Quadrate zweier ganzer Zahlen.*

Bemerkung: Anders gewendet sagt der Satz von Bruck-Ryser aus:

- *Ist $q \equiv 1 \pmod 4$ oder $q \equiv 2 \pmod 4$ und ist q nicht die Summe der Quadrate zweier ganzer Zahlen, so existiert keine projektive Ebene der Ordnung q.*

Für $q = 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, 19, 23, 25, 27, 29, 31, 32, 41, 43, 47, 49, 53, 59$ usw. existiert eine projektive Ebene der Ordnung q, da dies alles Primzahlen bzw. Primzahlpotenzen sind. Eine projektive Ebene der Ordnung $q = 6$ existiert wegen des Satzes von Bruck-Ryser nicht, da $6 \equiv 2 \pmod 4$ und 6 nicht die Summe der Quadrate zweier ganzer Zahlen ist. Anders ist dies für $q = 10$. Es ist $10 \equiv 2 \pmod 4$ und $10 = 3^2 + 1^2$. Der Satz von Bruck-Ryser liefert daher kein Ergebnis. Es wurde erwähnt,

dass durch einen Computerbeweis gezeigt wurde, dass es keine projektive Ebene der Ordnung 10 gibt. Fraglich ist, ob es projektive Ebenen der Ordnungen $12, 15, 18, 20, 24$ gibt, während der Satz von Bruck-Ryser die Nicht-Existenz projektiver Ebenen der Ordnungen $14, 21, 22, 30, 33, 38, 42, 46, 54, 57, 62$ usw. sichert. □

6.3.2 Zahlentheoretische Hilfsmittel

Wir beweisen zunächst die sogenannte *four-squares identity*:

Satz 6.7 (Four-Squares-Identity) *Für beliebige (r_1, r_2, r_3, r_4) und (x_1, x_2, x_3, x_4) ist*

$$y_1^2 + y_2^2 + y_3^2 + y_4^2 = (r_1^2 + r_2^2 + r_3^2 + r_4^2)(x_1^2 + x_2^2 + x_3^2 + x_4^2),$$

wobei

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} := \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \\ -r_2 & r_1 & r_4 & -r_3 \\ -r_3 & -r_4 & r_1 & r_2 \\ -r_4 & r_3 & -r_2 & r_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

Für ganzzahlige r_1, r_2, r_3, r_4 und x_1, x_2, x_3, x_4 sind also auch y_1, y_2, y_3, y_4 ganzzahlig.

Beweis: Man definiere

$$R := \left(\begin{array}{cc|cc} r_1 & r_2 & r_3 & r_4 \\ -r_2 & r_1 & r_4 & -r_3 \\ \hline -r_3 & -r_4 & r_1 & r_2 \\ -r_4 & r_3 & -r_2 & r_1 \end{array}\right) =: \left(\begin{array}{c|c} R_1 & R_2 \\ \hline -R_2 & R_1 \end{array}\right).$$

Mit $q := r_1^2 + r_2^2 + r_3^2 + r_4^2$ ist

$$R_1^T R_1 + R_2^T R_2 = \begin{pmatrix} r_1^2 + r_2^2 + r_3^2 + r_4^2 & 0 \\ 0 & r_1^2 + r_2^2 + r_3^2 + r_4^2 \end{pmatrix} = q \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

und

$$-R_1^T R_2 + R_2^T R_1 = 0,$$

folglich ist

$$R^T R = \begin{pmatrix} R_1^T R_1 + R_2^T R_2 & R_1^T R_2 - R_2^T R_1 \\ R_2^T R_1 - R_1^T R_2 & R_1^T R_1 + R_2^T R_2 \end{pmatrix} = q \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Daher ist

$$\begin{aligned} y_1^2+y_2^2+y_3^2+y_4^2 &= \left\| \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \right\|_2^2 \\ &= \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}^T R^T R \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \\ &= (r_1^2+r_2^2+r_3^2+r_4^2) \left\| \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \right\|_2^2 \\ &= (r_1^2+r_2^2+r_3^3+r_4^2)(x_1^2+x_2^2+x_3^2+x_4^2). \end{aligned}$$

Das war zu zeigen. □

Bemerkung: Es gilt entsprechend auch eine *two-squares identity*, nämlich:

- Für beliebige (r_1, r_2) und (x_1, x_2) ist

$$y_1^2+y_2^2 = (r_1^2+r_2^2)(x_1^2+x_2^2),$$

wobei

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := \begin{pmatrix} r_1 & r_2 \\ -r_2 & r_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Das kann natürlich ganz einfach nachgerechnet werden. □

Satz 6.8 *Es gilt:*

1. *Ist $p \in \mathbb{N}$ eine Primzahl und existieren ganze (o. B. d. A. nichtnegative) Zahlen x_1, x_2, die nicht beide verschwinden, mit $x_1^2+x_2^2 \equiv 0 \pmod p$, so ist p die Summe der Quadrate zweier ganzer Zahlen.*

2. *Ist $p \in \mathbb{N}$ eine Primzahl und existieren ganze (o. B. d. A. nichtnegative) Zahlen x_1, x_2, x_3, x_4, die nicht alle verschwinden, mit $x_1^2+x_2^2+x_3^2+x_4^2 \equiv 0 \pmod p$, so ist p die Summe der Quadrate von vier ganzen Zahlen.*

Beweis: Da $x_1^2+x_2^2$ eine natürliche Zahl ist und $x_1^2+x_2^2 \equiv 0 \pmod p$ gilt, existiert $r \in \mathbb{N}$ mit $rp = x_1^2+x_2^2$. Sei r minimal gewählt, d. h. ein kleineres Vielfaches von p als rp ist nicht Summe der Quadrate zweier ganzer Zahlen. Wir haben zu zeigen, dass $r = 1$. Angenommen, es sei $r \geq 2$. Man wähle ganze Zahlen u_1, u_2 mit

$$u_1 \equiv x_1 \pmod r, \quad u_2 \equiv x_2 \pmod r, \quad |u_i| \leq \frac{r}{2} \ \ (i = 1, 2).$$

Dies ist möglich. Denn $v_1 := x_1 \pmod r$ ist der Rest bei der Division von x_1 durch r und liegt daher in $[0, r-1]$. Dann setze man

$$u_1 := \begin{cases} v_1, & v_1 \le r/2, \\ v_1 - r, & v_1 > r/2. \end{cases}$$

Offensichtlich ist dann $u_1 - x_1$ ein ganzzahliges Vielfaches von r bzw. $u_1 \equiv x_1 \pmod r$ und $|u_1| \le r/2$. Entsprechend kann u_2 bestimmt werden. Mit $s_1, s_2 \in \mathbb{Z}$ ist dann

$$u_1 = x_1 + s_1 r, \qquad u_2 = x_2 + s_2 r$$

und folglich

$$u_1^2 + u_2^2 = x_1^2 + x_2^2 + 2r(x_1 s_1 + x_2 s_2) + r^2(s_1^2 + s_2^2).$$

Insbesondere ist

$$u_1^2 + u_2^2 \equiv x_1^2 + x_2^2 \equiv 0 \pmod r,$$

sodass eine nichtnegative ganze Zahl t mit $u_1^2 + u_2^2 = rt$ existiert. Hierbei ist $t < r$, denn wegen $|u_i| \le r/2$, $i = 1, 2$, ist

$$tr = u_1^2 + u_2^2 \le \frac{r^2}{4} + \frac{r^2}{4} = \frac{r^2}{2} < r^2.$$

Wäre $t = 0$, so wäre $u_1 = u_2 = 0$ und daher $x_1 \equiv x_2 \equiv 0 \pmod r$. Also sind x_1 und x_2 Vielfache von r und folglich $x_1^2 + x_2^2 = rp$ durch r^2 teilbar. Dies ist ein Widerspruch dazu, dass $r > 1$ und p eine Primzahl ist. Also ist $t \in \mathbb{N}$. Weiter ist

$$r^2 tp = (x_1^2 + x_2^2)(u_1^2 + u_2^2) = (x_1 u_1 + x_2 u_2)^2 + (x_1 u_2 - x_2 u_1)^2.$$

Wir zeigen, dass jeder der beiden Summanden auf der rechten Seite dieser Gleichung durch r^2 teilbar ist bzw.

$$(x_1 u_1 + x_2 u_2) \equiv (x_1 u_2 - x_2 u_1) \equiv 0 \pmod r$$

gilt. Dies ist aber klar, da

$$\begin{aligned} x_1 u_1 + x_2 u_2 &= x_1(x_1 + s_1 r) + x_2(x_2 + s_2 r) \\ &= x_1^2 + x_2^2 + (s_1 x_1 + s_2 x_2) r \\ &= r(p + s_1 x_1 + s_2 x_2) \end{aligned}$$

und entsprechend

$$\begin{aligned} x_1 u_2 - x_2 u_1 &= x_1(x_2 + s_2 r) - x_2(x_1 + s_1 r) \\ &= r(x_1 s_2 - x_2 s_1). \end{aligned}$$

Daher ist

$$\left(\frac{x_1 u_1 + x_2 u_2}{r}\right)^2 + \left(\frac{x_1 u_2 - x_2 u_1}{r}\right)^2 = tp$$

mit $1 \leq t < r$, ein Widerspruch dazu, dass rp das kleinste (natürliche) Vielfache von p ist, welches sich als Summe der Quadrate zweier ganzer Zahlen darstellen lässt. Damit ist der erste Teil des Satzes bewiesen.

Der zweite Teil des Satzes kann ganz ähnlich bewiesen werden. Sei r die kleinste natürliche Zahl mit der Eigenschaft, dass sich rp als Summe der Quadrate von vier ganzen Zahlen x_1, x_2, x_3, x_4 darstellen lässt. Wir nehmen an, es sei $r > 1$ und führen diese Annahme zum Widerspruch. Hierzu wähle man ganze Zahlen u_1, u_2, u_3, u_4 mit der Eigenschaft, dass

$$u_i \equiv x_i \pmod{r}, \quad |u_i| \leq \frac{r}{2}, \quad i = 1, 2, 3, 4.$$

Dies ist möglich, wie wir uns beim Beweis des Satzes von Bruck-Ryser überlegt haben. Dann ist

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 \equiv x_1^2 + x_2^2 + x_3^2 + x_4^2 \equiv 0 \pmod{r},$$

sodass eine ganze Zahl t mit $u_1^2 + u_2^2 + u_3^2 + u_4^2 = tr$ existiert. Wegen

$$(*) \qquad tr = u_1^2 + u_2^2 + u_3^2 + u_4^2 \leq \frac{r^2}{4} + \frac{r^2}{4} + \frac{r^2}{4} + \frac{r^2}{4} = r^2$$

ist $t \leq r$. Wäre $t = 0$, so wäre $u_1 = u_2 = u_3 = u_4 = 0$ und daher

$$x_1 \equiv x_2 \equiv x_3 \equiv x_4 \equiv 0 \pmod{r}$$

bzw. x_1, x_2, x_3, x_4 ganzzahlige Vielfache von r. Dann wäre aber

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = rp$$

durch r^2 teilbar. Dies ist ein Widerspruch dazu, dass $r > 1$ und p eine Primzahl ist. Also ist $t \geq 1$. Nun führen wir die Annahme $t = r$ zum Widerspruch. Aus $(*)$ folgt

$$u_1^2 = u_2^2 = u_3^2 = u_4^2 = \frac{r^2}{4},$$

insbesondere ist r gerade. Wegen $u_i \equiv x_i \pmod{r}$ existieren $s_i \in \mathbb{Z}$ mit $u_i = x_i + s_i r$, $i = 1, 2, 3, 4$. Daher ist

$$x_i = u_i - s_i r = \operatorname{sign}(u_i)\frac{r}{2} - s_i r = (\operatorname{sign}(u_i) - 2s_i)\frac{r}{2}, \quad i = 1, 2, 3, 4.$$

Hieraus liest man ab, dass $x_1^2+x_2^2+x_3^2+x_4^2 = rp$ durch r^2 teilbar ist, was ein Widerspruch dazu ist, dass $r > 1$ und p eine Primzahl ist. Also ist $1 \leq t < r$. Wegen der four-squares identity ist

$$\begin{aligned} r^2tp &= (tr)(rp) \\ &= (u_1^2 + u_2^2 + u_3^2 + u_4^2)(x_1^2 + x_2^2 + x_3^2 + x_4^2) \\ &= y_1^2 + y_2^2 + y_3^2 + y_4^2, \end{aligned}$$

wobei

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} := \begin{pmatrix} u_1 & u_2 & u_3 & u_4 \\ -u_2 & u_1 & u_4 & -u_3 \\ -u_3 & -u_4 & u_1 & u_2 \\ -u_4 & u_3 & -u_2 & u_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

Dann ist

$$\begin{aligned} y_1 &= u_1x_1 + u_2x_2 + u_3x_3 + u_4x_4 \\ &= x_1^2 + x_2^2 + x_3^2 + x_4^2 + (s_1x_1 + s_2x_2 + s_3x_3 + s_4x_4)r \\ &= (p + s_1x_1 + s_2x_2 + s_3x_3 + s_4x_4)r, \\ y_2 &= -u_2x_1 + u_1x_2 + u_4x_3 - u_3x_4 \\ &= (-s_2x_1 + s_1x_2 + s_4x_3 - s_3x_4)r, \\ y_3 &= -u_3x_1 - u_4x_2 + u_1x_3 + u_2x_4 \\ &= (-s_3x_1 - s_4x_2 + s_1x_3 + s_2x_4)r, \\ y_4 &= -u_4x_1 + u_3x_2 - u_2x_3 + u_1 2x_4 \\ &= (-s_4x_1 + s_3x_2 - s_2x_3 + s_1x_4)r. \end{aligned}$$

Also sind y_1, y_2, y_3, y_4 durch r und damit $y_1^2, y_2^2, y_3^2, y_4^2$ durch r^2 teilbar. Wegen

$$tp = \left(\frac{y_1}{r}\right)^2 + \left(\frac{y_2}{r}\right)^2 + \left(\frac{y_3}{r}\right)^2 + \left(\frac{y_4}{r}\right)^2$$

mit $t \in \mathbb{N}$ und $t < r$ hat man eine Darstellung von tp als Summe der Quadrate von vier ganzen Zahlen gefunden, was der Definition von r widerspricht. Damit ist auch der zweite Teil des Satzes bewiesen. □

Satz 6.9 *Ist $q \in \mathbb{N}$ die Summe der Quadrate zweier rationaler Zahlen, so ist q auch Summe der Quadrate zweier ganzer Zahlen.*

Beweis: Nach Voraussetzung gibt es $a, c \in \mathbb{N}$ und $b \in \mathbb{Z}$ mit $qc^2 = a^2 + b^2$. Zunächst nehmen wir an, q sei quadratfrei, also das Produkt paarweise verschiedener Primzahlen: $q = p_1p_2 \cdots p_k$. Dann ist $a^2 + b^2 = p_1p_2 \cdots p_kc^2$ ein Vielfaches von p_i bzw. $a^2 + b^2 \equiv 0 \pmod{p_i}$, $i = 1, \ldots, k$. Wegen des ersten Teiles von Satz **??** ist p_i die Summe der Quadrate zweier ganzer Zahlen, etwa $p_i = x_i^2 + y_i^2$, $i = 1, \ldots, k$. Dann ist

$$q = (x_1^2 + y_1^2)(x_2^2 + y_2^2) \cdots (x_k^2 + y_k^2).$$

Durch sukzessive $(k-1)$-malige Anwendung der two-squares identity (siehe die Bemerkung im Anschluss an Satz 6.7) erhalten wir, dass q die Summe der Quadrate zweier ganzer Zahlen ist. Nun nehmen wir an, q sei nicht quadratfrei, lasse sich also schreiben in der Form $q = nu^2$ mit einer quadratfreien natürlichen Zahl n und $u \in \mathbb{N}$. Dann ist $n(uc)^2 = a^2 + b^2$, sodass $n = r^2 + s^2$ mit ganzen Zahlen r, s. Insgesamt ist

$$q = nu^2 = (r^2 + s^2)u^2 = (ru)^2 + (su)^2,$$

also q auch in diesem Falle die Summe der Quadrate zweier ganzer Zahlen. □

Jetzt formulieren und beweisen wir noch einen berühmten Satz der Zahlentheorie.

Satz 6.10 (Lagrange) *Jede natürliche Zahl kann als Summe der Quadrate von vier ganzen Zahlen geschrieben werden.*

Beweis: Wegen der in Satz 6.7 formulierten four-squares identity ist mit zwei natürlichen Zahlen auch ihr Produkt Summe der Quadrate von vier ganzen Zahlen. Daher genügt es zu zeigen, dass jede Primzahl sich als Summe der Quadrate von vier ganzen Zahlen darstellen lässt. Wegen $2 = 1^2 + 1^2 + 0^2 + 0^2$ genügt es, ungerade Primzahlen zu behandeln. Jetzt zeigen wir:

- Sei p eine ungerade Primzahl. Dann existieren ganze Zahlen x_1, x_2 und m derart, dass $x_1^2 + x_2^2 + 1 = mp$ mit $0 < m < p$, insbesondere existieren ganze Zahlen x_1, x_2 mit $x_1^2 + x_2^2 + 1^2 + 0^2 \equiv 0 \pmod{p}$.

Denn: Sei

$$X_1 := \{x_1^2 : x_1 \in \{0, \ldots, (p-1)/2\}\}, \qquad X_2 := \{-x_2^2 - 1 : x_2 \in \{0, \ldots, (p-1)/2\}\}.$$

Dann sind die $(p+1)/2$ Elemente von X_1 modulo p sämtlich voneinander verschieden. Entsprechendes gilt für X_2. Denn sind $x_1, y_1 \in \{0, \ldots, (p-1)/2\}$ und $x_1^2 \equiv y_1^2 \pmod{p}$, so teilt p die ganze Zahl

$$x_1^2 - y_1^2 = (x_1 - y_1)(x_1 + y_1),$$

also $x_1 - y_1$ oder $x_1 + y_1$. Andererseits ist o. B. d. A.

$$0 \leq x_1 - y_1 \leq x_1 + y_1 \leq p - 1 < p,$$

daher ist $x_1 = y_1$. Entsprechend verläuft der Beweis für X_2. Die Mengen X_1 und X_2 sind disjunkt, da X_1 nur nichtnegative und X_2 nur negative Elemente enthält. Die Anzahl der Elemente von $X := X_1 \cup X_2$ ist $p + 1$. Wegen des Schubfachprinzips gibt es in X zwei verschiedene Elemente, die modulo p gleich sind. Diese können nicht beide in X_1 oder beide in X_2 liegen, wie wir gerade gezeigt haben. Daher existieren $x_1, x_2 \in \{0, \ldots, (p-1)/2\}$ mit $x_1^2 \pmod{p} = (-x_2^2 - 1) \pmod{p}$ bzw.

$$(x_1^2 + x_2^2 + 1) \equiv 0 \pmod{p}$$

und folglich $m \in \mathbb{Z}$ mit $x_1^2 + x_2^2 + 1 = mp$. Natürlich ist hier $m \in \mathbb{N}$ bzw. $0 < m$. Wegen

$$mp = x_1^2 + x_2^2 + 1 \leq \left(\frac{p-1}{2}\right)^2 + \left(\frac{p-1}{2}\right)^2 + 1 < p^2$$

ist $m < p$. Damit ist obige Zwischenbehauptung • bewiesen.

Nun ist der Rest des Beweises einfach. Wie wir gerade eben bewiesen haben, gibt es zu der ungeraden Primzahl p ganze Zahlen x_1, x_2 mit $(x_1^2 + x_2^2 + 1^2 + 0^2) \equiv 0 \pmod{P}$. Wegen des zweiten Teils von Satz **??** ist p die Summe der Quadrate von vier ganzen Zahlen. Damit ist der Satz von Lagrange bewiesen. □

Bemerkung: Z. B. ist $3 = 1^2+1^2+1^2$, $31 = 5^2+2^2+1^2+1^2$, $310 = 17^2+4^2+2^2+1^2$. Es können notwendige und hinreichende Bedingungen dafür angegeben werden, dass sich eine natürliche Zahl als Summe von zwei oder drei Quadraten ganzer Zahlen darstellen lässt. Darauf wollen wir nicht mehr eingehen. □

6.3.3 Der Beweis des Satzes von Bruck-Ryser

Beweis von Satz 6.6, dem Satz von Bruck-Ryser: Im ersten und entscheidenden Schritt zeigen wir:

- Ist $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ eine projektive Ebene der Ordnung $q \in \mathbb{N}$ und ist $q \equiv 1 \pmod 4$ oder $q \equiv 2 \pmod 4$, so existieren natürliche Zahlen a, c und eine ganze Zahl b mit $qc^2 = a^2 + b^2$, d. h. q ist Summe der Quadrate zweier rationaler Zahlen.

Mit $Q := q^2 + q + 1$ sei $A = (a_{ij}) \in \{0,1\}^{Q \times Q}$ eine Inzidenzmatrix der projektiven Ebene $(\mathcal{P}, \mathcal{G}, \mathcal{I})$ der Ordnung q. Bei einer gewissen Nummerierung der Punkte bzw. Geraden ist also $a_{ij} = 1$, falls der j-te Punkt auf der i-ten Geraden L_i liegt, andernfalls ist $a_{ij} = 0$. Sei $x = (x_1, \ldots, x_Q)^T$ zunächst beliebig und $z := A^T x$. Wegen Satz 6.5 ist $AA^T = qI + E$, wobei E die $Q \times Q$-Matrix ist, deren Einträge sämtlich gleich 1 sind, und $|\det(A)| = (q+1)q^{(q^2+q)/2}$, sodass A insbesondere nichtsingulär ist. Dann ist

$$\sum_{i=1}^{Q} z_i^2 = z^T z = x^T A A^T x = q x^T x + x^T E x = q \sum_{i=1}^{Q} x_i^2 + \left(\sum_{i=1}^{Q} x_i \right)^2.$$

Mit (noch unbestimmten) x_{Q+1} ist daher

$$\sum_{i=1}^{Q} z_i^2 + q x_{Q+1}^2 = q \sum_{i=1}^{Q+1} x_i^2 + \left(\sum_{i=1}^{Q} x_i \right)^2.$$

Wegen $q \equiv 1 \pmod 4$ oder $q \equiv 2 \pmod 4$ ist $Q + 1 = q^2 + q + 2$ ein Vielfaches von 4, wie eine einfache Rechnung zeigt. Daher ist

$$q \sum_{i=1}^{Q+1} x_i^2 = \sum_{i=0}^{(Q+1)/4-1} q(x_{4i+1}^2 + x_{4i+2}^2 + x_{4i+3}^2 + x_{4i+4}^2).$$

Nun wenden wir den *Vier-Quadrate-Satz von Lagrange* an, siehe Satz 6.10. Dieser besagt, dass jede natürliche Zahl als Summe der Quadrate von vier ganzen Zahlen geschrieben werden kann. Daher ist

$$q = r_1^2 + r_2^2 + r_3^2 + r_4^2$$

mit ganzen Zahlen r_1, r_2, r_3, r_4. Nun definiere man wie im Beweis der four-squares identity die Matric

$$R := \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \\ -r_2 & r_1 & r_4 & -r_3 \\ -r_3 & -r_4 & r_1 & r_2 \\ -r_4 & r_3 & -r_2 & r_1 \end{pmatrix}.$$

Wie wir im Beweis von Satz 6.7 gezeigt haben, ist $R^T R = qI$ und daher $\det(R)^2 = q^4$ und folglich R nichtsingulär. Sei weiter

$$\begin{pmatrix} y_{4i+1} \\ y_{4i+2} \\ y_{4i+3} \\ y_{4i+4} \end{pmatrix} := R \begin{pmatrix} x_{4i+1} \\ x_{4i+2} \\ x_{4i+3} \\ x_{4i+4} \end{pmatrix}, \qquad i = 0, \ldots, \frac{Q+1}{4} - 1.$$

Wegen der four-squares identity (siehe Satz 6.7) ist

$$y_{4i+1}^2 + y_{4i+2}^2 + y_{4i+3}^2 + y_{4i+4}^2 = (r_1^2 + r_2^2 + r_3^2 + r_4^2)(x_{4i+1}^2 + x_{4i+2}^2 + x_{4i+3}^2 + x_{4i+4}^2).$$

Für beliebige $x_1, \ldots, x_Q, x_{Q+1}$ ist daher

$$(*) \qquad \sum_{i=1}^{Q} z_i^2 + q x_{Q+1}^2 = \sum_{i=1}^{Q+1} y_i^2 + \left(\sum_{i=1}^{Q} x_i \right)^2,$$

wobei

$$\begin{pmatrix} z_1 \\ \vdots \\ z_Q \end{pmatrix} = A^T \begin{pmatrix} x_1 \\ \vdots \\ x_Q \end{pmatrix} = \begin{pmatrix} A^T & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_Q \\ x_{Q+1} \end{pmatrix}$$

und

$$\left(\begin{array}{c} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \hline \vdots \\ \hline y_{Q-2} \\ y_{Q-1} \\ y_Q \\ y_{Q+1} \end{array}\right) = \underbrace{\left(\begin{array}{c|c|c} R & \cdots & 0 \\ \hline \vdots & \ddots & \vdots \\ \hline 0 & \cdots & R \end{array}\right)}_{=:B} \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \hline \vdots \\ \hline x_{Q-2} \\ x_{Q-1} \\ x_Q \\ x_{Q+1} \end{array}\right)$$

bzw.

$$(**) \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_Q \\ x_{Q+1} \end{pmatrix} = B^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_Q \\ y_{Q+1} \end{pmatrix}, \qquad \begin{pmatrix} z_1 \\ \vdots \\ z_Q \end{pmatrix} = \underbrace{\begin{pmatrix} A^T & 0 \end{pmatrix}}_{=:C} \begin{pmatrix} y_1 \\ \vdots \\ y_Q \\ y_{Q+1} \end{pmatrix}.$$

Hierbei ist die Matrix $B \in \mathbb{Z}^{(Q+1)\times(Q+1)}$ nichtsingulär und folglich $B^{-1} \in \mathbb{Q}^{(Q+1)\times(Q+1)}$. Für beliebige $y_1, \ldots, y_{Q+1} \in \mathbb{Q}$ gilt also $(*)$, wobei $x_1, \ldots, x_{Q+1} \in \mathbb{Q}$ und $z_1, \ldots, z_Q \in \mathbb{Q}$ in Abhängigkeit von $y_1, \ldots, y_{Q+1}$ durch $(**)$ gegeben sind. Nun variieren wir nur noch $y_2, \ldots, y_{Q+1}$ und wählen y_1 in Abhängigkeit von $y_2, \ldots, y_{Q+1}$ so, dass $z_1^2 = y_1^2$. Dies kann folgendermaßen erreicht werden. Die erste Gleichung von

$$\begin{pmatrix} z_1 \\ \vdots \\ z_Q \end{pmatrix} = C \begin{pmatrix} y_1 \\ \vdots \\ y_{Q+1} \end{pmatrix}$$

laute

$$z_1 = c_{11} y_1 + c_{12} y_2 + \cdots + c_{1(Q+1)} y_{Q+1}.$$

Nun setze man

$$y_1 = y_1(y_2, \ldots, y_{Q+1}) := \begin{cases} \dfrac{1}{1-c_{11}}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}), & c_{11} \neq 1, \\ -\dfrac{1}{2}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}), & c_{11} = 1. \end{cases}$$

Für $c_{11} \neq 1$ und beliebige $y_2, \ldots, y_{Q+1}$ ist dann

$$\begin{aligned} z_1 &= \frac{c_{11}}{1-c_{11}}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) + c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1} \\ &= \left(\frac{c_{11}}{1-c_{11}}\right)(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) \\ &= \frac{1}{1-c_{11}}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) \\ &= y_1. \end{aligned}$$

Für $c_{11} = 1$ und beliebige $y_2, \ldots, y_{Q+1}$ ist dagegen

$$\begin{aligned} z_1 &= y_1 + c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1} \\ &= -\frac{1}{2}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) + (c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) \\ &= \frac{1}{2}(c_{12}y_2 + \cdots + c_{1(Q+1)}y_{Q+1}) \\ &= -y_1. \end{aligned}$$

Für beliebige $y_2, \ldots, y_{Q+1}$ ist daher

$$\sum_{i=2}^{Q} z_i^2 + qx_{Q+1}^2 = \sum_{i=2}^{Q+1} y_i^2 + w^2,$$

wobei $z_2, \ldots, z_Q, x_{Q+1}$ und w rationale Linearkombinationen (also Linearkombinationen mit rationalen Koeffizienten) von $y_2, \ldots, y_{Q+1}$ sind. In dieser Weise kann man fortfahren und erhält, dass

$$qx_{Q+1}^2 = y_{Q+1}^2 + w^2$$

für beliebige y_{Q+1}, wobei x_{Q+1} und w rationale Vielfache von y_{Q+1} sind. Wählt man nun für y_{Q+1} eine natürliche Zahl und denkt man sich die Gleichung $qx_{Q+1}^2 = y_{Q+1}^2 + w^2$ mit dem Quadrat einer geeigneten natürlichen Zahl multipliziert, so erhält man die Existenz natürlicher Zahlen a, c und $b \in \mathbb{Z}$ mit $qc^2 = a^2 + b^2$. D. h. q ist die Summe der Quadrate zweier rationaler Zahlen. Wegen Satz 6.9 ist q dann auch die Summe der Quadrate zweier ganzer Zahlen. Der Satz von Bruck-Ryser ist damit bewiesen. □

6.4 Die Sätze von Pappos und Desargues

Wir wollen in diesem Unterabschnitt zwei klassische Sätze, nämlich die Sätze von Pappos und Desargues, formulieren und beweisen. Hierbei gehen wir jeweils von einer projektiven Ebene $\mathbb{KP}^2$ über dem Körper $\mathbb{K}$ aus. Als Literatur nennen wir nur A. Beutelspacher, U. Rosenbaum (2004, S. 57 ff.).

Satz 6.11 (Pappos) *Gegeben sei die projektive Ebene $\mathbb{KP}^2$ über dem Körper $\mathbb{K}$. Seien x_1, y_1, z_1 verschiedene Punkte auf einer Geraden L_1 und x_2, y_2, z_2 verschiedene Punkte auf einer Geraden L_2, wobei $L_1 \neq L_2$ vorausgesetzt wird. Weiter nehmen wir an, dass x_1, y_1, z_1 sowie x_2, y_2, z_2 ungleich dem Schnittpunkt $a := L_1 \cap L_2$ der beiden Geraden L_1 und L_2 ist. Ist dann*

$$x := y_1 z_2 \cap z_1 y_2, \qquad y := x_1 z_2 \cap z_1 x_2, \qquad z := x_1 y_2 \cap y_1 x_2,$$

so sind x, y, z kollinear, liegen also auf einer Geraden L.

Beweis: In Abbildung 45 machen wir uns die Aussage des Satzes klar. Zunächst über-

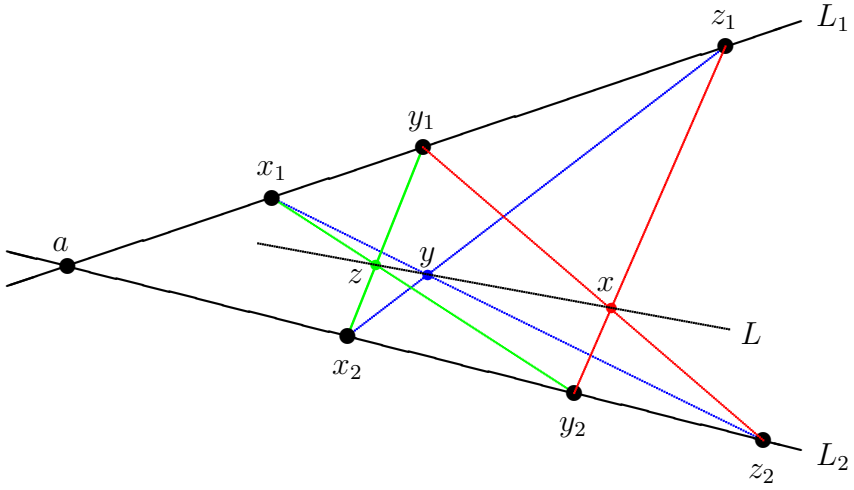


Abbildung 45: Der Satz von Pappos

legen wir uns, dass die Punkte x, y, z eindeutig bestimmt sind. Die Geraden $y_1 z_2$, $z_1 y_2$ usw., also Verbindungsgeraden zwischen Punkten aus L_1 und solchen aus L_2, sind eindeutig bestimmt. Denn wäre z. B. $y_1 = z_2$, so wäre $y_1 = z_2$ Schnittpunkt von L_1 und L_2, was wir ausgeschlossen haben. Weiter ist z. B. $y_1 z_2 \neq z_1 y_2$. Denn wäre $y_1 z_2 = z_1 y_2$, so gäbe es eine Gerade, auf der y_1, z_1, y_2, z_2 liegen. Durch y_1, z_1 ist die Gerade L_1, durch y_2, z_2 die Gerade L_2 bestimmt. Wir haben also einen Widerspruch dazu, dass $L_1 \neq L_2$ und die vorgegebenen Punkte von $a = L_1 \cap L_2$ verschieden sind. Die Punkte x, y und z sind daher als Schnittpunkte von jeweils zwei verschiedenen Geraden eindeutig bestimmt.

Sei $a = [\alpha]$ mit $\alpha \in \mathbb{K}^3 \setminus \{0\}$, wobei $[\alpha]$ den durch α erzeugten 1-dimensionalen Unterraum des $\mathbb{K}^3$ bezeichnet. Weiter sei $x_1 = [\xi_1]$ mit $\xi_1 \in \mathbb{K}^3 \setminus \{0\}$. Dann ist $L_1 = [\alpha, \xi_1]$ der von α und ξ_1 erzeugte 2-dimensionale Unterraum des $\mathbb{K}^3$ bzw. die Gerade durch a und x_1. Nachdem man notfalls ξ_1 mit einem Faktor multipliziert (wodurch sich $[\xi_1]$ nicht verändert) ist $y_1 = [\alpha + \xi_1]$ und $z_1 = [\alpha + \lambda\xi_1]$ mit $\lambda \in \mathbb{K} \setminus \{0, 1\}$. Entsprechend sei $y_2 = [\eta_2]$ mit $\eta_2 \in \mathbb{K}^3 \setminus \{0\}$ und damit $L_2 = [\alpha, \eta_2]$ sowie (nach Normierung) $x_2 = [\alpha + \eta_2]$ und $z_2 = [\alpha + \mu\eta_2]$ mit $\mu \in \mathbb{K} \setminus \{0, 1\}$. Insgesamt hat man also die Darstellungen

$$x_1 = [\xi_1], \qquad y_1 = [\alpha + \xi_1], \qquad z_1 = [\alpha + \lambda\xi_1]$$

und

$$x_2 = [\alpha + \eta_2], \qquad y_2 = [\eta_2], \qquad z_2 = [\alpha + \mu\eta_2],$$

wobei $\xi_1, \eta_2 \in \mathbb{K}^3\backslash\{0\}$, $\lambda, \mu \in \mathbb{K}\backslash\{0\}$ (da x_1, y_1, z_1 sowie x_2, y_2, z_2) paarweise verschieden sind). Dann ist

$$\begin{aligned} x &:= y_1 z_2 \cap z_1 y_2 \\ &= [\alpha + \xi_1, \alpha + \mu\eta_2] \cap [\alpha + \lambda\xi_1, \eta_2] \\ &= [\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2]. \end{aligned}$$

Um die letzte Gleichung einzusehen, beachten wir, dass

$$[\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2] = [\lambda(\alpha + \xi_1) + (1 - \lambda)(\alpha + \mu\eta_2)] \subset [[\alpha + \xi_1, \alpha + \mu\eta_2]$$

und

$$[\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2] = [\alpha + \lambda\xi_1) + (1 - \lambda)\mu\eta_2] \subset [\alpha + \lambda\xi_1, \eta_2].$$

Daher ist zunächst

$$[\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2] \subset [\alpha + \xi_1, \alpha + \mu\eta_2] \cap [\alpha + \lambda\xi_1, \eta_2].$$

Hier gilt sogar Gleichheit, da $\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2 \neq 0$. Weiter ist

$$\begin{aligned} y &:= x_1 z_2 \cap z_1 x_2 \\ &= [\xi_1, \alpha + \mu\eta_2] \cap [\alpha + \lambda\xi_1, \alpha + \eta_2] \\ &= [\alpha + \lambda\xi_1 + \mu\eta_2 - \mu\lambda\xi_1]. \end{aligned}$$

Die letzte Gleichung beweist man wie die entsprechende Stelle bei der Berechnung von x. Schließlich ist

$$\begin{aligned} z &:= x_1 y_2 \cap y_1 x_2 \\ &= [\xi_1, \eta_2] \cap [\alpha + \xi_1, \alpha + \eta_2] \\ &= [\xi_1 - \eta_2]. \end{aligned}$$

Hierzu beachte man, dass

$$[\xi_1 - \eta_2] \subset [\xi_1, \eta_2]$$

und

$$[\xi_1 - \eta_2] = [(\alpha + \xi_1) - (\alpha + \eta_2)] \subset [\alpha + \xi_1, \alpha + \eta_2].$$

Benutzen wir noch die Kommutativität von $\mathbb{K}$ bzw. $\mu\lambda = \lambda\mu$, so haben wir

$$x = [\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2], \quad y = [\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\xi_1], \quad z = [\xi_1 - \eta_2].$$

Daher ist

$$z = [\xi_1 - \eta_2] \subset xy = [\alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\eta_2, \alpha + \lambda\xi_1 + \mu\eta_2 - \lambda\mu\xi].$$

Also sind x, y und z kollinear, der Satz von Pappos ist bewiesen. □

Bemerkung: Wie wir gesehen haben, wird beim Beweis des Satzes von Pappos die Kommutativität des zugrunde liegenden Körpers $\mathbb{K}$ ausgenutzt. □

Satz 6.12 (Desargues) *In der projektiven Ebene $\mathbb{KP}^2$ über dem Körper $\mathbb{K}$ seien die beiden Dreiecke $\triangle x_1y_1x_1$ und $\triangle x_2y_2z_2$ mit der Eigenschaft gegeben, dass sich die Geraden x_1x_2, y_1y_2 und z_1z_2 in dem Punkt a schneiden*[24]*. Dann sind die Punkte*

$$x := y_1z_1 \cap y_2z_2, \qquad y := x_1z_1 \cap x_2z_2, \qquad z := x_1y_1 \cap x_2y_2$$

kollinear, liegen also auf einer Geraden.

Beweis: Zunächst veranschaulichen wir die Aussage des Satzes von Desargues in Abbildung 46. Sei $a = [\alpha]$ und

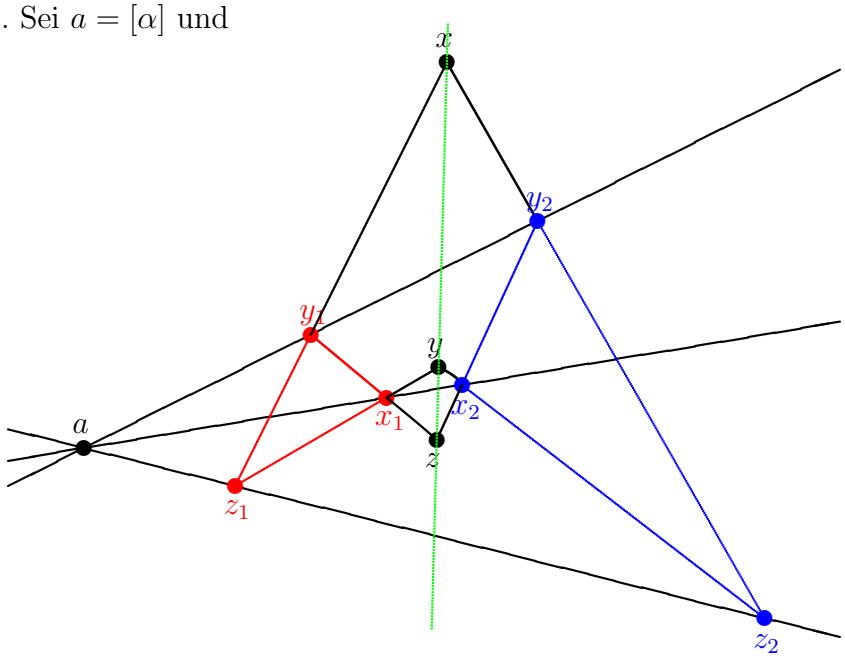


Abbildung 46: Der Satz von Desargues

$$x_i = [\xi_i], \quad y_i = [\eta_i], \quad z_i = [\zeta_i] \quad (i = 1, 2)$$

mit $\alpha \in \mathbb{K}^3 \setminus \{0\}$, $\xi_i, \eta_i, \zeta_i \in \mathbb{K}^3 \setminus \{0\}$, $i = 1, 2$. Da a, x_1, x_2 kollinear bzw. α, ξ_1, ξ_2 linear abhängig und ξ_1, ξ_2 wegen $x_1 \neq x_2$ linear unabhängig sind, existieren $\lambda_1^x, \lambda_2^x \in \mathbb{K}$ mit

$$\alpha = \lambda_1^x\xi_1 + \lambda_2^x\xi_2.$$

Hierbei sind $\lambda_1^x, \lambda_2^x \neq 0$, da x_1 und x_2 von a verschieden sind. Entsprechend existieren $\lambda_1^y, \lambda_2^y \in \mathbb{K} \setminus \{0\}$ sowie $\lambda_1^z, \lambda_2^z \in \mathbb{K} \setminus \{0\}$ mit

$$\alpha = \lambda_1^y\eta_1 + \lambda_2^y\eta_2 = \lambda_1^z\zeta_1 + \lambda_2^z\zeta_2.$$

Also ist

$$(*) \qquad \alpha = \lambda_1^x\xi_1 + \lambda_2^x\xi_2 = \lambda_1^y\eta_1 + \lambda_2^y\eta_2 = \lambda_1^z\zeta_1 + \lambda_2^z\zeta_2.$$

[24]Da wir von Dreiecken sprechen, ist implizit vorausgesetzt, dass die Punkte x_1, y_1, z_1 bzw. x_2, y_2, z_2 nicht kollinear sind. Weiter seien die Eckpunkte der beiden Dreiecke paarweise voneinander und von a verschieden.

Daher ist

$$\begin{aligned} x &:= y_1 z_1 \cap y_2 z_2 \\ &= [\eta_1, \zeta_1] \cap [\eta_2, \zeta_2] \\ &= [\lambda_1^y \eta_1 - \lambda_1^z \zeta_1]. \end{aligned}$$

Um die letzte Gleichung einzusehen beachten wir, dass $\lambda_1^y \eta_1 - \lambda_1^z \zeta_1 \neq 0$ (da $y_1 \neq z_1$), trivialerweise

$$[\lambda_1^y \eta_1 - \lambda_1^z \zeta_1] \subset [\eta_1, \zeta_1]$$

und wegen $(*)$

$$[\lambda_1^y \eta_1 - \lambda_1^z \zeta_1] = [\lambda_2^y \eta_2 - \lambda_2^z \zeta_2] \subset [\eta_2, \zeta_2]$$

gilt. Entsprechend ist

$$y = x_1 z_1 \cap x_2 z_2 = [\xi_1, \zeta_1] \cap [\xi_2, \zeta_2] = [\lambda_1^x \xi_1 - \lambda_1^z \zeta_1]$$

und

$$z = x_1 y_1 \cap x_2 y_2 = [\xi_1, \eta_1] \cap [\xi_2, \eta_2] = [\lambda_1^x \xi_1 - \lambda_1^y \eta_1].$$

Insgesamt ist also

$$x = [\lambda_1^y \eta_1 - \lambda_1^z \zeta_1], \qquad y = [\lambda_1^x \xi_1 - \lambda_1^z \zeta_1], \qquad z = [\lambda_1^x \xi_1 - \lambda_1^y \eta_1].$$

Wiederum wegen $(*)$ ist

$$x = [\lambda_1^y \eta_1 - \lambda_1^z \zeta_1] \subset [\lambda_1^y \eta_1 - \lambda_x \xi_1, \lambda_1^x \xi_1 - \lambda_1^z \zeta_1] = yz,$$

d. h. x, y und z sind kollinear. Der Satz von Desargues ist bewiesen. □

7 Steiner Systeme[25]

7.1 Kirkman's Schulmädchenproblem

Kirkman's schoolgirl problem bzw. das *Problem der 15 Schulmädchen* wurde 1850 von Thomas Kirkman formuliert. Es lautet:

- Fifteen young ladies in a school walk out three abreast[26] for seven days in succession: it is required to arrange them daily, so that no two shall walk twice abreast.

Bzw. (wir folgen der Übersetzung von W. Ahrens (1901, S. 274)):

[25] Diesen Abschnitt über Steiner Systeme widme ich meinem verstorbenen Kollegen H. L. de Vries. Ich bedaure es sehr, dass ich erst begonnen habe, mich für dieses Gebiet zu interessieren, als ich mit ihm nicht mehr darüber diskutieren konnte. Gespräche mit ihm hätten diesen Text ganz sicher origineller und damit besser gemacht.

[26] Seite an Seite, nebeneinander

- 15 Pensionatsmädchen gehen jeden Tag miteinander spazieren, je 3 in einer Reihe; wie sind die Anordnungen für die einzelnen Tage zu treffen, wenn im Laufe einer Woche jede gerade einmal mit jeder anderen gehen soll?

Da jedes Mädchen an jedem Tag mit 2 anderen Mädchen zusammen in einer Reihe geht, so kann es in den 7 Tagen der Woche mit 14 verschiedenen, d. h. mit allen anderen je einmal zusammentreffen. Es fragt sich nur, ob sich eine Anordnung treffen lässt, dass für alle 15 diese Forderung erfüllt ist.

Zu einer Menge S (von Schülerinnen) mit $|S| = 15$ hat man also $7 \cdot 5 = 35$ Tripel aus Elementen von S so zu bestimmen, dass jedes Paar von Elementen aus S in genau einem Tripel enthalten ist. Bei W. W. Rouse Ball (1905, S. 103 ff.) werden klassische Lösungsansätze beschrieben. Es gibt sieben nichtisomorphe Lösungen zu Kirkman's Problem. Wir geben eine Lösung an. Die 15 Mädchen seien mit $a_1, b_1, \ldots, g_1$, $a_2, b_2, \ldots, g_2$ und k bezeichnet. Eine mögliche Aufteilung wäre dann die folgende:

Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Sonnabend	Sonntag
a_1a_2k	b_1b_2k	c_1c_2k	d_1d_2k	e_1e_2k	f_1f_2k	g_1g_2k
$b_1e_2d_2$	$c_1f_2e_2$	$d_1g_2f_2$	$e_1a_2g_2$	$f_1b_2a_2$	$g_1c_2b_2$	$a_1d_2c_2$
$c_1b_2g_2$	$d_1c_2a_2$	$e_1d_2b_2$	$f_1e_2c_2$	$g_1f_2d_2$	$a_1g_2e_2$	$b_1a_2f_2$
$d_1f_1g_1$	$e_1a_1b_1$	$f_1a_1b_1$	$g_1b_1c_1$	$a_1c_1d_1$	$b_1d_1e_1$	$c_1e_1f_1$
$e_1f_2c_2$	$f_1g_2d_2$	$g_1a_2e_2$	$a_1b_2f_2$	$b_1c_2g_2$	$c_1d_2a_2$	$d_1e_2b_2$

Z. B. gehen (e_1, a_2) nur am Donnerstag nebeneinander. Diese Lösung können wir *zyklisch* nennen, da man die Kombination an einem bestimmten Wochentag dadurch erhält, dass man in der Kombination des vorherigen Wochentages a_1 durch b_1, b_1 durch c_1 usw., schließlich g_1 durch a_1 ersetzt, entsprechend a_2 durch b_2 usw., schließlich g_2 durch a_2. Eine Veranschaulichung dieser Lösung ist in Abbildung 47 angegeben. Wir haben in Abbildung 47 die Aufteilung der Schulmädchen am Montag, dem ersten Tag, links angegeben. Das erste Tripel a_1a_2k ist rot gezeichnet, das zweite Tripel $b_1e_2d_2$ blau, das dritte $c_1b_2g_2$ grün, das vierte $d_1f_1g_1$ gelb und das fünfte Tripel $e_1f_2c_2$ magenta. Die Aufteilung der Mädchen in Dreiergruppen am Dienstag erhält man, indem um den Winkel $2\pi/7$ im mathematisch positiven Sinn gedreht wird. Die entsprechende Aufteilung ist in Abbildung 47 rechts angegeben.

Das Problem der neun Schulmädchen ist etwas einfacher.

- Neun Schulmädchen spazieren vier Tage hintereinander in drei Dreiergruppen. Man teile sie täglich so ein, dass keine zwei Schulmädchen zweimal zusammen spazieren.

Eine Lösung hierzu geben wir in Abbildung 48 an. Die Mädchen seien von 1 bis 9 durchnummeriert. Die vier verschiedenen Tage entsprechen vier verschiedenen Farben, nämlich rot, blau, grün und schwarz. Eine weitere Visualisierung einer Lösung des Problems der neun Schulmädchen ist in Abbildung 49 angegeben, wir folgen hier A. Sodhi (2007). Die Mädchen 1 bis 8 finden sich, gleichmäßig verteilt, auf dem Rande der vier Kreise, das Mädchen 9 im Zentrum. Von Tag zu Tag wird der Kreis um den

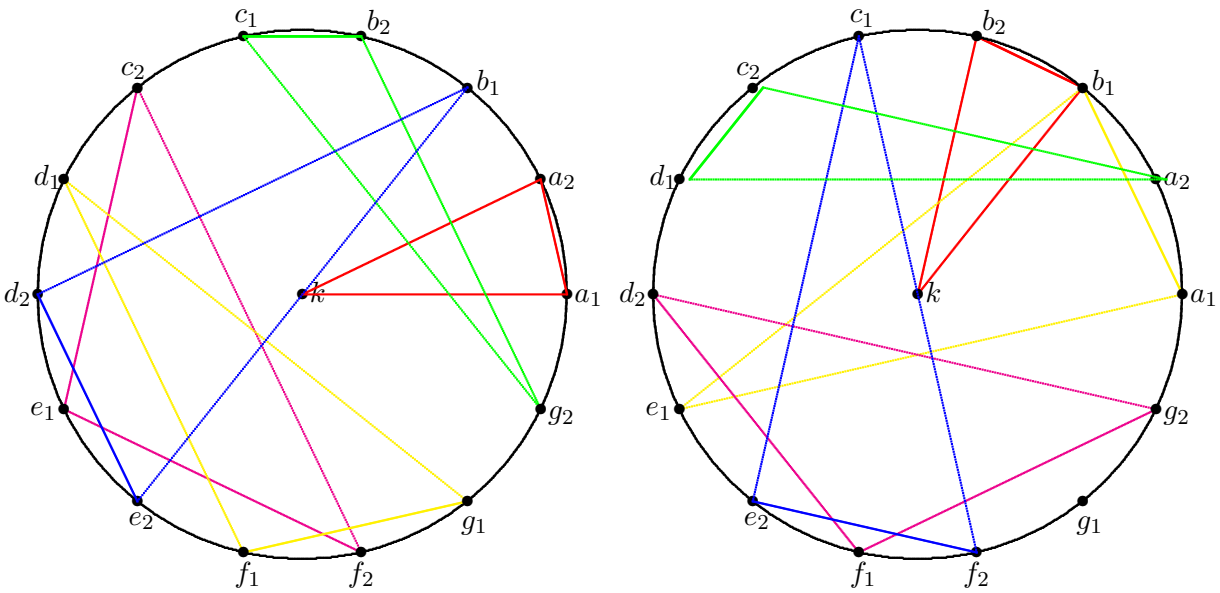


Abbildung 47: Eine zyklische Lösung von Kirkman's Schulmädchenproblem

Winkel 45° im Uhrzeigersinn gedreht. Damit sind in Abbildung 48 bzw. Abbildung 49 die folgenden "Aufstellungspläne" veranschaulicht:

Tag 1	Tag 2	Tag 3	Tag 4
1 2 3	1 4 7	1 5 9	1 6 8
4 5 6	2 5 8	2 6 7	2 4 9
7 8 9	3 6 9	3 4 8	3 5 7

Tag 1	Tag 2	Tag 3	Tag 4
1 5 9	1 6 7	1 2 4	1 3 8
2 7 8	2 3 5	3 7 9	2 6 9
3 4 6	4 8 9	5 6 8	4 5 7

Wir wollen uns überlegen, dass diese Lösungen im wesentlichen übereinstimmen. Was soll das aber heißen? Es ist nicht wesentlich, wie die Mädchen heißen, in welcher Reihenfolge (vorne, in der Mitte oder hinten) die drei Dreiergruppen an dem jeweiligen Tag spazieren und auch die Reihenfolge der Tage ist unwesentlich. Wir betrachten die folgende Permutation der Zahlen von 1 bis 9:

$$\Pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 1 & 5 & 7 & 2 & 8 & 3 & 4 & 6 \end{pmatrix}.$$

Dies bedeutet, dass den neun Mädchen neue Namen gegeben werden. Aus 1 wird 9, aus 2 wird 1 usw. Bei dieser Transformation erhält man aus der in Abbildung 48 dargestellten Lösung des Problems der neun Schulmädchen:

1 2 3	1 4 7	1 5 9	1 6 8
4 5 6	2 5 8	2 6 7	2 4 9
7 8 9	3 6 9	3 4 8	3 5 7

$\xrightarrow{\Pi}$

9 1 5	9 7 3	9 2 6	9 8 4
7 2 8	1 2 4	1 8 3	1 7 6
3 4 6	5 8 6	5 7 4	5 2 3

Berücksichtigt man nun noch, dass innerhalb der Dreiergruppen der Größe nach angeordnet werden kann, die Reihenfolge der Dreiergruppen sowie die Tage permutiert

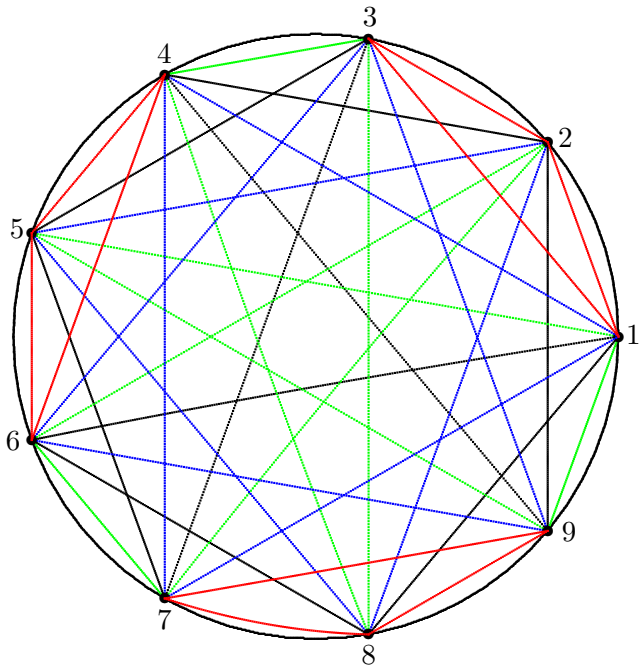


Abbildung 48: Lösung des Problems der neun Schulmädchen

werden können, so erhält man schließlich

9 1 5	9 7 3	9 2 6	9 8 4
7 2 8	1 2 4	1 8 3	1 7 6
3 4 6	5 8 6	5 7 4	5 2 3

$\equiv$

1 5 9	3 7 9	2 6 9	4 8 9
2 7 8	1 2 4	1 3 8	1 6 7
3 4 6	5 6 8	4 5 7	2 3 5

$\equiv$

1 5 9	1 2 4	1 3 8	1 6 7
2 7 8	3 7 9	2 6 9	2 3 5
3 4 6	5 6 8	4 5 7	4 8 9

$\equiv$

1 5 9	1 6 7	1 2 4	1 3 8
2 7 8	2 3 5	3 7 9	2 6 9
3 4 6	4 8 9	5 6 8	4 5 7

Daher sind die oben angegebenen Lösungen des Problems der neun Schulmädchen im wesentlichen gleich. Dies ist kein Wunder, denn eine Lösung dieses Problems ist (im wesentlichen) eindeutig (Beweis?).

Nun wollen wir noch eine Verallgemeinerung von Kirkman's Schulmädchenproblem angeben. Angenommen man hat n Schulmädchen, die jeden Tag in $n/3$ Dreiergruppen spazieren. Für welche Werte von n kann ein "Spazierplan" so aufgestellt werden, dass über eine gewisse Anzahl von Tagen jedes Paar von Schulmädchen in genau einer Dreiergruppe spaziert? Eine *notwendige* Bedingung an n für die Existenz eines solchen Spazierplans kann leicht angegeben werden. Zunächst sollte n natürlich durch 3 teilbar sein. Es gibt $\binom{n}{2} = \frac{1}{2}n(n-1)$ Paare von Schulmädchen und die Zahl der benötigten Tage

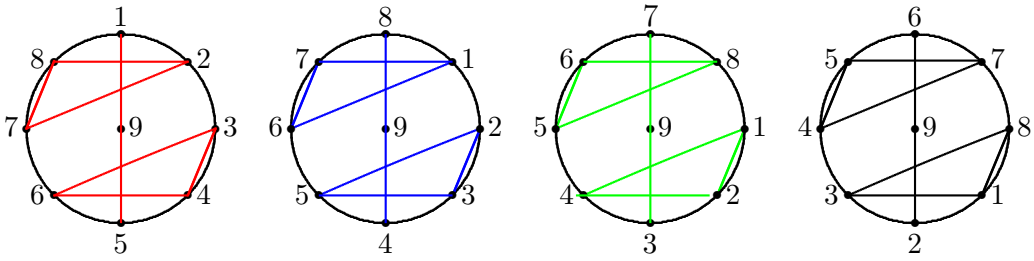


Abbildung 49: Eine weitere Lösung des Problems der neun Schulmädchen

ist $\frac{1}{2}(n-1)$. Daher muss n ungerade sein, insgesamt muss n kongruent 3 modulo 6 bzw. ein ungerades Vielfaches von 3 sein. Von D. K. Ray-Chaudhuri, R. M. Wilson (1971) wurde gezeigt, dass dies auch *hinreichend* für die Existenz eines zulässigen Spazierplans ist.

7.2 Definitionen, Beispiele

Von J. Steiner (1853) stammt ein kurzer Aufsatz, in dem am Anfang die folgende kombinatorische Aufgabe gestellt wird:

- Welche Zahl, N, von Elementen hat die Eigenschaft, daß sich die Elemente so zu dreien ordnen lassen, daß je zwei in *einer*, aber *nur in einer* Verbindung vorkommen? Wie viele wesentlich verschiedene Anordnungen, d. h. solche, die nicht durch eine bloße Permutation der Elemente auseinander vorgehen, giebt es bei jeder Zahl?

Wir definieren:

Definition 7.1 Ein *Steiner System* zu Parametern $t, k, n \in \mathbb{N}$ wird mit $S(t, k, n)$ bezeichnet und besteht aus einer Menge S mit $|S| = n$, der *Ordnung* des Steiner Systems, und einer Menge $\mathcal{B}$ von k-elementigen Teilmengen von S, die *Blöcke* genannt werden, mit der Eigenschaft, dass jede t-elementige Teilmenge von S in genau einem $B \in \mathcal{B}$, also in genau einem der Blöcke enthalten ist[27]. Ein Steiner System $S(2, 3, n)$ heißt ein *Steiner Tripel System* und wird mit $STS(n)$ bezeichnet. Ein Steiner System $S(3, 4, n)$ heißt ein *Steiner Quadrupel System* und wird mit $SQS(n)$ bezeichnet.

Bemerkung: Eng verwandt mit dem Begriff eines Steiner Tripel Systems ist der Begriff eines *balanced incomplete block designs* (BIBD). Sind n, k, λ natürliche Zahlen mit $n > k \geq 2$, so heißt ein Paar $(S, \mathcal{B})$ ein (n, k, λ)-BIBD, wenn S eine n-elementige

[27] Man beachte, dass nur Parameter $t, k, n \in \mathbb{N}$ mit $t < k < n$ für ein Steiner System $S(t, k, n)$ sinnvoll bzw. nichttrivial sind.

Menge, $\mathcal{B}$ eine Menge k-elementiger Teilmengen von S, den sogenannten *Blocks*, ist und dieses Paar die Eigenschaft besitzt, dass jedes Paar verschiedener Punkte aus S in genau λ Blocks enthalten ist[28]. Offenbar ist ein Steiner Tripel System der Ordnung n ein $(n, 3, 1)$-BIBD. □

Eine Lösung von Kirkman's Problem der 15 Schulmädchen ist ein Steiner Tripel System $STS(15)$. Denn die 35 Tripel (Blöcke), aus jeweils drei der 15 Schülerinnen bestehend, haben die Eigenschaft, dass je zwei Schülerinnen in genau einem Tripel vorkommen. Aber dieses Steiner Tripel System genügt noch der weiteren Bedingung, dass die 35 Blöcke in sieben disjunkte Komponenten zerlegt werden können, welche wiederum die Eigenschaft haben, dass jede der 15 Schülerinnen in ihnen genau einmal vorkommt. Das nennt man ein *Kirkman Tripel System*.

Beispiel: In Abbildung 50 geben wir ein Steiner Tripel System $STS(7)$ an. Es besteht

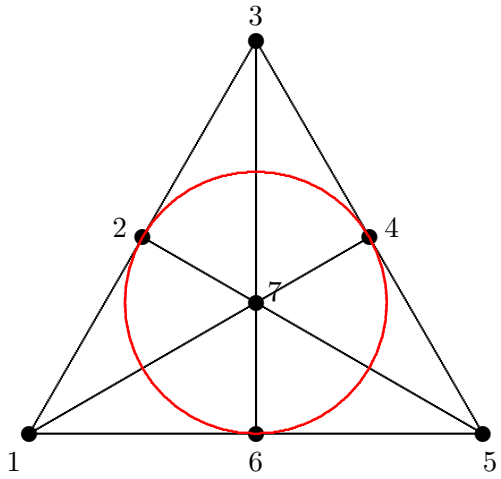


Abbildung 50: Ein $STS(7)$: Die Fano-Ebene

aus der Menge

$$S := \{1, 2, 3, 4, 5, 6, 7\}$$

und den sieben drei-elementigen Blöcken 123, 345, 156, 147, 257, 367 und 246, es ist also

$$\mathcal{B} := \{123, 345, 156, 147, 257, 367, 246\}.$$

Hierbei ist z. B. 123 eine Kurzschreibweise für $\{1, 2, 3\}$. Je zwei Elemente von S sind in genau einem der Blocks enthalten. Eine weitere Veranschaulichung eines Steiner Tripel Systems $STS(7)$ findet sich in Abbildung 51. Hier haben wir die Ecken des vollständigen Graphen K_7 auf einem Kreis platziert. In einem vollständigen Graphen ist jede Ecke mit jeder anderen durch eine Kante verbunden. Eine Zerlegung der Kanten in kantendisjunkte Dreiecke (bzw. vollständige Graphen K_3) liefert ein Steiner Tripel System $STS(7)$, da jede Kante in genau einem der Dreiecke enthalten ist.

[28]Für $\lambda > 1$ ist es durchaus erlaubt, dass ein Block z. B. zweimal in $\mathcal{B}$ enthalten ist. Daher ist in diesem Fall $\mathcal{B}$ eigentlich nicht eine Menge, sondern eine Sammlung (collection) von Teilmengen von S.

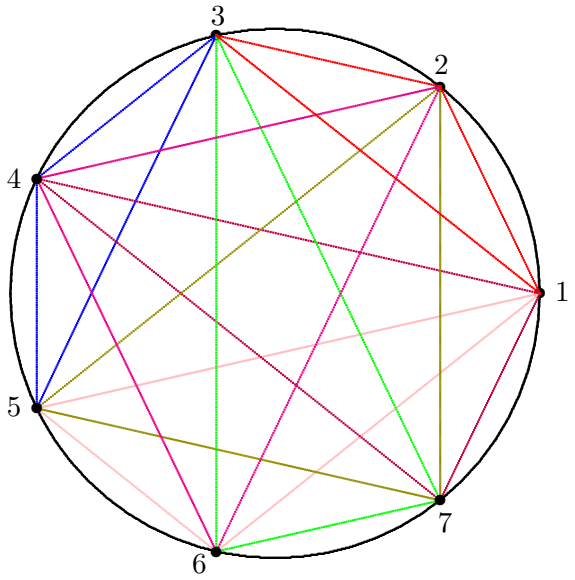


Abbildung 51: Eine kantendisjunkte Zerlegung des K_7 in Dreiecke

Jetzt wollen wir uns noch überlegen, dass ein $STS(7)$ im wesentlichen *eindeutig* bestimmt ist[29]. Dies bedeutet, dass man jedes andere $STS(7)$ mit den Daten $(S'\mathcal{B}')$ durch eine Umbenennung der Elemente der Grundmenge S und damit der der Blöcke in $\mathcal{B}$ erhalten kann. Genauer: Ist auch $(S', \mathcal{B}')$ ein $STS(7)$, so existiert eine bijektive Abbildung $\sigma\colon S \longrightarrow S'$ mit $\sigma(B) \in \mathcal{B}'$ für jedes $B \in \mathcal{B}$. Ist hierbei $B = xyz$, so ist natürlich $\sigma(B) := \sigma(x)\sigma(y)\sigma(z)$. Z. B. ist durch

$$S' := \{0, 1, 2, 3, 4, 6, 6\}$$

und

$$\mathcal{B}' := \{012, 034, 056, 235, 246, 136, 145\}$$

ebenfalls ein $STS(7)$ gegeben, siehe Abbildung 52. Definiert man $\sigma\colon S \longrightarrow S'$ durch

$$\sigma\colon \begin{array}{c|c|c|c|c|c|c} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

so gehen offenbar Blöcke aus $\mathcal{B}$ in Blöcke aus $\mathcal{B}'$ über. So ist z. B. $\sigma(123) = 012$. Nun wollen wir uns überlegen, dass ein durch $(S, \mathcal{B})$ gegebenes $STS(7)$ bis auf eine Umbenennung der Elemente von S eindeutig bestimmt ist. Wir benutzen hierbei, dass $|\mathcal{B}| = 7$ und jedes Element von S in genau drei Blöcken enthalten ist. Dies wird (wesentlich allgemeiner) in einer Bemerkung am Schluss des nächsten Unterabschnitts bewiesen. Ferner besteht der Durchschnitt aus zwei Blöcken aus genau einem Element. Denn der Durchschnitt zweier verschiedener Blöcke kann nicht aus zwei Elementen bestehen, da jedes Paar von Elementen in genau einem Block enthalten ist. Er kann aber auch nicht leer sein. Denn dann gäbe es ein Element, das in zwei der sieben Blöcke

[29] Beim Beweis dieser Tatsache tun wir uns vielleicht ein wenig schwer. In der Literatur wird die Eindeutigkeit eines $STS(7)$ (bis auf Isomorphie) eigentlich immer ohne Beweis angegeben. Einen Beweisansatz findet man bei W. D. Wallis (1988, S. 2)

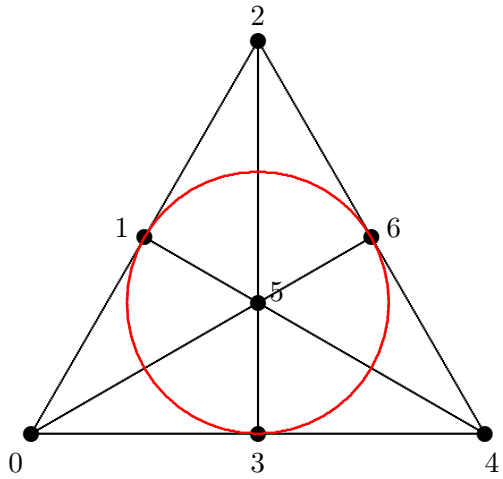


Abbildung 52: Ein weiteres $STS(7)$

nicht enthalten wäre. Die Kombination dieses Elementes mit einem der sechs restlichen Elemente wäre also in den übrigen fünf Blöcken enthalten, was bedeutet, dass ein Paar in zwei Blöcken enthalten wäre, ein Widerspruch. Wir gehen aus von einem $STS(7)$, das durch

$$S := \{1, 2, 3, 4, 5, 6, 7\}, \qquad \mathcal{B} := \{123, 147, 156, 246, 257, 345, 367\}$$

gegeben ist. Sei $(S, \mathcal{B}')$ ein weiteres $STS(7)$. Wir zeigen die Existenz einer Permutation $\sigma\colon \{1, \ldots, 7\} \longrightarrow \{1, \ldots, 7\}$ mit $\sigma(B) \in \mathcal{B}'$ für alle $B \in \mathcal{B}$. Es gibt genau drei Blöcke in $\mathcal{B}'$, die das Element 1 enthalten. Dies seien $1ab$, $1cd$ und $1ef$. Die Zahlen a, b, c, d, e, f sind aus $\{2, \ldots, 7\}$, ferner sind sie paarweise verschieden, da es sich jeweils um *Tripel* handelt (daher ist $a \neq b$, $c \neq d$, $e \neq f$) und der Durchschnitt zweier Blöcke nur aus der 1 bestehen kann (da jedes *Paar* von Elementen aus S in *genau einem* Block enthalten ist). Wir können annehmen, dass $a = 2$ ist, da wir den Block, der das Paar 12 enthält, natürlich $12b$ nennen können. Daher ist

$$\{b, c, d, e, f\} = \{3, 4, 5, 6, 7\}.$$

Auch das Element 2 gehört zu drei Blöcken in $\mathcal{B}'$, einer von diesen enthält auch 1 und ist $12b$ mit $b \in \{3, \ldots, 7\}$. Sei $2gh \in \mathcal{B}'$ mit $g, h \in \{3, \ldots, 7\}$ (und natürlich $g \neq h$) ein weiterer 2 (aber natürlich nicht 1) enthaltender Block aus $\mathcal{B}'$. Die Blöcke $12b$ und $2gh$ aus $\mathcal{B}'$ haben genau ein gemeinsames Element, nämlich 2. Daher ist $b \neq g$, $b \neq h$ bzw. $b \notin \{g, h\}$. Auch die Blöcke $1cd$ und $2gh$ haben genau ein gemeinsames Element. Dieses kann nicht 1 oder 2 sein. Daher ist

$$|\{g, h\} \cap \{c, d\}| = 1.$$

Ist $\{h\} = \{g, h\} \cap \{c, d\}$, so vertauschen wir g und h. Ist $\{d\} = \{g, h\} \cap \{c, d\}$, so vertauschen wir c und d. Daher können wir annehmen, dass $g = c$ und $h \notin \{c, d\}$. Ist $h = e$, so können wir e und f vertauschen und daher annehmen, dass $h = f$. Zwei der 2 enthaltenden Blöcke aus $\mathcal{B}'$ sind daher $12b$ und $2cf$. Der dritte 2 enthaltende Block kann nicht 1 enthalten (die Kombination 12 ist schon in $12b$ enthalten), nicht b (aus

demselben Grund) und auch nicht c und f (denn diese sind schon in $2ce$ enthalten). Der dritte 2 enthaltende Block aus $\mathcal{B}'$ ist daher notwendig $2de$. Nun betrachten wir die Blöcke in $\mathcal{B}'$, die b enthalten. Einer von ihnen ist $12b$. Ein weiterer sei bgh. Die Blöcke $12b$ und bgh aus $\mathcal{B}'$ haben genau ein gemeinsames Element, nämlich b. Daher ist $g, h \notin \{1, 2\}$. Die Blöcke $1cd$ und bgh enthalten ebenfalls genau ein gemeinsames Element, welches nicht gleich 1 oder b sein kann. Daher ist wieder

$$|\{g, h\} \cap \{c, d\}| = 1.$$

Wie oben kann man annehmen, dass $g = c$ und $h \notin \{c, d\}$. In bch ist dann notwendig $h = e$, weil das Paar cf schon in $2cf$ enthalten ist. Damit hat man mit $12b$ und bce zwei Blöcke in $\mathcal{B}'$ gefunden, die b enthalten. Der dritte ist notwendigerweise bdf. Nach eventuellen Umbenennungen der Elemente von S ist also $\mathcal{B}'$ gegeben durch

$$\mathcal{B}' = \{12b, 1cd, 1ef, 2cf, 2de, bce, bdf\},$$

wobei

$$\{b, c, d, e, f\} = \{3, 4, 5, 6, 7\}.$$

Damit ist die Eindeutigkeit (bis auf Umbenennungen) eines $STS(7)$ bewiesen. Eine Permutation, die

$$\mathcal{B} = \{123, 147, 156, 246, 257, 345, 367\}$$

in $\mathcal{B}'$ überführt, ist durch

$$\sigma\colon \begin{array}{c|c|c|c|c|c|c} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 2 & b & c & e & f & d \end{array}$$

gegeben. Denn

$$\begin{aligned} \sigma(\mathcal{B}) &= \sigma(\{123, 147, 156, 246, 257, 345, 367\}) \\ &= \{\sigma(123), \sigma(147), \sigma(156), \sigma(246), \sigma(257), \sigma(345), \sigma(367)\} \\ &= \{12b, 1cd, 1ef, 2cf, 2de, bce, bdf\} \\ &= \mathcal{B}'. \end{aligned}$$

Damit ist die Eindeutigkeit eines $STS(7)$ bewiesen. □

Dass durch die projektive Fano-Ebene $\mathbb{F}_2\mathbb{P}^2$ ein Steiner Tripel System gegeben ist, ist kein Wunder, wie der nächste Satz zeigt.

Satz 7.2 *Sei q eine Primzahlpotenz. Dann ist die projektive Ebene $\mathbb{F}_q\mathbb{P}^2$ ein Steiner-System $S(2, q+1, q^2+q+1)$ mit den q^2+q+1 Punkten von $\mathbb{F}_q\mathbb{P}^2$ als Grundmenge S und den zu den q^2+q+1 Geraden gehörenden $(q+1)$-elementigen Teilmengen von S als Blöcken.*

Beweis: Zu zeigen bleibt, dass jede 2-elementige Teilmenge von S in genau einem der Blöcke enthalten ist. Zu jeder 2-elementigen Teilmenge von S bzw. zwei verschiedenen Punkten x, y aus $\mathbb{F}_q\mathbb{P}^2$ gibt es genau eine Gerade L, die x und y enthält und damit auch genau eine $(q+1)$-elementige Teilmenge von Punkten, nämlich den $q+1$ auf L gelegenen Punkten. Damit ist der Satz bewiesen. □

Beispiel: Wir geben ein $STS(9)$ an, das wir schon aus dem vorigen Unterabschnitt kennen, siehe M. Granell, T. Griggs (1994). Sei $S := \{1,2,3,4,5,6,7,8,9\}$, die Blöcke mögen aus 123, 456, 789, 147, 258, 369, 159, 267, 348, 168, 249 und 357 bestehen, siehe Abbildung 48. Auch ein $STS(9)$ ist, genau wie ein $STS(7)$, bis auf Isomorphie bzw. eine Umbenennung der Elemente der Grundmenge S eindeutig bestimmt.

Steiner Systeme haben Anwendungen in der statistischen Versuchsplanung. Angenommen, es sollen neun verschiedene Müslimischungen miteinander verglichen werden. Eine einzelne Person ist nicht in der Lage, eine verlässliche Reihenfolge herzustellen, da sie beim Kosten der neunten Probe vergessen hat, wie die erste geschmeckt hat. Wir gehen daher davon aus, dass eine einzelne Person nur für drei Proben ein Ranking erstellen kann. Wegen des obigen $STS(9)$ mit 12 Blöcken, genügen zwölf Personen, die jeweils drei Proben begutachten derart, dass jedes Paar von Proben von genau einer Person getestet wird. □

Beispiel: Nun geben wir ein Steiner Quadrupel System $S(3,4,8)$ an. Sei

$$S := \{1,2,3,4,5,6,7,8\},$$

hierzu betrachte man die Menge $\mathcal{B}$ von vierzehn 4-elementigen Teilmengen bzw. Blöcke 1238, 1458, 1678, 2468, 2578, 3478, 3568, 4567, 2367, 2345, 1357, 1346, 1256 und 1247. Hierdurch ist ein Steiner Quadrupel System gegeben. Um dies einzusehen muss man zeigen, dass jede 3-elementige Teilmenge von S in genau einem der Blöcke enthalten ist. Es gibt $\binom{8}{3} = 56$ verschiedene 3-elementige Teilmengen der 8-elementigen Menge S. Durch einfaches Nachprüfen kann also die gemachte Aussage verifiziert werden. Z. B. ist 278 genau in 2578 enthalten. □

7.3 Einfache Ergebnisse zu Steiner Systemen

Hier wollen wir einige sehr einfache Fakten zu Steiner Systemen sammeln.

Satz 7.3 *Falls ein Steiner System $S(t,k,n)$ existiert, so existiert auch ein Steiner System $S(t-1,k-1,n-1)$. Die Existenz von $S(t-1,k-1,n-1)$ ist also eine notwendige Bedingung für die Existenz von $S(t,k,n)$.*

Beweis: Ein Steiner System $S(t,k,n)$ besteht aus einer Menge S mit $|S| = n$ sowie einer Menge $\mathcal{B}$ von k-elementigen Teilmengen von S, mit der Eigenschaft, dass jede t-elementige Teilmenge von S in genau einem $B \in \mathcal{B}$ enthalten ist. Man wähle $x \in S$ beliebig und setze

$$S_- := S \setminus \{x\}, \qquad \mathcal{B}_x := \{B \in \mathcal{B} : x \in B\}$$

sowie

$$\mathcal{B}_- := \{B \setminus \{x\} : B \in \mathcal{B}_x\}.$$

Dann besitzt S_- genau ein Element weniger als S, also $n-1$. Jeder Block $B \in \mathcal{B}_-$ ist $(k-1)$-elementig. Ist nun $T \subset S_-$ eine Teilmenge mit $|T| = t-1$, so ist $T \cup \{x\}$ eine Teilmenge von S, die in genau einem Block B aus $\mathcal{B}$ enthalten ist. Da $x \in T \cup \{x\}$, ist notwendigerweise $B \in \mathcal{B}_x$ und $T \in B \setminus \{x\} \in \mathcal{B}_-$. Also ist jede $(t-1)$-elementige

Teilmenge aus S_- in einem Block aus $\mathcal{B}_-$ enthalten, und dieser ist offenbar eindeutig. Damit ist der Satz bewiesen. □

Beispiel: Wie im letzten Beispiel im vorigen Unterabschnitt sei durch

$$S := \{1, 2, 3, 4, 5, 6, 7, 8\}$$

und

$$\mathcal{B} := \{1238, 1458, 1678, 2468, 2578, 3478, 3568, 4567, 2367, 2345, 1357, 1346, 1256, 1247\}$$

ein Steiner System $S(3, 4, 8)$ gegeben. Nach Satz 7.3 existiert dann auch ein Steiner System $S(2, 3, 7)$ und im Beweis des Satzes wird vorgeführt, wie ein solches konstruiert werden kann. Wählt man z. B. als das aus S zu entfernende Element $x := 4$, so ist

$$\mathcal{B}_x = \{1458, 2468, 3478, 4567, 2345, 1346, 1247\}$$

und damit

$$S_- = \{1, 2, 3, 5, 6, 7, 8\}, \qquad \mathcal{B}_- = \{158, 268, 378, 567, 235, 136, 127\}.$$

Durch $(S_-, \mathcal{B}_-)$ ist ein Steiner System $S(2, 3, 7)$ bzw. ein Steiner Tripel System $STS(7)$ gegeben. □

Bei der Formulierung und dem Beweis des folgenden einfachen Satzes folgen wir ziemlich genau P. J. Cameron (1994, S. 108).

Satz 7.4 *Sei S eine Menge mit $|S| = n$ und $\mathcal{B}$ eine Menge k-elementiger Teilmengen B von S. Dann gilt:*

1. *Ist jede t-elementige Teilmenge von S in höchstens einem $B \in \mathcal{B}$ enthalten ist, so ist*
$$|\mathcal{B}| \le \binom{n}{t} \Big/ \binom{k}{t}.$$
Hier gilt Gleichheit genau dann, wenn $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$ ist, also jede t-elementige Teilmenge von S in genau einem $B \in \mathcal{B}$ enthalten ist.

2. *Ist*
$$|\mathcal{B}| \le \binom{n}{t} \Big/ \binom{k}{t}$$
und ist jede t-elementige Teilmenge von S in mindestens einem $B \in \mathcal{B}$ enthalten, so ist $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$.

Beweis: Wir zählen die Anzahl der Elemente der Menge

$$M := \{(L, B) : B \in \mathcal{B},\ L \subset B,\ |L| = t\}.$$

Die Anzahl der t-elementigen Teilmengen L einer k-elementigen Menge $B \in \mathcal{B}$ ist $\binom{k}{t}$. Daher ist $|M| = |\mathcal{B}| \cdot \binom{k}{t}$. Auf der anderen Seite ist $\binom{n}{t}$ die Anzahl der t-elementigen

Teilmengen der Menge S mit $|S| = n$. Jede dieser Mengen liegt in *höchstens* einer Menge $B \in \mathcal{B}$. Daher ist $|M| \leq \binom{n}{t}$ und folglich

$$(*) \qquad |\mathcal{B}| \leq \binom{n}{t} \Big/ \binom{k}{t}.$$

Gleichheit in dieser Ungleichung tritt genau dann ein, wenn jede t-elementige Teilmenge von S in genau einem $B \in \mathcal{B}$ liegt bzw. $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$ ist.

Gilt nun $(*)$ und ist jede t-elementige Teilmenge von S in *mindestens* einem Block $B \in \mathcal{B}$ enthalten, so ist $|M| \geq \binom{n}{t}$ und daher

$$\binom{n}{t} \leq |M| = |\mathcal{B}|\binom{k}{t} \leq \binom{n}{t}.$$

Insgesamt ist

$$|M| = \binom{n}{t}, \qquad |\mathcal{B} = \binom{n}{t} \Big/ \binom{k}{t}.$$

Hieraus wiederum folgt, dass jede t-elementige Teilmenge von S in *genau* einem Block $B \in \mathcal{S}$ enthalten bzw. $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$ ist. □

Bemerkung: Insbesondere erhalten wir, dass die Ganzzahligkeit von $\binom{n}{t}/\binom{k}{t}$ eine *notwendige* Bedingung dafür ist, dass $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$ ist. Schon aus Satz 7.4 erhalten wir, dass kein Steiner System $S(2, 3, 8)$ existiert, da $\binom{3}{2} = 3$ nicht $\binom{8}{2} = 28$ teilt. Man kann aber weitere notwendige Bedingungen für die Existenz eines Steiner Systems $S(t, k, n)$ aufstellen. Denn sei $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$. Mit einem beliebigen $x \in S$ ist dann $(S_-, \mathcal{B}_-)$ mit

$$S_- := S \setminus \{x\}, \qquad \mathcal{B}_x := \{B \in \mathcal{B} : x \in B\}$$

sowie

$$\mathcal{B}_- := \{B \setminus \{x\} : B \in \mathcal{B}_x\}$$

ein Steiner System $S(t-1, k-1, n-1)$, wie wir im Beweis von Satz 7.3 nachgewiesen haben. Nach Satz 7.4 ist

$$|\mathcal{B}_-| = \binom{n-1}{t-1} \Big/ \binom{k-1}{t-1}$$

eine natürliche Zahl. Wegen

$$\binom{n}{t} \Big/ \binom{k}{t} = \frac{n}{k} \cdot \binom{n-1}{t-1} \Big/ \binom{k-1}{t-1}$$

ist daher $|\mathcal{B}|k = |\mathcal{B}_-|n$. Diese Argumentation kann fortgesetzt werden und man erhält insgesamt:

- Sei $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$. Dann ist $\binom{n-i}{t-i}/\binom{k-i}{t-i}$, $i = 0, \ldots, t-1$, eine natürliche Zahl. Ist dies der Fall, so sagen wir (t, k, n) sei eine *zulässige* Parametermenge.

Ist $t = 1$, so ist $(1, k, n)$ genau dann eine zulässige Parametermenge, wenn n durch k geteilt wird. Ist dies der Fall, so existiert ein Steiner System $S(1, k, n)$. Denn sei etwa $n = kp$ und $S := \{1, \ldots, n\}$, so nehme man $\mathcal{B} := \{B_1, \ldots, B_p\}$ mit

$$B_i := \{(i-1)k + 1, \ldots, ik\}, \quad i = 1, \ldots, p,$$

als Menge der k-elementigen Blöcke. Jedes Element von S bzw. jede 1-elementige Teilmenge von S ist dann in genau einem Block enthalten, so dass $S(1, k, n)$, wenn n durch k teilbar ist, trivialerweise ein Steiner System ist. Also ist $S(2, 3, n)$ der erste nichttriviale Fall eines Steiner Systems. Im nächsten Unterabschnitt werden wir uns damit beschäftigen, für welche $n \in \mathbb{N}$ ein Steiner System $S(2, 3, n)$ bzw. ein Steiner Tripel System $STS(n)$ existiert. □

Bemerkung: In einer Bemerkung im Anschluss an Definition 7.1 haben wir als Verallgemeinerung von Steiner Tripel Systemen definiert, was unter einem (n, k, λ)-BIBD verstanden wird. Hier wollen wir uns überlegen (siehe D. R. Stinson (2004, S. 4):

- In einem (n, k, λ)-BIBD $(S, \mathcal{B})$ ist jeder Punkt $x \in S$ in genau

$$r := \frac{\lambda(n-1)}{k-1}$$

  Blöcken enthalten.

Denn: Sei r_x die Anzahl der Blocks $B \in \mathcal{B}$ mit $x \in B$. Wir definieren

$$I := \{(y, B) \in (S \setminus \{x\}) \times \mathcal{B} : \{x, y\} \subset B\}$$

und berechnen $|I|$ auf zweierlei Weise. Es gibt $n - 1$ Möglichkeiten, $y \in S \setminus \{x\}$ zu wählen. Für jedes solches y gibt es λ Blocks $B \in \mathcal{B}$ mit $\{x, y\} \subset B$. Daher ist

$$|I| = \lambda(n-1).$$

Andererseits gibt es r_x Möglichkeiten, einen Block $B \in \mathcal{B}$ mit $x \in B$ auszuwählen. Für jedes solches B gibt es $k - 1$ Möglichkeiten, $y \in B \setminus \{x\}$ auszuwählen. Daher ist

$$|I| = r_x(k-1).$$

Kombiniert man die beiden Gleichungen, so erhalten wir

$$\lambda(n-1) = r_x(k-1).$$

Daher ist $r_x = \lambda(n-1)/(k-1)$ von x unabhängig und die Behauptung folgt.
Weiter gilt:

- In einem (n, k, λ)-BIBD $(S, \mathcal{B})$ ist

$$|\mathcal{B}| = \frac{\lambda(n^2 - n)}{k^2 - k}.$$

Denn: Diesmal definieren wir

$$I := \{(x, B) \in S \times \mathcal{B} : x \in B\}$$

und berechnen $|I|$ wieder auf zweierlei Weise. Zunächst gibt es n Möglichkeiten, ein $x \in S$ auszuwählen. Jedes solche x ist in $r := \lambda(n-1)/(k-1)$ Blöcken $B \in /calB$ enthalten. Daher ist

$$|I| = nr = \frac{\lambda n(n-1)}{k-1}.$$

Andererseits gibt es $|\mathcal{B}|$ Möglichkeiten, einen Block $B \in \mathcal{B}$ zu wählen. Zu jeder solchen Wahl B gibt es k Möglichkeiten $x \in B$ auszuwählen. Daher ist

$$|I| = |\mathcal{B}|k.$$

Kombiniert man beide Gleichungen, so erhält man

$$\frac{\lambda n(n-1)}{k-1} = |\mathcal{B}|k,$$

woraus die Behauptung folgt.
Mit $\lambda := 1$ und $k := 3$ ergibt sich für ein Steiner Tripel System der Ordnung n erneut, dass die Anzahl der Blöcke durch $n(n-1)/6$ gegeben ist. Ferner ist die Anzahl der Blöcke eines Steiner Tripel Systems der Ordnung n, in denen ein bestimmtes Element aus S enthalten ist, gleich $(n-1)/2$. □

7.4 Notwendige und hinreichende Bedingungen für die Existenz von Steiner Tripel Systemen der Ordnung n

Es wird sich herausstellen, dass es eine Bedingung an n gibt, die notwendig und hinreichend für die Existenz eines Steiner Tripel Systems $STS(n)$ ist. Da die *notwendige* Bedingung wesentlich einfacher einzusehen ist, werden wir sie zunächst behandeln. Als Literatur geben wir P. J. Cameron (1994, S. 109), C. C. Lindner, C. A. Rodger (1997, S. 3) und C. J. Colbourn, A. Rosa (1999, S. 23 ff.) an.

Satz 7.5 *Es existiere ein Steiner Tripel System $STS(n)$ mit $n \in \mathbb{N}$. Dann ist $n \equiv 1 \pmod 6$ oder $n \equiv 3 \pmod 6$.*

Beweis: Mit $STS(n)$ bzw. $S(2, 3, n)$ ist auch $S(1, 2, n-1)$ ein Steiner System. Wegen Satz 7.4 wird $\binom{n}{2} = n(n-1)/2$ durch $\binom{3}{2} = 3$ geteilt, außerdem wird $\binom{n-1}{1} = n-1$ durch $\binom{2}{1} = 2$ geteilt. Die zweite Bedingung sagt aus, dass n eine ungerade Zahl ist und damit notwendigerweise $n \equiv 1 \pmod 6$, $n \equiv 3 \pmod 6$ oder $n \equiv 5 \pmod 6$. Die erste Bedingung sagt aus, dass $n(n-1)/6$ eine natürliche Zahl ist. Angenommen, es wäre $n \equiv 5 \pmod 6$ bzw. $n = 6k+5$ mit einer nichtnegativen ganzen Zahl k. Dann wäre

$$\frac{n(n-1)}{6} = \frac{(6k+5)(3k+2)}{3}$$

keine natürliche Zahl, da weder $6k+5$ noch $3k+2$ durch 3 teilbar sind, ein Widerspruch. Damit ist $n \equiv 5 \pmod 6$ nicht möglich und der Satz ist bewiesen. □

Jetzt kommen wir zum schwierigeren Teil, dass nämlich die in Satz 7.5 angegebene Bedingung an n auch *hinreichend* für die Existent eines Steiner Tripel Systems $STS(n)$ ist. Im nächsten Satz wird sozusagen die Hälfte des Satzes bewiesen.

Satz 7.6 *Ist $n \equiv 3 \pmod 6$, so existiert ein Steiner Tripel System $STS(n)$.*

Beweis: Wegen $n \equiv 3 \pmod 6$ ist $n = 3m$ mit ungeradem m. Der Beweis ist konstruktiv, d. h. wir geben eine Menge S mit $|S| = n$ sowie eine Menge $\mathcal{B}$ von 3-elementigen Teilmengen B von S an, den Blöcken, die die Eigenschaft haben, dass jede 2-elementige Teilmenge von S in genau einem Block $B \in \mathcal{B}$ enthalten ist. Wir folgen ziemlich genau der Darstellung bei P. J. Cameron (1994, S. 110).

Wir erinnern an den Restklassenring $\mathbb{Z}/(m)$ (oder auch $\mathbb{Z}_m$ bzw. $\mathbb{Z}/m\mathbb{Z}$). Es ist

$$\mathbb{Z}/(m) := \{[i] : i = 0, \ldots, m-1\}.$$

Für $i \in \mathbb{Z}$ ist hierbei die *Restklasse* $[i]$ definiert als

$$[i] := i + m\mathbb{Z} = \{i + mk : k \in \mathbb{Z}\}.$$

Die dem Steiner Tripel System $STS(n) = STS(3m)$ zugrunde liegende Menge S ist durch drei Kopien von $\mathbb{Z}/(m)$ gegeben, es sei also

$$S := \{a_{[0]}, \ldots, a_{[m-1]}\} \cup \{b_{[0]}, \ldots, b_{[m-1]}\} \cup \{c_{[0]}, \ldots, c_{[m-1]}\}.$$

Blöcke sind Tripel aus Elementen von S, sie sind von zweierlei Art:

(a) Alle Tripel der Form $a_{[i]}a_{[j]}b_{[k]}$, $b_{[i]}b_{[j]}c_{[k]}$ oder $c_{[i]}c_{[j]}a_{[k]}$ mit $[i], [j], [k] \in \mathbb{Z}/(m)$, $[i] \neq [j]$ und $[i] + [j] = [2k]$,

(b) alle Tripel der Form $a_{[i]}b_{[i]}c_{[i]}$, $[i] \in \mathbb{Z}/(m)$.

Die Gesamtheit der Blöcke wird mit $\mathcal{B}$ bezeichnet. Wir werden gleich zeigen, dass

$$|\mathcal{B}| = \frac{n(n-1)}{6}$$

und je zwei Elemente aus S in höchstens einem der Blöcke aus $\mathcal{B}$ liegen. Wegen Satz 7.4 ist $(S, \mathcal{B})$ dann ein Steiner Tripel System.

Zunächst beachten wir:

- Seien $i, j \in \mathbb{Z}$ gegeben. Mit einem $k \in \mathbb{Z}$ gibt es dann eine eindeutige Restklasse $[k]$ mit $[i] + [j] = [2k]$.

Denn: Man definiere

$$k := \begin{cases} \dfrac{i+j}{2}, & \text{falls } i+j \text{ gerade,} \\ \dfrac{i+j+m}{2}, & \text{falls } i+j \text{ ungerade.} \end{cases}$$

Hierbei haben wir ausgenutzt, dass m ungerade ist. Offensichtlich ist

$$[i] + [j] = i + j + m\mathbb{Z} = 2k + m\mathbb{Z} = [2k].$$

Um die Eindeutigkeit von k bzw. der Restklasse $[k]$ zu beweisen, nehmen wir an, mit $k_1, k_2 \in \mathbb{Z}$ sei $[2k_1] = [2k_2]$. Mit einem $l \in \mathbb{Z}$ ist dann $2(k_1 - k_2) = ml$ eine gerade Zahl. Da m ungerade ist, ist l gerade. Also ist $k_1 - k_2 = ml/2$ und folglich $[k_1] = [k_2]$.

Nun zählen wir die Anzahl der Blöcke in $\mathcal{B}$. Es gibt $\binom{m}{2} = m(m-1)/2$ Möglichkeiten, die voneinander verschiedenen $[i]$ und $[j]$ aus der m-elementigen Menge $\mathbb{Z}/(m)$ auszuwählen. Nach obiger Überlegung gehört zu jedem Paar $([i], [j])$ eine eindeutige Restklasse $[k]$ mit $[i] + [j] = [2k]$. Insgesamt gibt es also

$$3\binom{m}{2} = \frac{3m(m-1)}{2}$$

Tripel bzw. Blöcke vom Typ (a). Da es ganz offensichtlich m Tripel vom Typ (b) gibt, ist

$$|\mathcal{B}| = \frac{3m(m-1)}{2} + m = \frac{3m(3m-1)}{6} = \frac{n(n-1)}{6},$$

wie behauptet.

Nun zeigen wir, dass jedes Paar von Elementen aus S in höchstens einem der Blöcke aus $\mathcal{B}$ liegt. Bei der Auswahl des Paares haben wir verschiedene Fälle zu unterscheiden.

1. $(a_{[i]}, a_{[j]})$ mit $[i] \neq [j]$.

   Ein Tripel, das dieses Paar[30] enthält, muss vom Typ (a) sein. Nach obiger Überlegung existiert eindeutig $[k]$ mit $[i]+[j] = [2k]$, so dass $(a_{[i]}, a_{[j]})$ in dem eindeutigen Tripel $a_{[i]}a_{[j]}b_{[k]}$ enthalten ist.

2. $(b_{[i]}, b_{[j]})$ oder $(c_{[i]}, c_{[j]})$ mit $[i] \neq [j]$.

   Diese beiden Fälle können wie Fall 1. erledigt werden.

3. $(a_{[i]}, b_{[i]})$.

   Dieses Paar liegt in genau einem Tripel vom Typ (b), nämlich in $a_{[i]}b_{[i]}c_{[i]}$. Weiter liegt es in keinem Tripel vom Typ (a). Denn es gibt kein Tripel $a_{[i]}a_{[j]}b_{[i]}$ mit $[i] \neq [j]$ und $[i] + [j] = [2i]$.

4. $(b_{[i]}, c_{[i]})$ oder $(a_{[i]}, c_{[i]})$.

   Diese beiden Fälle können wie in Fall 3. erledigt werden.

5. $(a_{[i]}, b_{[k]})$ mit $[i] \neq [k]$.

   Dieses Paar liegt in genau einem Block, nämlich einem Block vom Typ (a). Setzt man nämlich $j := 2k - i$, so ist $[i] + [j] = [2k]$ und $[i] \neq [j]$, sodass das Paar $(a_{[i]}, b_{[k]})$ genau in dem Block $a_{[i]}a_{[j]}b_{[k]}$ vom Typ (a) enthalten ist. Dass $(a_{[i]}, b_{[k]})$ mit $[i] \neq [k]$ in keinem Block vom Typ (b) enthalten sein kann, ist offensichtlich.

[30] Wir hätten auch $a_{[i]}a_{[j]}$ (oder auch $\{a_{[i]}, a_{[j]}\}$ statt $(a_{[i]}, a_{[j]})$ schreiben können.

6. $(b_{[i]}, c_{[k]})$ oder $(c_{[i]}, a_{[k]})$ mit $[i] \neq [k]$.

   Diese beiden Fälle können wie Fall 5. erledigt werden.

Damit ist gezeigt, dass durch $(S, \mathcal{B})$ ein Steiner Tripel System $STS(n)$ gegeben ist. □

Beispiel: Im Beweis von Satz 7.6 ist für $n \equiv 3 \pmod 6$ konstruktiv ein Steiner Tripel System $STS(n)$ bestimmt worden. Diese Konstruktion wollen wir $n = 9$ verdeutlichen. Mit $m = 3$ ist $n = 3m$. Wir unterscheiden im folgenden nicht zwischen i und der Restklasse $[i]$ in $\mathbb{Z}/(3)$. Die Grundmenge des Steiner Tripel Systems $STS(9)$ sei

$$S := \{a_0, a_1, a_2, b_0, b_1, b_2, c_0, c_1, c_2\},$$

siehe Abbildung 53. Zunächst haben wir dort in rot die drei Blöcke $a_0b_0c_0$, $a_1b_1c_1$ und

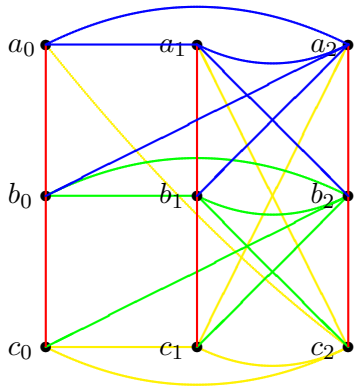


Abbildung 53: Das Steiner Tripel System $STS(9)$

$a_2b_2c_2$ vom Typ (b) eingetragen. Es bleiben neun Tripel vom Typ (a) zu bestimmen. In blau geben wir die Blöcke $a_0a_1b_2$, $a_0a_2b_1$ und $a_1a_2b_0$ an, in grün die Blöcke $b_0b_1c_2$, $b_0b_2c_1$ und $b_1b_2c_0$ und schließlich in gelb die Blöcke $c_0c_1a_2$, $c_0c_2a_1$ und $c_1c_2a_0$. □

Bemerkung: I. Allg. wird bei der Konstruktion eines Steiner Tripel Systems $STS(n)$ mit $n \equiv 3 \pmod 6$ auf die sogenannte *Bose Konstruktion*[31] verwiesen. Diese wollen wir jetzt in ihren Grundzügen skizzieren. Der folgende Begriff ist hierfür wichtig:

- Ein Paar $(Q, \circ)$ heißt eine *Quasigruppe* der Ordnung n, wenn Q eine n-elementige Menge und $\circ\colon Q \times Q \longrightarrow Q$ eine binäre Operation mit der Eigenschaft ist, dass für beliebige $a, b \in Q$ die Gleichungen $a \circ x = b$ und $y \circ a = b$ eindeutige Lösungen $x \in Q$ bzw. $y \in Q$ besitzen. Eine Quasigruppe $(Q, \circ)$ heißt *idempotent*, wenn $x \circ x = x$ für alle $x \in Q$, sie heißt *symmetrisch*, wenn $x \circ y = y \circ x$ für alle $x, y \in Q$.

Der Begriff einer Quasigruppe $(Q, \circ)$ steht in einem engen Zusammenhang mit dem eines *lateinischen Quadrats*. Denn ist $Q = \{x_1, \ldots, x_n\}$ und ist $L := (x_i \circ x_j)_{1 \leq i,j \leq n}$, so ist in jeder Zeile und jeder Spalte von L jedes Element von Q genau einmal enthalten. Und so eine quadratische Anordnung von n Symbolen, z. B. Zahlen von 1 bis n,

[31] In der Literatur wird in diesem Zusammenhang die Arbeit von R. C. Bose (1939) als grundlegend angegeben. Ich muss gestehen, dass ich in diesem langen Aufsatz die Bose-Konstruktion vergeblich gesucht habe.

nennt man ein *lateinisches Quadrat*. Dieses nennt man *idempotent*, wenn der Eintrag in (i,i) das i-te Symbol ist, es heißt *symmetrisch*, wenn an den Stellen (i,j) und (j,i) dasselbe Symbol steht. Offenbar sind Quasigruppe und lateinisches Quadrat nur zwei verschiedene Ausdrücke für dieselbe Sache: Die Operationstafel einer Quasigruppe ist ein lateinisches Quadrat, umgekehrt ist ein lateinisches Quadrat Operationstafel einer Quasigruppe. Zur Existenz symmetrischer, idempotenter Quasigruppen der Ordnung n überlegen wir uns:

- Es existiert genau dann eine symmetrische, idempotente Quasigruppe der Ordnung n, wenn n ungerade ist.

Denn: Sei $(Q,\circ)$ eine symmetrische, idempotente Quasigruppe der Ordnung n. Man wähle $z \in Q$ beliebig und definiere die Abbildung $T\colon Q\setminus\{z\} \longrightarrow Q\setminus\{z\}$ durch $x\circ T(x) = z$ für $x \in Q\setminus\{z\}$. Hierbei ist $T(x) \neq z$ und $T(x) \neq x$, da $(Q,\circ)$ idempotent. Da $(Q,\circ)$ symmetrisch ist, ist $T(x)\circ x = z$ bzw. $x = T(T(x))$ für alle $x \in Q\setminus\{z\}$. Folglich ist $Q\setminus\{z\}$ eine disjunkte Vereinigung von 2-elementigen Mengen der Form $\{x,T(x)\}$. Also ist $|Q|-1 = n-1$ gerade bzw. n ungerade. Nun zeigen wir, dass es zu jedem ungeraden n eine symmetrische, idempotente Quasigruppe gibt. Hierzu setzen wir

$$Q := \{0,1,\ldots,n-1\}$$

und definieren $\circ\colon Q\times Q \longrightarrow Q$ durch

$$x\circ y := \left(\frac{n+1}{2}\right)(x+y) \pmod n.$$

Da n als ungerade vorausgesetzt wurde, ist dies sinnvoll. Ganz offensichtlich ist dann $(Q,\circ)$ eine symmetrische, idempotente Quasigruppe.

Jetzt kommen wir zur Beschreibung der Bose-Konstruktion eines Steiner Tripel Systems $STS(n)$ bzw. $(S,\mathcal{B})$, falls $n \equiv 3 \pmod 6$, wobei wir im wesentlichen der Darstellung von C. C. Lindner, C. A. Rodger (1997, S. 6) folgen. Wie wir sehen werden, sind wir beim Beweis von Satz 7.6 ganz ähnlich vorgegangen.

Sei $n = 3m$ mit ungeradem m und $(Q,\circ)$ eine symmetrische, idempotente Gruppe der Ordnung m. Sei etwa $Q = \{0,\ldots,m-1\}$. Die Grundmenge S des zu konstruierenden Steiner Tripel Systems sei $S := Q\times\{1,2,3\}$, sie besteht also aus drei Kopien von Q. Die Tripel bzw. Blöcke aus $\mathcal{B}$ gehören zum Typ (a) oder (b), wobei

(a) Die Tripel $\{(i,1),(j,1),(i\circ j,2)\}$, $\{(i,2),(j,2),(i\circ j,3)\}$ und $\{(i,3),(j,3),(i\circ j,1)\}$ gehören zu $\mathcal{B}$, $0 \le i < j \le m-1$.

(b) Die Tripel $\{(i,1),(i,2),(i,3)\}$ gehören zu $\mathcal{B}$, $i = 0,\ldots,m-1$.

In Abbildung 54 veranschaulichen wir die Bose-Konstruktion. Links finden wir die Tripel vom Typ (a), rechts die vom Typ (b). Dies ist alles nur eine etwas andere Schreibweise als beim Beweis von Satz 7.6. Auch hier weist man leicht nach, dass $|\mathcal{B}| = n(n-1)/6$ und jedes Paar aus S zu genau einem Block aus $\mathcal{B}$ gehört. Damit ist ein zweites Mal, auf nicht wesentlich andere Art als im Beweis von Satz 7.6 gezeigt worden, dass im Falle $n \equiv 3 \pmod 6$ ein Steiner Tripel System $STS(n)$ existiert.

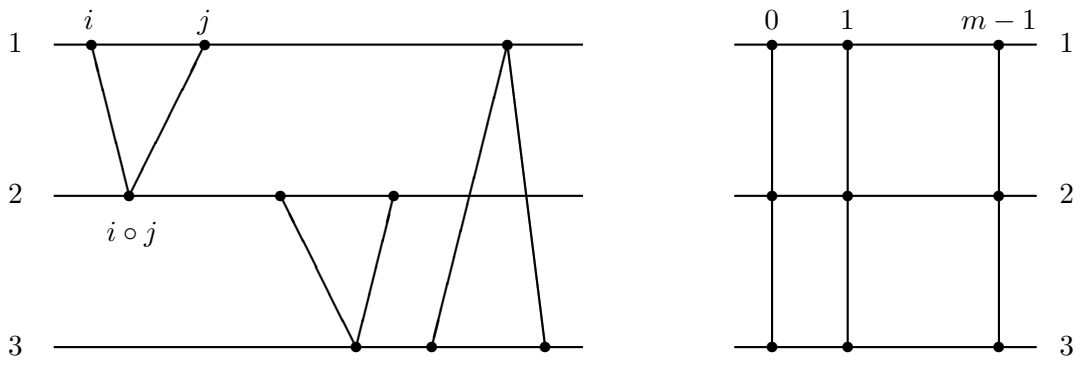


Abbildung 54: Die Bose-Konstruktion

Als Beispiel betrachten wir den Fall $n = 9$. Als symmetrische, idempotente Quasigruppe der Ordnung $m = 3$ nehmen wir

$\circ$	0	1	2
0	0	2	1
1	2	1	0
2	1	0	2

Bei dem zu konstruierenden Steiner Tripel System $STS(9) = (S, \mathcal{B})$ ist dann $S = \{0,1,2\} \times \{1,2,3\}$, während $\mathcal{B}$ die folgenden zwölf Tripel enthält, und zwar zunächst neun vom Typ (a) und dann drei vom Typ (b):

(a) $i = 0$, $j = 1$. Unter Berücksichtigung von $0 \circ 1 = 2$ erhalten wir:
$\{(0,1),(1,1),(2,2)\}$, $\{(0,2)(1,2),(2,3)\}$, $\{(0,3),(1,3),(2,1)\}$,
$i = 0$, $j = 2$. Unter Berücksichtigung von $0 \circ 2 = 1$ erhalten wir:
$\{(0,1),(2,1),(1,2)\}$, $\{(0,2),(2,2),(1,3)\}$, $\{(0,3),(2,3),(1,1)\}$,
$i = 1$, $j = 2$. Unter Berücksichtigung von $1 \circ 2 = 0$ erhalten wir:
$\{(1,1),(2,1),(0,1)\}$, $\{(1,2),(2,2),(0,3)\}$, $\{(1,3),(2,3),(0,1)\}$.

(b) $\{(0,1),(0,2),(0,3)\}$, $\{(1,1),(1,2),(1,3)\}$, $\{(2,1),(2,2),(2,3)\}$.

Damit sind unsere Bemerkungen zum Fall $n \equiv 3 \pmod 6$ abgeschlossen. □

Nachdem wir in Satz 7.6 für $n \equiv 3 \pmod 6$ die Existenz eines Steiner Tripel Systems nachgewiesen haben, folgt jetzt der zweite Teil der Existenzaussage.

Satz 7.7 *Ist $n \equiv 1 \pmod 6$, so existiert ein Steiner Tripel System $STS(n)$.*

Beweis: In der obigen Bemerkung haben wir die Bose-Konstruktion eines Steiner Tripel Systems $STS(n)$ im Falle $n \equiv 3 \pmod 6$ geschildert. Hier schildern wir nun die sogenannte *Skolem-Konstruktion* zur Bestimmung eines Steiner Tripel Systems für $n \equiv 1 \pmod 6$, die auf Th. Skolem (1958) zurückgeht. Ähnlich wie in der Bemerkung im Anschluss an Satz 7.6 ist es nötig, einige neue Begriffe bzw. Hilfsmittel einzuführen.

- Eine Quasigruppe $(Q, \circ)$ mit $Q = \{0, \ldots, n-1\}$ gerader Ordnung n heißt *halb-idempotent*, falls

$$x \circ x = \begin{cases} x, & \text{falls } x \in [0, n/2), \\ x - n/2, & \text{falls } x \in [n/2, n-1] \end{cases}$$

für alle $x \in Q$. Bei einer halb-idempotenten Quasigruppe mit gerader Ordnung n ist also die Diagonale der Operationstafel der Reihe nach mit

$$0, 1, \ldots, \frac{n}{2} - 1, 0, 1, \ldots, \frac{n}{2} - 1$$

besetzt.

Für ungerades n konnte die Existenz einer symmetrischen, idempotenten Quasigruppe der Ordnung n nachgewiesen werden. Entsprechend überlegen wir uns hier:

- Für gerades n existiert eine symmetrische, halb-idempotente Quasigruppe $(Q, \circ)$.

Denn (wir folgen D. R. Stinson (2004, S. 128 ff.)): Sei $Q := \{0, \ldots, n-1\}$. Wir definieren die Abbildung $\pi \colon Q \longrightarrow Q$ durch

$$\pi(x) := \begin{cases} \dfrac{x}{2}, & x \text{ gerade}, \\ \dfrac{x+n-1}{2}, & x \text{ ungerade}. \end{cases}$$

Offenbar ist π eine Permutation von Q, die Werte sind

x	0	1	2	3	$\cdots$	$n-2$	$n-1$
$\pi(x)$	0	$\frac{n}{2}$	1	$\frac{n}{2}+1$	$\cdots$	$\frac{n}{2}-1$	$n-1$

Definiert man nun die binäre Operation $\circ \colon Q \times Q \longrightarrow Q$ durch

$$x \circ y := \pi((x + y) \pmod n),$$

so ist $(Q, \circ)$ offenbar eine symmetrische, halb-idempotente Quasigruppe. Für $n = 6$ erhält man z. B. die folgende Operationstafel:

$\circ$	0	1	2	3	4	5
0	0	3	1	4	2	5
1	3	1	4	2	5	0
2	1	4	2	5	9	3
3	4	2	5	0	3	1
4	2	5	9	3	1	4
5	5	0	3	1	4	2

Jetzt schildern wir die Skolem-Konstruktion eines Steiner Tripel Systems $STS(n)$ für den Fall $n \equiv 1 \pmod 6$.

Sei $n = 6m+1$, $Q := \{0, 1, \ldots, 2m-1\}$ und $(Q, \circ)$ eine symmetrische, halb-idempotente Quasigruppe der Ordnung $2m$. Als Grundmenge S nehmen wir, wie bei der Bose-Konstruktion, drei Kopien von Q und einen zusätzlichen Punkt, den wir ∞ nennen. Es sei also

$$S := \{\infty\} \cup Q \times \{1, 2, 3\}.$$

Dann ist $|S| = 3 \cdot 2m + 1 = 6m + 1 = n$. Die Tripel bzw. Blöcke aus $\mathcal{B}$ gehören zum Typ (a), (b) oder (c).

(a) Die Tripel $\{(i,1),(j,1),(i\circ j,2)\}$, $\{(i,2),(j,2),(i\circ j,3)\}$ und $\{(i,3),(j,3),(i\circ j,1)\}$ gehören zu $\mathcal{B}$, $0 \le i < j \le 2m-1$,

(b) Die Tripel $\{\infty,(m+i,1),(i,2)\}$, $\{(\infty,(m+i,2),(i,3)\}$ und $\{\infty,(m+i,3),(i,1)\}$ gehören zu $\mathcal{B}$, $i = 0, \ldots, m-1$,

(c) Die Tripel $\{(i,1),(i,2),(i,3)\}$ gehören zu $\mathcal{B}$, $i = 0, \ldots, m-1$.

Tripel vom Typ (a) und (c) sind schon in Abbildung 54 veranschaulicht worden. Tripel vom Typ (b) findet man in Abbildung 55. Jetzt zählen wir die Anzahl der Blöcke

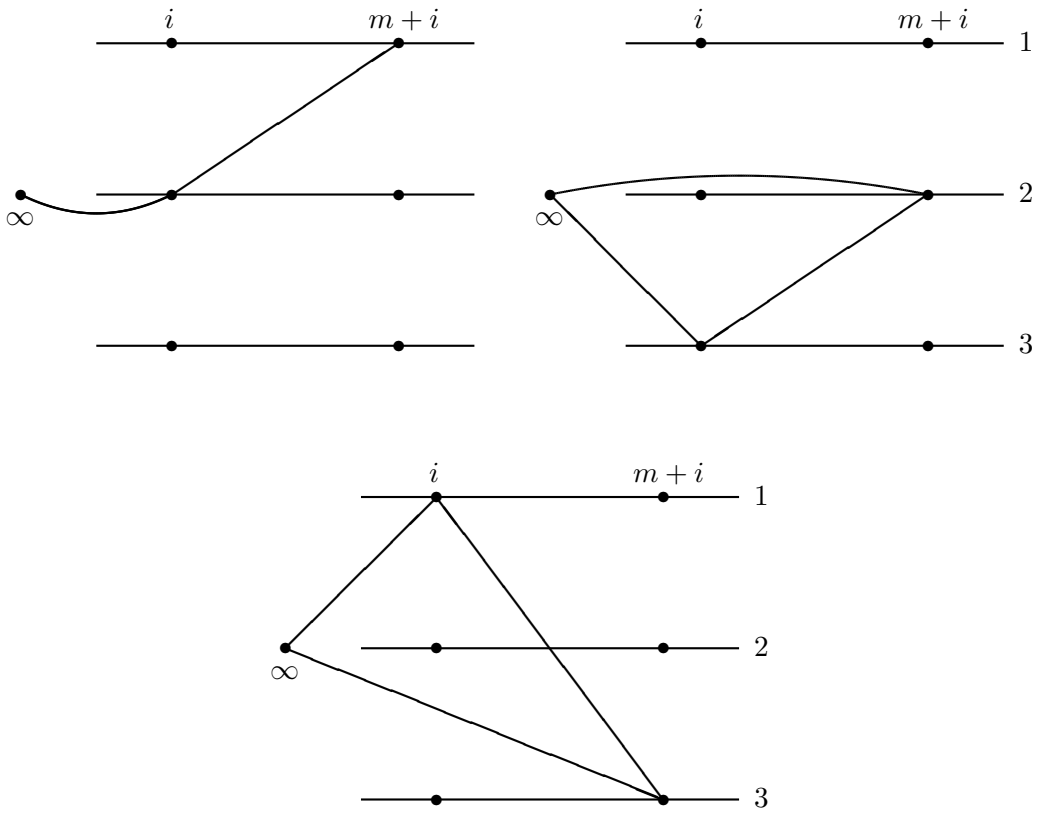


Abbildung 55: Die Skolem-Konstruktion: Blöcke vom Typ (b)

bzw. die Anzahl der Tripel vom Typ (a), (b) und (c). Vom Typ (a) gibt es $3\binom{2m}{2} =$

$3m(2m-1)$ Tripel, vom Typ (b) gibt es $3m$ Tripel und vom Typ (c) sind es m Tripel. Daher ist

$$|\mathcal{B}| = 3m(2m-1) + 3m + m = 6m^2 + m = \frac{n(n-1)}{6}.$$

Die Anzahl $|\mathcal{B}|$ der Blöcke ist also genau richtig.

Nun zeigen wir, dass jedes Paar von Elementen aus S in genau einem Block aus $\mathcal{B}$ enthalten ist. Wir nutzen aus, dass $(Q, \circ)$ mit $Q = \{0, \ldots, 2m-1\}$ eine symmetrische, halb-idempotente Quasigruppe ist. Wir haben verschiedene Fälle bei der Auswahl des Paares zu unterscheiden.

1. $((i,1),(j,1))$ mit $0 \leq i < j \leq 2m-1$.

   Dieses Paar kann nur in einem Tripel vom Typ (a) liegen, und zwar liegt es in $\{(i,1),(j,1),(i \circ j,2)\}$.

2. $((i,2),(j,2))$ oder $((i,3),(j,3))$ mit $0 \leq i < j \leq 2m-1$.

   Diese beiden Fälle können wie Fall 1. erledigt werden.

3. $((i,1),(i,2))$ mit $i \in \{0, \ldots, m-1\}$.

   Dieses Paar liegt in genau einem Tripel vom Typ (c), nämlich $\{(i,1),(i,2),(i,3)\}$. Ferner kann das Paar in keinem Tripel vom Typ (a) liegen. Denn andernfalls gäbe es ein j mit $0 \leq i < j \leq 2m-1$ mit $i \circ j = i$. Da $(Q, \circ)$ eine Quasigruppe ist, besitzt die Gleichung $i \circ j = i$ bei gegebenem $i \in \{0, \ldots, m-1\}$ eine eindeutige Lösung $j \in \{0, \ldots, 2m-1\}$. Da $(Q, \circ)$ eine halb-idempotente Quasigruppe der Ordnung $2m$ ist, ist $i \circ i = i$ wegen $i \in \{0, \ldots, m-1\}$. Also wäre $j = i$, ein Widerspruch zu $j \neq i$. Ferner kann das Paar $((i,1),(i,2))$ ganz offensichtlich auch nicht in einem Tripel vom Typ (b) liegen.

4. $((i,2),(i,3))$ oder $((i,3),(i,1))$ mit $i \in \{0, \ldots, m-1\}$.

   Diese beiden Fälle können wie Fall 3. erledigt werden.

5. $((i,1),(i,2))$ mit $i \in \{m, \ldots, 2m-1\}$.

   Dieses Paar liegt in genau einem Tripel vom Typ (a). Denn da $(Q, \circ)$ eine Quasigruppe ist, besitzt die Gleichung $i \circ j = i$ eine eindeutige Lösung $j \in Q$. Da ferner die Quasigruppe $(Q, \circ)$ halb-idempotent ist und $i \in \{m, \ldots, 2m-1\}$ gilt, ist $j \neq i$. Daher ist das Paar $((i,1),(i,2))$ mit $i \in \{m, \ldots, 2m-1\}$ in dem Block

   $$\{(i,1),(j,1),(i \circ j,2)\} = \{(i,1),(j,1),(i,2)\} = \{(i,1),(j,1),(j \circ i,2)\}$$

   enthalten, wobei wir o. B. d. A. annehmen können, dass $i < j$.

6. $((i,2),(i,3))$ oder $((i,3),(i,1))$ mit $i \in \{m, \ldots, 2m-1\}$.

   Diese beiden Fälle können wie Fall 5. erledigt werden.

7. $((i,1),(k,2))$ mit $i, k \in \{0, \ldots, 2m-1\}$, $i \neq k$.

   Dieses Paar, welches offensichtlich nicht in einem Block vom Typ (b) oder (c) liegen kann, liegt in genau einem Tripel vom Typ (a). Denn es existiert genau ein $j \in Q$ mit $i \circ j = k$. Daher liegt das Paar genau im Tripel $\{(i,1),(j,1),(i \circ j,2)\}$.

8. $((i,2),(k,3))$ oder $((i,3),(k,1))$ mit $i,k \in \{0,\ldots,2m-1\}$, $i \neq k$.

   Diese beiden Fälle können wie Fall 7. erledigt werden.

9. $((i,1),\infty)$ mit $i \in \{0,\ldots,m-1\}$.

   Dieses Paar kann nicht in einem Block vom Typ (a) oder (c) liegen, da diese ∞ nicht enthalten. Das Paar liegt aber in genau einem Block vom Typ (b), nämlich in $\{\infty,(m+i,3),(i,1)\}$ falls $i \in \{0,\ldots,m-1\}$ bzw. in $\{\infty,(i,1),(i-m,2)\}$ für $i \in \{m,\ldots,2m-1\}$.

10. $((i,2),\infty)$ oder $((i,3),\infty)$ mit $i \in \{0,\ldots,m-1\}$.

    Diese beiden Fälle können wie Fall 9. erledigt werden.

Jedes Paar von Elementen aus S liegt also in genau einem Block aus $\mathcal{B}$. Damit ist gezeigt, dass die Skolem-Konstruktion für $n \equiv 1 \pmod 6$ ein Steiner Tripel System $STS(n)$ bestimmt. Der Beweis des Satzes ist abgeschlossen. □

Beispiel: Wir wollen ein Steiner Tripel System $STS(19)$ konstruieren. Hierzu benutzen wir die symmetrische, halb-idempotente Quasigruppe $(Q,\circ)$ mit der Operationstafel (siehe oben)

$\circ$	0	1	2	3	4	5
0	0	3	1	4	2	5
1	3	1	4	2	5	0
2	1	4	2	5	9	3
3	4	2	5	0	3	1
4	2	5	9	3	1	4
5	5	0	3	1	4	2

Die Grundmenge eines $STS(19)$ ist

$$S = \{\infty\} \cup \{0,1,2,3,4,5\} \times \{1,2,3\}.$$

Zur Abkürzung werden Elemente aus S, die ∞ nicht enthalten, mit $01,02,03,11,\ldots,53$ bezeichnet. Als Blöcke vom Typ (a), (b) bzw. (c) erhalten wir die folgende Menge von

Tripeln:

$\{01, 11, 32\}$	$\{02, 12, 33\}$	$\{03, 13, 31\}$
$\{01, 21, 12\}$	$\{02, 22, 13\}$	$\{03, 13, 11\}$
$\{01, 31, 42\}$	$\{02, 32, 43\}$	$\{03, 33, 41\}$
$\{01, 41, 22\}$	$\{02, 42, 23\}$	$\{03, 43, 21\}$
$\{01, 51, 52\}$	$\{02, 52, 53\}$	$\{03, 53, 51\}$
$\{11, 21, 42\}$	$\{12, 22, 43\}$	$\{13, 23, 41\}$
$\{11, 31, 22\}$	$\{12, 32, 23\}$	$\{13, 33, 21\}$
$\{11, 41, 52\}$	$\{12, 42, 53\}$	$\{13, 43, 51\}$
$\{11, 51, 02\}$	$\{12, 52, 03\}$	$\{13, 53, 01\}$
$\{21, 31, 52\}$	$\{22, 32, 53\}$	$\{23, 33, 51\}$
$\{21, 41, 02\}$	$\{22, 42, 03\}$	$\{23, 43, 01\}$
$\{21, 51, 32\}$	$\{32, 42, 33\}$	$\{33, 43, 31\}$
$\{31, 41, 32\}$	$\{32, 42, 33\}$	$\{33, 43, 31\}$
$\{31, 51, 12\}$	$31, 52, 13\}$	$\{33, 53, 11\}$
$\{41, 51, 42\}$	$\{42, 52, 43\}$	$\{43, 53, 41\}$
$\{\infty, 31, 02\}$	$\{\infty, 32, 03\}$	$\{\infty, 33, 01\}$
$\{\infty, 41, 12\}$	$\{\infty, 42, 13\}$	$\{\infty, 43, 11\}$
$\{\infty, 51, 22\}$	$\{\infty, 52, 23\}$	$\{\infty, 53, 21\}$
$\{01, 02, 03)\}$	$\{11, 12, 13\}$	$\{21, 22, 23\}$

□

7.5 $n \equiv 5 \pmod 6$: Es existiert fast ein Steiner Tripel System der Ordnung n

Ist $n \equiv 5 \pmod 6$, so existiert zwar kein Steiner Tripel System, aber man kommt sehr nahe! Hierzu wird der Begriff eines Steiner Tripel Systems verallgemeinert.

Definition 7.8 Ein *pairwise balanced design* bzw. PBD ist ein Paar $(S, \mathcal{B})$, wobei S eine endliche Menge und $\mathcal{B}$ eine Menge von Teilmengen B von S ist, welche die Eigenschaft hat, dass je zwei Elemente aus S in genau einem sogenannten *Block* $B \in \mathcal{B}$ enthalten sind. Die Anzahl $|S|$ der Elemente von S heißt die *Ordnung* des PBD.

Ein PBD $(S, \mathcal{B})$ mit $|B| = 3$ für alle $B \in \mathcal{B}$ ist also ein Steiner Tripel System. Unser Ziel in diesem Unterabschnitt ist es, einen konstruktiven Beweis für den folgenden Beweis anzugeben. Wir halten uns an C. C. Lindner, C. A. Rodger (1997, S. 14 ff.).

Satz 7.9 *Sei $n \equiv 5 \pmod 6$. Dann existiert ein PBD $(S, \mathcal{B})$ der Ordnung n mit der Eigenschaft, dass $|B_0| = 5$ für ein $B_0 \in \mathcal{B}$ und $|B| = 3$ für alle $B \in \mathcal{B} \setminus B_0$.*

Beweis: Sei $n = 6m + 5$ mit $m \in \mathbb{N}$, $Q := \{0, \ldots, 2m\}$ und $(Q, \circ)$ eine symmetrische, idempotente Quasigruppe der Ordnung $2m + 1$. Sei $\pi \colon Q \longrightarrow Q$ die Permutation

$$\pi := \begin{pmatrix} 0 & 1 & 2 & \cdots & 2m-1 & 2m \\ 0 & 2 & 3 & \cdots & 2m & 1 \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & \cdots & 2m \end{pmatrix}.$$

Die Grundmenge S besteht aus drei Kopien von Q sowie zwei besonderen Symbolen, nämlich ∞_1 und ∞_2. Es sei also

$$S := \{\infty_1, \infty_2\} \cup Q \times \{1, 2, 3\}.$$

Dann ist

$$|S| = 2 + 3(2m + 1) = 6m + 5 = n.$$

Die Blöcke aus $\mathcal{B}$ gehören zum Typ (a), (b) bzw. (c), wobei die Blöcke vom Typ (a) oder (b) jeweils Tripel sind und nur (c) aus genau einem Quintupel besteht.

(a) Die Tripel

$$\{(i,1),(j,1),(i \circ j,2)\},\ \{(i,2),(j,2),(i \circ j,3)\},\ \{(i,3),(j,3),(\pi(i \circ j),1)\}$$

gehören zu $\mathcal{B}$, $0 \le i < j \le 2m$,

(b) Die Tripel

$$\begin{array}{ll} \{\infty_1,(2i+1,1),(2i+1,2)\}, & \{\infty_2,(\pi(2i+1),1),(\pi(2i+1),2)\}, \\ \{\infty_1,(\pi(2i+1),2),(\pi(2i+1),3)\}, & \{\infty_2,((2i+1,2),(2i+1,3)\}, \\ \{\infty_1,(2i+1,3),(\pi(2i+1),1)\}, & \{\infty_2,(\pi^{-1}(2i+1),3),(2i+1,1)\} \end{array}$$

gehören zu $\mathcal{B}$, $i \in \{0, \ldots, m-1\}$.

(c) Das Quintupel

$$\{\infty_1, \infty_2, (0,1), (0,2), (0,3)\}$$

gehört zu $\mathcal{B}$.

Nun zeigen wir, dass jedes Paar von Elementen aus S in genau einem Block aus $\mathcal{B}$ enthalten ist. Wieder haben wir verschiedene Fälle bei der Auswahl des Paares zu unterscheiden. Zunächst betrachten wir die Fälle, bei denen weder ∞_1 noch ∞_2 eines der beiden gewählten Elemente aus S ist.

1. $((i,1),(j,1))$ mit $0 \le i < j \le 2m$.

   Dieses Paar kann nur in einem Tripel vom Typ (a) liegen, und zwar liegt es genau in dem Tripel $\{(i,1),(j,1),(i \circ j,2)\}$.

2. $((i,2),(j,2))$ oder $((i,3),(j,3))$ mit $0 \le i < j \le 2m$.

   Diese beiden Fälle können wie Fall 1. erledigt werden.

3. $((i,1),(i,2))$ mit $i \in \{0, \ldots, 2m\}$.

   Hier müssen wir Fallunterscheidungen machen.

   (a) $i = 0$.

   Das Paar $((0,1),(0,2))$ liegt in dem Quintupel

   $$\{\infty_1, \infty_2, (0,1), (0,2), (0,3)\}$$

   und in keinem anderen Block.

(b) i ist ungerade: $i = 2k + 1$ mit $k \in \{0, \ldots, m-1\}$.
Das Paar $((2k+1, 1), (2k+1, 2))$ liegt in dem Block

$$\{\infty_1, (2k+1, 1), (2k+1, 2)\}$$

und in keinem anderen Block.

(c) $i > 0$ ist gerade: $i = 2k$ mit $k \in \{1, \ldots, m\}$.
Das Paar $((2k, 1), (2k, 2))$ liegt für $k \in \{1, \ldots, m\}$ wegen $\pi(2k-1) = 2k$ in dem Block

$$\{\infty_2, (\pi(2k-1), 1), (\pi(2k-1), 2)\}$$

und offenbar in keinem anderen Block.

4. $((i, 2), (i, 3))$ oder $((i, 3), (i, 1))$ mit $i \in \{0, \ldots, 2m\}$.

Diese beiden Fälle können weitgehend analog zu 3. behandelt werden.

(a) $i = 0$.
Das Paar $((0, 2), (0, 3))$ bzw. $((0, 3), (0, 1))$ liegen in dem Quintupel

$$\{\infty_1, \infty_2, (0, 1), (0, 2), (0, 3)\}$$

und in keinem anderen Block.

(b) i ist ungerade: $i = 2k + 1$ mit $k \in \{0, \ldots, m-1\}$.
Das Paar $((2k+1, 2), (2k+1, 3)$ liegt in dem Block

$$\{\infty_2, (2k+1, 2), (2k+1, 3)\}$$

vom Typ (b) und in keinem anderen Block. Etwas komplizierter ist es mit dem Paar $((2k+1, 3), (2k+1, 1))$. Offensichtlich liegt dieses Paar in keinem Block vom Typ (b) und natürlich auch nicht in dem Block vom Typ (c). Es liegt aber in genau einem Block vom Typ (a), wie wir uns jetzt überlegen. Da $(Q, \circ)$ eine Quasigruppe ist, existiert zu gegebenem $k \in \{0, \ldots, m-1\}$ genau ein $j \in Q$ mit

$$(2k+1) \circ j = \pi^{-1}(2k+1) = 2k.$$

Da die Quasigruppe $(Q, \circ)$ idempotent ist, ist $j \neq 2k + 1$. Bei dem Tripel

$$\{(i, 3), (j, 3), \pi(i \circ j), 1)\} = \{(2k+1, 3), (j, 3), (2k+1, 1)\}$$

können wir annehmen, dass $2k + 1 < j$, sodass das Paar

$$((2k+1, 3), (2k+1, 1))$$

in dem Block $\{(2k+1, 3), (j, 3), (2k+1, 1)\}$ enthalten ist.

(c) $i > 0$ ist gerade: $i = 2k$ mit $k \in \{1, \ldots, m\}$.
Das Paar $((2k, 2), (2k, 3))$ liegt wegen $\pi(2k-1) = 2k$ in dem Block

$$\{\infty_1, (\pi(2k-1), 2), (\pi(2k-1), 3)\}$$

und in keinem anderen Block. Mit dem Paar $((2k, 3), (2k, 1))$ ist es komplizierter. Offensichtlich liegt es in keinem Block vom Typ (b) und natürlich auch nicht in dem Quintupel (c). Zu vorgegebenem $k \in \{1, \ldots, m\}$ gibt es genau ein $j \in Q$ mit

$$(2k) \circ j = \pi^{-1}(2k) = 2k - 1.$$

Da die Quasigruppe $(Q, \circ)$ idempotent ist, ist $j \neq 2k$. Da wir $2k < j$ annehmen können, ist das Paar $((2k, 3), (2k, 1))$ in dem Block

$$\{(2k, 3), 8j, 3), (\pi((2k) \circ j), 1)\} = \{(2k, 3), 8j, 3), (2k, 1)\}$$

enthalten und sicher nicht in einem anderen Block vom Typ (a).

5. $(\infty_1, (i, 1))$ mit $i \in \{0, \ldots, 2m\}$.
Wieder müssen wir eine Fallunterscheidungen machen.

(a) $i = 0$.
Das Paar $(\infty_1, (0, 1)$ liegt im Quintupel $\{\infty_1, \infty_2, (0, 1), (0, 2), (0, 3)\}$ und keinem anderen Block.

(b) i ist ungerade: $i = 2k + 1$ mit $k \in \{0, \ldots, m-1\}$.
Das Paar $(\infty_1, (2k+1, 1))$ ist in dem Block $\{\infty_1, (2k+1, 1), (2k+1, 2)\}$ enthalten und in keinem anderen Block.

(c) $i > 0$ ist gerade: $i = 2k$ mit $k \in \{1, \ldots, m\}$.
Das Paar $(\infty_1, (2k, 1))$ ist in dem Block

$$\{\infty_1, (2k-1, 3), (\pi(2k-1), 1)\} = \{\infty_1, (2k-1, 3), (2k, 1)\}$$

und keinem anderen Block enthalten.

6. $(\infty_1, (i, 2))$ oder $(\infty_1, (i, 3))$ mit $i \in \{0, \ldots, 2m\}$.
Diese beiden Fälle können wie in 5. behandelt werden.

7. $(\infty_2, (i, 1))$, $(\infty_2, (i, 2))$ oder $(\infty_2, (i, 3))$ mit $i \in \{0, \ldots, 2m\}$.
Diese Fälle können wie in 5. bzw. 6. behandelt werden.

8. (∞_1, ∞_2).
Dieses Paar liegt genau in dem Block $\{\infty_1, \infty_2, (0, 1), (0, 2), (0, 3)\}$.

Damit haben wir gezeigt, dass jedes Paar von Elementen aus S in genau einem Block liegt. Der Satz ist bewiesen. □

Beispiel: Sei $n := 11$ und damit $n \equiv 5 \pmod 6$. Wir wollen ein PBD $(S, \mathcal{B})$ der Ordnung 11 bestimmen, wobei wir wie im Beweis von Satz 7.9 vorgehen werden. Mit $m := 1$ ist $n = 6m + 5$. Mit $Q := \{0, 1, 2\}$ nehmen wir als symmetrische, idempotente Quasigruppe $(Q, \circ)$ der Ordnung $3 = 2m + 1$ wieder

$\circ$	0	1	2
0	0	2	1
1	2	1	0
2	1	0	2

Wir definieren die Permutation $\pi\colon Q \longrightarrow Q$ durch

$$\pi := \begin{pmatrix} 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 2 & 1 \end{pmatrix}.$$

Die Grundmenge für das gesuchte PBD der Ordnung 11 ist

$$S := \{\infty_1, \infty_2\} \cup \{0, 1, 2\} \times \{1, 2, 3\}.$$

Die Blöcke vom Typ (a), (b) und (c) sind der Reihe nach

(a)	$\{(0,1),(1,1),(2,2)\}$	$\{(0,2),(1,2),(2,3)\}$	$\{(0,3),(1,3),(1,1)\}$
	$\{(0,1),(2,1),(1,2)\}$	$\{(0,2),(2,2),(1,3)\}$	$\{(0,3),(2,3),(2,1)\}$
	$\{(1,1),(2,1),(0,2)\}$	$\{(1,2),(2,2),(0,3)\}$	$\{(1,3),(2,3),(0,1)\}$
(b)	$\{\infty_1,(1,1),(1,2)\}$	$\{\infty_1,(2,2),(2,3)\}$	$\{\infty_1,(1,3),(2,1)\}$
	$\{\infty_2,(2,1),(2,2)\}$	$\{\infty_2,(1,2),(1,3)\}$	$\{\infty_2,(2,3),(1,1)\}$
(c)	$\{\infty_1,\infty_2,(0,1),(0,2),(0,3)\}$		

□

7.6 Zyklische Steiner Tripel Systeme

Wir halten uns in diesem Unterabschnitt eng an C. C. Lindner, C. A. Rodger (1997, S. 31 ff.).

Definition 7.10 Ein Steiner Tripel System $(S, \mathcal{B})$ der Ordnung n heißt *zyklisch*, wenn es eine aus nur einem Zyklus (der Länge n) bestehende Permutation $\pi\colon S \longrightarrow S$ gibt mit

$$\{x, y, z\} \in \mathcal{B} \Longrightarrow \{\pi(x), \pi(y), \pi(z)\} \in \mathcal{B}.$$

Beispiel: Sei (siehe das Beispiel im Anschluss an Definition 7.1)

$$S := \{1, 2, 3, 4, 5, 6, 7\}$$

und

$$\mathcal{B} := \{123, 345, 156, 147, 257, 367, 246\},$$

wobei hier wieder z. B. 123 eine Kurzschreibweise für die Menge $\{1,2,3\}$ ist. Dann ist $(S,\mathcal{B})$ ein $STS(7)$. Wir wollen zeigen, dass $(S,\mathcal{B})$ ein *zyklisches* $STS(7)$ ist. Hierzu definieren wir die Permutation $\pi\colon S \longrightarrow S$ durch $\pi := \begin{pmatrix} 3 & 5 & 2 & 1 & 6 & 4 & 7 \end{pmatrix}$. Dann ist

$$\begin{array}{lllllll} \pi(123) & = & \pi(1)\pi(2)\pi(3) & = & 615 & = & 156 \\ \pi(345) & = & \pi(3)\pi(4)\pi(5) & = & 572 & = & 257 \\ \pi(156) & = & \pi(1)\pi(5)\pi(6) & = & 624 & = & 246 \\ \pi(147) & = & \pi(1)\pi(4)\pi(7) & = & 673 & = & 367 \\ \pi(257) & = & \pi(2)\pi(5)\pi(7) & = & 123 & & \\ \pi(367) & = & \pi(3)\pi(6)\pi(7) & = & 543 & = & 345 \\ \pi(246) & = & \pi(2)\pi(4)\pi(6) & = & 174 & = & 147 \end{array}$$

Das Bild jedes Blocks $B \in \mathcal{B}$ unter π ist also wieder ein Block. Da π eine aus nur einem Zyklus bestehende Permutation ist, ist $(S,\mathcal{B})$ ein zyklisches Steiner Tripel System. □

Beispiel: Sei das $STS(9)$ durch

$$S := \{1,2,3,4,5,6,7,8,9\}$$

und

$$\mathcal{B} := \{123, 456, 789, 147, 258, 369, 159, 267, 348, 168, 249, 357\}$$

gegeben, siehe das Beispiel auf S. 145. Durch

$$\pi := \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} 4 & 5 & 6 \end{pmatrix}\begin{pmatrix} 7 & 8 & 9 \end{pmatrix}$$

ist eine Permutation von S gegeben, die jeden Block aus $\mathcal{B}$ in einen Block aus $\mathcal{B}$ überführt:

$$\begin{array}{lllllll} \pi(123) & = & \pi(1)\pi(2)\pi(3) & = & 231 & = & 123 \\ \pi(456) & = & \pi(4)\pi(5)\pi(6) & = & 564 & = & 456 \\ \pi(789) & = & \pi(7)\pi(8)\pi(9) & = & 897 & = & 789 \\ \pi(147) & = & \pi(1)\pi(4)\pi(7) & = & 258 & & \\ \pi(258) & = & \pi(2)\pi(5)\pi(8) & = & 369 & & \\ \pi(369) & = & \pi(3)\pi(6)\pi(9) & = & 147 & & \\ \pi(159) & = & \pi(1)\pi(5)\pi(9) & = & 267 & & \\ \pi(267) & = & \pi(2)\pi(6)\pi(7) & = & 348 & & \\ \pi(348) & = & \pi(3)\pi(4)\pi(8) & = & 159 & & \\ \pi(168) & = & \pi(1)\pi(6)\pi(8) & = & 249 & & \\ \pi(249) & = & \pi(2)\pi(4)\pi(9) & = & 357 & & \\ \pi(357) & = & \pi(3)\pi(5)\pi(7) & = & 168 & & \end{array}$$

Es gibt also eine Permutation von S, die jeden Block aus $\mathcal{B}$ in einen Block aus $\mathcal{B}$ überführt, aber es gibt keine zyklische Permutation mit der entsprechenden Eigenschaft. Daher gibt es kein zyklisches $STS(9)$. □

Wir stellen uns die Frage, für welche n ein zyklisches $STS(n)$ existiert. Die Antwort ist überraschend. Denn es stellt sich heraus, dass dies fast immer dann der Fall ist, wenn ein $STS(n)$ existiert.

Satz 7.11 *Ist* $n \neq 9$ *und* $n \equiv 1 \pmod 6$ *oder* $n \equiv 3 \pmod 6$, *so existiert ein zyklisches* $STS(n)$.

Aufbauend auf L. Heffter (1897) gab R. Peltesohn (1939) einen Beweis von Satz 7.11 an. Genauer wird die Existenz eines zyklischen Steiner Tripel Systems auf die Lösung von zwei sogenannten *Heffterschen Differenzenproblemen* zurückgeführt. Hierzu müssen einige Begriffe und Hilfsmittel eingeführt werden.

Sei $n \in \mathbb{N}$. Drei verschiedene natürliche Zahlen nennen wir ein *Differenzentripel modulo* n, wenn (i) ihre Summe $\equiv 0 \pmod{n}$ ist oder (ii) eines der drei Elemente gleich der Summe der beiden anderen $\pmod{n}$ ist. Bei gegebenem $m \in \mathbb{N}$ lauten die beiden *Heffterschen Differenzenprobleme* folgendermaßen:

1. Sei $n := 6m + 1$. Man bestimme eine Partition $HDP_1(m)$ der Menge

$$\{1, \ldots, 3m\}$$

   in m (verschiedene) Differenzentripel modulo n.

2. Sei $n := 6m + 3$. Man bestimme eine Partition $HDP_2(m)$ der Menge

$$\{1, \ldots, 3m + 1\} \setminus \{2m + 1\}$$

   in m (verschiedene) Differenzentripel modulo n.

Beispiele: 1. Es ist $HDP_1(1) = \{\{1, 2, 3\}\}$. Denn wegen $1 + 2 = 3$ ist $\{1, 2, 3\}$ ein Differenzentripel modulo 7.

2. Es ist $HDP_1(2) = \{\{1, 3, 4\}, \{2, 5, 6\}\}$. Denn einerseits sind $\{1, 3, 4\}$ und $\{2, 5, 6\}$ wegen $1 + 3 = 4$ und $2 + 5 + 6 = 13 \equiv 0 \pmod{13}$ Differenzentripel modulo 13, andererseits ist $\{\{1, 3, 4\}, \{2, 5, 6\}\}$ eine Partition von $\{1, 2, 3, 4, 5, 6\}$.

3. Es ist $HDP_1(3) = \{\{1, 5, 6\}, \{2, 8, 9\}, \{3, 4, 7\}\}$. Denn $\{1, 5, 6\}$ und $\{3, 4, 7\}$ sind Differenzentripel modulo 19, da $1 + 5 = 6$ und $3 + 4 = 7$. Ferner ist auch $\{2, 8, 9\}$ ein Differenzentripel modulo 19, da $2 + 8 + 9 = 19 \equiv 0 \pmod{19}$.

4. Das zweite Hefftersche Differenzenproblem besitzt für $m = 1$ (bzw. $n = 6 \cdot 1 + 3 = 9$) *keine* Lösung. Denn andernfalls müsste $\{1, 2, 4\}$ ein Differenzentripel modulo 9 sein, was nicht der Fall ist.

5. Es ist $HDP_2(2) = \{\{1, 3, 4\}, \{2, 6, 7\}\}$. Denn einerseits sind $\{1, 3, 4\}$ und $\{2, 6, 7\}$ Differenzentripel modulo 15, andererseits ist $HDP_2(2)$ eine Partition von $\{1, 2, 3, 4, 6, 7\}$.

6. Es ist $HDP_2(4) = \{\{1, 12, 13\}, \{2, 5, 7\}, \{3, 8, 11\}, \{4, 6, 10\}\}$, wie man unschwer nachrechnet. □

Ist $\{x, y, z\}$ ein Differenzentripel modulo n, so ist $x + y + z \equiv 0 \pmod{n}$ bzw. $x + y \equiv -z \pmod{n}$ oder $x + y \equiv z \pmod{n}$, bei entsprechender Bezeichnung der Elemente des Differenzentripels also $x + y \equiv \pm z \pmod{n}$. Das Tripel $\{0, x, x + y\}$ heißt dann ein *Basisblock* des Differentripels $\{x, y, z\}$ modulo n.

Beispiel: Wie wir eben gesehen haben, ist $HDP_1(2) = \{\{1, 3, 4\}, \{2, 5, 6\}\}$ eine Lösung des ersten Heffterschen Differenzenproblems für $m = 2$ bzw. $n = 6 \cdot 2 + 1 = 13$. Zu den Differenzentripeln $\{1, 3, 4\}$ bzw. $\{2, 5, 6\}$ gehörende Basisblocks sind $\{0, 1, 4\}$ bzw. $\{0, 2, 7\}$. Man beachte, dass ein Basisblock durch ein Differenzentripel nicht eindeutig bestimmt ist. So ist auch $\{0, 3, 4\}$ ein Basisblock für $\{1, 3, 4\}$ bzw. $\{0, 6, 11\}$ ein Basisblock zu $\{2, 5, 6\}$. □

Jetzt nehmen wir an, es sei schon bewiesen, dass das erste und das zweite Hefftersche Differenzenproblem für alle $m \in \mathbb{N}$ bzw. alle $m \in \mathbb{N}$ mit $m \geq 2$ (für das zweite Hefftersche Differenzenproblem) eine Lösung habe. Unter dieser *Prämisse* geben wir einen Beweis für Satz 7.11 an, wobei wir der Darstellung bei C. C. Lindner, C. A. Rodger (1997, S. 33) und C. J. Colbourn, A. Rosa (1999, S. 30 ff) folgen werden.

Beweis von Satz 7.11: Sei $D(n)$ eine Menge von Differenzentripeln, welche eine Lösung zu Heffters Differenzenproblem ist. Für $n = 6m+1$ sei also $D(n) = HDP_1(m)$, während $D(n) = HDP_2(m)$ für $n = 6m + 3$ und $m \geq 2$. Sei weiter $B(n)$ die Menge der Basisblocks zu den Differenzentripeln in $D(n)$. Wir definieren $(S, \mathcal{B})$ und damit ein (wie wir zeigen werden) zyklisches $STS(n)$, indem wir

$$S := \{0, \ldots, n-1\}$$

setzen, während bei der Definition der Blöcke aus $\mathcal{B}$ eine Fallunterscheidung gemacht wird. In jedem Fall sind alle Summen modulo n zu nehmen.

1. Ist $n \equiv 1 \pmod 6$ bzw. $n = 6m + 1$, so sei

$$\mathcal{B} := \{\{i, x+i, x+y+i\} \subset S : i \in \{0, \ldots, n-1\},\ \{0, x, x+y\} \in B(n)\}.$$

2. Ist $n \equiv 3 \pmod 6$ und $n \geq 15$ bzw. $n = 6m + 3$ mit $m \geq 2$, so sei

$$\begin{aligned}\mathcal{B} \ := \ & \{\{i, x+i, x+y+i\} \subset S : i \in \{0, \ldots, n-1\},\ \{0, x, x+y\} \in B(n)\} \\ & \cup \{\{i, 2m+1+i, 4m+1+i\} : i \in \{0, \ldots, 2m\}\}.\end{aligned}$$

   Die Menge der Tripel $\{\{i, 2m+1+i, 4m+2+i\} : i \in \{0, \ldots, 2m\}\}$ nennen wir eine *kurze Bahn*.

Dass $(S, \mathcal{B})$ ein $STS(n)$ ist, wird in zwei Schritten gezeigt. Wir zeigen nämlich:

(i) Es ist
$$|\mathcal{B}| \leq \frac{n(n-1)}{6}.$$

(ii) Sind $a, b \in S$ mit $a \neq b$, so existiert ein $B \in \mathcal{B}$ mit $\{a, b\} \subset B$.

Angenommen, (i) und (ii) seien schon bewiesen. Wegen des zweiten Teiles von Satz 7.4 ist dann $(S, \mathcal{B})$ ein $STS(n)$.

Nun müssen wir nachweisen, dass die Aussagen (i) und (ii) in unserem konkreten Fall erfüllt sind. Die Menge $D(n)$ der Differenzentripel besteht aus m Elementen, wobei $n = 6m + 1$ bzw. $n = 6m + 3$. Dasselbe gilt dann auch für die zugehörige Menge der Basisblocks. Für $n = 6m + 1$ ist offenbar

$$|\mathcal{B}| \leq n \cdot |B(n)| = nm = \frac{n(n-1)}{6}.$$

Für $n = 6m + 3$ ist dagegen

$$|\mathcal{B}| \leq n \cdot |B(n)| + 2m + 1 = nm + 2m + 1 = \frac{n(n-3)}{6} + \frac{n}{3} = \frac{n(n-1)}{6}.$$

Damit ist die Aussage (i) in unserem konkreten Fall bewiesen. Um (ii) nachzuweisen, geben wir uns $a, b \in S := \{0, \ldots, n-1\}$ mit $a \neq b$ vor und zeigen die Existenz eines Blocks $B \in \mathcal{B}$ mit $\{a, b\} \subset B$. Wir definieren $d := b - a$ modulo n. Wir können annehmen, dass $1 \leq d \leq (n-1)/2$, da man dies notfalls durch Vertauschen von a und b erreichen kann. Nun unterscheiden wir zwei Fälle:

(a) Es ist $n = 6m + 3$ und $d = 2m + 1$.

Wir wollen uns davon überzeugen, dass $\{a, b\}$ in einem Block enthalten ist, der zu einer kurzen Bahn gehört. Wir machen eine Fallunterscheidung. Ist $a \in \{0, \ldots, 2m\}$, so ist

$$\{a, b\} \subset \{a, b, 2d + a\} \equiv \{a, d + a, 2d + a\} \pmod{n},$$

also $\{a, b\}$ in einem Block einer kurzen Bahn enthalten. Für $a \in \{d, \ldots, d + 2m\}$ bzw. $a = d + i$ mit $i \in \{0, \ldots, 2m\}$ ist

$$\{a, b\} \subset \{i, a, b\} \equiv \{i, d + i, 2d + i\} \pmod{n}.$$

Ist schließlich $a \in \{2d, \ldots, 2d + 2m\}$ bzw. $a = 2d + i$ mit $i \in \{0, \ldots, 2m\}$, so ist

$$\begin{aligned} \{a, b\} &\subset \{b, d + i, a\} \\ &\equiv \{a + d, d + i, 2d + i\} \\ &\equiv \{3d + i, d + i, 2d + i\} \\ &\equiv \{i, d + i, 2d + i\} \pmod{n}. \end{aligned}$$

Also ist $\{a, b\}$ auch in diesem Fall in einem Block einer kurzen Bahn enthalten.

(b) Andernfalls ist d ein Element genau eines Differenzentripels $\{x, y, z\} \in D(n)$, da $D(n)$ eine Partition der Menge $\{1, \ldots, 3m\}$ bzw. $\{1, \ldots, 3m + 1\} \setminus \{2m + 1\}$ ist. Ist $d = x$, so ist

$$\{a, b\} \subset \{a, b, b + y\} \equiv \{a, x + a, x + y + a\} \pmod{n}.$$

Ist dagegen $d = y$, so bestimme man $i \in \{0, \ldots, n-1\}$ so, dass $x + i \equiv a \pmod{n}$. Dann ist

$$\{a, b\} \subset \{i, a, b\} \equiv \{i, x + i, x + y + i\} \pmod{n}.$$

Im Fall $d = z$ müssen wir zwei Fälle unterscheiden. Im ersten Fall ist $x + y \equiv z \pmod{n}$. Dann ist

$$\{a, b\} \subset \{a, x + a, b\} \equiv \{a, x + a, x + y + a\} \pmod{n}.$$

Im zweiten Fall ist $x + y + z \equiv 0 \pmod{n}$. Dann ist

$$\{a, b\} \subset \{b, x + b, a\} \equiv \{b, x + b, x + y + b\} \pmod{n}.$$

Damit ist gezeigt, dass durch $(S, \mathcal{B})$ ein $STS(n)$ gegeben ist. Dass dieses Steiner Tripel System auch zyklisch ist, ist einfach einzusehen. Man definiere nämlich die Permutation $\pi\colon \{0,\ldots,n-1\} \longrightarrow \{0,\ldots,n-1\}$ durch

$$\pi := \begin{pmatrix} 0 & 1 & \cdots & n-1 \end{pmatrix}.$$

Sei

$$B := \{i, x+i, x+y+i\}$$

mit $i \in \{0,\ldots,n-1\}$ und $\{0, x, x+y\} \in B(n)$. Dann ist (alle Summanden sind modulo n zu nehmen!)

$$\pi(B) = \{i+1, x+i+1, x+y+i+1\} \in \mathcal{B}.$$

Die entsprechende Aussage müssen wir uns im Falle $n \equiv 3 \pmod 6$ noch für die Blöcke einer kurzen Bahn überlegen. Für die ersten $2m$ Blöcke einer kurzen Bahn ist dies offensichtlich. Für den letzten Block ist es aber auch richtig, da

$$\begin{aligned} \pi(\{2m, 2m+1+2m, 4m+2+2m\}) &= \{2m+1, 4m+2, 6m+3\} \\ &\equiv \{2m+1, 4m+2, 0\} \pmod n, \end{aligned}$$

also auch das Bild des letzten Blocks einer kurzen Bahn ein Block dieser Bahn ist, nämlich der erste. Damit ist der Satz schließlich bewiesen, wenn man davon absieht, dass die Existenz von Lösungen der Heffterschen Differenzenprobleme noch nicht bewiesen ist. Dies werden wir im nächsten Unterabschnitt nachholen. □

Beispiel: Im Beweis zu Satz 7.11 haben wir die Blöcke eines zyklischen $STS(n)$, getrennt für $n \equiv 1 \pmod 6$ sowie $n \equiv 3 \pmod 6$, explizit angegeben. Für $n = 13$ wollen wir diese Konstruktion noch einmal durchführen. Wie wir uns oben überlegt haben, ist

$$D(13) = \{\{1,3,4\}, \{2,5,6\}\}.$$

Als zugehörige Basisblocks können wir

$$B(13) = \{\{0,1,4\}, \{0,2,7\}\}$$

nehmen. Die Grundmenge des zyklischen $STS(13)$ ist

$$S := \{0,\ldots,12\},$$

während die Menge der Blöcke durch

$$\mathcal{B} = \{\{i, 1+i, 4+i\}, \{i, 2+i, 7+i\} : i \in \{0,\ldots,12\}\}$$

gegeben ist, wobei natürlich wieder alle Summen modulo 13 zu verstehen sind. Genauer besteht $\mathcal{B}$ also aus den Blöcken

$$\{0,1,4\}, \{1,2,5\}, \{2,3,6\}, \{3,4,7\}, \{4,5,8\}, \{5,6,9\}, \{6,7,10\}, \{7,8,11\},$$
$$\{8,9,12\}, \{9,10,0\}, \{10,11,1\}, \{11,12,2\}, \{12,0,3\}$$

sowie

$$\{0,2,7\}, \{1,3,8\}, \{2,4,9\}, \{3,5,10\}, \{4,6,11\}, \{5,7,12\}, \{6,8,0\}, \{7,9,1\},$$
$$\{8,10,2\}, \{9,11,3\}, \{10,12,4\}, \{11,0,5\}, \{12,1,6\}.$$

Z. B. ist das Paar $\{2, 10\}$ genau im Tripel $\{8, 10, 2\}$ enthalten, das Paar $\{4, 9\}$ genau im Tripel $\{2, 4, 9\}$.

Für $n = 15$ ist

$$D(15) = \{\{1, 3, 4\}, \{2, 6, 7\}\}.$$

Mit

$$B(15) = \{\{0, 1, 4\}, \{2, 6, 7\}\}$$

erhalten wir

$$\begin{aligned}\mathcal{B} \;=\; &\{\{i, 1+i, 4+i\}, \{i, 2+i, 8+i\} : i \in \{0, \ldots, 14\}\} \\ &\cup \{\{i, 5+i, 10+i\} : i \in \{0, \ldots, 4\}\}.\end{aligned}$$

Daher besteht $\mathcal{B}$ aus den Blöcken

$$\{0, 1, 4\}, \{1, 2, 5\}, \{2, 3, 6\}, \{3, 4, 7\}, \{4, 5, 8\}, \{5, 6, 9\}, \{6, 7, 10\}, \{7, 8, 11\}, \{8, 9, 12\},$$
$$\{9, 10, 13\}, \{10, 11, 14\}, \{11, 12, 0\}, \{12, 13, 1\}, \{13, 14, 2\}, \{14, 0, 3\}$$

und

$$\{0, 2, 8\}, \{1, 3, 9\}, \{2, 4, 10\}, \{3, 5, 11\}, \{4, 6, 12\}, \{5, 7, 13\}, \{6, 8, 14\}, \{7, 9, 0\},$$
$$\{8, 10, 1\}, \{9, 11, 2\}, \{10, 12, 3\}, \{11, 13, 4\}, \{12, 14, 5\}, \{13, 0, 6\}, \{14, 1, 7\}$$

sowie

$$\{0, 5, 10\}, \{1, 6, 11\}, \{2, 7, 12\}, \{3, 8, 13\}, \{4, 9, 14\}.$$

Z. B. ist das Paar $\{7, 12\}$ genau im Tripel $\{2, 7, 12\}$ enthalten, das Paar $\{1, 13\}$ genau in dem Tripel $\{12, 13, 1\}$. □

7.7 Eine Lösung der Heffterschen Differenzenprobleme

In diesem Unterabschnitt wollen wir eine Lücke schließen, die im Beweis von Satz 7.11 blieb. Wir wollen nämlich den Nachweis dafür nachholen, dass die beiden Heffterschen Differenzenprobleme eine Lösung besitzen. Die Formulierung dieser Probleme wiederholen wir hier. Bei gegebenen $n \in \mathbb{N}$ nennen wir drei verschiedene natürliche Zahlen ein *Differenzentripel modulo* n, wenn (i) ihre Summe $\equiv 0 \pmod{n}$ ist oder (ii) eines der drei Elemente gleich der Summe der beiden anderen $\pmod{n}$ ist. Bei gegebenem $m \in \mathbb{N}$ lauten die beiden *Heffterschen Differenzenprobleme* folgendermaßen:

1. Sei $n := 6m + 1$. Man bestimme eine Partition $HDP_1(m)$ der Menge

   $$\{1, \ldots, 3m\}$$

   in m (verschiedene) Differenzentripel modulo n.

2. Sei $n := 6m + 3$. Man bestimme eine Partition $HDP_2(m)$ der Menge

   $$\{1, \ldots, 3m + 1\} \setminus \{2m + 1\}$$

   in m (verschiedene) Differenzentripel modulo n.

Die folgenden beiden Sätze stammen von R. Peltesohn (1939).

Satz 7.12 *Für alle $m \in \mathbb{N}$ existiert eine Lösung des ersten Heffterschen Differenzenproblems, also eine Partition $HDP_1(m)$ von $\{1, \ldots, 3m\}$ in m Differenzentripel modulo n, wobei $n := 6m + 1$.*

Beweis: Der Beweis erfolgt durch Angabe der entsprechenden Partitionen. Zunächst geben wir die Lösung für $m = 1, 2, 3$ an. Wie man leicht nachrechnet (das haben wir in einem früheren Beispiel schon getan), ist

$$\begin{aligned} HDP_1(1) &= \{\{1,2,3\}\}, \\ HDP_1(2) &= \{\{1,3,4\},\{2,5,6\}\}, \\ HDP_1(3) &= \{\{1,5,6\},\{2,8,9\},\{3,4,7\}\}. \end{aligned}$$

(1) Für $k \in \mathbb{N}$ mit $k \geq 2$ und $m := 3k$ besteht eine Lösung $HDP_1(m)$ des ersten Heffterschen Differenzenproblems aus den $3m$ Tripeln

$$\begin{array}{ll} \{3j+1, 4k-j+1, 4k+2j+2\}, & j = 0, \ldots, k-1, \\ \{3j+2, 8k-j, g(k,j)\}, & j = 0, \ldots, k-1, \\ \{3j+3, 6k-2j-1, 6k+j+2\}, & j = 0, \ldots, k-2, \\ \multicolumn{2}{c}{\{3k, 3k+1, 6k+1\}.} \end{array}$$

Hierbei[32] ist

$$g(k,j) := \begin{cases} 8k+2j+2, & j = 0, \ldots, \lfloor (k-2)/2 \rfloor, \\ 10k-2j-1, & j = \lfloor (k-2)/2 \rfloor + 1, \ldots, k-1. \end{cases}$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1, 4k+1, 4k+2\}$			$\{2, 8k, 8k+2\}$			$\{3, 6k-1, 6k+2\}$		
$\{4, 4k, 4k+4\}$			$\{5, 8k-1, 8k+4\}$			$\{6, 6k-3, 6k+3\}$		
$\{7, 4k-1, 4k+6\}$			$\{8, 8k-2, 8k+6\}$			$\{9, 6k-5, 6k+4\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-8, 3k+4, 6k-4\}$			$\{3k-7, 7k+3, 8k+5\}$			$\{3k-6, 4k+5, 7k-1\}$		
$\{3k-5, 3k+3, 6k-2\}$			$\{3k-4, 7k+2, 8k+3\}$			$\{3k-3, 4k+3, 7k\}$		
$\{3k-2, 3k+2, 6k\}$			$\{3k-1, 7k+1, 8k+1\}$					

sowie

$$\{\alpha, \beta, \gamma\} := \{3k, 3k+1, 6k+1\}.$$

[32] Hier ist die Darstellung bei C. C. Lindner, C. A. Rodger (1997, S. 187) und C. J. Colbourn, A. Rosa (1999, S. 31) m. E. nicht vollständig. Korrekt ist die Darstellung bei R. Peltesohn (1939), der wir im wesentlichen folgen.

Hierdurch ist eine Partition der Zahlen $\{1,\dots,3m\} = \{1,\dots,9k\}$ gegeben. Die Verteilung der Zahlen aus $\{1,\dots,9k\}$ ergibt sich aus

$$\begin{array}{c|c} 1,\dots,3k-1 & S_1, S_4, S_7 \\ 3k & \alpha \\ 3k+1 & \beta \\ 3k+2,\dots,4k+1 & S_2 \\ 4k+2,\dots,6k & S_3 S_8 \\ 6k+1 & \gamma \\ 6k+2,\dots,7k & S_9 \\ 7_{k+1},\dots,8k & S_5 \\ 8k+1,\dots,9k & S_6 \end{array}$$

Wie man leicht nachrechnet, sind alle angegebenen Tripel Differenzentripel modulo $n = 18k+1$. Damit ist nachgewiesen, dass das erste Hefftersche Differenzenproblem für $m \equiv 0 \pmod 3$ eine Lösung besitzt.

(2) Für $k \in \mathbb{N}$ und $m := 3k+1$ besteht eine Lösung $HDP_1(m)$ des ersten Heffterschen Differenzenproblems aus den $3m$ Tripeln

$$\begin{array}{ll} \{3j+1, 8k-j+3, g(k,j)\}, & j = 0,\dots,k-1, \\ \{3j+2, 6k-2j+1, 6k+j+3\}, & j = 0,\dots,k-1, \\ \{3j+3, 4k-j+1, 4k+2j+4\}, & j = 0,\dots,k-1, \\ \multicolumn{2}{c}{\{3k+1, 4k+2, 7k+3\}.} \end{array}$$

Hierbei ist

$$g(k,j) := \begin{cases} 8k+2j+4, & j = 0,\dots \lfloor (k-1)/2 \rfloor, \\ 10k-2j+3, & j = \lfloor (k-1)/2 \rfloor + 1,\dots,k-1. \end{cases}$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1, 8k+3, 8k+4\}$			$\{2, 6k+1, 6k+3\}$			$\{3, 4k+1, 4k+4\}$		
$\{4, 8k+2, 8k+6\}$			$\{5, 6k-1, 6k+4\}$			$\{6, 4k, 4k+6\}$		
$\{7, 8k+1, 8k+8\}$			$\{8, 6k-3, 6k+5\}$			$\{9, 4k-1, 4k+8\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-5, 7k+5, 8k+7\}$			$\{3k-4, 4k+5, 7k+1\}$			$\{3k-3, 3k+3, 6k\}$		
$\{3k-2, 7k+4, 8k+5\}$			$\{3k-1, 4k+3, 7k+2\}$			$\{3k, 3k+2, 6k+2\}$		

sowie

$$\{\alpha, \beta, \gamma\} := \{3k+1, 4k+2, 7k+3\}.$$

Hierdurch ist eine Partition der Zahlen $\{1,\dots,3m\} = \{1,\dots,9k+3\}$ gegeben. Die

Verteilung der Zahlen aus $\{1,\dots,9k+3\}$ ergibt sich aus

$1,\dots,3k$	S_1, S_4, S_7
$3k+1$	α
$3k+2,\dots 4k+1$	S_8
$4k+2$	β
$4k+3,\dots,6k+2$	S_5, S_9
$6k+3,\dots,7k+2$	S_6
$7k+3$	γ
$7k+4,\dots,8k+3$	S_2
$8k+4,\dots,9k+3$	S_9

Wie man leicht nachrechnet, sind alle angegebenen Tripel Differenzentripel modulo $n = 18k+7$. Damit ist nachgewiesen, dass das erste Heffterschе Differenzenproblem für $m \equiv 1 \pmod 3$ eine Lösung besitzt.

(3) Für $k \in \mathbb{N}$ und $m := 3k+2$ besteht eine Lösung $HDP_1(m)$ des ersten Heffterschen Differenzenproblems aus den $3m$ Tripeln

$$\begin{array}{ll}\{3j+1, 4k-j+3, 4k+2j+4\}, & j=0,\dots,k,\\ \{3j+2, 6k-2j+3, 6k+j+5\}, & j=0,\dots,k-1,\\ \{3j+3, 8k-j+5, g(k,j)\}, & j=0,\dots,k-1,\\ \multicolumn{2}{c}{\{3k+2, 7k+5, 8k+6\}.}\end{array}$$

Hierbei ist

$$g(k,j) := \begin{cases} 8k+2j+8, & j=0,\dots,\lfloor (k-2)/2\rfloor,\\ 10k-2j+5, & j=\lfloor (k-2)/2\rfloor+1,\dots,k-1.\end{cases}$$

Z. B. ist

$$g(1,0)=15, \quad g(2,0)=24, \quad g(2,1)=23.$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1,4k+3,4k+4\}$			$\{2,6k+3,6k+5\}$			$\{3,8k+5,8k+8\}$		
$\{4,4k+2,4k+6\}$			$\{5,6k+1,6k+6\}$			$\{6,8k+4,8k+10\}$		
$\{7,4k+1,4k+8\}$			$\{8,6k-1,6k+7\}$			$\{9,8k+3,8k+12\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-5,3k+5,6k\}$			$\{3k-4,4k+7,7k+3\}$			$\{3k-3,7k+7,8k+9\}$		
$\{3k-2,3k+4,6k+2\}$			$\{3k-1,4k+5,7k+4\}$			$\{3k,7k+6,8k+7\}$		
$\{3k+1,3k+3,6k+4\}$								

sowie

$$\{\alpha,\beta,\gamma\} := \{3k+2, 7k+5, 8k+6\}.$$

Hierdurch ist eine Partition der Zahlen $\{1,\ldots,3m\} = \{1,\ldots,9k+7\}$ gegeben. Die Verteilung der Zahlen aus $\{1,\ldots,9k+7\}$ ergibt sich aus

$$\begin{array}{c|c} 1,\ldots,3k+1 & S_1, S_4, S_7 \\ 3k+2 & \alpha \\ 3k+3,\ldots,4k+3 & S_2 \\ 4k+4,\ldots,6k+4 & S_3, S_5 \\ 6k+5,\ldots,7k+4 & S_6 \\ 7k+5 & \beta \\ 7k+6,\ldots,8k+5 & S_8 \\ 8k+6 & \gamma \\ 8k+7,\ldots,9k+6 & S_9 \end{array}$$

Da man wieder leicht nachrechnet, dass die angegebenen Tripel Differenztripel modulo $n = 18k+13$ sind, ist die Behauptung für $m \equiv 2 \pmod 3$ bewiesen.
Insgesamt ist der Satz vollständig bewiesen. □

Satz 7.13 *Für alle $m \in \mathbb{N}$ mit $m \geq 2$ existiert eine Lösung des zweiten Heffterschen Differenzenproblems, also eine Partition $HDP_2(m)$ von $\{1,\ldots,3m+1\} \setminus \{2m+1\}$ in m Differenzentripel modulo n, wobei $n := 6m+3$.*

Beweis: Der Beweis erfolgt durch Angabe der entsprechenden Partitionen.
(1) Sei $m \equiv 0 \pmod 3$, also $m := 3k$ mit $k \in \mathbb{N}$. Dann besteht eine Lösung $HDP_2(m)$ des zweiten Heffterschen Differenzenproblems aus den $3m$ Tripeln

$$\begin{array}{ll} \{3j+1, 8k-j+1, g(k,j)\}, & j = 0,\ldots,k-1, \\ \{3j+2, 4k-j, 4k+2j+2\}, & j = 0,\ldots,k-1, \\ \{3j+3, 6k-2j-1, 6k+j+2\}, & j = 0,\ldots,k-1. \end{array}$$

Hierbei ist

$$g(k,j) := \begin{cases} 8k+2j+2, & j = 0,\ldots,\lfloor (k-1)/2 \rfloor, \\ 10k-2j-1, & j = \lfloor (k-1)/2 \rfloor + 1,\ldots,k-1. \end{cases}$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1, 8k+1, 8k+2\}$			$\{2, 4k, 4k+2\}$			$\{3, 6k-1, 6k+2\}$		
$\{4, 8k, 8k+4\}$			$\{5, 4k-1, 4k+4\}$			$\{6, 6k-3, 6k+3\}$		
$\{7, 8k-1, 8k+6\}$			$\{8, 4k-2, 4k+6\}$			$\{9, 6k-5, 6k+4\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-5, 7k+3, 8k+5\}$			$\{3k-4, 3k+2, 6k-2\}$			$\{3k-3, 4k+3, 7k\}$		
$\{3k-2, 7k+2, 8k+3\}$			$\{3k-1, 3k+1, 6k\}$			$\{3k, 4k+1, 7k+1\}$		

Hierdurch ist eine Partition von $\{1,\ldots,3m+1\} \setminus \{2m+1\} = \{1,\ldots,9k+1\} \setminus \{6k+1\}$

gegeben. Die Verteilung dieser Zahlen auf die Spalten $S_1, \ldots, S_9$ ergibt sich aus

$1, \ldots, 3k$	S_1, S_4, S_7
$3k+1, \ldots, 4k$	S_5
$4k+1, \ldots, 6k$	S_8, S_5
$6k+1$	fehlt
$6k+2, \ldots, 7k+1$	S_9
$7k+2, \ldots, 8k+1$	S_2
$8k+2, \ldots, 9k+1$	S_3

Da die angegebenen Tripel Differenzentripel modulo $n = 18k+3$ sind, ist die Behauptung für $m \equiv 0 \pmod 3$ bewiesen.
(2) Sei $m \equiv 1 \pmod 3$, also $m = 3k+1$ mit $k \in \mathbb{N}$. Für $k = 1, 2, 3$ erhalten wir die folgenden Lösungen des zweiten Heffterschen Differenzenproblems:

k	m	n	$HDP_2(m)$
1	4	27	$\{1,12,13\},\{2,5,\},\{3,8,11\},\{4,6,10\}$
2	7	45	$\{1,11,12\},\{2,17,19\},\{3,20,22\},\{4,10,14\},$ $\{5,8,13\},\{6,18,21\},\{7,9,16\}$
3	10	63	$\{1,15,16\},\{2,27,29\},\{3,25,28\},\{4,14,18\},\{5,26,31\}$ $\{6,17,23\},\{7,13,20\},\{8,11,19\},\{9,24,30\},\{10,12,22\}$

Für $k \in \mathbb{N}$ mit $k \geq 4$ und $m := 3k+1$ besteht eine Lösung $HDP_2(m)$ des zweiten Heffterschen Differenzenproblems aus den $3m$ Tripeln

$$\begin{array}{ll} \{3j+1, 4k-j+3, 4k+2j+4\}, & j = 0, \ldots, k, \\ \{3j+2, 8k-j+2, g(k,j)\}, & j = 2, \ldots, k-2, \\ \{3j+3, 6k-2j+1, 6k+j+4\}, & j = 1, \ldots, k-1, \end{array}$$

sowie

$$\{2, 8k+3, 8k+5\}, \quad \{3, 8k+1, 8k+4\}, \quad \{8k+2, 8k+7\}$$

und

$$\{3k-1, 3k+2, 6k+1\}, \quad \{3k, 7k+3, 8k+6\}.$$

Hierbei ist

$$g(k,j) := \begin{cases} 8k+2j+4, & j = 2, \ldots, \lfloor (k-1)/2 \rfloor, \\ 10k-2j+5, & j = \lfloor (k-1)/2 \rfloor + 1, \ldots, k-1. \end{cases}$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1, 4k+3, 4k+4\}$			$\{6, 6k-1, 6k+5\}$			$\{8, 8k, 8k+8\}$		
$\{4, 4k+2, 4k+6\}$			$\{9, 6k-3, 6k+6\}$			$\{11, 8k-1, 8k+10\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-14, 3k+8, 6k-6\}$			$\{3k-9, 4k+9, 7k\}$			$\{3k-7, 7k+5, 8k+11\}$		
$\{3k-11, 3k+7, 6k-4\}$			$\{3k-6, 4k+7, 7k+1\}$			$\{3k-4, 7k+4, 8k+9\}$		
$\{3k-8, 3k+6, 6k-2\}$			$\{3k-3, 4k+5, 7k+2\}$					
$\{3k-5, 3k+5, 6k\}$								
$\{3k-2, 3k+4, 6k+2\}$								
$\{3k+1, 3k+3, 6k+4\}$								

und

$$\begin{aligned}
\{\alpha_1, \beta_1, \gamma_1\} &:= \{3k-1, 3k+2, 6k+1\},\\
\{\alpha_2, \beta_2, \gamma_2\} &:= \{3k, 7k+3, 8k+6\},\\
\{\alpha_3, \beta_3, \gamma_3\} &:= \{3, 8k+1, 8k+4\},\\
\{\alpha_4, \beta_4, \gamma_4\} &:= \{2, 8k+3, 8k+5\},\\
\{\alpha_5, \beta_5, \gamma_5\} &:= \{5, 8k+2, 8k+7\}.
\end{aligned}$$

Hierdurch ist eine Partition von

$$\{1, \ldots, 3m+1\} \setminus \{2m+1\} = \{1, \ldots, 9k+4\} \setminus \{6k+3\}$$

gegeben. Die Verteilung dieser Zahlen auf die Spalten $S_1, \ldots, S_9$ ergibt sich aus

$1, \ldots, 3k+1$	$S_1, S_4, S_7, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$
$3k+2$	β_1
$3k+3, \ldots, 4k+3$	S_2
$4k+4, \ldots, 6k+4$ ohne $6k+3$	S_3, S_5, γ_1
$6k+3$	fehlt
$6k+5, \ldots, 7k+2$	S_6
$7k+3$	β_2
$7k+4, \ldots, 8k$	S_8
$8k+1, \ldots, 8k+7$	$\beta_3, \beta_4, \beta_5, \gamma_2, \gamma_3, \gamma_4, \gamma_5$
$8k+8, \ldots, 9k+4$	S_9

Da die angegebenen Tripel Differenzentripel modulo $n = 18k+9$ sind, ist die Behauptung für $m \equiv 1 \pmod 3$ bewiesen.

(3) Sei $m \equiv 2 \pmod 3$, also $m = 3k+2$ mit $k \in \mathbb{N}$. Eine Lösung $HDP_2(m)$ des zweiten Heffterschen Differenzenproblems besteht aus den $3m$ Tripeln

$$\begin{aligned}
&\{3j+1, 4k-j+3, 4k+2j+4\}, && j = 0, \ldots, k,\\
&\{3j+2, 8k-j+6, g(k,j)\}, && j = 0, \ldots, k,\\
&\{3j+3, 6k-2j+3, 6k+j+6\}, && j = 0, \ldots, k-1,
\end{aligned}$$

wobei

$$g(k,j) := \begin{cases} 8k+2j+8, & j = 0, \ldots, \lfloor (k-1)/2 \rfloor,\\ 10k-2j+7, & j = \lfloor (k-1)/2 \rfloor + 1, \ldots, k. \end{cases}$$

Als Tripel hat man also

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
$\{1, 4k+3, 4k+4\}$			$\{2, 8k+6, 8k+8\}$			$\{3, 6k+3, 6k+6\}$		
$\{4, 4k+2, 4k+6\}$			$\{5, 8k+5, 8k+10\}$			$\{6, 6k+1, 6k+7\}$		
$\{7, 4k+1, 4k+8\}$			$\{8, 8k+4, 8k+12\}$			$\{9, 6k-1, 6k+8\}$		
$\vdots$			$\vdots$			$\vdots$		
$\{3k-2, 3k+4, 6k+2\}$			$\{3k-1, 7k+7, 8k+9\}$			$\{3k, 4k+5, 7k+5\}$		
$\{3k+1, 3k+3, 6k+4\}$			$\{3k+2, 7k+6, 8k+7\}$					

Hierdurch ist eine Partition von

$$\{1,\dots,3m+1\}\setminus\{2m+1\}=\{1,\dots,9k+7\}\setminus\{6k+5\}$$

gegeben. Die Verteilung dieser Zahlen auf die Spalten $S_1,\dots,S_9$ ergibt sich aus

$1,\dots,3k+2$	S_1,S_4,S_7
$3k+3,\dots,4k+3$	S_2
$4k+4,\dots,6k+4$	S_3,S_8
$6k+5$	fehlt
$6k+6,\dots,7k+5$	S_9
$7k+6,\dots,8k+6$	S_5
$8k+7,\dots,9k+7$	S_6

Da die angegebenen Tripel Differenzentripel modulo $n=18k+15$ sind, ist die Behauptung für $m\equiv 2 \pmod 3$ bewiesen.
Insgesamt ist der Satz bewiesen. □

7.8 Steiner Quadrupel Systeme

Unter einem *Steiner Quadrupel System* $(S,\mathcal{B})$ versteht man bekanntlich (siehe Definition 7.1) eine endliche Menge S (deren Anzahl $|S|$ heißt die *Ordnung* des Quadrupel Systems) und eine Menge $\mathcal{B}$ von 4-elementigen Teilmengen von S, den sogenannten *Blocks*, mit der Eigenschaft, dass jede 3-elementige Teilmenge von S in genau einem Block enthalten ist. Ein Steiner Quadrupel System der Ordnung n wird auch mit $SQS(n)$ bezeichnet. Als Literatur nennen wir zunächst den Übersichtsartikel von C. C. Lindner, A. Rosa (1973) und C. C. Lindner, C. A. Rodger (1997, S. 145 ff.).
Wegen des ersten Teils von Satz 7.4 ist die Anzahl der Blocks in einem Steiner Quadrupel System $(S,\mathcal{B})$ durch

$$(*)\qquad |\mathcal{B}|=\binom{n}{3}\Big/\binom{4}{3}=\frac{n(n-1)(n-2)}{24}$$

gegeben. Durch den zweiten Teil von Satz 7.4 ist nachgewiesen, dass ein Paar $(S,\mathcal{B})$, bestehend aus einer n-elementigen Menge S und einer Menge $\mathcal{B}$ von 4-elementigen Teilmengen von S ein Steiner Quadrupel System ist, falls $(*)$ gilt und jede 3-elementige Teilmenge von S in mindestens einem Block $B\in\mathcal{B}$ enthalten ist. Wir sind natürlich vor allem an notwendigen und hinreichenden Bedingungen dafür interessiert, dass ein $SQS(n)$ existiert.
Beispiel: Wir überlegen uns, dass ein $SQS(2^k)$ für jedes $k\in\mathbb{N}$ existiert. Hierzu setzen wir $S:=\mathbb{Z}_2^k$, wobei $\mathbb{Z}_2^k$ die Menge aller binären Vektoren der Länge k ist, d. h. $\mathbb{Z}_2^k$ besteht aus Vektoren $a\in\{0,1\}^k$. Dann ist offenbar $|S|=2^k$. Die Summe $a+b$ zweier Elemente $a,b\in\mathbb{Z}_2^k$ ist komponentenweise erklärt, wobei jede Summe modulo 2 reduziert wird. (Z. B. ist $(0,1,0,1)+(1,1,1,1)=(1,0,1,0)$, abgekürzt $0101+1111=1010$.) Nun definieren wir

$$\mathcal{B}:=\{\{a,b,c,d\}: a,b,c,d\in\mathbb{Z}_2^k \text{ sind paarweise verschieden mit } a+b+c+d=0\}.$$

Wir überlegen uns, dass $(S, \mathcal{B})$ ein $SQS(2^k)$ ist. Hierzu ist zu zeigen, dass jedes Tripel $\{a, b, c\} \subset S$ (mit drei *verschiedenen* $a, b, c \in \mathbb{Z}_2^k$) in genau einem Block $\{a, b.c, d\} \in \mathcal{B}$ enthalten ist. Wegen $a + b + c + d = 0$ ist notwendigerweise $d := a + b + c$. Zu zeigen bleibt, dass d von a, b, c verschieden ist. Angenommen, es wäre etwa $d = a$. dann wäre $b + c = 0$ bzw. $c = b$, ein Widerspruch. Da $d = b$ und $d = c$ entsprechend behandelt werden können, ist schließlich gezeigt, dass $(S, \mathcal{B})$ ein Steiner Quadrupel System der Ordnung 2^k ist. Ist z. B. $k = 3$, so besteht S aus allen binären Vektoren der Länge 3, d. h. es ist

$$S = \{000, 001, 010, 011, 100, 101, 110, 111\}.$$

In einem Steiner Quadrupel System $(S, \mathcal{B})$ der Ordnung 8 ist $|\mathcal{B}| = 8 \cdot 7 \cdot 6/24 = 14$. Als Blöcke erhält man:

$$\begin{array}{ll}
\{000, 001, 010, 011\}, & \{100, 101, 110, 111\}, \\
\{000, 001, 100, 101\}, & \{010, 011, 110, 111\}, \\
\{000, 001, 110, 111\}, & \{010, 011, 100, 101\}, \\
\{000, 010, 100, 110\}, & \{001, 011, 101, 111\}, \\
\{000, 010, 101, 111\}, & \{001, 011, 100, 110\}, \\
\{000, 011, 100, 111\}, & \{001, 010, 101, 110\}, \\
\{000, 011, 101, 110\}, & \{001, 010, 100, 111\}.
\end{array}$$

Wir werden später sehen, dass aus der Existenz eines $SQS(n)$ die eines $SQS(2n)$ folgt. Dann kann auch unabhängig von obigen Überlegungen aus der Existenz eines $SQS(8)$ die Existenz eines $SQS(2^k)$ für alle $k \geq 3$ gefolgert werden. □

Beispiel: Wir wollen einen $SQS(10)$ konstruieren, halten uns hierbei sehr eng an C. C. Lindner, C. A. Rodger (1997, S. 147). Hierzu sei S die Menge der Kanten des vollständigen Graphen K_5 mit den fünf Ecken $1, \ldots, 5$, also

$$S := \{12, 13, 14, 15, 23, 24, 25, 34, 35, 45\},$$

wobei wieder z. B. 12 eine Kurzschreibweise für $\{1, 2\}$ ist. In Abbildung 56 veranschaulichen wir den vollständigen Graphen K_5. Als Menge $\mathcal{B}$ der Blöcke nehmen wir die Menge von vier Kanten, die einen Graphen bilden, welcher isomorph zu einem der folgenden drei Graphen ist: Dem vollständigen bipartiten Graphen $K_{1,4}$, dem Kreis C_4 mit vier Ecken und $K_2 + K_3$, der eckendisjunkten Vereinigung von K_2 und K_3. In Abbildung 57 sind diese Graphen exemplarisch dargestellt. Als Elemente von $\mathcal{B}$ erhält man dann zunächst fünf Blöcke, die zu den fünf Graphen vom Typ $K_{1,4}$ gehören (je nach dem, welche der fünf Ecken den Grad vier besitzt), nämlich

$$\{12, 13, 14, 15\},\ \{12, 23, 24, 25\},\ \{13, 23, 34, 35\},\ \{14, 24, 34, 45\},\ \{15, 25, 35, 45\}.$$

Vom Typ C_4 gibt es, wenn man sich entschieden hat, welche der fünf Ecken isoliert ist, drei verschiedene Graphen. Mit der isolierten Ecke 5 geben wir diese in Abbildung 58 an. Vom Typ C_4 gibt es daher $5 \cdot 3 = 15$ Graphen, ebenso viele Blöcke gehören zu $\mathcal{B}$. Indem man der Reihe nach die Ecken 1 bis 5 als isolierte Ecken wählt erhält man die

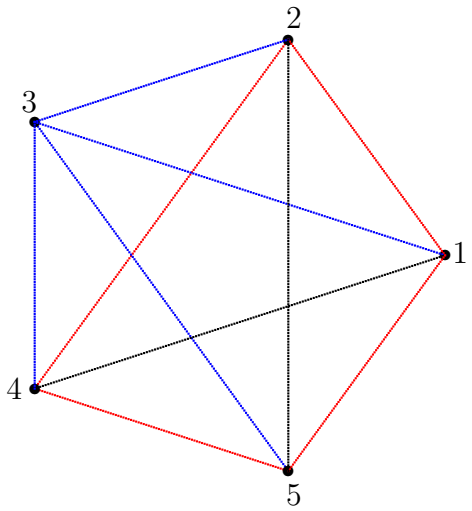
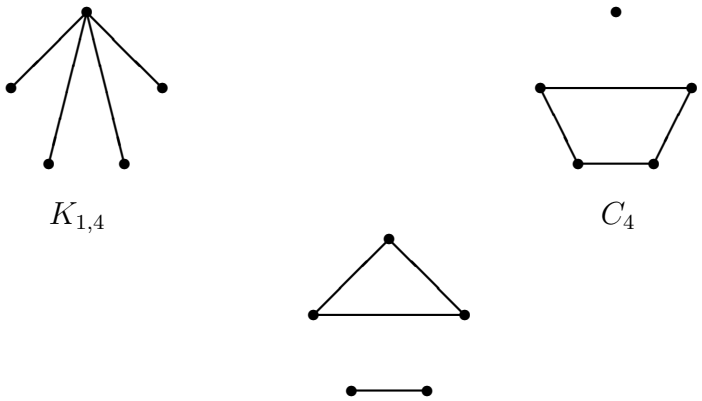


Abbildung 56: Der vollständige Graph K_5, ein $K_{1,4}$ und ein C_4



$K_{1,4}$ C_4

$K_2 + K_3$

Abbildung 57: Die Graphen $K_{1,4}$, C_4 und $K_2 + K_3$

Blöcke

$$\begin{array}{lll} \{23,24,35,45\}, & \{23,25,34,45\}, & \{24,25,34,35\}, \\ \{13,14,35,45\}, & \{13,15,34,45\}, & \{14,15,34,35\}, \\ \{12,14,25,45\}, & \{12,15,24,45\}, & \{14,15,24,25\}, \\ \{12,13,25,35\}, & \{12,15,23,35\}, & \{13,15,23,25\}, \\ \{12,13,24,34\}, & \{12,14,23,34\}, & \{13,14,23,24\}. \end{array}$$

Von den Graphen vom Typ K_2+K_3 gibt es so viele, wie man drei Ecken aus fünf Ecken auswählen kann. Ihre Anzahl ist also $\binom{5}{3} = 10$. Die zu den Graphen vom Typ K_2+K_3 gehörenden Blöcke sind

$$\{12,13,23,45\},\ \{12,14,24,35\},\ \{12,15,25,34\},\ \{13,14,34,25\},\ \{13,15,35,24\}$$

sowie

$$\{14,15,45,23\},\ \{23,24,34,15\},\ \{23,25,35,14\},\{24,25,45,13\},\ \{34,35,45,12\}.$$

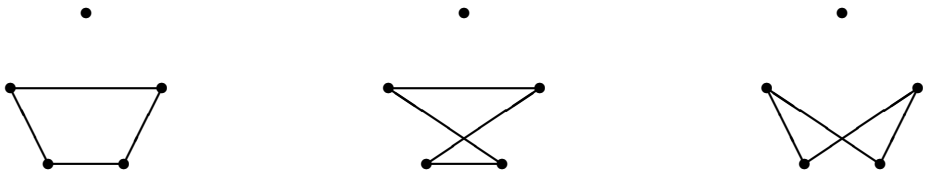


Abbildung 58: Drei Graphen vom Typ C_4

Die Anzahl der Blöcke in $\mathcal{B}$ ist $5 + 15 + 10 = 30 = 10 \cdot 9 \cdot 8/24$, also hat $\mathcal{B}$ die richtige Anzahl an Blöcken. Weshalb ist aber $(S, \mathcal{B})$ ein $STS(10)$? Hierzu ist zu zeigen: Gibt man sich ein Tripel aus S vor, also drei Kanten des vollständigen Graphen K_5, so gibt es (genau) einen Block aus $\mathcal{B}$, in dem dieses Tripel enthalten ist. Wegen des Handshaking-Lemmas ist die Summe der Grade der mit den drei Kanten inzidierenden Ecken gleich 6. Gibt es in dem von diesen drei Kanten erzeugten Untergraphen des K_5 eine Ecke v mit dem Grad 3, gibt es also eine Ecke v mit drei Nachbarn, so haben diese sämtlich den Grad 1. Fügt man daher die Kante von v zu der noch nicht inzidierenden Ecke hinzu, so hat man einen eindeutigen Untergraphen des K_5 vom Typ $K_{1,4}$, der die vorgegebenen drei Kanten enthält, und damit auch den gesuchten Block. Ist z. B. das Tripel $\{23, 34, 35\}$ (hier wäre $v = 3$) gegeben, so wird die Kante 13 hinzugefügt und man erhält den Block $\{13, 23, 34, 35\}$, in Abbildung 56 haben wir diesen Block blau gezeichnet. Im zweiten Fall gibt es in dem von den drei vorgegebenen Kanten erzeugten Untergraphen des K_5 genau zwei Ecken mit dem Grad 2, die beiden übrigen Ecken haben den Grad 1. Verbindet man diese beiden Ecken durch eine Kante, so erhält man einen Untergraphen des K_5 vom Typ C_4. Ist z. B. das Tripel $\{12, 15, 45\}$ gegeben, so fügt man die Kante 24 hinzu und erhält den, in Abbildung 56 rot gezeichneten Block $\{12, 15, 24, 45\}$. Im letzten Fall gibt es genau eine Ecke vom Grad 2, die anderen vier Ecken haben den Grad 1. Verbindet man die beiden Nachbarn der Ecke vom Grad 2 durch eine Kante, so erhalten wir eine Graphen vom Typ $K_2 + K_3$, in dem das vorgegebene Tripel enthalten ist. Ist z. B. das Tripel $\{12, 23, 45\}$ vorgegeben, so fügt man die Kante 13 hinzu und erhält den Block $\{12, 13, 23, 45\}$. □

Im folgenden Satz wird eine *notwendige* Bedingung für die Existenz eines $SQS(n)$ formuliert und bewiesen, von der H. Hanani (1960) nachwies, dass sie auch *hinreichend* für die Existenz eines $SQS(n)$ ist.

Satz 7.14 *Existiert mit $n \in \mathbb{N}$ und $n \geq 4$ ein $SQS(n)$, so ist $n \equiv 2 \pmod 6$ oder $n \equiv 4 \pmod 6$.*

Beweis: Sei $(S, \mathcal{B})$ ein $SQS(n)$. Mit einem beliebigen $p \in S$ definiere man

$$\mathcal{B}(p) := \{B \setminus \{p\} : B \in \mathcal{B}, p \in B\}.$$

Dann besteht $\mathcal{B}(p)$ aus in S enthaltenen Tripeln, die dadurch entstehen, dass aus Blöcken (Quadrupeln) in $\mathcal{B}$, die p enthalten, das Element p entfernt wird. Wir überlegen uns, dass $(S \setminus \{p\}, \mathcal{B}(p))$ ein $STS(n-1)$ ist. Denn sind x, y zwei verschiedene Elemente von $S \setminus \{p\}$, so liegt $\{p, x, y\}$ in genau einem Block $B \in \mathcal{B}$, da $(S, \mathcal{B})$ ein Steiner

Quadrupel System ist. Folglich liegt $\{x, y\}$ in genau einem Block von $\mathcal{B}(p)$. Daher ist $(S \setminus \{p\}, \mathcal{B}(p))$ ein $STS(n-1)$. Wegen Satz 7.5 ist $n-1 \equiv 1 \pmod 6$ oder $n-1 \equiv 3 \pmod 6$ bzw. $n \equiv 2 \pmod 6$ oder $n \equiv 4 \pmod 6$. Damit ist der Satz bewiesen. □

Bemerkung: In einer Bemerkung im Anschluss an Satz 7.4 haben wir nachgewiesen:

- Sei $(S, \mathcal{B})$ ein Steiner System $S(t, k, n)$. Dann ist $\binom{n-i}{t-i} / \binom{k-i}{t-i}$, $i = 0, \ldots, t-1$, eine natürliche Zahl.

Wir wollen uns überlegen, dass wir die Aussage von Satz 7.14 auch hieraus ableiten können. Für $t = 3$ und $k = 4$ erhalten wir:

- Ist $(S, \mathcal{B})$ ein $SQS(n)$, so sind $n(n-1)(n-2)/24$, $(n-1)(n-2)/6$ und $(n-2)/2$ ganzzahlig.

Existiert also ein $SQS(n)$, so ist notwendigerweise $(n-2)/2$ ganzzahlig und daher $n = 2m$ mit einem $m \in \mathbb{N}$ gerade. Da

$$\frac{(n-1)(n-2)}{6} = \frac{(2m-1)(m-1)}{3}$$

ganzzahlig ist, ist $m-1$ oder $2m-1$ durch 3 teilbar. Im ersten Fall ist $m-1 = 3k$ mit $k \in \mathbb{N}$ und daher $n = 2m = 2 + 6k$ bzw. $n \equiv 2 \pmod 6$. Nehmen wir also an, $m-1$ sei nicht durch 3 teilbar. Dann muss $2m-1$ durch 3 teilbar sein, es ist also $n-1 = 2m-1 = 3k$ mit $k \in \mathbb{N}$. Da n gerade ist, ist k ungerade. Dann ist aber

$$n = 1 + 3k = 4 + 6\left(\frac{k-1}{2}\right)$$

und folglich $n \equiv 4 \pmod 6$. Damit haben wir einen zweiten Beweis für Satz 7.14 erhalten. Andererseits zeigt der oben angegebene Beweis, dass man zu jedem $SQS(n)$ ein $STS(n-1)$, ein sogenanntes *abgeleitetes* (*derived*) Steiner Tripel System angeben kann. Eine bisher unbewiesene Vermutung besagt, dass *jedes* $STS(n)$ sich ableiten lässt aus einem $SQS(n+1)$. Für $n \leq 15$ ist diese Vermutung richtig, siehe I. Diener, E. Schmitt, H. L. de Vries (1985). □

Wir erwähnten schon, dass die in Satz 7.14 angegebene notwendige Bedingung für die Existenz eines $SQS(n)$ auch hinreichend ist. Ein Beweis dieser Tatsache liegt außerhalb unserer Möglichkeiten. Wir verweisen lediglich auf die Originalarbeit H. Hanani (1960) sowie auf C. C. Lindner, C. A. Rodger (1997).

Aber ein Ergebnis, das wir oben schon erwähnt hatten, wollen wir doch noch beweisen, weil beim Beweis ein hübsches Konzept der Graphentheorie eine Rolle spielt, nämlich die 1-*Faktorisierung* eines Graphen. Bei dem Beweis der Aussage, dass aus der Existenz eines $SQS(n)$ die eines $SQS(2n)$ folgt, halten wir uns im wesentlichen an C. C. Lindner, C. A. Rodger (1997, S. 153) und P. J. Cameron (1994, S. 121). Zunächst machen wir einige Vorbemerkungen.

Angenommen, n Mannschaften sind Teilnehmer an einem Turnier, bei dem jede Mannschaft gegen jede andere anzutreten hat. Insgesamt gibt es bei n Mannschaften, wenn jede Mannschaft gegen jede andere anzutreten hat, $\binom{n}{2} = n(n-1)/2$ Spiele. Ist n

gerade, so können an jedem Spieltag $n/2$ Spiele (zwischen jeweils zwei Mannschaften) stattfinden, sodass man für das Turnier der n Mannschaften mindestens $n-1$ Spieltage benötigt. Ist n ungerade, so können nur $(n-1)/2$ Spiele pro Spieltag stattfinden, eine Mannschaft hat jeweils einen Ruhetag und man benötigt mindestens n Spieltage. Jetzt müssen wir uns aber noch überlegen, dass es einen zulässigen Turnierplan (engl.: tournament schedule) mit der minimalen Anzahl an Spieltagen gibt. Zunächst betrachten wir den Fall, dass n ungerade ist. Wir zeichnen ein regelmäßiges n-gon in die Ebene, wobei wir die Ecken mit $0, \ldots, n-1$ nummerieren, siehe Abbildung 59 für $n = 9$. Jede

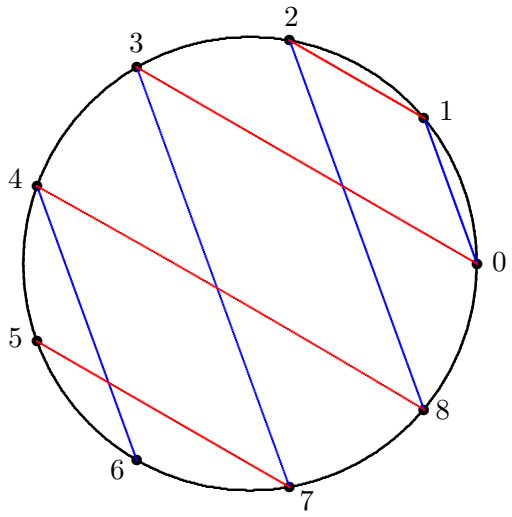


Abbildung 59: Ein Turnierplan bei $n = 9$ Mannschaften

Ecke repräsentiert eine Mannschaft. Zu jeder Kante des n-gons gibt es $(n-3)/2$ parallele Diagonalen. Gehen wir z. B. für $n = 9$ aus von der Kante 01, so hat man die hierzu parallelen Diagonalen 28, 37 und 46. Dies ergibt schon den blau eingetragenen Spieltag, an dem die Mannschaft 5 einen Ruhetag hat. Rot eingetragen sind in Abbildung 59 noch die Ansetzungen eines Spieltages, an dem die Mannschaften 1 und 2 gegen einander spielen. An diesem Spieltag hat die Mannschaft 6 einen Ruhetag. So kann man fortfahren und man erhält einen zulässigen Spielplan mit ebenso vielen Spieltagen, wie es Kanten im regelmäßigen n-gon gibt, also n. Ist dagegen n gerade, so denke man sich etwa die n-te Mannschaft zunächst aus dem Turnier entfernt. Für die verbliebenen $n-1$ Mannschaften gibt es einen Turnierplan mit $n-1$ Spieltagen. Man erhält einen zulässigen Turnierplan für alle n Mannschaften, indem an jedem der $n-1$ Spieltage die zunächst nicht berücksichtigte n-te Mannschaft gegen die Mannschaft antritt, die eigentlich (bei $n-1$ Mannschaften) einen Ruhetag hat. Damit ist für gerades n gezeigt, dass es einen zulässigen Turnierplan mit $n-1$ Spieltagen gibt.

Bei geradem n ist ein Turnierplan für n Mannschaften nichts anderes als eine 1-Faktorisierung des vollständigen Graphen K_n, wobei die Ansetzungen der $n-1$ Spieltage jeweils einem 1-Faktor entsprechen. Genauer definieren wir:

Definition 7.15 Sei $G = (V, E)$ ein Graph. Ein 1-*Faktor* von G ist eine Menge $F \subset E$ von Kanten mit der Eigenschaft, dass jede Ecke $v \in V$ mit genau einer Kante aus F inzidiert. Eine 1-*Faktorisierung* des Graphen G ist eine Partition (disjunkte Vereinigung)

der Menge E der Kanten von G in 1-Faktoren. Unter einer 1-Faktorisierung einer n-elementigen Menge verstehen wir eine 1-Faktorisierung des vollständigen Graphen K_n mit den Elementen der n-elementigen Menge als Ecken.

Beispiel: In Abbildung 60 geben wir den vollständigen Graphen K_6 und drei verschiedene 1-Faktoren des K_6 an. Die drei angegebenen 1-Faktoren haben keine ge-

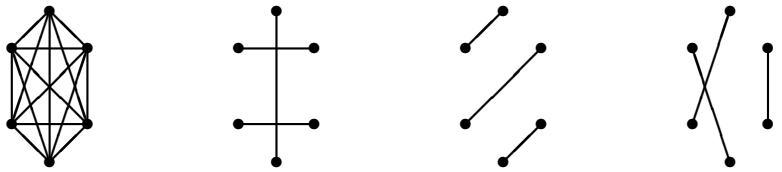


Abbildung 60: Der K_6 und drei verschiedene 1-Faktoren

meinsame Kante und es stellt sich naheliegenderweise die Frage, ob diese sich zu einer 1-Faktorisierung des K_6 erweitern lassen. Entfernt man aus dem K_6 die in diesen drei 1-Faktoren enthaltenen Kanten, so erhält man den in Abbildung 61 angegebenen Graphen. Rechts daneben haben wir zwei 1-Faktoren dieses Graphen angegeben, womit

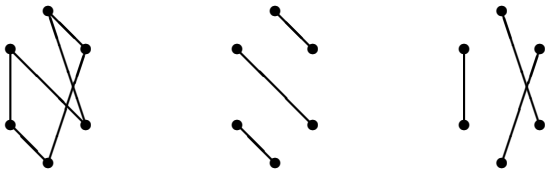


Abbildung 61: Zwei weitere 1-Faktoren des K_6

wir insgesamt eine 1-Faktorisierung des K_6 erhalten haben. □

Satz 7.16 *Existiert ein $SQS(n)$ für ein $n \geq 4$, so existiert auch ein $SQS(2n)$.*

Beweis: Sei $(S, \mathcal{B})$ ein $SQS(n)$. Wegen Satz 7.14 ist $n \equiv 2 \pmod 6$ oder $n \equiv 4 \pmod 6$, insbesondere ist n gerade. Daher existiert eine 1-Faktorisierung $\{F_1, \ldots, F_{n-1}\}$ des vollständigen Graphen K_n mit den n Elementen von S als Ecken. Jeder 1-Faktor F_i, $i = 1, \ldots, n-1$, besteht also aus $n/2$ Kanten $\{a, b\}$, wobei a, b zwei verschiedene Elemente von S sind. Wir setzen $S' := S \times \{1, 2\}$. Dann besteht S' aus zwei Kopien der Menge S und es ist folglich $|S'| = 2n$. Nun definieren wir die Menge $\mathcal{B}'$ von 4-elementigen Teilmengen der Menge S' auf die folgende Weise:

(a) Es ist $\mathcal{B} \times \{1, 2\} \subset \mathcal{B}'$, d. h. für jedes Quadrupel $B = \{a, b, c, d\} \in \mathcal{B}$ ist $\{(a, i), (b, i), (c, i), (d, i)\} \in \mathcal{B}'$, $i \in \{1, 2\}$.

(b) Es ist $\{(a, 1), (b, 1), (c, 2), (d, 2)\} \in \mathcal{B}'$, falls $\{a, b\} \in F_i$ und $\{c, d\} \in F_i$, $i = 1, \ldots, n-1$.

Um zu zeigen, dass $(S', \mathcal{B}')$ ein $SQS(2n)$ ist, ist nachzuweisen, dass jedes Tripel aus $S' = S \times \{1, 2\}$ in genau einem Quadrupel aus $\mathcal{B}'$ enthalten ist. Wegen des zweiten Teils von Satz 7.4 (mit $t = 3$, $k = 4$) genügt es zu zeigen, dass

(i) Es ist $|\mathcal{B}'| \leq 2n(2n-1)(2n-2)/24$,

(ii) Jedes Tripel aus S' ist in mindestens einem Quadrupel aus $\mathcal{B}'$ enthalten.

Zum Nachweis von (i) beachten wir, dass

$$\begin{aligned} |\mathcal{B}'| &\leq 2|\mathcal{B}| + (n-1)\frac{n^2}{4} \\ &= \frac{n(n-1)(n-2)}{12} + \frac{(n-1)n^2}{4} \\ &= \frac{2n(2n-1)(2n-2)}{24}, \end{aligned}$$

wobei wir bei der Anwendung von (a) ausgenutzt haben, dass

$$|\mathcal{B}| = \frac{n(n-1)(n-2)}{24},$$

und bei der Anwendung von (b), dass in jedem der 1-Faktoren F_i, $i = 1, \ldots, n-1$, unabhängig voneinander zweimal eine Kante aus $n/2$ Kanten gewählt werden. Damit ist (i) nachgewiesen. Zum Nachweis von (ii) geben wir uns ein Tripel aus S' vor und unterscheiden die folgenden Fälle:

1. Alle drei Elemente des vorgegebenen Tripels aus S' haben dieselbe zweite Komponente, es sei also durch $\{(x, i), (y, i), (z, i)\}$ mit $i \in \{1, 2\}$ gegeben, wobei $\{x, y, z\} \subset S$. Da $(S, \mathcal{B})$ ein $SQS(n)$ ist, existiert genau ein $B = \{w, x, y, z\} \in \mathcal{B}$. Daher ist $\{(w, i), (x, i), (y, i), (z, i)\} \in \mathcal{B}'$ ein das gegebene Tripel aus S' enthaltendes Quadrupel aus $\mathcal{B}'$.

2. Wenn 1. nicht erfüllt ist, dann sind genau zwei der zweiten Komponenten gleich, d. h. das Tripel ist durch $\{(x, 1), (y, 1)), (z, 2)\}$ oder durch $\{(x, 2), (y, 2), (z, 1)\}$ gegeben. Da $\{F_1, \ldots, F_{n-1}\}$ eine 1-Faktorisierung des K_n mit den n Elementen von S als Ecken ist, ist die Kante $\{x, y\}$ in genau einem 1-Faktor F_i enthalten, d. h. es gibt genau ein $i \in \{1, \ldots, n-1\}$ mit $\{x, y\} \subset F_i$. Hier haben wir nur ausgenutzt, dass die Elemente einer 1-Faktorisierung eine disjunkte Vereinigung der Kanten des K_n bilden. Da weiter F_i ein 1-Faktor des K_n (mit den n Elementen von S als Ecken) ist, gibt es zu $z \in S$ eine Kante $\{w, z\} \in F_i$. Dann sind $\{(x, 1), (y, 1), (z, 2), (w, 2)\} \in \mathcal{B}'$ bzw. $\{(x, 2), (y, 2), (z, 1), (w, 1)\} \in \mathcal{B}'$ Quadrupel aus $\mathcal{B}'$, welche die Tripel $\{(x, 1), (y, 1), (z, 2)\}$ bzw. $\{(x, 2), (y, 2), (z, 1)\}$ enthalten.

Damit ist der Satz bewiesen. □

Beispiel: Durch $(S, \mathcal{B}) := (\{1, 2, 3, 4\}, \{\{1, 2, 3, 4\}\})$ ist ein $SQS(4)$ gegeben. Wir wollen mit Hilfe der im Beweis von Satz 7.16 angegebenen Methode ein $SQS(8)$ konstruieren. In Abbildung 62 geben wir den K_4 und eine 1-Faktorisierung des K_4 durch

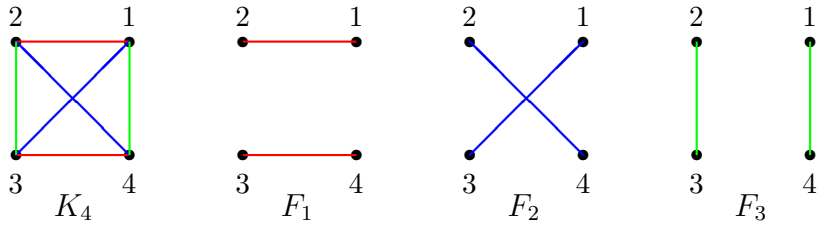


Abbildung 62: Der K_4 und eine 1-Faktorisierung

$\{F_1, F_2, F_3\}$ mit

$$F_1 := \{12, 34\}, \qquad F_2 := \{13, 24\}, \qquad F_3 := \{14, 23\}$$

an. Hier steht einmal wieder z. B. 12 für $\{1, 2\}$. Die Grundmenge S' des zu konstruierenden $SQS(8)$ ist

$$S' := \{1, 2, 3, 4\} \times \{1, 2\} = \{(1, 1), (2, 1), (3, 1), (4, 1), (1, 2), (2, 2), (3, 2), (4, 2)\}.$$

Als Menge $\mathcal{B}'$ der Blöcke erhalten wir:

(a) $\mathcal{B} \times \{1, 2\}$ gehört zu $\mathcal{B}'$ bzw.

$$\{(1, 1), (2, 1), (3, 1), (4, 1)\}, \quad \{(1, 2), (2, 2), (3, 2), (4, 2)\}$$

sind Blöcke in $\mathcal{B}'$.

(b) Sind $\{a, b\}, \{c, d\} \in F_i$, $i = 1, 2, 3$, so gehören $\{(a, 1), (b, 1), (c, 2), (d, 2)\}$ zu $\mathcal{B}'$ bzw.

$i = 1$	$i = 2$
$\{(1, 1), (2, 1), (1, 2), (2, 2)\}$	$\{(1, 1), (3, 1), (1, 2), (3, 2)\}$
$\{(1, 1), (2, 1), (3, 2), (4, 2)\}$	$\{(1, 1), (3, 1), (2, 2), (4, 2)\}$
$\{(3, 1), (4, 1), (1, 2), (2, 2)\}$	$\{(2, 1), (4, 1), (1, 2), (3, 2)\}$
$\{(3, 1), (4, 1), (3, 2), (4, 2)\}$	$\{(2, 1), (4, 1), (2, 2), (4, 2)\}$

sowie

$i = 3$
$\{(1, 1), (4, 1), (1, 2), (4, 2)\}$
$\{(1, 1), (4, 1), (2, 2), (3, 2)\}$
$\{(2, 1), (3, 1), (1, 2), (4, 2)\}$
$\{(2, 1), (4, 1), (2, 2), (4, 2)\}$.

Eine Umbenennung der Elemente von S' durch

$(1, 1)$	$(2, 1)$	$(3, 1)$	$(4, 1)$	$(1, 2)$	$(2, 2)$	$(3, 2)$	$(4, 2)$
1	2	3	4	5	6	7	8

ergibt die Menge der Blöcke $\mathcal{B}'$ bestehend aus

$$1234, \qquad 5678$$

sowie

1256	1357	1458
1278	1368	1467
3456	2457	2358
3478	2468	2367.

Bei C. C. Lindner, A. Rosa (1978, S. 150) wird die folgende Menge von 14 Quadrupeln als Menge $\mathcal{B}$ von Blöcken eines $SQS(8)$ angegeben:

$$(*) \qquad \mathcal{B}: \begin{array}{ll} 1248 & 3567 \\ 2358 & 1467 \\ 3468 & 1257 \\ 4578 & 1236 \\ 1568 & 2347 \\ 2678 & 1345 \\ 1378 & 2456. \end{array}$$

Eine weitere Menge $\mathcal{B}$ von möglichen Blöcken eines $SQS(8)$ hatten wir schon in einem Beispiel am Schluss des Unterabschnitts 7.2 angegeben. Nun ist ein $SQS(8)$ bis auf Isomorphie, d. h. bis auf eine Umbenennung der Elemente der Grundmenge S, eindeutig bestimmt. Hierzu findet man bei C. C. Lindner, A. Rosa (1978, S. 151) die folgende Bemerkung:

- *To show the uniqueness of the $SQS(8)$ it is sufficient to observe that it is completely determined by any of its derived triple systems.*

Das wollen wir uns genauer überlegen. Sei $(S, \mathcal{B})$ ein $SQS(8)$, wobei $S := \{1, \ldots, 8\}$. Man wähle ein beliebiges $p \in S$ und bilde das abgeleitete $STS(7)$, welches durch $(S \setminus \{p\}, \mathcal{B}(p))$ mit

$$\mathcal{B}(p) := \{B \setminus \{p\} : B \in \mathcal{B}, p \in B\}$$

gegeben ist. Elemente von $\mathcal{B}(p)$ sind also Tripel, die aus p enthaltenden Quadrupeln aus $\mathcal{B}$ durch Streichen von p entstehen. Jetzt stellt sich die Frage, wie man aus der Kenntnis von p und $\mathcal{B}(p)$ die Menge $\mathcal{B}$ der Blöcke eines $SQS(8)$ zurückgewinnen kann. Hierbei ist $|\mathcal{B}| = 14$ und $|\mathcal{B}(p)| = 7$. Sieben Blöcke von $\mathcal{B}$ erhalten wir, indem wir zu jedem Tripel aus $\mathcal{B}(p)$ die Zahl p hinzufügen. Diese Menge nennen wir $\mathcal{B}(p) \cup \{p\}$, natürlich ist

$$\mathcal{B}(p) \cup \{p\} = \{B \in \mathcal{B} : p \in B\}$$

genau die Menge aller Blöcke aus $\mathcal{B}$, die p enthalten. Aus einem Quadrupel $B \in \mathcal{B}$ gewinnen wir das Quadrupel $B_c := \{1, \ldots, 8\} \setminus B$. In B_c sind also genau die Zahlen aus S enthalten, die in B *nicht* vorkommen. Die Menge $(\mathcal{B}(p) \cup \{p\})_c$ ist sinngemäß zu verstehen. Wir machen uns diese Operationen durch ein Beispiel klar und betrachten die durch $(*)$ gegebene Blockmenge eines $SQS(8)$. Wählen wir z. B. $p = 4$, so erhalten

wir

$$\begin{array}{rlrlrl}
 & 128 & & 1248 & & 3567 \\
 & 368 & & 3468 & & 1257 \\
 & 578 & & 4578 & & 1236 \\
\mathcal{B}(4): & 167 & \mathcal{B}(4)\cup\{4\}: & 1467 & (\mathcal{B}(4)\cup\{4\})_c: & 2358 \\
 & 237 & & 2347 & & 1568 \\
 & 135 & & 1345 & & 2678 \\
 & 256, & & 2456, & & 1378.
\end{array}$$

Und zumindest in diesem Fall stellt man leicht fest, dass $(\mathcal{B}(4)\cup\{4\})\cup(\mathcal{B}(4)\cup\{4\})_c = \mathcal{B}$, dass also das $SQS(8)$ vollständig bestimmt ist durch ein abgeleitetes Tripel System $STS(7)$. Wir wollen unsere Vermutung auch noch für $p = 8$ nachprüfen und erhalten

$$\begin{array}{rlrlrl}
 & 124 & & 1248 & & 3567 \\
 & 235 & & 2358 & & 1467 \\
 & 346 & & 3468 & & 1257 \\
\mathcal{B}(8): & 457 & \mathcal{B}(8)\cup\{8\}: & 4578 & (\mathcal{B}(8)\cup\{8\})_c: & 1236 \\
 & 156 & & 1568 & & 2347 \\
 & 267 & & 2678 & & 1345 \\
 & 137, & & 1378, & & 2456.
\end{array}$$

Auch hier stimmt die Vermutung. Daher beweisen wir jetzt die folgende Aussage.

- *Sei* $S := \{1,\dots,8\}$ *und* $(S,\mathcal{B})$ *ein* $SQS(8)$. *Sei* $p \in S$ *beliebig und*

$$\mathcal{B}(p) := \{B \setminus \{p\} : B \in \mathcal{B}, p \in B\}.$$

  Dann ist

$$\mathcal{B} = (\mathcal{B}(p)\cup\{p\})\cup(\mathcal{B}(p)\cup\{p\})_c,$$

  wobei

$$(\mathcal{B}(p)\cup\{p\})_c := \{S \setminus C : C \in \mathcal{B}(p)\cup\{p\}\}.$$

Denn: Völlig klar ist, dass $\mathcal{B}(p)\cup\{p\} \subset \mathcal{B}$. Es genügt daher nachzuweisen, dass $(\mathcal{B}(p)\cup\{p\})_c \subset \mathcal{B}$. Sei hierzu $B \in \mathcal{B}(p)\cup\{p\}$ gegeben und $B_c := S \setminus B$. Wir haben zu zeigen, dass $B_c \in \mathcal{B}$. Hierzu nutzen wir aus:

(a) $(\{1,\dots,8\},\mathcal{B})$ ist ein $SQS(8)$, d. h. jede 3-elementige Teilmenge von $\{1,\dots,8\}$ ist in genau einem Quadrupel $B \in \mathcal{B}$ enthalten.

(b) $(\{1,\dots,8\}\setminus\{p\},\mathcal{B}(p))$ ist ein $STS(7)$, d. h. jede 2-elementige Teilmenge von $\{1,\dots,8\}\setminus\{p\}$ ist in genau einem Tripel enthalten, welches aus einem p enthaltenden Quadrupel $B \in \mathcal{B}$ durch Weglassen des Elementes p entsteht.

(c) Der Durchschnitt zweier Blöcke eines $STS(7)$ besteht aus genau einem Element.

Sei $B = abcp \in \mathcal{B}(p) \cup \{p\}$ und $B_c = wxyz$. Zum Tripel xyz gibt es genau ein xyz enthaltendes Quadrupel $B_0 = w_0xyz \in \mathcal{B}$. Hier ist $w_0 \neq p$, denn andernfalls wäre $B_0 \in \mathcal{B}(p) \cup \{p\}$ bzw. $xyz \in \mathcal{B}(p)$ und $abc \cap xyz = \emptyset$, ein Widerspruch zu (c). w_0 ist aber auch von a, b und c verschieden. Denn angenommen, es wäre z. B. $w_0 = a$. Dann wären $bcp \in \mathcal{B}(a)$ und $xyz \in \mathcal{B}(a)$, was wiederum wegen $bcp \cap xyz = \emptyset$ ein Widerspruch zu (e) ist. Da w_0 von a, b, c, p und natürlich x, y, z verschieden ist, ist $w_0 = w$ und $B_c = wxyz \in \mathcal{B}$. Damit ist nachgewiesen, dass ein $SQS(8)$ vollständig bestimmt (completely determined) durch eines seiner abgeleiteten Tripel Systeme $SQS(7)$ ist. Da diese im wesentlichen eindeutig sind, gilt dies auch für die Quadrupel Systeme $SQS(8)$. Da auch ein $STS(9)$ bis auf Isomorphie eindeutig bestimmt ist (das haben wir *nicht* bewiesen), erhält man auf gleichem Wege wie oben auch die Eindeutigkeit eines $SQS(10)$. □

Bemerkung: Die Konstruktion eines $SQS(2n)$ aus einem $SQS(n)$ im Beweis von Satz 7.16 kann verallgemeinert werden, siehe z. B. C. C. Lindner, A. Rosa (1978) und K. T. Phelps (1976). Sind nämlich $(S_1, \mathcal{B}_1)$ und $(S_2, \mathcal{B}_2)$ zwei disjunkte $SQS(n)$ (also $|S_1| = |S_2| = n$ und $S_1 \cap S_2 = \emptyset$), so erhält man auf die folgende Weise durch $(S, \mathcal{B})$ ein $SQS(2n)$:

- Seien $\{F_1^{(1)}, \ldots, F_{n-1}^{(1)}\}$ und $\{F_1^{(2)}, \ldots, F_{n-1}^{(2)}\}$ jeweils eine 1-Faktorisierung des vollständigen Graphen K_n mit S_1 bzw. S_2 als Menge der Ecken. Sei ferner

  $$\pi \colon \{1, \ldots, n-1\} \longrightarrow \{1, \ldots, n-1\}$$

  eine Permutation. Sei $S := S_1 \cup S_2$ und $\mathcal{B} \subset S$ die folgende Menge von Quadrupeln:

  (a) Es ist $\mathcal{B}_1 \cup \mathcal{B}_2 \subset \mathcal{B}$, alle Quadrupel aus $\mathcal{B}_1$ und $\mathcal{B}_2$ gehören also zu $\mathcal{B}$,

  (b) Es ist $\{a, b, c, d\} \in \mathcal{B}$, falls $\{a, b\} \in F_i^{(1)}$, $\{c, d\} \in F_{\pi(i)}^{(2)}$, $i = 1, \ldots, n-1$.

Der Beweis stimmt natürlich praktisch wörtlich mit dem von Satz 7.16 überein. □

Die in der Bemerkung gemachte Beobachtung wird von K. T. Phelps (1976, Theorem 2) (siehe auch C. C. Lindner, A. Rosa (1978, Theorem 7.1) und C. J. Colbourn, A. Rosa (1999, Theorem 22.18)) zum Beweis des folgenden Satzes benutzt. C. C. Lindner, A. Rosa (1978) schreiben hierzu:

- The proof of Phelp's theorem that an $STS(2n+1)$ with a derived subsystem of order n is itself derived, though elementary, is elegant enough to deserve a proof.

Satz 7.17 *Sei $(S, \mathcal{B})$ ein $STS(2n+1)$ und $(T, \mathcal{C})$ mit $T \subset S$ und $\mathcal{C} \subset \mathcal{B}$ ein $STS(n)$, welches aus einem $SQS(n+1)$ abgeleitet ist. Dann ist $(S, \mathcal{B})$ selbst ein (aus einem $SQS(2n+2)$) abgeleitetes Tripel System.*

Beweis: Die Idee des Beweises ist einfach. Da $(T, \mathcal{C})$ ein abgeleitetes $STS(n)$ ist, existiert ein $SQS(n+1)$ mit $(T, \mathcal{C})$ als abgeleitetem Tripel System. Mit Hilfe der in obiger Bemerkung geschilderten Konstruktion bilde man (auf geeignete Weise) ein $SQS(2n+2)$ und zeige, dass $(S, \mathcal{B})$ hiervon ein abgeleitetes $STS(2n+1)$ ist.

Zunächst überlegen wir uns:

- Jedes Tripel aus $\mathcal{B}$ ist entweder vollständig in T enthalten oder enthält genau ein Element aus T.

Denn: $\mathcal{C}$ besteht genau aus den Tripeln aus $\mathcal{B}$, die vollständig in T enthalten sind. Da $(T, \mathcal{C})$ ein $STS(n)$ und $(S, \mathcal{B})$ ein $STS(2n+1)$, ist

$$|\mathcal{C}| = \frac{n(n-1)}{6}, \qquad |\mathcal{B}| = \frac{(2n+1)2n}{6}$$

und folglich

$$|\mathcal{B} \setminus \mathcal{C}| = \frac{(2n+1)2n}{6} - \frac{n(n-1)}{6} = \frac{n(n+1)}{2}$$

die Anzahl der Tripel aus $\mathcal{B}$, die nicht vollständig in T enthalten sind. Andererseits ist die Anzahl der Tripel aus $\mathcal{B}$, die genau ein Element aus der n-elementigen Menge T (und zwei weitere aus der $(n+1)$-elementigen Menge $S \setminus T$) enthalten, ebenfalls gleich $n(n+1)/2$. Damit ist die Zwischenbehauptung bewiesen.

Im weiteren nehmen wir o. B. d. A. an, dass

$$T := \{1, \ldots, n\}, \qquad S := \{1, \ldots, 2n+1\}.$$

Eine 1-Faktorisierung $\{B_1, \ldots, B_n\}$ der $(n+1)$-elementigen Menge $S \setminus T$ (siehe Definition 7.15) erhalten wir, indem wir definieren:

$$B_i := \{\{x, y\} \subset S \setminus T : \{i, x, y\} \in \mathcal{B}\}, \qquad i \in T.$$

Denn B_i, $i = 1, \ldots, n$, ist ein 1-Faktor des vollständigen Graphen K_{n+1} mit den Elementen von $S \setminus T$ als Ecken, da es für jedes $x \in S \setminus T$ genau einen Block aus $\mathcal{B}$ gibt, der das Paar $\{i, x\}$ enthält. Weiter ist $\{B_1, \ldots, B_n\}$ eine 1-Faktorisierung des K_{n+1}, da offensichtlich durch diese Mengen eine Partition der Kanten des K_{n+1} bilden.

Da $(T, \mathcal{C})$ ein abgeleitetes $STS(n)$ ist, gibt es ein $p \notin S$ und eine Menge

$$\mathcal{C}^* \subset T^* := T \cup \{p\}$$

von Quadrupeln derart, dass $(T^*, \mathcal{C}^*)$ ein $SQS(n+1)$ mit

$$(T, \mathcal{C}) = (T^* \setminus \{p\}, \mathcal{C}^*(p))$$

ist, wobei

$$\mathcal{C}^*(p) := \{C^* \setminus \{p\} : C^* \in \mathcal{C}^*, p \in C^*\}.$$

Sei $(S \setminus T, \mathcal{D})$ ein beliebiges $SQS(n+1)$ (ein solches existiert!) und $\{A_1, \ldots, A_n\}$ eine beliebige 1-Faktorisierung der $(n+1)$-elementigen Menge $T^* = T \cup \{p\}$, wobei der 1-Faktor A_i die "Kante" $\{i, p\}$ enthalte, $i = 1, \ldots, n$. Die Existenz einer solchen 1-Faktorisierung ist klar, denn wegen der vorausgesetzten Existenz eines $STS(n)$ ist $n+1$ gerade. Weiter sei

$$\pi \colon \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$$

die Identität. Nun wenden wir die Aussage in obiger Bemerkung mit

$$(S_1, \mathcal{B}_1) := (T \cup \{p\}, \mathcal{C}^*), \qquad (S_2, \mathcal{B}_2) := (S \setminus T, \mathcal{D})$$

an. Wegen $p \notin S$ ist $S_1 \cap S_2 = \emptyset$, ferner ist $S_1 \cup S_2 = S \cup \{p\}$. Weiter ist $\{A_1, \ldots, A_n\}$ eine 1-Faktorisierung von $S_1 = T \cup \{p\}$ und $\{B_1, \ldots, B_n\}$ eine 1-Faktorisierung von $S_2 = S \setminus T$. Wir erhalten nach obiger Bemerkung ein Quadrupel System $(S \cup \{p\}, \mathcal{E})$ der Ordnung $2(n+1)$, wobei die Menge $\mathcal{E}$ der Quadrupel gegeben ist durch:

(a) Alle Quadrupel aus $\mathcal{C}^*$ und $\mathcal{D}$ gehören zu $\mathcal{E}$,

(b) Ein Quadrupel $\{a, b, c, d\} \in S \cup \{p\}$ liegt in $\mathcal{E}$, falls $\{a, b\} \in A_i$ und $\{c, d\} \in B_i$, $i = 1, \ldots, n$.

Dann ist $(S, \mathcal{B})$ ein aus dem Quadrupel System $(S \cup \{p\}, \mathcal{E})$ abgeleitetes Tripel System! Hierzu genügt es zu zeigen, dass $\mathcal{B} = \mathcal{E}(p)$, wobei

$$\mathcal{E}(p) = \{B \setminus \{p\} : B \in \mathcal{E}, p \in B\}.$$

Die Blöcke aus $\mathcal{D}$ sind in $S \setminus T$ enthalten und enthalten daher $p = 0$ *nicht*. Lässt man in $p = 0$ enthaltenden Blöcken aus $\mathcal{C}^*$ dieses Element fort, so erhält man $\mathcal{C}^*(p) = \mathcal{C}$, also genau die Blöcke des Untersystems $(T, \mathcal{C})$. Damit haben wir die nach (a) gebildeten Quadrupel aus $\mathcal{E}$ erfasst. In einem nach (b) gebildeten Quadrupel $\{a, b, c, d\} \in \mathcal{E}$ ist $\{a, b\} \in A_i$, $\{c, d\} \in B_i$, $i = 1, \ldots, n$. In der 1-Faktorisierung $\{B_1, \ldots, B_n\}$ von $S \setminus T$ enthalten die 1-Faktoren B_i das Element B_i nicht. Dagegen enthalten die 1-Faktoren A_i einer 1-Faktorisierung $\{A_1, \ldots, A_n\}$ von $T \cup \{p\}$ als einzige $p = 0$ enthaltende Elemente die "Kante" $\{0, i\}$. Insgesamt ist

$$\mathcal{E}(p) = C \cup \bigcup_{i=1}^{n} \{\{i, b, c\} : \{b, c\} \in B_i\}.$$

Nach Definition von B_i, $i = 1, \ldots, n$, ist $\mathcal{E}(p) \subset \mathcal{B}$. Andererseits ist

$$|\mathcal{E}(p)| = |\mathcal{C}| + \sum_{i=1} n|B_i| = \frac{n(n-1)}{6} + n\frac{n+1}{2} = \frac{2n^2+n}{3} = \frac{(2n+1)2n}{6} = |\mathcal{B}|$$

und damit $\mathcal{E}(p) = \mathcal{B}$. Der Satz ist bewiesen. □

Beispiel: Mit geringfügigen Modifikationen geben wir ein Beispiel bei K. T. Phelps (1976) zu Satz 7.17 wieder. Sei $n = 7$ und daher

$$S := \{1, \ldots, 15\}, \qquad T := \{1, \ldots, 7\}.$$

Die Menge der Blöcke $\mathcal{B} \subset S$ bzw. $\mathcal{C} \subset T$ mit $\mathcal{C} \subset \mathcal{B}$ seien gegeben durch

$$\underbrace{\underbrace{\begin{array}{l}\{1,2,3\}\\ \{1,4,5\}\\ \{1,6,7\}\\ \{2,4,6\}\\ \{2,5,7\}\\ \{3,4,7\}\\ \{3,5,6\}\end{array}}_{=\mathcal{C}} \begin{array}{l}\{1,8,9\}\\ \{1,10,11\}\\ \{1,12,13\}\\ \{1,14,15\}\\ \{2,8,10\}\\ \{2,9,11\}\\ \{2,12,14\}\end{array} \begin{array}{l}\{2,13,15\}\\ \{3,8,11\}\\ \{3,9,12\}\\ \{3,10,15\}\\ \{3,13,14\}\\ \{4,8,13\}\\ \{4,9,10\}\end{array} \begin{array}{l}\{4,11,14\}\\ \{4,12,15\}\\ \{5,8,15\}\\ \{5,9,14\}\\ \{5,10,13\}\\ \{5,11,12\}\\ \{6,8,14\}\end{array} \begin{array}{l}\{6,9,13\}\\ \{6,10,12\}\\ \{6,11,15\}\\ \{7,8,12\}\\ \{7,9,15\}\\ \{7,10,14\}\\ \{7,11,13\}.\end{array}}_{=\mathcal{B}}$$

Eine 1-Faktorisierung $\{B_1, \ldots, B_7\}$ von $S \setminus T = \{8, \ldots, 15\}$ erhält man durch

$$B_i := \{\{x, y\} \subset \{8, 15\} : \{i, x, y\} \in \mathcal{B}\}, \quad i = 1, \ldots, 7.$$

Es ist

$$\begin{aligned}
B_1 &= \{\{8,9\},\{10,11\},\{12,13\},\{14,15\}\},\\
B_2 &= \{\{8,10\},\{9,11\},\{12,14\},\{13,15\}\},\\
B_3 &= \{\{8,11\},\{9,12\},\{10,15\},\{13,14\}\},\\
B_4 &= \{\{8,13\},\{9,10\},\{11,14\},\{12,15\}\},\\
B_5 &= \{\{8,15\},\{9,14\},\{10,13\},\{11,12\}\},\\
B_6 &= \{\{8,14\},\{9,13\},\{10,12\},\{11,15\}\},\\
B_7 &= \{\{8,12\},\{9,15\},\{10,14\},\{11,13\}\}.
\end{aligned}$$

Sei $p := 0$. Dann ist $(T \cup \{p\}, \mathcal{C}^*)$ mit

$$\mathcal{C}^* := \left\{ \begin{array}{ll}
\{0,1,2,3\}, & \{1,2,4,7\},\\
\{0,1,4,5\}, & \{1,2,5,6\},\\
\{0,1,6,7\}, & \{1,3,4,6\},\\
\{0,2,4,6\}, & \{1,3,5,7\},\\
\{0,2,5,7\}, & \{2,3,4,5\},\\
\{0,3,4,7\}, & \{2,3,6,7\},\\
\{0,3,5,6\}, & \{4,5,6,7\}
\end{array} \right\}$$

ein $SQS(8)$ zur Grundmenge $T \cup \{p\} = \{0, 1, \ldots, 7\}$, welches $(T, \mathcal{C})$ als abgeleitetes Tripel System besitzt. Denn einerseits stimmt $\mathcal{C}^*$ mit der Menge der Blöcke $\mathcal{B}'$ auf Seite 184 überein, wenn man dort für den Übergang von $\{1, \ldots, 8\}$ zu $\{0, \ldots, 7\}$ von jeder Ziffer 1 abzieht, andererseits ist offensichtlich

$$\mathcal{C}^*(p) = \{C^* \setminus \{p\} : C^* \in \mathcal{C}^*, p \in C^*\} = \mathcal{C}.$$

Gehen wir weiter im Beweis von Satz 7.17, so müssen wir jetzt ein beliebiges $SQS(8)$ auf $S \setminus T = \{8, \ldots, 15\}$ wählen. Die Blöcke seien z. B.

$$\mathcal{D} := \left\{ \begin{array}{ll}
\{8,9,10,11\}, & \{9,10,12,15\},\\
\{8,9,12,13\}, & \{9,10,13,14\},\\
\{8,9,14,15\}, & \{9,11,12,14\},\\
\{8,10,12,14\}, & \{9,11,13,15\},\\
\{8,10,13,15\}, & \{10,11,12,13\},\\
\{8,11,12,15\}, & \{10,11,14,15\},\\
\{8,11,13,14\}, & \{12,13,14,15\}.
\end{array} \right\}$$

Hier entsteht $\mathcal{D}$ offenbar dadurch, dass 8 auf jedes Element der Menge $\mathcal{C}^*$ addiert wird. Weiter benötigen wir eine beliebige 1-Faktorisierung $\{A_1, \ldots, A_7\}$ von $T \cup \{p\} = \{0, 1, \ldots, 7\}$, wobei der 1-Faktor A_i die "Kante" $\{i, p\}$ enthalte, $i = 1, \ldots, 7$. Eine

solche ist gegeben durch

$$\begin{aligned}
A_1 &= \{\{0,1\},\{2,3\},\{4,5\},\{6,7\}\},\\
A_2 &= \{\{0,2\},\{1,3\},\{4,6\},\{5,7\}\},\\
A_3 &= \{\{0,3\},\{1,2\},\{4,7\},\{5,6\}\},\\
A_4 &= \{\{0,4\},\{1,5\},\{2,6\},\{3,7\}\},\\
A_5 &= \{\{0,5\},\{1,4\},\{2,7\},\{3,6\}\},\\
A_6 &= \{\{0,6\},\{1,7\},\{2,4\},\{3,5\}\},\\
A_7 &= \{\{0,7\},\{1,6\},\{2,5\},\{3,4\}\}.
\end{aligned}$$

Hiermit erhalten wir ein Quadrupel System zur Grundmenge $S \cup \{p\} = \{0, 1, \ldots, 15\}$ mit der Menge der Blöcke $\mathcal{E}$, welche besteht aus

(a) allen Blöcken (insgesamt 28) aus $\mathcal{C}^*$ und $\mathcal{D}$

sowie

(b) Quadrupel $\{a, b, c, d\} \in \{0, 1, \ldots, 15\}$ mit $\{a, b\} \in A_i$ und $\{c, d\} \in B_i$, $i = 1, \ldots, 7$.

Dies ergibt insgesamt 112 weitere Blöcke:

$i = 1$:

$$\begin{array}{llll}
\{0,1,8,9\}, & \{0,1,10,11\}, & \{0,1,12,13\}, & \{0,1,14,15\},\\
\{2,3,8,9\}, & \{2,3,10,11\}, & \{2,3,12,13\}, & \{2,3,14,15\},\\
\{4,5,8,9\}, & \{4,5,10,11\}, & \{4,5,12,13\}, & \{4,5,14,15\},\\
\{6,7,8,9\}, & \{6,7,10,11\}, & \{6,7,12,13\}, & \{6,7,14,15\}.
\end{array}$$

$i = 2$:

$$\begin{array}{llll}
\{0,2,8,10\}, & \{0,2,9,11\}, & \{0,2,12,14\}, & \{0,2,13,15\},\\
\{1,3,8,10\}, & \{1,3,9,11\}, & \{1,3,12,14\}, & \{1,3,13,15\},\\
\{4,6,8,10\}, & \{4,6,9,11\}, & \{4,6,12,14\}, & \{4,6,13,15\},\\
\{5,7,8,10\}, & \{5,7,9,11\}, & \{5,7,12,14\}, & \{5,7,13,15\}.
\end{array}$$

$i = 3$:

$$\begin{array}{llll}
\{0,3,8,11\}, & \{0,3,9,12\}, & \{0,3,10,15\}, & \{0,3,13,14\},\\
\{1,2,8,11\}, & \{1,2,9,12\}, & \{1,2,10,15\}, & \{1,2,13,14\},\\
\{4,7,8,11\}, & \{4,7,9,12\}, & \{4,7,10,15\}, & \{4,7,13,14\},\\
\{5,6,8,11\}, & \{5,7,9,12\}, & \{5,7,10,15\}, & \{5,7,13,14\}.
\end{array}$$

$i = 4$:

$$\begin{array}{llll}
\{0,4,8,13\}, & \{0,4,9,10\}, & \{0,4,11,14\}, & \{0,4,12,15\},\\
\{1,5,8,13\}, & \{1,5,9,10\}, & \{1,5,11,14\}, & \{1,5,12,15\},\\
\{2,6,8,13\}, & \{2,6,9,10\}, & \{2,6,11,14\}, & \{2,6,12,15\},\\
\{3,7,8,13\}, & \{3,7,9,10\}, & \{3,7,11,14\}, & \{3,7,12,15\}.
\end{array}$$

$i = 5$:

$$\begin{array}{llll}
\{0,5,8,15\}, & \{0,5,9,14\}, & \{0,5,10,13\}, & \{0,5,11,12\},\\
\{1,4,8,15\}, & \{1,4,9,14\}, & \{1,4,10,13\}, & \{1,4,11,12\},\\
\{2,7,8,15\}, & \{2,7,9,14\}, & \{2,7,10,13\}, & \{2,7,11,12\},\\
\{3,6,8,15\}, & \{3,6,9,14\}, & \{3,6,10,13\}, & \{3,6,11,12\}.
\end{array}$$

$i = 6$:

$$\begin{array}{llll} \{0,6,8,14\}, & \{0,6,9,13\}, & \{0,6,10,12\}, & \{0,6,11,15\}, \\ \{1,7,8,14\}, & \{1,7,9,13\}, & \{1,7,10,12\}, & \{1,7,11,15\}, \\ \{2,4,8,14\}, & \{2,4,9,13\}, & \{2,4,10,12\}, & \{2,4,11,15\}, \\ \{3,5,8,14\}, & \{3,5,9,13\}, & \{3,5,10,12\}, & \{3,5,11,15\}. \end{array}$$

$i = 7$:

$$\begin{array}{llll} \{0,7,8,12\}, & \{0,7,9,15\}, & \{0,7,10,14\}, & \{0,7,11,13\}, \\ \{1,6,8,12\}, & \{1,6,9,15\}, & \{1,6,10,14\}, & \{1,6,11,13\}, \\ \{2,5,8,12\}, & \{2,5,9,15\}, & \{2,5,10,14\}, & \{2,5,11,13\}, \\ \{3,4,8,12\}, & \{3,4,9,15\}, & \{3,4,10,14\}, & \{3,4,11,13\}. \end{array}$$

Die Menge $\mathcal{E}(p)$ von Tripeln erhält man dadurch, dass man alle $p = 0$ enthaltenden Quadrupel aus $\mathcal{E}$ betrachtet und bei ihnen $p = 0$ entfernt. Man erhält

$$\mathcal{E}(p) = \left\{ \begin{array}{lllll} \{1,2,3\}, & \{1,8,9\}, & \{3,8,11\}, & \{5,8,15\}, & \\ \{1,4,5\}, & \{1,10,11\}, & \{3,9,12\}, & \{5,9,14\}, & \\ \{1,6,7\}, & \{1,12,13\}, & \{3,10,15\}, & \{5,10,13\}, & \{7,8,12\}, \\ \{2,4,6\}, & \{1,14,15\}, & \{3,13,14\}, & \{5,11,12\}, & \{7,9,15\}, \\ \{2,5,7\}, & \{2,8,10\}, & \{4,8,13\}, & \{6,8,14\}, & \{7,10,14\}, \\ \{3,4,7\}, & \{2,9,11\}, & \{4,9,10\}, & \{6,9,13\}, & \{7,11,13\}. \\ \{3,5,6\}, & \{2,12,14\}, & \{4,11,14\}, & \{6,10,12\}, & \\ & \{2,13,15\}, & \{4,12,15\}, & \{6,11,15\} & \end{array} \right\}$$

Offenbar ist $\mathcal{E}(p) = \mathcal{B}$, womit die Gültigkeit des Satzes in einem Spezialfall verifiziert ist. □

8 Die Permanenten-Vermutung von van der Waerden und ihr Beweis

8.1 Definitionen und Formulierung der Vermutung

Die *Determinante* einer $n \times n$-Matrix $A = (a_{ij})$ kann bekanntlich durch

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

definiert werden. Hierbei bedeutet S_n die *symmetrische Gruppe der Permutationen* von $\{1, \ldots, n\}$ und $\operatorname{sgn}(\sigma)$ steht für das *Signum* der Permutation $\sigma \in S_n$, siehe auch den Beweis zu 5.5, dem Satz von Binet-Cauchy. Dagegen ist die *Permanente* einer $n \times n$-Matrix $A = (a_{ij})$ definiert durch

$$\operatorname{per}(A) := \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}.$$

Die Permanente kann man als *scheinbar* einfachen Zwilling der Determinante ansehen. Scheinbar, weil sie in mancher Beziehung schwieriger ist, z. B. der Berechnung. Trotzdem haben die Zwillinge einige gemeinsame Eigenschaften wie die Multilinearität oder die Entwicklung nach Zeilen oder Spalten.

Von B. L. VAN DER WAERDEN (1926) stammt die folgende Aufgabe. Wir geben wörtlich die Formulierung von van der Waerden wieder.

- Die Funktion
$$Q = \sum_{(k_1,\ldots,k_n)} \alpha_{1k_1}\alpha_{2k_2}\cdots\alpha_{nk_n}$$
(Summation über alle Permutationen der Ziffern $1,\ldots,n$) ist unter den Nebenbedingungen
$$\begin{cases} \alpha_{ik} \geq 0 & \\ \sum_{i=1}^{n} \alpha_{ik} = 1 & (k = 1,\ldots,n) \\ \sum_{k=1}^{n} \alpha_{ik} = 1 & (i = 1,\ldots,n) \end{cases}$$
nur positiver Werte fähig (Dénes König, Graphen und ihre Anwendungen § 2, Math. Annal. 77, S. 457). Das Minimum der Funktion ist (unter den genannten Nebenbedingungen) zu bestimmen.

Matrizen, die den genannten Nebenbedingungen genügen, deren Einträge also sämtlich nichtnegativ sind und deren Zeilen- und Spaltensummen gleich 1 sind, nennt man *doppelt stachastisch*. Die Menge der doppelt stochastischen $n \times n$-Matrizen bezeichnen wir mit Ω_n. Offenbar ist Ω_n eine kompakte Menge in $\mathbb{R}^{n\times n}$. Da die Permanente eine stetige Funktion ihres Arguments ist, gibt es eine doppelt stochastische Matrix mit minimaler Permanente, d. h. die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \operatorname{per}(A), \quad A \in \Omega_n,$$

besitzt eine Lösung. Eine Lösung von (P) nennen wir eine *minimale* (*minimizing* oder *minimal*) Matrix. Von B. L. VAN DER WAERDEN (1926) wurde explizit keine Vermutung geäußert, aber natürlich besteht diese darin, dass das Minimum genau für $J_n := \frac{1}{n}E$ angenommen wird. Hierbei ist E die $n \times n$-Matrix, deren Einträge sämtlich gleich 1 sind. Unser Ziel in diesem Abschnitt besteht folglich darin, den folgenden Satz zu beweisen.

Satz 8.1 *Sei A eine doppelt stochastische $n \times n$-Matrix. Dann ist*

$$\operatorname{per}(A) \geq \frac{n!}{n^n}$$

und Gleichheit gilt genau dann, wenn $A = J_n := \frac{1}{n}E$, wobei E die $n \times n$-Matrix ist, deren Einträge sämtlich gleich 1 sind. Oder anders gesagt: J_n ist die eindeutige Lösung der obigen Optimierungsaufgabe (P).

Dieser Satz wurde etwa zur gleichen Zeit von G. P. Egorychev (1981) und D. I. Falikman (1981) bewiesen. In einer note added in proof zu der Arbeit D. E. Knuth (1981) findet man die folgende Bemerkung:

- It was recently learned that part of Egorychev's result was antcipated a year earlier by D. I. Falikman, whose elegant proof appears in Matematicheski Zametki (June 1981). Falikman's paper, which was received for publication on May 14, 1979, establishes the minimal value of doubly stochastic permanents but does not show that this value is uniquely attained.

Bemerkung: Natürlich folgt aus Satz 8.1, dass die Permanente einer doppelt stochastischen Matrix positiv ist. Dieses Ergebnis (mit einer wesentlich schlechteren unteren Schranke) kann man auch mit Hilfe des Satzes von Birkhoff-Neumann (siehe z. B. Unterabschnitt 29.1 bei J. Werner (2013)) erhalten. Dieser sagt aus:

- *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 $\sum_{k=1}^{l} \lambda_k = 1$ und $A = \sum_{k=1}^{l} \lambda_k P_k$ gibt.*

Da es $n!$ Permutationen der Zahlen $\{1, \ldots, n\}$ gibt, ist $l \leq n!$ im Satz von Birkhoff-Neumann und daher o. B. d. A. $l = n!$. Wegen des Satzes von Carathéodory, diesen haben wir schon beim Beweis des Satzes 2.4 von F. John auf Seite 14 benutzt, könnten wir ein wesentlich kleineres l wählen. Dann existiert ein $k_0 \in \{1, \ldots, l\}$ mit $\lambda_{k_0} \geq 1/n!$. Denn in der Summe endlich vieler nichtnegativer Zahlen ist wenigstens ein Summand größer oder gleich dem Mittelwert dieser Zahlen. Die $n \times n$-Permutationsmatrix P_{k_0} enthält in jeder Zeile und jeder Spalte genau eine 1 und sonst nur Nullen. Daher ist durch

$$P_{k_0} = \begin{pmatrix} e_{\sigma_0(1)}^T \\ \vdots \\ e_{\sigma_0(n)}^T \end{pmatrix}$$

eine Permutation $\sigma_0 \in S_n$ definiert. Hierbei bedeutet e_i natürlich den i-ten Einheitsvektor im $\mathbb{R}^n$. Dann ist

$$\begin{aligned} \operatorname{per}(A) &= \sum_{\sigma\in S_n} \prod_{i=1}^{n} a_{i\sigma(i)} \\ &\geq \prod_{i=1}^{n} a_{i\sigma_0(i)} \\ &\geq \prod_{i=1}^{n} [\lambda_{k_0} \underbrace{(P_{k_0})_{i\sigma_0(i)}}_{=1}] \\ &= \lambda_{k_0}^n \\ &\geq \frac{1}{(n!)^n}. \end{aligned}$$

Insbesondere ist die Permanente einer doppelt stochastischen Matrix positiv. □

Bemerkung: Bei M. Marcus, M. Newman (1959) wird bewiesen, dass

$$\operatorname{per}(A) \geq (n^2 - n + 1)^{1-n} \quad \textit{für jedes } A \in \Omega_n.$$

Der Beweis hierfür ist hübsch. Wir wollen nach dem gleichen Muster ein etwas schlechteres Ergebnis beweisen.

- Für nichtnegative Matrizen $A, B \in \mathbb{R}^{n\times n}$ ist $\operatorname{per}(A+B) \geq \operatorname{per}(A) + \operatorname{per}(B)$.

  Denn: Dies ist offensichtlich, wenn man beachtet, dass jeder in

  $$\sum_{\sigma\in S_n} \prod_{i=1}^{n} a_{i\sigma(i)} \quad \text{bzw.} \quad \sum_{\sigma\in S_n} \prod_{i=1}^{n} b_{i\sigma(i)}$$

  auftretende Term auch in

  $$\sum_{\sigma\in S_n} \prod_{i=1}^{n} (a_{i\sigma(i)} + b_{i\sigma(i)})$$

  vorkommt.

- Für jedes $A \in \Omega_n$ ist $\operatorname{per}(A) \geq (n^2+1)^{1-n}$.

  Denn: Wegen des Satzes von Birkhoff-Neumann und des Satzes von Carathédory exisistiert ein $l \in \{1, \ldots, n^2+1\}$, positive Zahlen $\lambda_1, \ldots, \lambda_l$ mit $\sum_{k=1}^{l} \lambda_k = 1$ und $n \times n$-Permutationsmatrizen $P_1, \ldots, P_l$ mit $A = \sum_{k=1}^{l} \lambda_k P_k$. Wegen der gerade eben bewiesenen Ungleichung erhalten wir

  $$\begin{aligned}
  \operatorname{per}(A) &= \operatorname{per}\left(\sum_{k=1}^{l} \lambda_k P_k\right) \\
  &\geq \sum_{k=1}^{l} \operatorname{per}(\lambda_k P_k) \\
  &= \sum_{k=1}^{l} \lambda_k^n \underbrace{\operatorname{per}(P_k)}_{=1} \\
  &\geq \sum_{k=1}^{l} \left(\frac{1}{l}\right)^n \\
  &= l^{1-n} \\
  &\geq (n^2+1)^{1-n}.
  \end{aligned}$$

Hierbei haben wir ausgenutzt, dass die Aufgabe

$$\text{Minimiere} \quad \sum_{k=1}^{l} x_k^n \quad \text{auf} \quad M := \{x \in \mathbb{R}^l : x \geq 0,\ e^T x = 1\}$$

für $n > 1$ eine eindeutige Lösung besitzt, welche durch $x^* := (1/n)e$ gegeben ist. Hierbei ist e der Vektor des $\mathbb{R}^l$, dessen Komponenten sämtlich gleich 1 sind. Bei M. MARCUS, M. NEWMAN (1959) wird ein Ergebnis von Birkhoff benutzt, dass sich nämlich jede doppelt stochastische $n \times n$-Matrix als Konvexkombination von höchstens n^2-n+1 Permutationsmatrizen darstellen lässt. Das ergibt dann die bessere Abschätzung $\operatorname{per}(A) \geq (n^2 - n + 1)^{1-n}$. Einen kleinen Eindruck über die erwähnten unteren Schranken erhält man durch die folgende Tabelle.

n	$1/(n!)^n$	$(n^2+1)^{1-n}$	$(n^2-n+1)^{1-n}$	$n!/n^n$
2	0.2500	0.2000	0.3333	0.5000
3	0.0046	0.0100	0.0204	0.2222

□

Beispiel: Eine doppelt stochastische 2×2-Matrix hat die Form

$$A_\lambda = \begin{pmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{pmatrix}$$

mit $\lambda \in [0,1]$. Dann ist

$$\operatorname{per}(A_\lambda) = \lambda^2 + (1-\lambda)^2.$$

Offensichtlich ist

$$\min_{\lambda \in [0,1]} \operatorname{per}(A_\lambda) = \operatorname{per}(A_{1/2}) = 2\left(\frac{1}{2}\right)^2 = \frac{1}{2} = \frac{2!}{2^2},$$

für $n = 2$ ist Satz 8.1 also richtig. Wesentlich komplizierter wird die Situation schon für $n = 3$. Einen Beweis der van der Waerden Vermutung in diesem Fall findet man bei M. MARCUS, M. NEWMAN (1959, p. 71), mit einem ad hoc Beweis bei H. TVERBERG (1963) und auch bei P. J. EBERLEIN, G. S. MUDHOLKAR (1968). Deren Ansatz wollen wir hier kurz schildern.

1. Für $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ definieren wir die symmetrischen Funktionen

$$e_1 \colon \mathbb{R}^3 \longrightarrow \mathbb{R}, \qquad e_2 \colon \mathbb{R}^3 \longrightarrow \mathbb{R}, \qquad e_3 \colon \mathbb{R}^3 \longrightarrow \mathbb{R}$$

durch

$$e_1(x) := x_1 + x_2 + x_3, \qquad e_2(x) := x_1x_2 + x_2x_3 + x_3x_1, \qquad e_3(x) := x_1x_2x_3.$$

Sei

$$A = \begin{pmatrix} a^1 & a^2 & a^3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

eine doppelt stochastische 3×3-Matrix. Dann ist

$$\operatorname{per}(A) = 1 + \sum_{j=1}^{3} (-e_2 + 2e_3)(a^j).$$

Denn: Wir definieren

$$T_r(A) := \{a^{i_1} + a^{i_2} + a^{i_3} : \{i_1, \ldots, i_r\} \subset \{1,2,3\}\}, \quad r = 1,2,3.$$

Dann ist

$$\begin{aligned} T_3(A) &= \{a^1 + a^2 + a^3\}, \\ T_2(A) &= \{a^1 + a^2, a^2 + a^3, a^3 + a^1\}, \\ T_1(A) &= \{a^1, a^2, a^3\}. \end{aligned}$$

Nach leichter Rechnung (wobei wir noch nicht benutzen, dass A doppelt stochastisch ist) erhalten wir

$$\begin{aligned} & \sum_{x \in T_3(A)} e_3(x) - \sum_{x \in T_2(A)} e_3(x) + \sum_{x \in T_1(A)} e_3(x) \\ = \; & (a_{11} + a_{12} + a_{13})(a_{21} + a_{22} + a_{23})(a_{31} + a_{32} + a_{33}) \\ & \quad - (a_{11} + a_{12})(a_{21} + a_{22})(a_{31} + a_{32}) \\ & \quad - (a_{12} + a_{13})(a_{22} + a_{23})(a_{32} + a_{33}) \\ & \quad - (a_{13} + a_{11})(a_{23} + a_{21})(a_{33} + a_{31}) \\ & \quad + a_{11}a_{21}a_{31} + a_{12}a_{22}a_{32} + a_{13}a_{23}a_{33} \\ = \; & a_{11}a_{22}a_{33} + a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} + a_{13}a_{22}a_{31} \\ = \; & \mathrm{per}(A). \end{aligned}$$

Nun nutzen wir aus, dass A doppelt stochastisch ist. Mit $e := (1,1,1)^T$ ist

$$T_3(A) = \{e\}, \quad T_2(A) = \{e - a^3, e - a^1, e - a^2\}, \quad T_1(A) = \{a^1, a^2, a^3\}.$$

Weiter ist

$$e_3(e - x) = (1 - x_1)(1 - x_2)(1 - x_3) = 1 - e_1(x) + e_2(x) - e_3(x)$$

und daher

$$\begin{aligned} \mathrm{per}(A) &= \sum_{x \in T_3(A)} e_3(x) - \sum_{x \in T_2(A)} e_3(x) + \sum_{x \in T_1(A)} e_3(x) \\ &= e_3(e) - \sum_{j=1}^{3} e_3(e - a^j) + \sum_{j=1}^{3} e_3(a^j) \\ &= 1 + \sum_{j=1}^{3} (\underbrace{-1 + e_1(a^j)}_{=0} - e_2(a^j) + 2e_3(a^j)) \\ &= 1 + \sum_{j=1}^{3} (-e_2 + 2e_3)(a^j). \end{aligned}$$

2. Die Optimierungsaufgabe

$$\begin{cases} \text{Minimiere} \quad f(x) := (-e_2 + 2e_3)(x) = -(x_1x_2 + x_2x_3 + x_3x_1) + 2x_1x_2x_3 \\ \qquad\qquad \text{auf} \quad M := \{x \in \mathbb{R}^3 : x \geq 0,\ e^T x = 1\} \end{cases}$$

besitzt die eindeutige Lösung $x^* := (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$.

Denn: Sei $x^* \in M$ eine Lösung. Angenommen, es ist $x^* \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$. Dann existiert ein $i \in \{1, 2, 3\}$ mit $x_i^* < \frac{1}{3}$. O. B. d. A. ist $i = 3$, also $x_3^* < \frac{1}{3}$. Mit $x^* = (x_1^*, x_2^*, x_3^*)^T \in M$ ist auch $x^{**} := (\frac{1}{2}(x_1^* + x_2^*), \frac{1}{2}(x_1^* + x_2^*), x_3^*)^T \in M$. Da x^* eine Lösung obiger Optimierungsaufgabe ist, ist

$$\begin{aligned} 0 &\leq f(x^{**}) - f(x^*) \\ &= f(\tfrac{1}{2}(x_1^* + x_2^*), \tfrac{1}{2}(x_1^* + x_2^*), x_3^*) - f(x_1^*, x_2^*, x_3^*) \\ &= -[\tfrac{1}{4}(x_1^* + x_2^*)^2 + (x_1^* + x_2^*)x_3^*] + \tfrac{1}{2}(x_1^* + x_2^*)^2 x_3^* \\ &\qquad + [x_1^*x_2^* + x_2^*x_3^* + x_3^*x_1^*] - x_1^*x_2^*x_3^* \\ &= -\tfrac{1}{4}(x_1^*)^2 - \tfrac{1}{2}x_1^*x_2^* - \tfrac{1}{4}(x_2^*)^2 - x_1^*x_3^* - x_2^*x_3^* + \tfrac{1}{2}(x_1^*)^2 x_3^* + x_1^*x_2^*x_3^* + \tfrac{1}{2}(x_2^*)^2 x_3^* \\ &\qquad + [x_1^*x_2^* + x_2^*x_3^* + x_3^*x_1^*] - 2x_1^*x_2^*x_3^* \\ &= -\tfrac{1}{4}(x_1^*)^2 + \tfrac{1}{2}x_1^*x_2^* - \tfrac{1}{4}(x_2^*)^2 + \tfrac{1}{2}(x_1^*)^2 x_3^* + \tfrac{1}{2}(x_2^*)^2 x_3^* - x_1^*x_2^*x_3^* \\ &= \tfrac{1}{2}(x_1^* - x_2^*)^2 (\underbrace{x_3^* - \tfrac{1}{2}}_{<0}) \\ &\leq 0. \end{aligned}$$

Ist also in einer Lösung x^* eine Komponente, z. B. die dritte, kleiner als $\frac{1}{3}$, so stimmen die beiden anderen überein, es ist also etwa $x_1^* = x_2^*$ und $x_3^* = 1 - 2x_1^*$. Wegen $0 \leq x_3^* < \frac{1}{3}$ ist $\frac{1}{3} < x_1^* \leq \frac{1}{2}$. Weiter ist

$$\begin{aligned} f(x_1^*, x_2^*, x_3^*) &= f(x_1^*, x_1^*, 1 - 2x_3^*) \\ &= -[(x_1^*)^2 + 2x_1^*(1 - 2x_1^*)] + 2(x_1^*)^2(1 - 2x_1^*) \\ &= -2x_1^* + 5(x_1^*)^2 - 4(x_1^*)^3. \end{aligned}$$

Definiert man

$$\phi(t) := f(t, t, 1 - 2t) = -2t + 5t^2 - 4t^3,$$

so ist

$$\phi'(t) = -2 + 10t - 12t^2 = -2(1 - 3t)(1 - 2t), \qquad \phi''(t) = 10 - 24t.$$

Daher nimmt die Funktion $\phi(\cdot)$ ihr Minimum auf $[\frac{1}{3}, \frac{1}{2}]$ in $\frac{1}{3}$ an, ein Widerspruch dazu, dass das Minimum in $x_1^* \in (\frac{1}{3}, \frac{1}{2}]$ eintritt.

3. Unter Benutzung von 1. und 2. erhalten wir für eine beliebige doppelt stochasti-

sche 3×3-Matrix $A = \begin{pmatrix} a^1 & a^2 & a^3 \end{pmatrix}$, dass

$$\begin{aligned} \operatorname{per}(A) &= 1 + \sum_{j=1}^{3} (-e_2 + 2e_3)(a^j) \\ &\geq 1 + \sum_{j=1}^{3} (-e_2 + 2e_3)(\tfrac{1}{3}e) \\ &= 1 - 3 \cdot \frac{7}{27} \\ &= \frac{3!}{3^3}, \end{aligned}$$

wobei Gleichheit genau dann gilt, wenn $a^j = \frac{1}{3}e$, $j = 1, 2, 3$, bzw. $A = \frac{1}{3}E$. Daher ist die van der Waerden Vermutung für $n = 3$ richtig.

Auch für $n = 4$ konnte von P. J. Eberlein, G. S. Mudholkar (1968, Theorem 4) nachgewiesen werden, dass die van der Waerden Vermutung richtig ist. P. J. Eberlein (1969, Theorem 2) bewies die Gültigkeit der van der Waerden Vermutung für $n = 5$. □

Bemerkung: Bis zu dem Zeitpunkt im Jahre 1981, als die van der Waerden Vermutung bewiesen wurde, sind zahlreiche Aufsätze zu Spezialfällen der van der Waerden Vermutung erschienen, Wir wollen hier nur zwei dieser Ergebnisse nennen. Ihre Beweise sind alles andere als trivial, was darauf hindeutet, dass auch ein Beweis der (allgemeinen) van der Waerden Vermutung nicht ganz einfach ist. Das erste Ergebnis stammt von M. Marcus, M.Newman (1959, Theorem 3) und sagt aus:

- Besitzt die Optimierungsaufgabe

  (P) $\qquad\qquad$ Minimiere $\quad \operatorname{per}(A), \quad A \in \Omega_n,$

  eine Lösung $A^* > 0$, so ist $J_n := \frac{1}{n}E$ die eindeutige Lösung von (P) und daher $\operatorname{per}(A) \geq n!/n^n$ für alle doppelt stochastischen Matrizen $A \in \Omega_n$.

Dieses Ergebnis (mit Beweis) findet man auch in dem Lehrbuch H. Minc (1978, S. 76 ff). Das zweite Ergebnis, das wir hier erwähnen wollen, stammt von M. Marcus, H. Minc (1962) und sagt aus:

- Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch, positiv semidefinit und doppelt stochastisch, so ist

  $$\operatorname{per}(A) \geq \frac{n!}{n^n}.$$

  In dieser Ungleichung gilt Gleichheit genau dann, wenn $A = \frac{1}{n}E$.

Übersichten zur Geschichte der van der Waerden Vermutung findet man bei H. Minc (1978, Chapter 5) oder H. Minc (1983). □

8.2 Der Beweis von Egorychev

In diesem Unterabschnitt wollen wir den Beweis der van der Waerden Vermutung durch G. P. Egorychev (1981) schildern, wobei die Ausarbeitungen von J. H. van Lint (1981,1982) und vor allem von D. E. Knuth (1981) (siehe auch M. Hall (1986, S. 58 ff.)) benutzt werden.

8.2.1 Quadratische Formen

Ist $F = (f_{ij}) \in \mathbb{R}^{n\times n}$, so nennen wir die durch $f(x) := x^T Fx$ definierte Abbildung $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ eine *quadratische Form in n Variablen*. Hier können wir o. B. d. A. annehmen, dass F symmetrisch ist bzw. $f_{ij} = f_{ji}$ für alle $1 \le i, j \le n$ gilt, da man andernfalls F durch $\frac{1}{2}(F + F^T)$ ersetzen kann. Zwei durch symmetrische Matrizen $F, G \in \mathbb{R}^{n\times n}$ gegebene quadratische Formen f, g heißen *äquivalent*, wenn eine nichtsinguläre Matrix $T \in \mathbb{R}^{n\times n}$ mit $G = T^T FT$ existiert. Mit $x = Ty$ ist dann

$$f(x) = x^T Fx = y^T T^T FTy = y^T Gy = g(y).$$

Ziel ist es nun, zu einer gegebenen quadratischen Form eine "einfachere" äquivalente quadratische Form zu bestimmen. Dies geschieht durch die folgenden beiden Lemmata.

Lemma 8.2 *Sei $f(x) = x^T Fx$ mit symmetrischem $F = (f_{ij}) \in \mathbb{R}^{n\times n}$ eine quadratische Form in n Variablen und $a = (a_i) \in \mathbb{R}^n$ ein Vektor mit $a_1 \neq 0$ und $c := f(a) \neq 0$. Die Matrix $T \in \mathbb{R}^{n\times n}$ sei definiert durch*

$$T := \begin{pmatrix} a_1 & -a_1(Fa)_2/c & \cdots & -a_1(Fa)_j/c & \cdots & -a_1(Fa)_n/c \\ a_2 & 1 - a_2(Fa)_2/c & \cdots & -a_2(Fa)_j/c & \cdots & -a_2(Fa)_n/c \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ a_j & -a_j(Fa)_2/c & \cdots & 1 - a_j(Fa)_j/c & \cdots & -a_j(Fa)_n/c \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ a_n & -a_n(Fa)_2/c & \cdots & -a_n(Fa)_j/c & \cdots & 1 - a_n(Fa)_n/c \end{pmatrix}.$$

Dann ist T nichtsingulär und es ist

$$T^T FT = \begin{pmatrix} c & 0_{n-1}^T \\ 0_{n-1} & G_{n-1} \end{pmatrix},$$

wobei 0_{n-1} der Nullvektor im $\mathbb{R}^{n-1}$ und $G_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ symmetrisch ist. Mit $x = Ty$ ist also

$$f(x) = cy_1^2 + \underbrace{\begin{pmatrix} y_2 \\ \vdots \\ y_n \end{pmatrix}^T G_{n-1} \begin{pmatrix} y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=:g(y_2,\ldots,y_n)} = cy_1^2 + g(y_2, \ldots, y_n),$$

wobei g eine quadratische Form in den $n-1$ Variablen $y_2, \ldots, y_n$ ist.

Beweis: Wir definieren

$$S := \begin{pmatrix} (Fa)_1/c & (Fa)_2/c & \cdots & (Fa)_n/c \\ -a_2/a_1 & & & \\ \vdots & & I_{n-1} & \\ -a_n/a_1 & & & \end{pmatrix},$$

wobei I_{n-1} die $(n-1) \times (n-1)$-Einheitsmatrix bedeutet und zeigen $TS = I$ bzw. $S = T^{-1}$. Hierzu weisen wir nach, dass $(TS)_{ij} = e_i^T TS e_j = \delta_{ij}$ ist, wobei e_i bzw. e_j den i-ten bzw. j-ten Einheitsvektor im $\mathbb{R}^n$ darstellt und δ_{ij} das Kronecker-Symbol ist. Es ist

$$(TS)_{11} = \begin{pmatrix} a_1 \\ -a_1(Fa)_2/c \\ \vdots \\ -a_1(Fa)_n/c \end{pmatrix}^T \begin{pmatrix} (Fa)_1/c \\ -a_2/a_1 \\ \vdots \\ -a_n/a_1 \end{pmatrix} = \frac{1}{c} a^T Fa = 1.$$

Für $i = 2, \ldots, n$ ist

$$(TS)_{i1} = a_1 \begin{pmatrix} a_i \\ -a_i(Fa)_2/c \\ \vdots \\ 1 - a_i(Fa)_i/c \\ \vdots \\ -a_i(Fa)_n/c \end{pmatrix}^T \begin{pmatrix} (Fa)_1/c \\ -a_2/a_1 \\ \vdots \\ -a_i/a_1 \\ \vdots \\ -a_n/a_1 \end{pmatrix} = \frac{a_i}{a_2} \left(-1 + \frac{1}{c} \sum_{k=1}^{n} a_k (Fa)_k \right) = 0.$$

Für $j = 2, \ldots, n$ ist entsprechend

$$(TS)_{1j} = \begin{pmatrix} a_1 \\ -a_1(Fa)_2/c \\ \vdots \\ -a_1(Fa)_j/c \\ \vdots \\ -a_1(Fa)_n/c \end{pmatrix}^T \begin{pmatrix} (Fa)_j/c \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

Für $i = 2, \ldots, n$ ist

$$(TS)_{ii} = \begin{pmatrix} a_i \\ -a_i(Fa)_2/c \\ \vdots \\ 1 - a_i(Fa)_i/c \\ \vdots \\ -a_i(Fa)_n/c \end{pmatrix}^T \begin{pmatrix} (Fa)_i/c \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = 1.$$

Schließlich ist für $i, j = 2, \ldots, n$ mit $i \neq j$ offensichtlich

$$(TS)_{ij} = a_i(Fa)_j/c - a_i(Fa)_j/c = 0.$$

Insgesamt ist $TS = I$, also T nichtsingulär und $S = T^{-1}$. Weiter ist

$$(T^T FT)_{11} = (Te_1)^T FTe_1 = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}^T F \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = f(a) = c.$$

Für $j = 2, \ldots, n$ ist schließlich

$$(T^T FT)_{1j} = \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix}^T F \begin{pmatrix} -a_1(Fa)_j/c \\ \vdots \\ 1 - a_j(Fa)_j/c \\ \vdots \\ -a_n(Fa)_j/c \end{pmatrix} = a^T Fe_j - \frac{(Fa)_j}{c} a^T Fa = 0.$$

Damit ist das Lemma bewiesen. □

Der eben geschilderte Übergang von einer quadratischen Form $f(x) = f(x_1, \ldots, x_n)$ in n Variablen zu einer äquivalenten quadratischen Form $cy_1^2 + g(y_2, \ldots, y_n)$, wobei $g(y_2, \ldots, y_n)$ eine quadratische Form in $n-1$ Variablen ist, kann offenbar fortgesetzt werden, wodurch das folgende Lemma erhalten wird.

Lemma 8.3 *Jede quadratische Form $f(x) = f(x_1, \ldots, x_n) = x^T Fx$ ist äquivalent zu einer einfachen quadratischen Form*

$$g(y) = g(y_1, \ldots, y_n) = y_1^2 + \cdots + y_p^2 - y_{p+1}^2 - \cdots - y_r^2$$

mit gewissen $0 \le p \le r \le n$.

Beweis: Ist $f(a) = 0$ für alle $a \in \mathbb{R}^n$, so ist die Behauptung mit $p = r = 0$ richtig. Andernfalls gibt es ein $a \neq 0$ mit $c = f(a) \neq 0$. Da man notfalls Variable permutieren kann, ist o. B. d. A. $a_1 \neq 0$. Wegen Lemma 8.2 wissen wir die Existenz einer nichtsingulären Matrix $T \in T^{n \times n}$ mit

$$T^T FT = \begin{pmatrix} c & 0_{n-1}^T \\ 0_{n-1} & G_{n-1} \end{pmatrix}.$$

Mit

$$D := \begin{pmatrix} 1/|c|^{1/2} & 0_{n-1}^T \\ 0_{n-1} & I_{n-1} \end{pmatrix}$$

ist dann

$$(TD)^T FTD = \begin{pmatrix} \operatorname{sign}(c) & 0_{n-1}^T \\ 0_{n-1} & G_{n-1} \end{pmatrix},$$

also die quadratische Form f äquivalent zu $\operatorname{sign}(c) y_1^2 + g(y_2, \ldots, y_n)$. Eine Fortsetzung dieses Prozesses (und eine eventuelle Permutation der Variablen) ergibt die Behauptung. □

Wir geben ein Beispiel bei D. E. Knuth (1981) an.

Beispiel: Gegeben sei die quadratische Form

$$\begin{aligned} f(x_1,x_2,x_3) &:= x_1^2+x_2^2+x_3^2-2x_1x_2-2x_1x_3-2x_2x_3 \\ &= \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^T \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \end{aligned}$$

wir wenden Lemma 8.2 mit $a := e_1$ (erster Einheitsvektor im $\mathbb{R}^3$) an. Dann ist $c := f(a) = 1$ und

$$T = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad T^TFT = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & -2 & 0 \end{pmatrix}.$$

Mit $x = Ty$ bzw. $x_1 = y_1 + y_2 + y_3$, $x_2 = y_2$ und $x_3 = y_3$ ist also

$$f(x_1,x_2,x_3) = y_1^2 - 4y_2y_3.$$

Denselben Prozess wenden wir nun auf

$$g(y_2,y_3) = -4y_2y_3 = \begin{pmatrix} y_2 \\ y_3 \end{pmatrix}^T \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix}$$

an. Wegen $g(1,1) = -4$ erhalten wir mit Hilfe der Transformation

$$\begin{pmatrix} y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}z_1 - \frac{1}{2}z_2 \\ \frac{1}{2}z_1 + \frac{1}{2}z_2 \end{pmatrix}$$

die Darstellung $-4y_2y_3 = -z_1^2 + z_2^2$. Daher ist

$$f(x_1,x_2,x_3) = y_1^2 - 4y_2y_3 = y_1^2 + z_2^2 - z_1^2 = (x_1 - x_2 - x_3)^2 + (x_3 - x_2)^2 - (x_2 + x_3)^2.$$

□

Lemma 8.4 *Die Zahlen p und r in Lemma 8.3 sind eindeutig bestimmt. Hat man also zwei äquivalente quadratische Formen in n Variablen mit*

$$y_1^2 + \cdots + y_p^2 - y_{p+1}^2 - \cdots - y_r^2 = z_1^2 + \cdots + z_q^2 - z_{q+1}^2 - \cdots - z_s^2$$

mit $0 \le p \le r \le n$ und $0 \le q \le s \le n$, so ist $p = q$ und $r = s$.

Beweis: Da die beiden angegebenen quadratischen Formen äquivalent sind, existiert eine nichtsinguläre Matrix $T = (t_{ij}) \in \mathbb{R}^{n\times n}$ mit $y = Tz$ bzw. $y_i = \sum_{j=1}^n t_{ij}z_j$, $i = 1,\ldots,n$. Wir nehmen an, es sei $p < q$. Es gibt $z_1^*,\ldots,z_q^*$, nicht alle gleich Null, mit

$$\sum_{j=1}^{q} t_{ij}z_j^* = 0, i = 1,\ldots,p,$$

da es sich hier um ein homogenes lineares Gleichungssystem mit mehr Unbekannten als Gleichungen handelt. Weiter setze man $z_j^* := 0$, $j = q+1, \ldots, n$, und

$$y_i^* := \sum_{j=1}^{n} t_{ij} z_j^* = \sum_{j=1}^{q} t_{ij} z_j^*, \quad i = 1, \ldots, n.$$

Dann ist $y^* = Tz^*$ und folglich

$$\underbrace{y_1^{*2} + \cdots + y_p^{*2}}_{=0} - y_{p+1}^{*2} - \cdots - y_r^{*2} = z_1^{*2} + \cdots + z_q^{*2} - \underbrace{z_{q+1}^{*2} - \cdots - z_s^{*2}}_{=0}$$

bzw.

$$-y_{p+1}^{*2} - \cdots - y_r^{*2} = z_1^{*2} + \cdots + z_q^{*2}.$$

Hieraus folgt $z_1^* = \cdots = z_q^* = 0$, ein Widerspruch zur Wahl von $z_1^*, \ldots, z_q^*$. Daher ist $p \geq q$, aus Symmetriegründen $q \geq p$ und insgesamt $p = q$. Die Annahme $r < s$ führen wir ähnlich zum Widerspruch, wobei wir die schon bewiesene Beziehung $p = q$ benutzen. Wir können $z_{p+1}^*, \ldots, z_s^*$, nicht alle gleich Null, finden mit

$$\sum_{j=p+1}^{s} t_{ij} z_j^* = 0, \quad i = p+1, \ldots, r.$$

Denn hier handelt es sich wieder um ein homogenes lineares Gleichungssystem mit mehr Unbekannten als Gleichungen. Weiter setze man $z_j^* = 0$, $j = 1, \ldots, p$, $j = s+1, \ldots, n$, und

$$y_i^* := \sum_{j=1}^{n} t_{ij} z_j^* = \sum_{j=p+1}^{s} t_{ij} z_j^*, \quad i = 1, \ldots, n.$$

Dann ist $y^* = Tz^*$ und daher (wir nutzen wieder aus, dass $p = q$)

$$y_1^{*2} + \cdots + y_p^{*2} - \underbrace{y_{p+1}^{*2} - \cdots - y_r^{*2}}_{=0} = \underbrace{z_1^{*2} + \cdots + z_p^{*2}}_{=0} - z_{p+1}^{*2} - \cdots - z_s^{*2}$$

bzw.

$$y_1^{*2} + \cdots + y_p^{*2} = -z_{p+1}^{*2} - \cdots - z_s^{*2}.$$

Hieraus folgt $z_{p+1}^* = \cdots = z_s^* = 0$, ein Widerspruch. Damit ist die Annahme $r < s$ zum Widerspruch geführt worden.Aus Symmetriegründen ist $r = s$. Das Lemma ist bewiesen. □

Wegen Lemma 8.3 und Lemma 8.4 können jeder quadratischen Form f (in n Variablen) die Invarianten $p = p(f)$ und $r = r(f)$ zugeordnet werden. Ist $f(x) = x^T F x$ mit symmetrischem $F \in \mathbb{R}^{n \times n}$, so existiert eine nichtsinguläre Matrix $T \in \mathbb{R}^{n \times n}$ mit

$$T^T F T = \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_{r-p} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hier gibt p die Anzahl der positiven und $r - p$ die Anzahl der negativen Eigenwerte von F an. Der Rang von F ist r. Ist $r = n$, so ist F nichtsingulär und besitzt daher nur von Null verschiedene Eigenwerte.

Lemma 8.5 *Seien $F_0, F_1 \in \mathbb{R}^{n\times n}$ symmetrisch und $f_0(x) := x^T F_0 x$, $f_1(x) := x^T F_1 x$ die zugehörigen quadratischen Formen. Sei*

$$F_\theta := (1-\theta)F_0 + \theta F_1$$

und

$$f_\theta := x^T F_\theta x = (1-\theta)f_0(x) + \theta f_1(x)$$

die zugehörige quadratische Form. Ist $r(f_\theta) = n$ für alle $\theta \in [0,1]$, so ist $p(f_0) = p(f_1)$, d. h. F_0 und F_1 haben dieselbe Anzahl positiver Eigenwerte.

Beweis: Bekanntlich hängen die Nullstellen eines Polynoms stetig von dessen Koeffizienten ab. Daher sind die (reellen) Eigenwerte einer symmetrischen Matrix stetige Funktionen ihrer Koeffizienten. Insbesondere hängen die Eigenwerte von F_θ stetig von θ ab. Wegen der Voraussetzung, dass $r(f_\theta) = n$ für alle $\theta \in [0,1]$ ist, sind für $\theta \in [0,1]$ alle Eigenwerte von F_θ entweder positiv oder negativ. Daher existiert ein $\epsilon > 0$ mit der Eigenschaft, dass $\epsilon \leq \lambda$ für alle positiven Eigenwerte von F_θ, $\theta \in [0,1]$. Weiter ist $p(f_\theta)$ die Anzahl positiver Eigenwerte von F_θ. Wir definieren

$$I := \{\theta \in [0,1] : p(f_\theta) = p(f_0)\}$$

und zeigen, dass $1 \in I$ bzw. $p(f_0) = p(f_1)$ gilt. Wegen $0 \in I$ ist I natürlich nichtleer. Sei $\theta_0 := \sup\{\theta : \theta \in I\}$. Wir überlegen uns, dass $\theta_0 \in I$. Seien $\lambda_1(\theta_0), \ldots, \lambda_n(\theta_0)$ die Eigenwerte von F_{θ_0}. Wegen der stetigen Abhängigkeit der Eigenwerte der Matrix F_θ von θ existiert zu $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ derart, dass es zu jedem $\theta \in [0,1]$ mit $|\theta - \theta_0| \leq \delta$ eine Nummerierung der Eigenwerte $\lambda_1(\theta), \ldots, \lambda_n(\theta)$ von F_θ mit $|\lambda_i(\theta) - \lambda_i(\theta_0)| \leq \frac{1}{2}\epsilon$, $i = 1, \ldots, n$, gibt. Nach Definition von θ_0 existiert ein $\theta \in [\theta_0 - \delta, \theta_0] \cap I$. Nach Definition von I hat F_θ genau $p_0 := p(f_0)$ positive Eigenwerte, dies seien etwa $\lambda_1(\theta), \ldots, \lambda_{p_0}(\theta)$. Wir zeigen, dass auch F_{θ_0} genau p_0 positive Eigenwerte besitzt bzw. $\theta_0 \in I$ gilt. Für $i = 1, \ldots, p_0$ ist

$$\lambda_i(\theta_0) = \underbrace{\lambda_i(\theta)}_{\geq \epsilon} + \lambda_i(\theta_0) - \lambda_i(\theta) \geq \epsilon - |\lambda_i(\theta) - \lambda_i(\theta_0)| \geq \tfrac{1}{2}\epsilon.$$

Daher sind auch $\lambda_1(\theta_0), \ldots, \lambda_{p_0}(\theta_0)$ positiv, es ist also $p(f_{\theta_0}) \geq p_0$. Angenommen, es gibt ein $i \in \{p_0 + 1, \ldots, n\}$ mit $\lambda_i(\theta_0) > 0$. Dann wäre

$$\epsilon \leq \lambda_i(\theta_0) = \lambda_i(\theta) + \lambda_i(\theta_0) - \lambda_i(\theta) \leq \lambda_i(\theta) + \tfrac{1}{2}\epsilon,$$

also auch $\lambda_i(\theta) > 0$, ein Widerspruch. Angenommen, es sei $\theta_0 < 1$. Dann gibt es ein $\theta \in (\theta_0, \theta_0 + \delta] \cap [0,1]$. Wie eben zeigt man, dass auch $\theta \in I$, da F_θ genau die positiven Eigenwerte $\lambda : 1(\theta), \ldots, \lambda_{p_0}(\theta)$ besitzt. Nun ist aber $\theta \in I$ ein Widerspruch zur Definition von θ_0. Damit ist das Lemma schließlich bewiesen. □

8.2.2 Quadratische Formen und Permanenten

Sind im folgenden $c_1, \ldots, c_n \in \mathbb{R}^n$, so sei $\mathrm{per}(c_1, \ldots, c_n)$ die Permanente derjenigen Matrix, die $c_1^T, \ldots, c_n^T$ als Zeilen besitzt[33]. Ist also $c_i = (c_{i1}, \ldots, c_{in})^T$), $i = 1, \ldots, n$, so

[33]Ist $A \in \mathbb{R}^{n\times n}$, so ist offenbar $\mathrm{per}(A) = \mathrm{per}(A^T)$.

sei

$$\mathrm{per}(c_1,\ldots,c_n) := \mathrm{per}\begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}.$$

Im folgenden Lemma steckt die Hauptarbeit.

Lemma 8.6 *Seien $a_1,\ldots,a_{n-1} \in \mathbb{R}^n$ gegeben mit der Eigenschaft, dass $a_i \geq 0$ (komponentenweise) und wenigstens $n+1-i$ Komponenten von a_i positiv sind, $i = 1,\ldots,n-1$.*

1. *Ist $b \in \mathbb{R}^n$ und $\mathrm{per}(a_1,\ldots,a_{n-1},b) = 0$, so ist $\mathrm{per}(a_1,\ldots,a_{n-2},b,b) \leq 0$ und $\mathrm{per}(a_1,\ldots,a_{n-2},b,b) = 0$ genau dann, wenn $b = 0$.*
2. *Die quadratische Form f in n Variablen sei definiert durch*

$$f(x) := \mathrm{per}(a_1,\ldots,a_{n-2},x,x).$$

Dann ist $r(f) = n$ und $p(f) = 1$.

Beweis: Der Beweis erfolgt durch vollständige Induktion nach n, und zwar für beide Teile zusammen. D. h. für den Induktionsanfang zeigen wir, dass die Aussagen 1. und 2. für $n = 2$ richtig sind, nehmen anschließend an, beide Teile seien für $n-1$ richtig und beweisen danach, dass zunächst der zweite und dann auch der erste Teil für n richtig ist.

Sei $n = 2$,

$$a_1 = \begin{pmatrix} a_{11} & a_{12} \end{pmatrix}^T \in \mathbb{R}^2$$

mit $a_{11} > 0$ und $a_{21} > 0$ und

$$b = \begin{pmatrix} b_1 & b_2 \end{pmatrix}^T \in \mathbb{R}^2$$

mit

$$\mathrm{per}(a_1,b) = \mathrm{per}\begin{pmatrix} a_{11} & a_{12} \\ b_1 & b_2 \end{pmatrix} = a_{11}b_2 + a_{21}b_1 = 0$$

seien gegeben. Dann ist

$$\mathrm{per}(b,b) = \mathrm{per}\begin{pmatrix} b_1 & b_2 \\ b_1 & b_2 \end{pmatrix} = 2b_1b_2 = -\frac{a_{12}}{a_{11}}b_1^2 \leq 0.$$

Offensichtlich ist $\mathrm{per}(b,b) = 0$ genau dann wenn $b = 0$. Weiter ist

$$f(x) := \mathrm{per}(x,x) = \mathrm{per}\begin{pmatrix} x_1 & x_2 \\ x_1 & x_2 \end{pmatrix} = 2x_1x_2 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{=:F} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Die 2×2-Matrix F hat die Eigenwerte ± 1, sie ist also nichtsingulär und besitzt genau einen positiven Eigenwert, es ist also $r(f) = 2$ und $p(f) = 1$. Damit ist der Induktionsanfang gelegt.

Nun nehmen wir an, die beiden Aussagen würden für $n-1$ gelten. Zunächst zeigen wir die zweite Aussage für n. Sei

$$f(x) := \mathrm{per}(a_1, \ldots, a_{n-2}, x, x) = \mathrm{per} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n-2,1} & \cdots & a_{n-2,n} \\ x_1 & \cdots & x_n \\ x_1 & \cdots & x_n \end{pmatrix}.$$

Dann ist

$$(*) \qquad f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} x_i x_j$$

mit

$$(**) \qquad f_{ij} := \begin{cases} \mathrm{per}((a_1, \ldots, a_{n-2})(i,j)), & i \neq j, \\ 0, & i = j. \end{cases}$$

Hierbei ist $(a_1, \ldots, a_{n-2})(i,j) \in \mathbb{R}^{(n-2)\times(n-2)}$ die Matrix, die aus $(a_1, \ldots, a_{n-2}) \in \mathbb{R}^{(n-2)\times n}$ durch Streichen der i-ten und j-ten Spalte hervorgeht. Dies erhält man mit Hilfe der Laplace-Entwicklung der Permanente. Für $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ ist nämlich (Entwicklung nach der i-ten Zeile)

$$\mathrm{per}(A) = \sum_{j=1}^{n} a_{ij} \mathrm{per}(A_{ij}), \quad i = 1, \ldots, n.$$

Hierbei ist $A_{ij} \in \mathbb{R}^{(n-1)\times(n-1)}$ die Matrix, die man aus A erhält, indem man die i-te Zeile und die j-te Spalte streicht. Entwickelt man $f(x) = \mathrm{per}(a_1, \ldots, a_{n-2}, x, x)$ nach der letzten Zeile, so erhält man

$$f(x) = \sum_{i=1}^{n} x_i \mathrm{per}((a_1, \ldots, a_{n-2}, x)(i)),$$

wobei $(a_1, \ldots, a_{n-2}, x)(i) \in \mathbb{R}^{(n-1)\times(n-1)}$ aus $(a_1, \ldots, a_{n-2}, x) \in \mathbb{R}^{(n-1)\times n}$ durch Streichen der i-ten Spalte hervorgeht. Entsprechend ist

$$\mathrm{per}((a_1, \ldots, a_{n-2}, x)(i)) = \sum_{\substack{j=1 \\ j\neq i}}^{n} x_j \mathrm{per}((a_1, \ldots, a_{n-2})(i,j)), \quad i = 1, \ldots, n$$

woraus wir durch Einsetzen die Darstellung $(*)$, $(**)$ erhalten. Im Gegensatz zur Behauptung nehmen wir an, es sei $r(f) < n$. Dann ist die zur quadratischen Form f gehörende Matrix $F = (f_{ij})$ singulär, es existiert also $c = (c_1, \ldots, c_n)^T \neq 0$ mit $Fc = 0$

bzw. $\sum_{j=1}^{n} f_{ij}c_j = 0$, $i = 1, \ldots, n$. Für alle $x \in \mathbb{R}^n$ ist dann

$$\begin{aligned}
\operatorname{per}(a_1, \ldots, a_{n-2}, c, x) &= \sum_{j=1}^{n} c_j \operatorname{per}((a_1, \ldots, a_{n-2}, x)(j)) \\
&= \sum_{j=1}^{n} c_j \sum_{\substack{i=1 \\ i \neq j}} x_i \operatorname{per}((a_1, \ldots, a_{n-2})(i,j)) \\
&= \sum_{j=1}^{n} c_j \left(\sum_{i=1}^{n} x_i f_{ij} \right) \\
&= \sum_{i=1}^{n} x_i \Bigg(\underbrace{\sum_{j=1}^{n} f_{ij} c_j}_{=0} \Bigg) \\
&= 0.
\end{aligned}$$

Insbesondere ist $\operatorname{per}(a_1, \ldots, a_{n-2}, c, c) = 0$. Für $j = 1, \ldots, n$ ist weiter

$$\begin{aligned}
\operatorname{per}((a_1, \ldots, a_{n-2}, c)(j)) &= \sum_{\substack{i=1 \\ i \neq j}}^{n} c_i \operatorname{per}((a_1, \ldots, a_{n-2})(i,j)) \\
&= \sum_{i=1}^{n} c_i f_{ij} \\
&= 0.
\end{aligned}$$

Hierbei geht $(a_1, \ldots, a_{n-2}, c)(j) \in \mathbb{R}^{(n-1)\times(n-1)}$ aus $(a_1, \ldots, a_{n-2}, c) \in \mathbb{R}^{(n-1)\times n}$ durch Streichen der j-ten Spalte hervor. Nach Induktionsannahme ist

$$\operatorname{per}((a_1, \ldots, a_{n-3}, c, c)(j)) \leq 0, \quad j = 1, \ldots, n,$$

wobei Gleichheit genau dann gilt, wenn $c_k = 0$ für alle $k \in \{1, \ldots, n\} \setminus \{j\}$. Nun ist

$$\begin{aligned}
0 &= \operatorname{per}((a_1, \ldots, a_{n-2}, c, c)) \\
&= \sum_{j=1}^{n} \underbrace{a_{n-2,j}}_{\geq 0} \underbrace{\operatorname{per}((a_1, \ldots, a_{n-3}, c, c)(j))}_{\leq 0} \\
&\leq 0.
\end{aligned}$$

Also ist $\operatorname{per}((a_1, \ldots, a_{n-3}, c, c)(j)) = 0$, falls $a_{n-2,j} > 0$. Nun hat a_{n-2} nach Voraussetzung mindestens zwei (sogar mindestens drei, was aber hier nicht benötigt wird) positive Komponenten, etwa an den Stellen $j_1 \neq j_2$. Also ist $c_k = 0$ für alle $k \in \{1, \ldots, n\} \setminus \{j_1\}$ und $c_k = 0$ für alle $k \in \{1, \ldots, n\} \setminus \{j_2\}$. Insgesamt ist $c_k = 0$ für alle $k \in \{1, \ldots, n\}$ bzw. $c = 0$. Hierdurch haben wir einen Widerspruch zu $c \neq 0$ erhalten und $r(f) = n$ nachgewiesen. Damit ist die erste Hälfte des zweiten Teils bewiesen. Zum Beweis der zweiten Hälfte definieren wir die quadratischen Formen f_0 und f_1 in n Variablen durch

$$f_0(x) := \operatorname{per}(a_1, a_2, \ldots, a_{n-2}, x, x), \qquad f_1(x) := \operatorname{per}(e, a_2, \ldots, a_{n-2}, x, x),$$

wobei $e := (1, \ldots, 1)^T \in \mathbb{R}^n$. Anschließend definieren wir für $\theta \in [0,1]$ die quadratische Form f_θ durch

$$\begin{aligned} f_\theta(x) &:= (1-\theta) f_0(x) + \theta f_1(x) \\ &= \operatorname{per}((1-\theta)a_1, a_2, \ldots, a_{n-2}, x, x) + \operatorname{per}(\theta e, a_2, \ldots, a_{n-2}, x, x) \\ &= \operatorname{per}((1-\theta)a_1 + \theta e, a_2, \ldots, a_{n-2}, x, x), \end{aligned}$$

wobei wir ausgenutzt haben, dass die Permanente multilinear ist. Aus der schon bewiesenen ersten Hälfte des zweiten Teils folgt, dass $r(f_\theta) = n$ für alle $\theta \in [0,1]$. Wegen Lemma 8.5 ist $p(f_0) = p(f_1)$. Setzt man jetzt

$$f_2(x) := \operatorname{per}(e, e, a_3, \ldots, a_{n-2}, x, x)$$

und

$$g_\theta(x) := (1-\theta) f_1(x) + \theta f_2(x),$$

so ist

$$g_\theta(x) = \operatorname{per}(e, (1-\theta)a_2 + \theta e, \ldots, a_{n-2}, x, x),$$

und damit $r(g_\theta) = n$ für alle $\theta \in [0,1]$ und $p(f_1) = p(f_2)$. Also ist $p(f_0) = p(f_2)$. In dieser Weise kann man fortfahren. Mit

$$f_{n-2}(x) := \operatorname{per}(e, e, \ldots, e, x, x)$$

erhält man $p(f_0) = p(f_{n-2})$. Wegen $\operatorname{per}((e, \ldots, e)(i,j)) = (n-2)!$ für $i \neq j$ und der Darstellung $(*), (**)$ ist

$$\frac{1}{(n-2)!} f_{n-2}(x) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i x_j = x^T F x,$$

wobei F die $n \times n$-Matrix ist, die in der Diagonalen 0 und außerhalb 1 als Eintrag besitzt. Es ist also $F = ee^T - I$. Daher besitzt F den positiven Eigenwert $n-1$ mit zugehörigem Eigenvektor e und den $(n-1)$-fachen Eigenwert -1 mit zugehörigen Eigenvektoren aus dem $(n-1)$-dimensionalen linearen Raum span $(e)^\perp$. Also ist $p(f_0) = p(f_{n-2}) = 1$ und der Induktionsbeweis für den zweiten Teil des Lemmas ist abgeschlossen.

Jetzt zeigen wir die erste Aussage für n. Die Voraussetzungen an die Anzahl positiver Komponenten von $a_1, \ldots, a_{n-1}$ implizieren, dass $c := f(a_{n-1}) > 0$, wobei

$$f(x) := \operatorname{per}(a_1, \ldots, a_{n-2}, x, x) = x^T F x = \sum_{i,j=1}^{n} f_{ij} x_i x_j.$$

Hierzu müssen wir uns die Existenz einer Permutation $\sigma \in S_n$ mit $a_{i\sigma(i)} > 0$, $i = 1, \ldots, n-1$, und $a_{n-1,\sigma(n)} > 0$ überlegen. Nun hat a_{n-1} mindestens zwei positive Komponenten. Man setze $\sigma(n)$ gleich dem Index einer der positiven Komponenten und $\sigma(n-1)$ gleich dem anderen. Der Vektor a_{n-2} hat mindestens drei positive Komponenten, man setze $\sigma(n-2)$ gleich einem dieser Indizes, der noch nicht aufgetreten ist.

In dieser Weise kann man fortfahren und erhält $f(a_{n-1}) > 0$. Da eine Permanente sich bei einer Permutation der Spalten (oder Zeilen) nicht ändert, können wir annehmen, dass $a_{n-1,1} > 0$. Jetzt wenden wir Lemma 8.2 mit $a := a_{n-1}$ an. Hiernach ist

$$f(x_1, \ldots, x_n) = cy_1^2 + g(y_2, \ldots, y_n),$$

wobei (siehe den Beweis von Lemma 8.2)

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} := \begin{pmatrix} (Fa)_1/c & (Fa)_2/c & \cdots & (Fa)_n/c \\ -a_2/a_1 & & & \\ \vdots & & I_{n-1} & \\ -a_n/a_1 & & & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Wegen der schon bewiesenen Aussage $p(f) = 1$ ist

$$g(y_2, \ldots, y_n) \leq 0 \quad \text{für alle } y_2, \ldots, y_n,$$

wobei Gleichheit genau dann gilt, wenn $y_2 = \cdots = y_n = 0$. Insbesondere ist

$$\begin{aligned} y_1 &= \frac{1}{c} \sum_{i=1}^{n} (Fa)_i x_i \\ &= \frac{1}{c} \sum_{i=1}^{n} x_i \sum_{j=1}^{n} f_{ij} a_{n-1,j} \\ &= \frac{1}{c} \sum_{i=1}^{n} x_i \sum_{\substack{j=1 \\ j \neq i}} \operatorname{per}((a_1, \ldots, a_{n-2})(i,j)) a_{n-1,j} \\ &= \frac{1}{c} \sum_{i=1}^{n} x_i \operatorname{per}((a_1, \ldots, a_{n-2}, a_{n-1})(i)) \\ &= \frac{1}{c} \operatorname{per}(a_1, \ldots, a_{n-1}, x). \end{aligned}$$

Jetzt kommen wir zum Schluss des Induktionsbeweises für den ersten Teil. Sei $b \in \mathbb{R}^n$ mit $\operatorname{per}(a_1, \ldots, a_{n-1}, b) = 0$. Dann ist wegen der gerade eben bewiesenen Aussage $f(b) = \operatorname{per}(a_1, \ldots, a_{n-2}, b, b) \leq 0$, wobei Gleichheit genau dann gilt, wenn $b = 0$.
Damit ist der Beweis des Lemmas abgeschlossen. □

Lemma 8.6 ist das wichtigste Hilfsmittel zum Beweis des folgenden Satzes, der von A. D. Aleksandrov (1938) stammt.

Satz 8.7 *Seien $a_1, \ldots, a_{n-1} \in \mathbb{R}^n$ gegeben mit der Eigenschaft, dass $a_i \geq 0$ (komponentenweise) und wenigstens $n+1-i$ Komponenten von a_i positiv sind, $i = 1, \ldots, n-1$, weiter sei $a_n \in \mathbb{R}^n$. Dann ist*

$$(*) \quad \operatorname{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_n)^2 \geq \operatorname{per}(a_1, \ldots, a_{n-1}, a_{n-1}) \cdot \operatorname{per}(a_1, \ldots, a_{n-2}, a_n, a_n).$$

In der Ungleichung $(*)$ gilt Gleichheit genau dann, wenn $a_n = \lambda a_{n-1}$ mit $\lambda \in \mathbb{R}$.

Beweis: Wieder ist wegen der Voraussetzung über die Anzahl positiver Komponenten von $a_1, \ldots, a_{n-1}$ gesichert, dass $\mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1}) > 0$. Daher ist

$$\lambda := \frac{\mathrm{per}(a_1, \ldots, a_{n-1}, a_n)}{\mathrm{per}(a_1, \ldots, a_{n-1}, a_{n-1})}$$

wohldefiniert. Wir setzen $b := a_n - \lambda a_{n-1}$. Wegen der Multilinearität der Permanente ist $\mathrm{per}(a_1, \ldots, a_{n-1}, b) = 0$. Wegen des ersten Teils von Lemma 8.6 ist

$$\begin{aligned}
0 &\geq \mathrm{per}(a_1, \ldots, a_{n-2}, b) \\
&= \mathrm{per}(a_1, \ldots, a_{n-2}, b, a_n) - \lambda \mathrm{per}(a_1, \ldots, a_{n-2}, b, a_{n-1}) \\
&= \mathrm{per}(a_1, \ldots, a_{n-2}, a_n, a_n) - 2\lambda \mathrm{per}(a_1, \ldots, a_n) \\
&\qquad + \lambda^2 \mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1}) \\
&= \mathrm{per}(a_1, \ldots, a_{n-2}, a_n, a_n) - 2\lambda \mathrm{per}(a_1, \ldots, a_{n-1}, b + \lambda a_{n-1}) \\
&\qquad + \lambda^2 \mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1}) \\
&= \mathrm{per}(a_1, \ldots, a_{n-2}, a_n, a_n) - \lambda^2 \mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1}) \\
&= \mathrm{per}(a_1, \ldots, a_{n-2}, a_n, a_n) - \frac{\mathrm{per}(a_1, \ldots, a_n)^2}{\mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1})}.
\end{aligned}$$

Dies ist genau die Ungleichung $(*)$ und Gleichheit gilt wegen des ersten Teils von Lemma 8.6 genau dann wenn $b = 0$ bzw. $a_n = \lambda a_{n-1}$. □

Mit Hilfe eines Grenzprozesses befreien wir uns der Voraussetzung über die Anzahl positiver Komponenten von $a_1, \ldots, a_{n-1}$.

Korollar 8.8 *Seien $a_1, \ldots, a_{n-1} \in \mathbb{R}^n$ nichtnegative Vektoren und $a_n \in \mathbb{R}^n$ beliebig. Dann gilt die Ungleichung $(*)$ in Satz 8.7.*

Beweis: Man ersetze a_i durch $a_i + \epsilon e$, $i = 1, \ldots, n-1$, mit $\epsilon > 0$ und $e := (1, \ldots, 1)^T$, wende Satz 8.7 an und mache den Grenzübergang $\epsilon \to 0+$. □

8.2.3 Doppelt stochastische Matrizen, Heiratssatz

Eine $n \times n$-Matrix heißt bekanntlich *doppelt stochastisch*, wenn alle ihrer Einträge nichtnegativ sind und sämtliche Zeilen- und Spaltensummen gleich 1 sind.

Definition 8.9 Eine doppelt stochastische Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ besitzt den *Gleichgewichtspunkt* $x = (x_j) \in \mathbb{R}^n$, wenn $Ax = x$ bzw. $\sum_{j=1}^n a_{ij} x_j = x_i$, $i = 1, \ldots, n$. Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ heißt *zerlegbar* (*decomposable*), wenn $\{1, \ldots, n\}$ durch nichtleere Teilmengen S und T so partitioniert werden kann, dass $a_{ij} = 0$ für alle $(i, j) \in S \times T$.

Dann gilt:

Lemma 8.10 *Sei $A \in \mathbb{R}^{n \times n}$ doppelt stochastisch. Besitzt A einen Gleichgewichtspunkt $x = (x_j) \in \mathbb{R}^n$, dessen Komponenten nicht alle gleich sind, so ist A zerlegbar.*

Beweis: Ist x ein Gleichgewichtspunkt für A, so ist auch $x + ce$ mit $c \in \mathbb{R}$ und $e := (1, \ldots, 1)^T$ ein Gleichgewichtspunkt für A, da

$$A(x + ce) = Ax + cAe = x + ce,$$

wobei wir nur ausgenutzt haben, dass $\sum_{j=1}^{n} a_{ij} = 1$, $i = 1, \ldots, n$. Daher können wir annehmen, x sei nichtnegativ (aber nicht der Nullvektor). Man setze

$$S := \{i \in \{1, \ldots, n\} : x_i = 0\}, \qquad T := \{i \in \{1, \ldots, n\} : x_i > 0\}.$$

Aus $\sum_{j=1}^{n} a_{ij} x_j = x_i$, $i = 1, \ldots, n$, folgt dann $a_{ij} = 0$ für $(i, j) \in S \times T$, d. h. A ist zerlegbar. □

Lemma 8.11 *Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ eine doppelt stochastische Matrix. Sei A zerlegbar, durch die nichtleeren Teilmengen S, T sei eine Partition von $\{1, \ldots, n\}$ mit $a_{ij} = 0$ für $(i, j) \in S \times T$ gegeben. Dann gilt:*

$$a_{ij} > 0 \Longrightarrow i, j \in S \text{ oder } i, j \in T.$$

Beweis: Es ist

$$\begin{aligned}
\sum_{i \in S} \sum_{j \in S} a_{ij} &= \sum_{i \in S} \underbrace{\sum_{j=1}^{n} a_{ij}}_{=1} \\
&= \sum_{i \in S} 1 \\
&= \sum_{j \in S} 1 \\
&= \sum_{j \in S} \sum_{i=1}^{n} a_{ij} \\
&= \sum_{j \in S} \sum_{i \in S} a_{ij} + \sum_{j \in S} \sum_{i \in T} a_{ij}
\end{aligned}$$

und folglich

$$\sum_{i \in T} \sum_{j \in S} a_{ij} = 0 \quad \text{bzw.} \quad a_{ij} = 0 \text{ für } (i, j) \in T \times S.$$

Hieraus folgt die Behauptung. □

Jetzt erinnern wir an den *Heiratssatz*, auf den wir in J. Werner (2013, Abschnitt 29) eingegangen sind. Gestellt ist das folgende Problem:

- Gegeben sei eine Menge U von m Damen und eine Menge W von n 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 beide einer gegenseitigen

langfristigen Beziehung, z. B. einer Heirat, zustimmen würden[34]. 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.

Dann gilt der folgende

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.*

Einen Beweis findet man z. B. bei J. Werner (2013). Als Spezialfall mit $m = n$ erhält man das folgende Lemma (siehe D. E. Knuth (1981, Lemma 3.2), wir haben lediglich die Rollen der Damen und Herren vertauscht).

Lemma 8.12 *Man betrachte eine Menge von U von n Damen und eine Menge V von n Herren. Für jedes Paar $(u, v) \in U \times V$ von Herren und Frauen sei bekannt, ob sie befreundet oder nicht befreundet (compatible or incompatible) sind. Wenn es keinen zulässigen Heiratsplan für alle n Damen und n Herren gibt, wenn es also keine bijektive Abbildung $\sigma\colon U \longrightarrow V$ gibt derart, dass das Paar $(u, \sigma(u)) \in U \times V$ für alle $u \in U$ miteinander befreundet ist, so gibt es eine Menge $S \subset U$ von $k = |S|$ Damen, die mit lediglich $k - 1$ Herren befreundet ist.*

Beweis: Die Aussage des Lemmas folgt offenbar sofort aus dem Heiratssatz. □

Bei D. E. Knuth (1981) wird Lemma 8.12 dazu benutzt, den Satz von Birkhoff-Neumann zu beweisen, dass nämlich eine $n \times n$-Matrix genau dann doppelt stochastisch ist, wenn sie eine Konvexkombination von Permutationsmatrizen ist. Hierauf wollen wir nicht mehr eingehen, da der Beweis bei J. Werner (2013, Abschnitt 29) im wesentlichen hiermit übereinstimmt.

8.2.4 Minimale Matrizen

Eine Lösung der Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \operatorname{per}(A), \quad A \in \Omega_n,$$

wobei Ω_n die Menge der doppelt stochastischen $n \times n$-Matrizen ist, nennen wir eine *minimale Matrix*. Da Ω_n in $\mathbb{R}^{n \times n}$ kompakt und die Permanente stetig von ihrem Argument abhängt, ist die Existenz einer minimalen Matrix gesichert. Die Permanente einer doppelt stochastischen Matrix ist positiv, wie wir uns zu Beginn des Abschnitts überlegt haben. Der (Optimal-) Wert der Aufgabe (P) ist also positiv.

[34] 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.

Schon wiederholt wurde die Laplace-Entwicklung einer Matrix bezüglich einer Zeile oder Spalte ausgenutzt. Ist $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ so bezeichne $A_{ij} \in \mathbb{R}^{(n-1)\times(n-1)}$ die Matrix, die man aus A durch Streichen der i-ten Zeile und der j-ten Spalte erhält. Dann gilt:

$$\mathrm{per}(A) = \sum_{j=1}^{n} a_{ij}\mathrm{per}(A_{ij}), \quad i = 1, \ldots, n,$$

und

$$\mathrm{per}(A) = \sum_{i=1}^{n} a_{ij}\mathrm{per}(A_{ij}), \quad j = 1, \ldots, n.$$

Stört man eine Matrix A durch eine kleine Störung ϵB so wirkt sich das folgendermaßen auf die Permanente aus:

$$\mathrm{per}(A + \epsilon B) = \mathrm{per}(A) + \epsilon \sum_{i,j=1}^{n} b_{ij}\mathrm{per}(A_{ij}) + O(\epsilon^2).$$

Hierbei ist $O(\epsilon^2) = \epsilon^2 p(\epsilon, A, B)$, wobei p ein Polynom in ϵ und den Werten von A und B ist. Insbesondere ist

$$\lim_{\epsilon\to 0+} \frac{\mathrm{per}(A + \epsilon B) - \mathrm{per}(A)}{\epsilon} = \sum_{i,j=1}^{n} b_{ij}\mathrm{per}(A_{ij}).$$

Ist $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine doppelt stochastische Matrix, so heißt $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ eine *gültige Modifikation* (*valid modification*) für A, wenn die Zeilen- und Spaltensummen von B alle gleich Null sind und $b_{ij} \geq 0$ falls $a_{ij} = 0$. Dies ist gleichbedeutend damit, dass mit A auch $A + \epsilon B$ für alle hinreichend kleinen $\epsilon > 0$ doppelt stochastisch ist[35].

Lemma 8.13 *Ist $A \in \mathbb{R}^{n\times n}$ eine minimale Matrix und $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ eine gültige Modifikation für A, so ist*

$$\sum_{i,j=1}^{n} b_{ij}\mathrm{per}(A_{ij}) \geq 0.$$

Beweis: Für alle hinreichend kleinen $\epsilon > 0$ ist

$$\frac{\mathrm{per}(A + \epsilon B) - \mathrm{per}(A)}{\epsilon} \geq 0.$$

Mit $\epsilon \to 0+$ folgt die Behauptung. □

Lemma 8.14 *Eine minimale Matrix ist nicht zerlegbar.*

Beweis: Sei $A \in \mathbb{R}^{n\times n}$ eine minimale Matrix. Im Gegensatz zur Behauptung nehmen wir an, dass A zerlegbar sei, es also eine Partition von $\{1, \ldots, n\}$ durch nichtleere Teilmengen S, T mit $a_{ij} = 0$ für $(i, j) \in S \times T$ gibt. Wir wissen, dass die Permanente

[35] In der Optimierung würde man statt von einer gültigen Modifikation der doppelt stochastischen Matrix A von einer *zulässigen Richtung* in A bezüglich der Menge Ω_n der doppelt stochastischen Matrizen sprechen.

einer doppelt stochastischen Matrix positiv ist. Daher existiert eine Permutation $\pi \in S_n$ mit

$$a_{i\pi(i)} > 0, \quad i = 1, \ldots, n.$$

Man wähle $(s,t) \in S \times T$ beliebig und definiere $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ als eine Matrix deren Einträge sämtlich gleich Null sind außer an den Positionen $(s, \pi(s))$, $(t, \pi(t))$, $(s, \pi(t))$ sowie $(t, \pi(s)$. Dort seien die Einträge gegeben durch

$$b_{s\pi(s)} = b_{t\pi(t)} = -1, \qquad b_{s\pi(t)} = b_{t\pi(s)} = 1.$$

Wegen $a_{s\pi(s)} > 0$ und $s \in S$ ist $\pi(s) \in S$. Entsprechend ist $\pi(t) \in T$. Daher ist B eine gültige Modifikation für A. Eine Anwendung von Lemma 8.13 ergibt

$$-\mathrm{per}(A_{s\pi(s)}) - \mathrm{per}(A_{t\pi(t)}) + \mathrm{per}(A_{s\pi(t)}) + \mathrm{per}(A_{t\pi(s)}) \geq 0.$$

Das ergibt einen Widerspruch. Denn $\mathrm{per}(A_{s\pi(s)})$ und $\mathrm{per}(A_{t\pi(t)})$ sind positiv, während $\mathrm{per}(A_{s\pi(t)}) = \mathrm{per}(A_{t\pi(s)}) = 0$ ist. Denn wegen $a_{i\pi(i)} > 0$, $i = 1, \ldots, n$, ist insbesondere $\prod_{\substack{i=1\\ i\neq s}}^{n} a_{i\pi(i)} > 0$. Folglich ist mindestens einer der (nichtnegativen) Summanden zur Berechnung von $\mathrm{per}(A_{s\pi(s)})$ sogar positiv und daher $\mathrm{per}(A_{s\pi(s)}) > 0$. Entsprechend ist auch $\mathrm{per}(A_{t\pi(t)}) > 0$. Nun zeigen wir, dass $\mathrm{per}(A_{t\pi(s)}) = 0$. Sei $k := |S|$ und $i \in S$. Dann ist $(a_{ij})_{j\neq\pi(s)}$ eine Zeile in $A_{t\pi(s)}$ und es gilt

$$j \in \{1, \ldots, n\} \setminus \{\pi(s)\},\ a_{ij} > 0 \Longrightarrow j \in S \setminus \{\pi(s)\}.$$

Also gibt es zu den k zu S gehörenden Zeilen von $A_{t\pi(s)}$ jeweils höchstens $k-1$ von Null verschiedene Komponenten. Hieraus folgt, dass $\mathrm{per}(A_{t\pi(s)}) = 0$. Denn angenommen, es gibt eine bijektive Abbildung

$$\tau\colon \{1, \ldots, n\} \setminus \{t\} \longrightarrow \{1, \ldots, n\} \setminus \{\pi(s)\}$$

mit $\prod_{\substack{i=1\\ i\neq t}}^{n} a_{i\tau(i)} > 0$. Dann ist auch $\prod_{i\in S} a_{i\tau(i)} > 0$. Da es aber zu jeder der k Zeilen $i \in S$ höchstens $k-1$ positive Komponenten gibt, ist dies ein Widerspruch. Der Beweis für $\mathrm{per}(A_{s\pi(t)}) = 0$ kann entsprechend geführt werden. Das Lemma ist bewiesen. □

Bemerkung: Eine unmittelbare und nützliche Folgerung aus Lemma 8.14 ist, dass eine minimale Matrix keine 1 enthalten kann. Denn ist dies doch der Fall, so können wir o. B. d. A. (nach eventuellen Permutationen von Zeilen und Spalten) annehmen, es sei $a_{11} = 1$. Dann hätte A die Form

$$A = \begin{pmatrix} 1 & 0^T \\ 0 & A_{n-1} \end{pmatrix}$$

mit $A_{n-1} \in \Omega_{n-1}$. Mit $S := \{1\}$ und $T := \{2, \ldots, n\}$ wäre $a_{ij} = 0$ für alle $(i,j) \in S \times T$. Dies ist ein Widerspruch dazu, dass A nicht zerlegbar ist. In einer minimalen Matrix hat also jede Zeile und jede Spalte mindestens zwei positive Einträge. □

Lemma 8.15 *Ist $A \in \mathbb{R}^{n\times n}$ eine minimale Matrix, so ist $\mathrm{per}(A_{ij}) > 0$ für alle i, j.*

Beweis: Gegeben sei $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$. Angenommen, es ist im Gegensatz zur Behauptung $\operatorname{per}(A_{ij}) = 0$. Wir wollen Lemma 8.12, das Korollar zum Heiratssatz, benutzen und setzen

$$U := \{1,\ldots,n\} \setminus \{i\}, \qquad V := \{1,\ldots,n\} \setminus \{j\}$$

für die Menge der Damen bzw. Herren. Weiter sagen wir ein Paar $(u,v) \in U \times V$ sei miteinander befreundet, wenn $a_{uv} > 0$. Wegen der Annahme $\operatorname{per}(A_{ij}) = 0$ gibt es keine bijektive Abbildung $\sigma\colon U \longrightarrow V$ mit $a_{u\sigma(u)} > 0$ für alle $u \in U$, d. h. es gibt keinen zulässigen Heiratsplan für die $n-1$ Damen und die $n-1$ Herren. Aus Lemma 8.12 folgt, dass es eine k-elementige Menge $S \subset U$ von Damen (bzw. Zeilen) gibt derart, dass diese nur mit $k-1$ Herren befreundet sind bzw. alle von Null verschiedenen Einträge in den zu S gehörenden Zeilen von A_{ij} in nur $k-1$ Spalten auftreten. Man setze $T := \{1,\ldots,n\} \setminus S$. Man beachte, dass $T \neq \emptyset$ da $i \notin S$ bzw. $i \in T$. Jetzt können wir die Zeilen von A so permutieren (hierbei verändert die Permanente von A sich nicht), dass die zu S gehörenden Zeilen die ersten k Zeilen und die zu T gehörenden Zeilen die restlichen $n-k$ Zeilen sind. Anschließend permutiere man die Spalten der so entstandenden Matrix so, dass in die ersten k Spalten die Spalte j sowie die $k-1$ Spalten stehen, in denen die Zeilen aus S von Null verschiedene Elemente besitzen. Die so erhaltene Matrix nennen wir A'. Da A' ebenfalls doppelt stochastisch ist und $\operatorname{per}(A') = \operatorname{per}(A)$ gilt, ist A' ebenfalls eine minimale Matrix. Da A' die Form

$$A' \;=\; \begin{array}{r} k\,\{ \\ n-k\,\{ \end{array} \left(\begin{array}{c|c} * & 0 \\ \hline * & * \end{array}\right) \qquad \begin{array}{cc} \underbrace{}_{k} & \underbrace{}_{n-k} \end{array}$$

hat, ist A' zerlegbar. Dies ist ein Widerspruch zu Lemma 8.14. □

Von dem folgenden Satz schreibt D. E. Knuth (1981):

- The next property of minimal matrices is the key to everything that follows; it was first proved by Marcus and Newman in 1959.

Einen Beweis findet man auch bei H. Minc (1978, S. 74). Wir werden dem Beweis von D. London (1971, Theorem 1) folgen.

Satz 8.16 *Ist $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine minimale Matrix, so ist*

$$\operatorname{per}(A_{ij}) = \operatorname{per}(A) \quad \textit{für alle } (i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\} \textit{ mit } a_{ij} > 0$$

und

$$\operatorname{per}(A_{ij}) \geq \operatorname{per}(A) \quad \textit{für alle } (i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\} \textit{ mit } a_{ij} = 0.$$

Beweis: Die Permanente einer doppelt stochastischen Matrix ist positiv und ändert sich nicht durch Permutieren der Spalten. Daher können wir o. B. d. A. annehmen, dass die Diagonalelemente a_{ii} von A, $i = 1,\ldots,n$, sämtlich positiv sind. Dies wird erst ziemlich zum Schluss des Beweises benutzt.

Wir definieren $f\colon \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$ durch $f(X) := \mathrm{per}(X)$. Da A eine minimale Matrix ist und die Menge Ω_n der doppelt stochastischen $n \times n$-Matrizen konvex ist, ist

$$f(A + \epsilon(B - a)) - f(A) \geq 0 \quad \text{für alle } B \in \Omega_n \text{ und alle } \epsilon \in [0, 1]$$

und daher

$$\begin{aligned} f'(A; B - A) &:= \lim_{\epsilon \to 0+} \frac{f(A + \epsilon(B - A)) - f(A)}{\epsilon} \\ &= \sum_{i,j=1}^{n} (b_{ij} - a_{ij})\mathrm{per}(A_{ij}) \\ &= \sum_{i,j=1}^{n} b_{ij}\mathrm{per}(A_{ij}) - \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{n} a_{ij}\mathrm{per}(A_{ij})}_{=\mathrm{per}(A)} \\ &= \sum_{i,j=1}^{n} b_{ij}\mathrm{per}(A_{ij}) - n\mathrm{per}(A) \\ &\geq 0 \end{aligned}$$

für jedes $B \in \Omega_n$. Für $B = A$ gilt hier das Gleichheitszeichen. Folglich ist A Lösung der linearen Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\mathrm{per}(A_{ij}) \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad b_{ij} \geq 0 \quad (i, j = 1, \ldots, n), \\ \displaystyle\sum_{j=1}^{n} b_{ij} = 1 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} b_{ij} = 1 \quad (j = 1, \ldots, n) \end{cases}$$

und der zugehörige Minimalwert ist $\min(\mathrm{P}) = n\mathrm{per}(A)$. Dies ist eine lineare Optimierungsaufgabe vom Typ des Transportproblems. Die zugehörige *duale lineare Optimierungsaufgabe* ist (siehe z. B. J. Werner (2000, S. 12))

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \displaystyle\sum_{i=1}^{n} \lambda_i + \sum_{j=1}^{n} \mu_j \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad \lambda_i + \mu_j \leq \mathrm{per}(A_{ij}), \quad (i, j = 1, \ldots, n). \end{cases}$$

Die Dualitätstheorie für lineare Optimierungsaufgaben liefert uns, dass mit (P) auch die duale Aufgabe (D) lösbar ist und die Optimalwerte übereinstimmen. D. h. es existiert ein Paar $(\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^n$ mit

$$\lambda_i + \mu_j \leq \mathrm{per}(A_{ij}), \quad (i, j = 1, \ldots, n)$$

und $\lambda^T e + \mu^T e = n\mathrm{per}(A)$, wobei die Komponenten des Vektors $e \in \mathbb{R}^n$ sämtlich gleich Eins sind. Nun ist

$$n\mathrm{per}(A) = \sum_{i=1}^{n} a_{ij}\mathrm{per}(A_{ij}) \geq \sum_{i,j=1}^{n} a_{ij}(\lambda_i + \mu_j) = \lambda^T e + \mu^T e = n\mathrm{per}(A).$$

Hieraus folgt $\lambda_i + \mu_j = \mathrm{per}(A_{ij})$ für alle $(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\}$ mit $a_{ij} > 0$. In der Sprache der Optimierung ist dies die Gleichgewichtsbedingung (complementary slackness condition). Offenbar gilt

$$a_{ij}\mathrm{per}(A_{ij}) = a_{ij}(\lambda_i + \mu_j), \quad i,j = 1,\dots,n.$$

Durch Aufsummieren über j bzw. i erhalten wir

$$\mathrm{per}(A) = \sum_{j=1}^{n} a_{ij}\mathrm{per}(A_{ij}) = \sum_{j=1}^{n} a_{ij}(\lambda_i + \mu_j) = \lambda_i + \sum_{j=1}^{n} a_{ij}\mu_j, \quad i = 1,\dots,n,$$

bzw.

$$\mathrm{per}(A) = \sum_{i=1}^{n} a_{ij}\mathrm{per}(A_{ij}) = \sum_{i=1}^{n} a_{ij}(\lambda_i + \mu_j) = \mu_j + \sum_{i=1}^{n} a_{ij}\lambda_i, \quad j = 1,\dots,n,$$

wobei wir jeweils ausgenutzt haben, dass $A = (a_{ij})$ doppelt stochastisch ist. Mit dem Vektor $e \in \mathbb{R}^n$, dessen Komponenten sämtlich gleich Eins sind, ist dann

$$\mathrm{per}(A)e = \lambda + A\mu, \qquad \mathrm{per}(A)e = \mu + A^T\lambda.$$

Eine Multiplikation der ersten Gleichung mit A^T liefert unter Benutzung von $A^Te = e$, dass

$$\mathrm{per}(A)e = A^T\lambda + A^TA\mu.$$

Mit Hilfe der zweiten Gleichung erhält man

$$A^TA\mu = \mu.$$

Entsprechend ist

$$AA^T\lambda = \lambda.$$

Wegen Lemma 8.14 ist die doppelt stochastische Matrix A als minimale Matrix nicht zerlegbar. Dann sind auch AA^T und A^TA doppelt stochastisch und nicht zerlegbar. Ersteres erkennt man an

$$\sum_{j=1}^{n}(AA^T)_{ij} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{ik}a_{jk} = \sum_{k=1}^{n} a_{ik} \underbrace{\sum_{j=1}^{n} a_{jk}}_{=1} = \sum_{k=1}^{n} a_{ik} = 1, \quad i = 1,\dots,n,$$

und entsprechenden Rechnungen. Um nachzuweisen, dass mit A auch AA^T nicht zerlegbar ist, nehmen wir das Gegenteil an, dass es also nichtleere Teilmengen S, T von $\{1,\dots,n\}$ mit $S \cup T = \{1,\dots,n\}$ und $(AA^T)_{ij} = 0$ für alle $(i,j) \in S \times T$ gibt. Wegen

$$(AA^T)_{ij} = \sum_{k=1}^{n} a_{ik}a_{jk}$$

ist für $(i,j) \in S \times T$ dann

$$a_{ik}a_{jk} = 0, \quad k = 1,\dots,n.$$

Da wir zu Beginn des Beweises nachgewiesen haben, dass o. B. d. A. alle Diagonalelemente von A positiv sind, ist $a_{ij} = 0$ für alle $(i,j) \in S \times T$. Dies ist ein Widerspruch zur Unzerlegbarkeit von A. Ähnlich kann nachgewiesen werden, dass auch $A^T A$ nicht zerlegbar ist. Wegen

$$AA^T\lambda = \lambda, \quad A^T A\mu = \mu$$

sind λ bzw. μ Gleichgewichtspunkte der nicht zerlegbaren Matrizen AA^T bzw. $A^T A$. Das Lemma 8.17 impliziert, dass die Komponenten von Gleichgewichtspunkten unzerlegbarer Matrizen alle gleich sein müssen. Daher existieren $c, d \in \mathbb{R}$ mit $\lambda = ce$ und $\mu = de$. Wegen $\lambda_i + \mu_j \leq \text{per}(A_{ij})$ für alle (i,j) und $\lambda_i + \mu_j = \text{per}(A_{ij})$, falls $a_{ij} > 0$, ist

$$c + d = \text{per}(A_{ij}) \text{ falls } a_{ij} > 0, \quad c + d \leq \text{per}(A_{ij}) \text{ falls } a_{ij} = 0.$$

Daher ist

$$\text{per}(A) = \sum_{j=1}^{n} a_{ij}\text{per}(A_{ij}) = \sum_{j=1}^{n} a_{ij}(c+d) = c + d.$$

Der Satz ist bewiesen. □

8.2.5 Egorychev's Theorem

In dem folgenden Lemma (siehe G. P. Egorychev (1981, Theorem 2) wird zum ersten Mal Satz 8.7 bzw. das zugehörige Korollar angewandt.

Lemma 8.17 *Ist $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine minimale Matrix, so ist* $\text{per}(A_{ij}) = \text{per}(A)$ *für alle* $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\}$.

Beweis: Da wir Zeilen und Spalten vertauschen können, ist o. B. d. A. $(i,j) = (n,n)$. Wegen Satz 8.16 ist $\text{per}(A_{nn}) \geq \text{per}(A)$. Im Widerspruch zur Behauptung nehmen wir an, es sei $\text{per}(A_{nn}) > \text{per}(A)$. Dies ist wegen Satz 8.16 nur möglich, wenn $a_{nn} = 0$. Die n-te Spalte von A muss einen positiven Eintrag enthalten, o. B. d. A. sei $a_{(n-1)n} > 0$. Bezeichne $a_j^T \in \mathbb{R}^n$ die j-te Zeile von A, $j = 1,\ldots,n$. Wegen des Korollars zu Satz 8.7 ist dann

$$\begin{aligned} \text{per}(A)^2 &= \text{per}(a_1,\ldots,a_{n-1},a_n)^2 \\ &\geq \text{per}(a_1,\ldots,a_{n-2},a_{n-1},a_{n-1}) \cdot \text{per}(a_1,\ldots,a_{n-2},a_n,a_n). \end{aligned}$$

Es ist aber

$$\text{per}(a_1,\ldots,a_{n-2},a_{n-1},a_{n-1}) = \sum_{j=1}^{n} a_{(n-1)j}\text{per}(A_{nj}) > \sum_{j=1}^{n} a_{(n-1)j}\text{per}(A) = \text{per}(A),$$

da

$$\begin{aligned} a_{(n-1)n}\text{per}(A_{nn}) &> a_{(n-1)n}\text{per}(A), \\ a_{(n-1)j}\text{per}(A_{nj}) &\geq a_{(n-1)j}\text{per}(A), \quad j = 1,\ldots,n-1. \end{aligned}$$

Weiter ist

$$\text{per}(a_1,\ldots,a_{n-2},a_n,a_n) = \sum_{j=1}^{n} a_{nj}\text{per}(A_{nj}) \geq \sum_{j=1}^{n} a_{nj}\text{per}(A) = \text{per}(A),$$

da

$$a_{nj}\mathrm{per}(A_{nj}) \geq a_{nj}\mathrm{per}(A), \quad j = 1, \ldots, n.$$

Insgesamt erhalten wir den Widerspruch $\mathrm{per}(A)^2 > \mathrm{per}(A)^2$, womit $\mathrm{per}(A_{nn}) = \mathrm{per}(A)$ und das ganze Lemma bewiesen ist. □

Wir brauchen nur noch ein weiteres Lemma.

Lemma 8.18 *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine minimale Matrix, bei der alle Einträge a_{ij}, eventuell bis auf die für $i = n$ (also die letzte Zeile), positiv sind. Dann ist $a_{ij} = 1/n$ für alle $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$.*

Beweis: Wegen Lemma 8.17 ist

$$\mathrm{per}(a_1, \ldots, a_{n-2}, a_n, a_n) = \sum_{j=1}^{n} a_{nj}\mathrm{per}(A) = \mathrm{per}(A)$$

und entsprechend

$$\mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_{n-1}) = \sum_{j=1}^{n} a_{(n-1)j}\mathrm{per}(A) = \mathrm{per}(A).$$

Wegen

$$\mathrm{per}(A) = \mathrm{per}(a_1, \ldots, a_{n-2}, a_{n-1}, a_n)$$

gilt in der Ungleichung $(*)$ in Satz 8.7 (diesen Satz können wir wegen der vorausgesetzten Positivität der Einträge von A, eventuell bis auf die letzte Zeile, anwenden) sogar Gleichheit. Daher ist $a_n = \lambda a_{n-1}$ mit einem gewissen $\lambda \in \mathbb{R}$. Da A doppelt stochastisch ist, ist $\lambda = 1$ bzw. $a_n = a_{n-1}$. Die letzten beiden Zeilen von A stimmen also überein. Insbesondere hat auch die n-te Zeile von A nur positive Einträge. In dieser Weise kann man fortfahren. Genauer könnte man zu Beginn die Zeilen a_{n-2}^T und a_{n-1}^T miteinander vertauschen und erhält wie eben $a_{n-2} = a_n$ und damit sukzessive, dass alle Zeilen von A gleich sind. Dann bestehen auch alle Spalten von A aus identischen Elementen. Daher sind alle Einträge von A gleich $1/n$. □

Jetzt kommt das Finale, nämlich der Beweis von Satz 8.1. Wir müssen uns noch von der Positivitätsvoraussetzung in Lemma 8.18 befreien.

Satz 8.19 (Egorychev) *Ist $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine minimale Matrix, so ist $a_{ij} = 1/n$ für alle $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ und daher $\mathrm{per}(A) = n!/n^n$.*

Beweis: Sei $B \in \mathbb{R}^{n\times n}$ aus A dadurch gewonnen, dass eine Zeile a_i^T durch eine andere Zeile a_k^T ersetzt wird. Entsprechend sei C aus A dadurch erhalten, dass die Zeile a_k^T durch a_i^T ersetzt werde. Mit

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix}$$

sei also

$$B = \begin{pmatrix} a_1^T \\ \vdots \\ a_{i-1}^T \\ a_k^T \\ a_{i+1}^T \\ \vdots \\ a_n^T \end{pmatrix}, \qquad C = \begin{pmatrix} a_1^T \\ \vdots \\ a_{k-1}^T \\ a_i^T \\ a_{k+1}^T \\ \vdots \\ a_n^T \end{pmatrix}.$$

Dann ist

$$\operatorname{per}(B) = \sum_{j=1}^{n} a_{kj} \underbrace{\operatorname{per}(A_{ij})}_{=\operatorname{per}(A)} = \operatorname{per}(A) \underbrace{\sum_{j=1}^{n} a_{kj}}_{=1} = \operatorname{per}(A),$$

wobei wir Lemma 8.17 ausgenutzt haben. Entsprechend ist

$$\operatorname{per}(C) = \sum_{j=1}^{n} a_{ij} \underbrace{\operatorname{per}(A_{kj})}_{=\operatorname{per}(A)} = \operatorname{per}(A) \underbrace{\sum_{j=1}^{n} a_{ij}}_{=1} = \operatorname{per}(A).$$

Nun definiere man

$$D := \frac{1}{2}(B + C).$$

Dann ist

$$D^T = \begin{pmatrix} a_1 & \cdots & a_{i-1} & \frac{1}{2}(a_i + a_k) & a_{i+1} & \cdots & a_{k-1} & \frac{1}{2}(a_i + a_k) & a_{k+1} & \cdots & a_n \end{pmatrix}.$$

Die Matrizen D bzw. D^T sind offensichtlich doppelt stochastisch. Wegen der Multilinearität der Permanente ist ferner

$$\begin{aligned}
\operatorname{per}(D) &= \operatorname{per}(a_1, \ldots, a_{i-1}, \tfrac{1}{2}(a_i + a_k), a_{i+1}, \ldots, a_{k-1}, \tfrac{1}{2}(a_i + a_k), a_{k+1}, \ldots, a_n) \\
&= \tfrac{1}{2}\operatorname{per}(a_1, \ldots, a_{i-1}, a_i + a_k, a_{i+1}, \ldots, a_{k-1}, \tfrac{1}{2}(a_i + a_k), a_{k+1}, \ldots, a_n) \\
&= \tfrac{1}{2}[\operatorname{per}(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_{k-1}, \tfrac{1}{2}(a_i + a_k), a_{k+1}, \ldots, a_n) \\
&\qquad + \operatorname{per}(a_1, \ldots, a_{i-1}, a_k, a_{i+1}, \ldots, a_{k-1}, \tfrac{1}{2}(a_i + a_k), a_{k+1}, \ldots, a_n)] \\
&= \tfrac{1}{4}[\operatorname{per}(a_1, \ldots, a_{k-1}, a_i + a_k, a_{k+1}, \ldots, a_n) \\
&\qquad \operatorname{per}(a_1, \ldots, a_{i-1}, a_k, a_{i+1}, \ldots, a_{k-1}, a_i + a_k, a_{k+1}, \ldots, a_n)] \\
&= \tfrac{1}{4}[\operatorname{per}(C) + \operatorname{per}(A) + \operatorname{per}(B) + \operatorname{per}(A)] \\
&= \operatorname{per}(A).
\end{aligned}$$

Also ist mit A auch D eine minimale Matrix. Unser Ziel besteht darin, durch eine Folge solcher Mittelungen je zweier aus den ersten $n-1$ Zeilen eine minimale Matrix E zu bestimmen, deren Einträge e_{ij} bis auf eventuell $i = n$, also die letzte Zeile, sämtlich positiv sind und deren letzte Zeile mit der von A übereinstimmt. . Hierbei

ist die Bemerkung wichtig, die wir im Anschluss an Lemma 8.14 gemacht haben, dass nämlich eine minimale Matrix keine 1 als Eintrag enthalten kann und daher in jeder Zeile und jeder Spalte mindestens zwei positive Einträge vorkommen. Hat eine Zeile an einer gewissen Position eine 0, so mittele man sie mit einer Zeile, die an dieser Position einen positiven Eintrag besitzt. Hierdurch erhöht sich die Anzahl positiver Einträge und nach endlich vielen Schritten hat man das Ziel erreicht. Aus Lemma 8.18 erhalten wir, dass $e_{ij} = 1/n$ für alle $(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\}$ und insbesondere die letzte Zeile von E, die ja mit der letzten Zeile von A übereinstimmt, gleich $(1/n,\dots,1/n)^T$ ist. Entsprechend kann man zeigen, dass *jede* Zeile von A gleich $(1/n,\dots,1/n)^T$ ist. Der Satz ist damit bewiesen. □

8.3 Der Beweis von Gurvits

Von L. Gurvits (2008) stammt ein weiterer Beweis der Vermutung von van der Waerden. Bei der Darstellung halten wir uns vor allem an M. Laurent, A. Schrijver (2010) und M. Aigner, G. M. Ziegler (2018, Kapitel 24). Genauer wird von Gurvits ein Satz formuliert und bewiesen, aus dem u. a. der Satz von Egorychev bzw. die Richtigkeit der van der Waerden Permanenten-Vermutung folgt.

8.3.1 Definitionen, Formulierung des Satzes von Gurvits

Ein *Polynom* $p \in \mathbb{R}[x_1,\dots,x_n]$ in den n Variablen $x_1,\dots,x_n$ ist eine endliche reelle Linearkombination von *Monomen* $x_1^{\alpha_1}\cdots x_n^{\alpha_n}$, wobei $\alpha_1,\dots,\alpha_n$ nichtnegative ganze Zahlen sind. Ein Polynom $p \in \mathbb{R}[x_1,\dots,x_n]$ heißt *homogen vom Grad* n, wenn jedes in p auftretende Monom $x_1^{\alpha_1}\cdots x_n^{\alpha_n}$ den Totalgrad $n = \alpha_1 + \cdots + \alpha_n$ besitzt. Für $p \in \mathbb{R}[x_1,\dots,x_n]$ sei die *Ableitung* $p' \in \mathbb{R}[x_1,\dots,x_{n-1}]$ *in* x_n definiert durch

$$p'(x_1,\dots,x_{n-1}) := \frac{\partial p(x)}{\partial x_n}\,\Big|_{x_n=0}.$$

Die Ableitung eines homogenen Polynoms vom Grad n ist offenbar ein homogenes Polynom vom Grad $n-1$. In p' treten nämlich genau die Monome von p auf, die linear in x_n sind. Mit $\mathbb{R}_+$ bezeichnen wir die Menge der nichtnegativen reellen Zahlen. Entsprechend ist $\mathbb{R}_+^n$ die Menge der (komponentenweisen) nichtnegativen Vektoren mit n reellen Komponenten. Sei weiter $\mathbb{R}_{++}$ die Menge der positiven reellen Zahlen und $\mathbb{R}_{++}^n$ die Menge der n-Vektoren mit positiven Komponenten. Für eine komplexe Zahl $z \in \mathbb{C}$ sei $\mathrm{Re}(z)$ der Real- und $\mathrm{Im}(z)$ der Imaginärteil von z. Sei $\mathbb{C}_+ := \{z \in \mathbb{C} : \mathrm{Re}(z) \geq 0\}$ die rechte komplexe Halbebene und $\mathbb{C}_{++} := \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$ deren Inneres. Die Mengen $\mathbb{C}_+^n$ und $\mathbb{C}_{++}^n$ seien naliegend definiert.

Die folgende Definition ist für die Formulierung des Satzes von Gurvits wichtig.

Definition 8.20 1. Die *Kapazität* $\mathrm{cap}(p)$ von $p \in \mathbb{R}[x_1,\dots,x_n]$ ist definiert als

$$\mathrm{cap}(p) := \inf\Big\{p(x) : x \in \mathbb{R}_+^n,\ \prod_{j=1}^n x_j = 1\Big\}.$$

2. Das Polynom $p \in \mathbb{R}[x_1, \ldots, x_n]$ heißt *H-stabil*[36], wenn p keine Nullstelle in C^n_{++} besitzt.

Jetzt können wir den Satz von Gurvits formulieren.

Satz 8.21 (Gurvits) *Das Polynom $p \in \mathbb{R}_+[x_1, \ldots, x_n]$ mit nichtnegativen Koeffizienten sei homogen vom Grad n und H-stabil. Dann gilt $p' \equiv 0$ oder p' ist H-stabil (und homogen vom Grad $n-1$). In beiden Fällen ist*

$$(*) \qquad \operatorname{cap}(p') \geq \operatorname{cap}(p) \cdot g(k).$$

Hierbei ist $k := \deg_{x_n}(p)$ der Grad von x_n in p (also die höchste Potenz, in der x_n in einem der zu p gehörenden Monome auftritt) und $g\colon \mathbb{N}_0 \longrightarrow \mathbb{R}$ definiert durch

$$g(k) := \begin{cases} 1, & k = 0, \\ \left(\dfrac{k-1}{k}\right)^{k-1}, & k \in \mathbb{N}. \end{cases}$$

Hierbei bezeichnet $\mathbb{N}_0$ die Menge der nichtnegativen ganzen Zahlen.

8.3.2 Der Satz von Egorychev folgt aus dem Satz von Gurvits

Für eine Matrix

$$A = (a_{ij}) = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \in \mathbb{R}^{n \times n}$$

definieren wir das Polynom $p_A \in \mathbb{R}[x_1, \ldots, x_n]$ durch

$$p_A(x_1, \ldots, x_n) = p_A(x) := \prod_{i=1}^{n} a_i^T x = \prod_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_j.$$

Bevor wir den Satz von Egorychev mit Hilfe des Satzes von Gurvits beweisen, formulieren und beweisen wir zwei Lemmata.

Lemma 8.22 *Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ doppelt stochastisch. Dann ist* $\operatorname{cap}(p_A) = 1$.

Beweis: Wir erinnern an die Ungleichung vom geometrisch-arithmetischem Mittel (siehe z.B. J. Werner (2013, Abschnitt 27)).

- Sind $x_1, \ldots, x_n \in \mathbb{R}_+$, so ist

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Eine Variante der Ungleichung vom geometrisch-arithmetischen Mittel ist:

[36] Von M. Laurent, A. Schrijver (2010) wird im Anschluss an die Definition der H-Stabilität gesagt: Here "H" stands for "half-plane". Ich bin mir aber sicher, dass "H" für **H**urwitz steht.

- Sind $x_1, \ldots, x_n \in \mathbb{R}_+$ und $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ mit $\sum_{i=1}^n \lambda_i = 1$, so gilt

$$\text{(GM} - \text{AM)} \qquad \prod_{i=1}^n x_i^{\lambda_i} \leq \sum_{i=1}^n \lambda_i x_i.$$

Denn: O. B. d. A. können wir annehmen, dass $x_1, \ldots, x_n$ und $\lambda_1, \ldots, \lambda_n$ positiv, also aus $\mathbb{R}_{++}$ sind. Wegen der Konkavität des Logarithmus auf $\mathbb{R}_{++}$ ist

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

Eine Anwendung der (monoton wachsenden) Exponentialfunktion auf beide Seiten dieser Ungleichung gibt einen Beweis der angegebenen Variante der Ungleichung vom geometrisch-arithmetischen Mittel. Diese erhält man wiederum im Spezialfall $\lambda_1 = \cdots = \lambda_n = 1/n$.
Für $x \in \mathbb{R}_+^n$ mit $\prod_{j=1}^n x_j = 1$ ist

$$p_A(x) = \prod_{i=1}^n \sum_{j=1}^n a_{ij} x_j \geq \prod_{i=1}^n \prod_{j=1}^n x_j^{a_{ij}} = \prod_{j=1}^n \prod_{i=1}^n x_j^{a_{ij}} = \prod_{j=1}^n x_j^{\sum_{i=1}^n a_{ij}} = \prod_{j=1}^n x_j = 1.$$

Daher ist $\operatorname{cap}(p_A) \geq 1$. Wegen $p_A(e) = 1$ mit $e := (1, \ldots, 1)^T$, ist $\operatorname{cap}(p_A) = 1$. □
In den Beweis des zweiten Lemma geht der Satz von Gurvits ein, er ist also erst dann vollständig bewiesen, wenn der Satz von Gurvits bewiesen ist.

Lemma 8.23 *Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ eine nichtnegative Matrix. Dann ist*

$$\operatorname{per}(A) \geq \operatorname{cap}(p_A) \prod_{i=1}^n g(\min(i, \lambda_A(i))),$$

wobei $\lambda_A(i)$ die Anzahl der von Null verschiedenen Einträge in der i-ten Spalte von A und $g(\cdot)$ wie im Satz 8.21 von Gurvits definiert ist.

Beweis: Wir können annehmen, dass A keine Nullzeile enthält, da andernfalls die Aussage trivialerweise richtig ist. Da A nach Voraussetzung eine nichtnegative Matrix ist, ist $p_A \in \mathbb{R}_+[x_1, \ldots, x_n]$. Das Polynom p_A ist ferner H-stabil, besitzt also keine Nullstelle in der offenen rechten Halbebene. Denn ist $p_A(x) = 0$, so ist $\sum_{j=1}^n a_{kj} x_j = 0$ für ein $k \in \{1, \ldots, n\}$ und damit auch $\sum_{j=1}^n a_{kj} \operatorname{Re}(x_j) = 0$. Da aber ein $l \in \{1, \ldots, n\}$ mit $a_{kl} > 0$ existiert, ist $p_A(x) \neq 0$ für alle $x \in \mathbb{C}_{++}^n$. Weiter definieren wir $q_i \in \mathbb{R}_+[x_1, \ldots, x_i]$ durch

$$q_i(x_1, \ldots, x_i) := \frac{\partial^{n-i}(x_1, \ldots, x_n)}{\partial x_{i+1} \cdots \partial x_n} \Bigg|_{x_{i+1} = \cdots = x_n = 0}, \qquad i = 0, \ldots, n.$$

Dann ist

$$q_0 = \frac{\partial^n p_A(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n} = \operatorname{per}(A)$$

der Koeffizient von $x_1 \cdots x_n$ in der Darstellung von p_A und

$$q_n(x_1, \ldots, x_n) = p_A(x_1, \ldots, x_n).$$

Offenbar ist q_i homogen vom Grad i und $q_i' = q_{i-1}$, $i = 1, \ldots, n$. Einen Beweis des obigen Lemmas erhalten wir durch sukzessive Anwendung des Satzes von Gurvits auf q_i, $i = n, \ldots, 1$. Denn mit $q_n = p_A$ sind alle q_i H-stabil, ferner

$$\operatorname{cap}(q_{i-1}) \geq \operatorname{cap}(q_i) \cdot g(\deg_{x_i}(q_i)) \geq \operatorname{cap}(q_i) \cdot g(\min(i, \lambda_A(i))), \quad i = 1, \ldots, n,$$

da $\deg_{x_i}(q_i) \leq \min(i, \lambda_A(i))$ und $g(\cdot)$ monoton nicht wachsend ist. Wegen $\operatorname{cap}(q_0) = q_0 = \operatorname{per}(A)$ und $q_n = p_A$ erhalten wir die Aussage des Lemmas. □

Jetzt bekommen wir sehr schnell bis auf die Eindeutigkeit die Aussage des Satzes von Egorychev.

- Ist $A \in \mathbb{R}^{n \times n}$ doppelt stochastisch, so ist

$$\operatorname{per}(A) \geq \frac{n!}{n^n}.$$

Denn: Aus Lemma 8.22 und Lemma 8.23 erhalten wir (wieder benutzen wir, dass $g(\cdot)$ monoton nicht wachsend ist)

$$\operatorname{per}(A) \geq \underbrace{\operatorname{cap}(p_A)}_{=1} \cdot \prod_{i=1}^{n} g(\min(i, \lambda_A(i))) \geq \prod_{i=1}^{n} \left(\frac{i-1}{i}\right)^{i-1} = \prod_{i=1}^{n} i \frac{(i-1)^{i-1}}{i^i} = \frac{n!}{n^n}.$$

□

Aufwendiger ist es, die Eindeutigkeitsaussage im Satz von Egorychev zu beweisen.

- Ist $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ doppelt stochastisch und

$$\operatorname{per}(A) = \frac{n!}{n^n},$$

so ist $a_{ij} = 1/n$ für alle $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$.

Denn: Da wir in der Ungleichung

$$\operatorname{per}(A) \geq \prod_{i=1}^{n} g(\min(i, \lambda_A(i))) \geq \prod_{i=1}^{n} g(i) = \frac{n!}{n^n}$$

Gleichheit haben, ist $i \leq \lambda_A(i)$, $i = 1, \ldots, n$, und insbesondere $n = \lambda_A(n)$. In der letzten Spalte sind also alle Einträge von 0 verschieden. Da man Spalten von A vertauschen kann ohne die Permanente zu ändern, können wir annehmen, dass *alle* Einträge von A von 0 verschieden sind. Aus dem selben Grund genügt es zu zeigen, dass alle Einträge der *letzten* Spalte von A gleich $1/n$ sind. Weiter besteht auch in der Ungleichung

$$\operatorname{cap}(p_A') \geq \underbrace{\operatorname{per}(p_A)}_{=1} \cdot g(n, \lambda_A(n)) = g(n) = \left(\frac{n-1}{n}\right)^{n-1}$$

Gleichheit. Also ist

$$\operatorname{cap}(p'_A) = \inf\left\{p'_A(y) : y \in \mathbb{R}_+^{n-1},\ \prod_{j=1}^{n-1} y_j = 1\right\} = \left(\frac{n-1}{n}\right)^{n-1}.$$

Sei $y \in \mathbb{R}_+^{n-1}$ mit $\prod_{j=1}^{n-1} y_j = 1$ vorgegeben. Aus

$$p_A(x) = \prod_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}x_j\right)$$

erhalten wir wegen

$$p'_A(y) = \frac{\partial p_A(x)}{\partial x_n}\ \Big|_{x=(y,0)}$$

mit Hilfe der Produktregel

$$(f_1 \cdots f_n)' = \sum_{k=1}^{n} f'_k \prod_{\substack{i=1\\ i\neq k}}^{n} f_i$$

die folgende Gleichungs-Ungleichungskette:

$$\begin{aligned}
p'_A(y) &= \sum_{k=1}^{n} a_{kn} \prod_{\substack{i=1\\ i\neq k}}^{n}\left(\sum_{j=1}^{n-1} a_{ij}y_j\right)\\
&\geq \prod_{k=1}^{n}\prod_{\substack{i=1\\ i\neq k}}^{n}\left(\sum_{j=1}^{n-1} a_{ij}y_j\right)^{a_{kn}}\\
&\qquad \text{(GM-AM)}\\
&= \prod_{i=1}^{n}\prod_{\substack{k=1\\ k\neq i}}^{n}\left(\sum_{j=1}^{n-1} a_{ij}y_j\right)^{a_{kn}}\\
&= \prod_{i=1}^{n}\left(\sum_{j=1}^{n-1} a_{ij}y_j\right)^{1-a_{in}}\\
&\qquad \text{(wegen } \textstyle\sum_{\substack{k=1\\ k\neq i}}^{n} a_{kn} = 1 - a_{in}\text{)}\\
&= \prod_{i=1}^{n}\left[(1-a_{in})\sum_{j=1}^{n-1}\frac{a_{ij}}{1-a_{in}}y_j\right]^{1-a_{in}}\\
&\qquad \text{(Es ist } a_{in} \neq 1,\ i = 1,\ldots,n\text{)}\\
&\geq \prod_{i=1}^{n}(1-a_{in})^{1-a_{in}} \prod_{j=1}^{n-1} y_j^{a_{ij}}\\
&\qquad \text{(GM-AM)}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\prod_{i=1}^{n}(1-a_{in})^{1-a_{in}}\right)\left(\prod_{j=1}^{n-1}\prod_{i=1}^{n} y_j^{a_{ij}}\right)\\
&= \left(\prod_{i=1}^{n}(1-a_{in})^{1-a_{in}}\right)\left(\prod_{j=1}^{n-1} y_j^{\sum_{i=1}^{n} a_{ij}}\right)\\
&= \left(\prod_{i=1}^{n}(1-a_{in})^{1-a_{in}}\right)\underbrace{\prod_{j=1}^{n-1} y_j}_{=1}\\
&= \prod_{i=1}^{n}(1-a_{in})^{1-a_{in}}\\
&\geq \left(\frac{n-1}{n}\right)^{n-1}.
\end{aligned}
$$

Die letzte Ungleichung sieht man folgendermaßen ein. Sei $f\colon \mathbb{R}_{++} \longrightarrow \mathbb{R}$ definiert durch $f(x) := \ln x^x$ auf $\mathbb{R}_{++}$ konvex, für $x_1, \ldots, x_n \in \mathbb{R}_{++}$ ist daher

$$
f\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{1}{n}(f(x_1) + \cdots + f(x_n))
$$

bzw. nach leichter Umformung

$$
\ln\left(\frac{x_1 + \cdots + x_n}{n}\right)^{\sum_{i=1}^{n} x_i} \leq \ln(x_1^{x_1} \cdots x_n^{x_n}).
$$

Wendet man auf beide Seiten die Exponentialfunktion an, so erhält man

$$
x_1^{x_1} \cdots x_n^{x_n} \geq \left(\frac{x_1 + \cdots + x_n}{n}\right)^{\sum_{i=1}^{n}}.
$$

Gleichheit gilt hier genau dann, wenn $x_1 = \cdots = x_n$. Dies wenden wir mit $x_i := 1 - a_{in}$ an. Dann ist $\sum_{i=1}^{n} x_i = n - 1$, womit die letzte Ungleichung bewiesen ist. Da aber

$$
\operatorname{cap}(p'_A) = \inf\left\{p'_A(y) : y \in \mathbb{R}_+^{n-1},\ \prod_{j=1}^{n-1} y_j = 1\right\} = \left(\frac{n-1}{n}\right)^{n-1},
$$

besteht auch in der letzten Ungleichung in der obigen Gleichung-Ungleichungskette sogar Gleichheit. Folglich ist $1 - a_{1n} = \cdots = 1 - a_{in}$ und folglich $a_{1n} = \cdots = a_{nn} = 1/n$. Damit ist auch obige Eindeutigkeitsaussage bewiesen. In der Tat folgt also der Satz von Egorychev aus dem Satz von Gurvits. □

8.3.3 Der Beweis des Satzes von Gurvits

Wir formulieren und beweisen zunächst zwei Lemmata. Für $x = (x_j) \in \mathbb{C}^n$ sei hierbei $\operatorname{Re}(x) = (\operatorname{Re}(x_j))$.

Lemma 8.24 *Sei $p \in \mathbb{R}_+[x_1, \ldots, x_n]$ homogen vom Grad n und H-stabil. Dann ist*

$$|p(x)| \geq |p(\mathrm{Re}(x))| \qquad \textit{für alle } x \in \mathbb{C}_+^n.$$

Beweis: Aus Stetigkeitsgründen genügt es, die behauptete Ungleichung für alle $x \in \mathbb{C}_{++}^n$ zu beweisen. Sei $x \in \mathbb{C}_{++}^n$ fest, also $\mathrm{Re}(x) > 0$. Wir definieren $f\colon \mathbb{C} \longrightarrow \mathbb{C}$ durch

$$f(s) := p(x + s\mathrm{Re}(x)).$$

Da p nach Voraussetzung homogen vom Grad n ist, ist f in s ein Polynom vom Grad n, besitzt also n Nullstellen in $\mathbb{C}$. Daher existieren $b_1, \ldots, b_n \in \mathbb{C}$ mit

$$p(x + s\mathrm{Re}(x)) = p(\mathrm{Re}(x)) \prod_{i=1}^{n} (s - b_i) \qquad \text{für alle } s \in \mathbb{C}.$$

Da $p(x + b_i\mathrm{Re}(x)) = 0$ und p nach Voraussetzung H-stabil ist, ist $x + b_i\mathrm{Re}(x) \notin \mathbb{C}_{++}^n$ bzw.

$$\mathrm{Re}(x + b_i\mathrm{Re}(x)) = \underbrace{\mathrm{Re}(x)}_{>0}(1 + \mathrm{Re}(b_i)) \leq 0, \quad i = 1, \ldots, n.$$

Also ist $\mathrm{Re}(b_i) \leq -1$ und damit $|b_i| \geq 1$, $i = 1, \ldots, n$. Damit erhalten wir

$$|p(x)| = |p(x + 0 \cdot \mathrm{Re}(x))| = |p(\mathrm{Re}(x))| \prod_{i=1}^{n} \underbrace{|b_i|}_{\geq 1} \geq |p(\mathrm{Re}(x))|.$$

Das Lemma ist bewiesen. □

Lemma 8.25 *Sei $p \in \mathbb{R}_+[x_1, \ldots, x_n]$ homogen vom Grad n und H-stabil. Für jedes $y \in \mathbb{C}_{++}^n$ mit $\prod_{j=1}^{n-1} \mathrm{Re}(y_j) = 1$ ist dann*

$$\mathrm{cap}(p) \leq \frac{p(\mathrm{Re}(y), t)}{t} \quad \textit{für jedes } t > 0.$$

Beweis: Sei $t > 0$ beliebig fest. Wir setzen

$$x := t^{-1/n}(\mathrm{Re}(y), t).$$

Dann ist $x \in \mathbb{R}_{++}^n$ und

$$\prod_{i=1}^{n} x_i = t^{-1} \Bigg(\underbrace{\prod_{j=1}^{n-1} \mathrm{Re}(y_j)}_{=1} \Bigg) t = 1.$$

Da p homogen vom Grad n ist daher

$$\mathrm{cap}(p) \leq p(x) = (t^{-1/n})^n p(\mathrm{Re}(y), t) = \frac{p(\mathrm{Re}(y), t)}{t},$$

und das war zu zeigen.

□

Bemerkung: In den beiden zuletzt bewiesenen Lemmata wird nicht ausgenutzt, dass p ein Polynom mit nichtnegativen Koeffizienten ist, außerdem genügt es vorauszusetzen, dass p homogen (mit einem beliebigen Grad) ist. □

Es folgt jetzt der

Beweis von Satz 8.21, dem Satz von Gurvits: Wir zeigen:

- Sei $p \in \mathbb{R}_+[x_1, \ldots, x_n]$ homogen vom Grad n und H-stabil. Ein $y \in \mathbb{C}^n_{++}$ mit $\prod_{j=1}^{n-1} \mathrm{Re}(y_j) = 1$ sei gegeben. Dann gelten die folgenden beiden Aussagen:
  (i) Falls $p'(y) = 0$, so ist $p' \equiv 0$.
  (ii) Ist y reell, also $y \in \mathbb{R}^n_{++}$, so ist $p'(y) \geq \mathrm{cap}(p) \cdot g(\deg_{x_n}(p))$.

Denn: Es ist $p(y,t)$ ein Polynom vom Grad $k := \deg_{x_n}(p)$ in t. Wir unterscheiden drei Fälle.

(1) Es ist $p(y,0) = 0$.

Wegen Lemma 8.24 ist $p(\mathrm{Re}(y), 0) = 0$ und daher

$$p'(y) = \frac{\partial p(x)}{\partial x_n}\bigg|_{x=(y,0)} = \lim_{t \searrow 0} \frac{p(y,t) - p(y,0)}{t} = \lim_{t \searrow 0} \frac{p(y,t)}{t}$$

und ebenso

$$p'(\mathrm{Re}(y)) = \lim_{t \searrow 0} \frac{p(\mathrm{Re}(y), t)}{t}.$$

Wegen $p(\mathrm{Re}(y), t) \leq |p(\mathrm{Re}(y), t)| \leq |p(y,t)|$ (Lemma 8.24) und der Ungleichung $(*)$ in Lemma 8.25 ist

$$\mathrm{cap}(p) \leq \lim_{t \searrow 0} \frac{p(\mathrm{Re}(y), t)}{t} = p'(\mathrm{Re}(y)) \leq \lim_{t \searrow 0} \frac{|p(y,t)|}{t} = |p'(y)|,$$

wobei wir wiederholt ausgenutzt haben, dass $p(y,0) = 0$ und $p(\mathrm{Re}(y), 0) = 0$. Nun ist es einfach, die Aussagen (i) und (ii) nachzuweisen. Falls nämlich $p'(y) = 0$, so ist $p'(\mathrm{Re}(y)) = 0$. Da aber p' ein Polynom mit nichtnegativen Koeffizienten und $\mathrm{Re}(y) \in \mathbb{R}^n_{++}$ ist, ist $p' \equiv 0$. Also gilt (i). Ist y reell, so ist wegen obiger Gleichungs-Ungleichungskette und wegen $g(k) \leq 1$ für alle $k \in \mathbb{N}_0$, dass

$$g(\deg_{x_n}(p)) \cdot \mathrm{cap}(p) \leq \mathrm{cap}(p) \leq p'(y).$$

Damit gilt auch (ii).

(2) Es ist $p(y,t)$ ein Polynom in t mit einem Grad, der kleiner oder gleich 1 ist.

Da $p(\mathrm{Re}(y), t) \leq |p(y,t)|$ für alle $t > 0$ wegen Lemma 8.24, hat auch $p(\mathrm{Re}(y), t)$ als Polynom in t höchstens den Grad 1. Da dies der Fall ist, ist

$$p'(y) = \lim_{t \to \infty} \frac{p(y,t)}{t}, \qquad p'(\mathrm{Re}(y)) = \lim_{t \o \infty} \frac{p(\mathrm{Re}(y), t)}{t}.$$

Aus obiger Ungleichung $(*)$ zu Beginn des Beweises erhalten wir

$$\operatorname{cap}(p) \leq \lim_{t\to\infty} \frac{p(\operatorname{Re}(y),t)}{t} = p'(\operatorname{Re}(y)) \leq \lim_{t\to\infty} \frac{|p(y,t)|}{t} = |p'(y)|.$$

Wie im ersten Fall schließen wir auf die Gültigkeit von (i) und (ii).

(3) Es ist $p(y,0) \neq 0$ und $p(y,t)$ hat als Polynom in t einen Grad, der größer oder gleich 2 ist.

Dann ist $k := \deg_{x_n}(p) \geq 2$ der Grad des Polynoms $p(y,t)$ in t, sodass

$$p(y,t) = p(y,0)\prod_{i=1}^{k}(1+a_i t)$$

mit gewissen $a_1,\ldots,a_k \in \mathbb{C}$. Dann ist

$$p'(y) = p(y,0)\sum_{i=1}^{k} a_i.$$

Da $p(y,t)$ als Polynom in t mindestens den Grad 2 hat, sind nicht alle a_i gleich Null. Nun kommt eine wichtige weitere Zwischenbehauptung.

– Ist $a_i \neq 0$, so ist a_i^{-1} eine nichtnegative (reelle) Linearkombination von $y_1,\ldots,y_{n-1}$.

Denn: Sei $a_i \neq 0$. Angenommen, die Aussage sei nicht richtig. Dann besitzt das Gleichungssystem $Ax = b$ mit

$$A := \begin{pmatrix} \operatorname{Re}(y_1) & \cdots & \operatorname{Re}(y_{n-1}) \\ \operatorname{Im}(y_1) & \cdots & \operatorname{Im}(y_{n-1}) \end{pmatrix}, \qquad b := \begin{pmatrix} \operatorname{Re}(a_i^{-1}) \\ \operatorname{Im}(a_i^{-1}) \end{pmatrix}$$

keine Lösung $x \in \mathbb{R}_+^{n-1}$. Wegen des Farkas Lemmas (siehe z. B. J. Werner (2013, Abschnitt 45)) existiert $z = (c,d)^T \in \mathbb{R}^2$ derart, dass $A^T z \in \mathbb{R}_+^{n-1}$ und $b^T z < 0$ gilt. Genauer ist

$$\operatorname{Re}(y_j)c + \operatorname{Im}(y_j)d \geq 0, \quad j = 1,\ldots,n-1,$$

und

$$\operatorname{Re}(a_i^{-1})c + \operatorname{Im}(a_i^{-1})d < 0.$$

Wir bestimmen ein $\lambda \in \mathbb{C}$ derart, dass $(\lambda y, -\lambda a_i^{-1}) \in \mathbb{C}_{++}^n$. Hierzu setzen wir $\lambda := c + \epsilon - id$ mit noch unbestimmtem $\epsilon > 0$. Dann ist

$$\operatorname{Re}(\lambda y_j) = \underbrace{c\operatorname{Re}(y_j) + d\operatorname{Im}(y_j)}_{\geq 0} + \underbrace{\epsilon\operatorname{Re}(y_j)}_{>0} > 0, \quad j = 1,\ldots,n-1,$$

und

$$\operatorname{Re}(-\lambda a_i^{-1}) = \underbrace{-(c\operatorname{Re}(a_i^{-1}) + d\operatorname{Im}(a_i^{-1}))}_{>0} - \epsilon\operatorname{Re}(a_i^{-1}) > 0$$

falls $\epsilon > 0$ hinreichend klein ist. Wenn dies der Fall ist, ist also $(\lambda y, -\lambda a_i^{-1}) \in \mathbb{C}_{++}^n$. Da aber $p(\lambda y, -\lambda a_i^{-1}) = 0$ ist das ein Widerspruch zur H-Stabilität von p. Damit ist die Zwischenbehauptung bewiesen.

Jetzt müssen wir noch in dem vorliegenden Fall (3) die Gültigkeit der Aussagen (i) und (ii) nachweisen. Wegen der Zwischenbehauptung ist $\mathrm{Re}(a_i^{-1}) > 0$ und daher auch $\mathrm{Re}(a_i) > 0$ für alle i mit $a_i \neq 0$. Da nicht alle a_i verschwinden, ist folglich $\sum_{i=1}^k a_i \neq 0$ und daher

$$p'(y) = p(y,0) \sum_{i=1}^k a_i \neq 0.$$

Also ist die Aussage (i) richtig. Um (ii) zu beweisen nehmen wir an, y sei reell, es sei also $y \in \mathbb{R}_{++}^{n-1}$ mit $\prod_{j=1}^{n-1} y_j = 1$. Wegen der obigen Zwischenbehauptung sind alle von 0 verschiedenen a_i reell und positiv und folglich insbesondere $\sum_{i=1}^k a_i > 0$. Außerdem ist

$$\frac{p'(y)}{p(y,0)} = \sum_{i=1}^k a_i > 0.$$

Wir erinnern daran, dass $k := \deg_{x_n}(p) \geq 2$ und setzen

$$t := \frac{k}{k-1} \frac{p(y,0)}{p'(y)}.$$

Mit Hilfe der Ungleichung vom geometrisch-arithmetischen Mittel erhalten wir

$$\begin{aligned}
\frac{p(y,t)}{p(y,0)} &= \prod_{i=1}^k (1 + a_i t) \\
&\leq \left[\frac{1}{k} \sum_{i=1}^k (1 + a_i t) \right]^k \\
&= \left[\frac{1}{k} \left(1 + \frac{p'(y)}{p(y,0)} t \right) \right]^k \\
&= \left[\frac{1}{k} \left(k + \frac{k}{k-1} \right)]^k \right. \\
&= \left(\frac{k}{k-1} \right)^k.
\end{aligned}$$

Mit dem angegebenen $t > 0$ erhalten wir aus Lemma 8.25, dass

$$\begin{aligned}
\mathrm{cap}(p) &\leq \frac{p(y,t)}{t} \\
&= p'(y) \frac{k-1}{k} \frac{p(y,t)}{p(y,0)} \\
&\leq p'(y) \frac{k-1}{k} \left(\frac{k}{k-1} \right)^k
\end{aligned}$$

$$
\begin{aligned}
&= p'(y)\left(\frac{k}{k-1}\right)^{k-1} \\
&= \frac{p'(y)}{g(k)},
\end{aligned}
$$

oder $p'(y) \geq \mathrm{cap}(p) \cdot g(k)$, womit auch im Fall (3) die Aussage (ii) bewiesen ist.

Jetzt müssen wir uns noch überlegen, dass mit dem Beweis der beiden Aussagen (i) und (ii) alles bewiesen ist. Das ist einfach. Denn ist $p' \not\equiv 0$, so ist $p'(y) \neq 0$ für alle $y \in \mathbb{C}_{++}^{n-1}$ mit $\prod_{j=1}^{n-1} \mathrm{Re}(y_j) = 1$ wegen der Aussage (i). Da p' homogen vom Grad $n-1$ ist, kann p' auf $\mathbb{C}_{++}^{n-1}$ keine Nullstelle besitzen, ist also H-stabil. Aus der Aussage (ii) folgt sofort, dass $\mathrm{cap}(p') \geq \mathrm{cap}(p) \cdot g(\deg_{x_n}(p))$. Damit ist der Satz von Gurvits bewiesen. □

8.3.4 Weitere Folgerungen aus dem Satz von Gurvits

Für $k, n \in \mathbb{N}$ sei Λ_n^k die Menge der $n \times n$-Matrizen mit nichtnegativen *ganzzahligen* Einträgen, deren Zeilen- und Spaltensummen sämtlich gleich k sind. Weiter definieren wir

$$\lambda_k(n) := \min\{\mathrm{per}(A) : A \in \Lambda_n^k\}, \qquad \theta_k := \inf\{\lambda_k(n)^{1/n} : n \in \mathbb{N}\}.$$

Als Folgerung aus Lemma 8.23 erhalten wir

Satz 8.26 *Ist $A \in \Lambda_n^k$, also A eine nichtnegative ganzzahlige $n \times n$-Matrix, deren Zeilen- und Spaltensummen sämtlich gleich k sind, so ist*

$$\mathrm{per}(A) \geq \left(\frac{(k-1)^{k-1}}{k^{k-2}}\right)^n$$

und damit

$$\lambda_k(n) \geq \left(\frac{(k-1)^{k-1}}{k^{k-2}}\right)^n, \qquad \theta_k \geq \frac{(k-1)^{k-1}}{k^{k-2}}.$$

Beweis: Sei $B := (1/k)A$. Dann ist B doppelt stochastisch, ferner ist die Anzahl $\lambda_B(k)$ der von 0 verschiedenen Einträge in der k-ten Spalte höchstens gleich k, da andernfalls die entsprechende Spaltensumme in A größer oder gleich $k+1$ wäre. Aus Lemma 8.23 unter Benutzung von Lemma 8.22 und der im Satz von Gurvits definierten monoton nicht wachsenden Funktion $g(\cdot)$ folgt daher

$$\mathrm{per}(B) \geq \underbrace{\mathrm{cap}(p_B)}_{=1} \prod_{k=1}^{n} g(\min(k, \lambda_B(k))) = \prod_{k=1}^{n} g(\lambda_B(k)) \geq \prod_{k=1}^{n} g(k) = \left(\frac{k-1}{k}\right)^{(k-1)n}.$$

Daher ist

$$\mathrm{per}(A) = \frac{1}{k^n}\mathrm{per}(B) \geq \frac{1}{k^n}\left(\frac{k-1}{k}\right)^{(k-1)n} = \left(\frac{(k-1)^{k-1}}{k^{k-2}}\right)^n$$

und das sollte gezeigt werden. □

Bemerkung: Von A. Schrijver, W. G. Valiant (1980, Corollary 1) (siehe auch M. Hall Jr. (1986, S. 71)) wurde bewiesen, dass $\theta_k \leq (k-1)^{k-1}/k^{k-2}$ und außerdem die Vermutung geäußert, dass hier Gleichheit gilt. Dass diese Vermutung für $k = 3$ richtig ist, wurde schon von A. Schrijver, W. G. Valiant (1980, Corollary 2) bemerkt, da die entsprechende untere Schranke von M. Voorhoeve (1979) (siehe auch M. Hall Jr. (1986, S. 70)) bewiesen wurde. Die Vermutung ist von A. Schrijver (1998) bewiesen worden, wobei der Autor zugibt, dass der Zugang recht kompliziert ist. Durch den eben bewiesenen Satz ist die Vermutung einfacher bewiesen worden. □

9 Der Primzahlsatz

Der Primzahlsatz sagt aus:

- Die Anzahl $\pi(x)$ der Primzahlen $\leq x$ ist asymptotisch gleich $x/\log x$ (wofür wir $\pi(x) \sim x/\log x$ schreiben). Genauer heißt dies, dass

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\log x} = 1.$$

Dieser Satz wurde unabhängig von einander 1896 von Hadamard und de la Vallée Poussin bewiesen. Wir wollen hier einen Beweis dieses bemerkenswerten Satzes präsentieren, der auf D. J. Newman (1980) zurückgeht[37]. Zur Veranschaulichung geben wir an:

x	$\pi(x)$	$\pi(x)/(x/\log x)$
10	4	0.9210
100	25	1.1512
1 000	168	1.1605
10 000	1 229	1.1320
100 000	9 592	1.1043
1 000 000	78 498	1.0845

Bei P. Bundschuh (2002) wird $\pi(10^i)$, $i = 1, \ldots, 18$, angegeben.

9.1 Die Riemannsche ζ-Funktion

Jeder Beweis des Primzahlsatzes mit funktionentheoretischen Methoden benutzt die Riemannsche ζ-Funktion. Diese ist für komplexes[38] $s = \sigma + it$ mit $\mathrm{Re}(s) = \sigma > 1$ durch die *Dirichlet*-Reihe

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}$$

definiert. Wir zeigen:

[37] D. J. Newman beginnt seinen Artikel mit den Worten: The magnificent prime number theorem has received much attention and many proofs througout the past century.

[38] Wir halten uns an die vorherrschende Notation. Komplexe Variable werden fast immer mit s bezeichnet, der Realteil von s mit σ und der Imaginärteil mit t. Ich habe nicht herausbekommen, weshalb der Imaginärteil von s nicht mit τ bezeichnet wird.

Lemma 9.1 *Die Riemannsche ζ-Funktion $\zeta(s)$ ist durch ihre Dirichlet-Reihe auf der Halbebene*

$$H(1) := \{s \in \mathbb{C} : \mathrm{Re}(s) > 1\}$$

wohldefiniert. Genauer gilt:

1. *Die Dirichlet-Reihe ist für jedes $s \in H(1)$ absolut konvergent und auf $H(1)$ lokal gleichmäßig konvergent.*
2. *Die durch die Dirichlet-Reihe auf $H(1)$ definierte Funktion ζ ist dort holomorph*[39]*.*

Beweis: Sei $s = \sigma + it \in H(1)$, also $\sigma > 1$. Es ist

$$n^s = n^{\sigma+it} = e^{(\sigma+it)\log n} = e^{\sigma \log n} e^{it \log n} = n^\sigma e^{it \log n}$$

und daher

$$\left|\frac{1}{n^s}\right| = \frac{1}{n^\sigma}.$$

Die Reihe $\sum_{n=1}^\infty 1/n^\sigma$ konvergiert wegen des *Integralkriteriums für Reihen*. Dieses sagt aus:

- Sei $f\colon [1,\infty) \longrightarrow \mathbb{R}$ eine stetige, positive, monoton fallende Funktion. Dann konvergiert $\sum_{n=1}^\infty f(n)$ genau dann, wenn $\int_1^\infty f(x)\,dx$ existiert.

Das Integralkriterium wenden wir mit $f(x) := 1/x^\sigma$ an. Dann existiert $\int_1^\infty f(x)\,dx$, da

$$\int_1^N f(x)\,dx = \int_1^N \frac{1}{x^\sigma}\,dx = \frac{1 - N^{1-\sigma}}{\sigma - 1} \overset{N\to\infty}{\to} \frac{1}{\sigma - 1}.$$

Aus

$$\left|\sum_{n=1}^N \frac{1}{n^s}\right| \le \sum_{n=1}^N \left|\frac{1}{n^s}\right| = \sum_{n=1}^N \left|\frac{1}{n^{\sigma+it}}\right| = \sum_{n=1}^N \frac{1}{n^\sigma}$$

für jedes $N \in \mathbb{N}$ und der Konvergenz von $\sum_{n=1}^\infty 1/n^\sigma$ folgt die absolute Konvergenz der die ζ-Funktion für $\mathrm{Re}(s) > 1$ definierenden Dirichlet-Reihe. Die die ζ-Funktion auf $H(1)$ definierende Dirichlet-Reihe konvergiert auf $H(1)$ lokal gleichmäßig, wenn es zu jedem $s_0 \in H(1)$ eine Umgebung U_0 von s_0 gibt, auf der die Folge $\{\sum_{n=1}^N 1/n^s\}_{N\in\mathbb{N}}$ der Partialsummen gleichmäßig gegen $\zeta(s) = \sum_{n=1}^\infty 1/n^s$ konvergiert. Um dies nachzuweisen, sei $s_0 \in H(1)$ bzw. $\sigma_0 := \mathrm{Re}(s_0) > 1$. Man wähle ein $\delta_0 > 0$ mit $\sigma_0 - \delta_0 > 1$ und setze

$$U_0 := \{s \in \mathbb{C} : \mathrm{Re}(s) \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]\}.$$

Dann ist U_0 eine Umgebung von s_0 und für jedes $s \in U_0$ ist

$$\left|\frac{1}{n^s}\right| = \left|\frac{1}{n^{\mathrm{Re}(s)}}\right| \le \frac{1}{n^{\sigma_0-\delta_0}}, \qquad \sum_{n=1}^\infty \frac{1}{n^{\sigma_0-\delta_0}} < \infty.$$

[39] Eine komplexwertige Funktion f heißt auf einer Menge $D \subset \mathbb{C}$ *holomorph*, wenn es zu jedem $z_0 \in D$ eine offene Umgebung von z_0 gibt, auf der f definiert und komplex differenzierbar ist.

Aus dem *Majorantenkriterium von Weierstraß* folgt die auf U_0 gleichmäßige Konvergenz der Folge der Partialsummen $\{\sum_{n=1}^{N} 1/n^s\}_{N\in\mathbb{N}}$ gegen $\zeta(s)$. Aus dem *Konvergenzsatz von Weierstraß* folgt, dass die Riemannsche ζ-Funktion auf $H(1)$ holomorph ist. Das war zu zeigen. □

Sei $\{p_k\}_{k\in\mathbb{N}}$ die Folge der der Größe nach angeordneten Primzahlen, also $p_1 = 2$, $p_2 = 3$ usw. Unser nächstes Ziel besteht darin, die *Eulersche Produktdarstellung* der Riemannschen ζ-Funktion zu beweisen. Für reelles $s > 1$ erkannte Euler schon die Gültigkeit der Beziehung

$$1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots = \zeta(s) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-s}} = \frac{1}{1-2^{-s}} \cdot \frac{1}{1-3^{-s}} \cdot \frac{1}{1-5^{-s}} \cdot \cdots .$$

Hier wollen wir uns dieses sogenannte Euler-Produkt *plausibel* machen, was etwas anderes ist als es zu *beweisen*. Aus

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \cdots , \qquad \frac{1}{2^s}\zeta(s) = \frac{1}{2^s} + \frac{1}{4^s} + \frac{1}{6^s} + \frac{1}{8^s} + \cdots$$

erhält man durch Subtraktion

$$\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1 + \frac{1}{3^s} + \frac{1}{5^s} + \frac{1}{7^s} + \frac{1}{9^s} + \cdots ,$$

d. h. es fallen Terme weg, in denen im Nenner der Faktor 2 vorkommt. Weiter ist

$$\frac{1}{3^s}\left(1 - \frac{1}{2^s}\right)\zeta(s) = \frac{1}{3^s} + \frac{1}{9^s} + \frac{1}{15^s} + \frac{1}{21^s} + \cdot$$

und daher

$$\left(1 - \frac{1}{3^s}\right)\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1 + \frac{1}{5^s} + \frac{1}{7^s} + \frac{1}{11^s} + \cdots .$$

Führt man dies fort, so erhält man, lax formuliert,

$$\cdots \left(1 - \frac{1}{11^s}\right)\left(1 - \frac{1}{7^s}\right)\left(1 - \frac{1}{5^s}\right)\left(1 - \frac{1}{3^s}\right)\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1$$

und hieraus

$$\begin{aligned} \zeta(s) &= \frac{1}{\left(1 - \frac{1}{2^s}\right)\left(1 - \frac{1}{3^s}\right)\left(1 - \frac{1}{5^s}\right)\left(1 - \frac{1}{7^s}\right)\left(1 - \frac{1}{11^s}\right)\cdots} \\ &= \frac{1}{1-2^{-s}} \cdot \frac{1}{1-3^{-s}} \cdot \frac{1}{1-5^{-s}} \cdot \frac{1}{1-7^{-s}} \cdot \frac{1}{1-11^{-s}} \cdots . \end{aligned}$$

Für einen *Beweis* der Eulerschen Produktdarstellung der ζ-Funktion müssen wir Konvergenzbegriffe für ein unendliches Produkt komplexer Zahlen bzw. holomorpher Funktionen klären. Würde man ein unendliches Produkt komplexer Zahlen konvergent nennen, wenn die Folge der Partialprodukte konvergiert, so wäre ein solches Produkt konvergent, wenn nur einer der Faktoren verschwindet, unabhängig von der Größe der anderen Faktoren. Auch könnte ein unendliches Produkt komplexer Zahlen den Wert 0 haben, ohne dass einer der Faktoren verschwindet. Beides ist unerwünscht. Daher definieren wir:

- Sei $\{a_k\}_{k\in\mathbb{N}}$ eine Folge komplexer Zahlen, bei der nur endlich viele Folgenglieder verschwinden. Der Folgenindex $j \in \mathbb{N}$ sei minimal mit der Eigenschaft, dass $a_k \neq 0$ für alle $k \geq j$. Dann nennen wir das unendliche Produkt $\prod_{k=1}^{\infty} a_k$ *konvergent*, wenn die Folge $\{\prod_{k=j}^{n} a_k\}_{n\in\mathbb{N}}$ der Partialprodukte einen Grenzwert $\hat{a} \neq 0$ besitzt. Der Wert des (konvergenten) Produktes ist durch

$$\prod_{k=1}^{\infty} a_k = \begin{cases} \hat{a}, & j = 1, \\ 0, & j > 1 \end{cases}$$

  definiert.

Dann ist gesichert, dass ein konvergentes unendliches Produkt genau dann verschwindet, wenn wenigstens einer der Faktoren gleich 0 ist. Weiter ist $\lim_{k\to\infty} a_k = 1$ offenbar eine notwendige Bedingung für die Konvergenz des unendlichen Produkts $\prod_{k=1}^{\infty} a_k$.

Jetzt kommen wir zu unendlichen Produkten holomorpher Funktionen.

- Sei $S \subset \mathbb{C}$ ein Gebiet (nichtleer, offen und zusammenhängend) und $f_k \colon S \longrightarrow \mathbb{C}$ holomorphe Funktionen, $k \in \mathbb{N}$. Wir sagen, das Produkt $\prod_{k=1}^{\infty}(1+f_k)$ *konvergiert normal auf* S, wenn für jede kompakte Teilmenge $K \subset S$ die unendliche Reihe $\sum_{k=1}^{\infty} \|f_k\|_K$ konvergiert, wobei

$$\|f\|_K := \sup_{z\in K} |f(z)|.$$

Für die weiteren Überlegungen müssen wir uns einige Tatsachen zum *komplexen Logarithmus* ins Gedächtnis zurückrufen. Bei gegebenem $z \in \mathbb{C} \setminus \{0\}$ heißt die Menge $\mathrm{Log}(z) := \{w \in \mathbb{C} : \exp(w) = z\}$ der *komplexe Logarithmus* von z. Hat z die Polardarstellung $z = |z|e^{i\phi}$, wobei $\phi = \arg(z)$ nur bis auf ein Vielfaches von 2π festgelegt ist, so ist

$$\mathrm{Log}(z) = \{\log(|z|) + i(\arg(z) + 2k\pi) : k \in \mathbb{Z}\},$$

wobei $\log(|z|)$ der natürliche Logarithmus der positiven Zahl $|z|$ ist. Mit $\mathbb{C}^-$ bezeichnen wir die komplexe Ebene $\mathbb{C}$, bei der die nichtpositive Halbachse herausgeschnitten ist, d. h. es ist

$$\mathbb{C}^- := \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\}.$$

Der *Hauptzweig des Logarithmus* $\mathrm{Log} \colon \mathbb{C}^- \longrightarrow \mathbb{C}$ ist dann definiert durch

$$\mathrm{Log}(z) := \log(|z|) + i\arg(z) \quad \text{mit} \quad \arg(z) \in (-\pi, \pi).$$

Auf der positiven reellen Halbachse stimmt der Hauptzweig des komplexen Logarithmus also mit dem reellen Logarithmus überein.

Nun können wir das folgende Lemma beweisen.

Lemma 9.2 *Das unendliche Produkt $\prod_{k=1}^{\infty}(1+f_k)$ konvergiere normal auf dem Gebiet $S \subset \mathbb{C}$, wobei $f_k \colon S \longrightarrow \mathbb{C}$ holomorph sind, $k \in \mathbb{N}$. Dann konvergiert die Folge der Partialprodukte $\{\prod_{k=1}^{n}(1+f_k)\}_{n\in\mathbb{N}}$ auf jeder kompakten Teilmenge $K \subset S$ gleichmäßig und damit auch auf S lokal gleichmäßig.*

Beweis: Sei $K \subset S$ kompakt. Da $\prod_{k=1}^{\infty}(1+f_k)$ auf S normal konvergent ist, ist $\sum_{k=1}^{\infty}\|f_k\|_K$ konvergent. Insbesondere ist die Folge $\{f_k\}$ auf K gleichmäßig konvergent gegen die Nullfunktion. Für alle hinreichend großen und daher o. B. d. A. für *alle* k ist also

$$1 + f_k(s) \in \mathbb{C}^- \quad \text{für alle } s \in K,$$

wobei wie oben

$$\mathbb{C}^- := \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\}.$$

Im nächsten Schritt zeigen wir:

- Die mit dem Hauptzweig des Logarithmus gebildete Reihe $\sum_{k=1}^{\infty}\operatorname{Log}(1+f_k)$ ist auf K gleichmäßig konvergent.

Denn: Für $|w| < 1$ ist $1 + w \in \mathbb{C}^-$ und es gilt die Reihenentwicklung

$$\operatorname{Log}(1+w) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} w^k$$

für den Hauptzweig des Logarithmus. Für alle hinreichend großen k ist $\|f_k\|_K \leq \frac{1}{2}$. Für diese k ist und alle $s \in K$ ist daher

$$\begin{aligned} |\operatorname{Log}(1+f_k(s))| &\leq |f_k(s)|(1 + |f_k(s)| + |f_k(s)|^2 + \cdots) \\ &\leq \|f_k\|_K \cdot \sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k \\ &= 2\|f_k\|_K. \end{aligned}$$

Wegen des Majorantenkriteriums von Weierstraß konvergiert $\sum_{k=1}^{\infty}\operatorname{Log}(1+f_k)$ gleichmäßig auf K und das war für die Zwischenbehauptung zu zeigen. Im zweiten Schritt zeigen wir:

- Die Folge der Partialprodukte $\{\prod_{k=1}^{n}(1+f_k)\}_{n\in\mathbb{N}}$ ist auf K gleichmäßig konvergent.

Denn: Sei

$$P_n(s) := \sum_{k=1}^{n} \operatorname{Log}(1+f_k(s)), \qquad P(s) := \lim_{n\to\infty} P_n(s).$$

Wegen der ersten Zwischenbehauptung konvergiert die Folge $\{P_n\}$ auf K gleichmäßig gegen P. Als gleichmäßiger Limes stetiger Funktionen ist P auf K stetig. Daher ist mit K auch $P(K)$ kompakt. Es existiert $n_0 \in \mathbb{N}$ mit $\|P_n - P\|_K \leq 1$ für alle $n \geq n_0$. Daher gibt es eine kompakte Menge $W \subset \mathbb{C}$, die $P_n(K)$ und $P(K)$ für alle $n \geq n_0$ enthält. Da die Exponentialfunktion auf der kompakten Menge W gleichmäßig stetig ist, gibt es zu vorgegebenem $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ mit

$$v, w \in W,\ |v - w| \leq \delta \Longrightarrow |\exp(v) - \exp(w)| \leq \epsilon.$$

Wegen der auf K gleichmäßigen Konvergenz der Folge $\{P_n\}$ gegen P existiert ein $n_1 \geq n_0$ mit $\|P_n - P\|_K \leq \delta$ für alle $n \geq n_1$. Für diese n ist $|\exp(P_n(s)) - \exp(P(s))| \leq \epsilon$ für alle $s \in K$ bzw. $\|\exp(P_n) - \exp(P)\|_K \leq \epsilon$. Wegen

$$\exp(P_n) = \exp\left(\sum_{k=1}^{n} \mathrm{Log}(1+f_k)\right) = \prod_{k=1}^{n}(1+f_k)$$

konvergiert die Folge $\{\prod_{k=1}^{n}(1+f_k)\}_{n\in\mathbb{N}}$ auf K gleichmäßig gegen $\exp(P)$, womit das Lemma bewiesen ist. □

Satz 9.3 (Euler) *Für alle $s \in \mathbb{C}$ mit $\mathrm{Re}(s) > 1$ ist*

$$\zeta(s) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-s}}.$$

Hierbei ist das rechts stehende unendliche Produkt normal konvergent auf der Menge $H(1) := \{s \in \mathbb{C} : \mathrm{Re}(s) > 1\}$. Insbesondere ist $\zeta(s) \neq 0$ für alle $s \in H(1)$.

Beweis: Zunächst zeigen wir, dass das rechts stehende unendliche Produkt normal auf $H(1)$ konvergiert. Sei hierzu $K \subset H(1)$ kompakt. Es existiert ein $\delta > 0$ mit $\mathrm{Re}(s) \geq 1+\delta$ für alle $s \in K$. Wegen

$$\frac{1}{1 - p_k^{-s}} = 1 + f_k(s) \quad \text{mit} \quad f_k(s) := \frac{p_k^{-s}}{1 - p_k^{-s}}$$

haben wir $\|f_k\|_K$ abzuschätzen und anschließend $\sum_{k=1}^{\infty} \|f_k\|_K < \infty$ nachzuweisen. Für $s \in K$ und $\sigma := \mathrm{Re}(s)$ ist

$$|p_k^{-s}| = \frac{1}{p_k^{\sigma}} \leq \frac{1}{p_k^{1+\delta}} \leq \frac{1}{(k+1)^{1+\delta}} \leq \frac{1}{2}$$

und daher

$$|f_k(s)| = \frac{|p_k^{-s}|}{|1 - p_k^{-s}|} \leq \frac{|p_k^{-s}|}{1 - |p_k^{-s}|} \leq 2 \cdot |p_k^{-s}| \leq \frac{2}{(k+1)^{1+\delta}}.$$

Folglich ist

$$\sum_{k=1}^{\infty} \|f_k\|_K \leq 2 \cdot \sum_{k=2} \frac{1}{k^{1+\delta}} < \infty,$$

womit die normale Konvergenz des im Satz rechts stehenden Euler-Produkts nachgewiesen ist. Wegen Lemma 9.2 konvergiert die Folge $\{\prod_{k=1}^{n} 1/(1 - p_k^{-s})\}_{n\in\mathbb{N}}$ auf jeder kompakten Teilmenge von $H(1)$ gleichmäßig.

Jetzt müssen wir noch zeigen, dass der Limes der Folge $\{\prod_{k=1}^{n} 1/(1 - p_k^{-s})\}_{n\in\mathbb{N}}$ für jedes $s \in H(1)$ gerade $\zeta(s)$ ist. Sei hierzu $n \in \mathbb{N}$ beliebig und $\mathcal{P}_n := \{p_1, \ldots, p_n\}$ die Menge der ersten n Primzahlen. Sei $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$ die Menge der nichtnegativen ganzen Zahlen. Mit

$$\mathbb{N}(\mathcal{P}_n) := \{N = p_1^{\alpha_1} \cdots p_n^{\alpha_n} : \alpha_i \in \mathbb{N}_0,\ i = 1, \ldots, n\}$$

bezeichnen wir die Menge derjenigen natürlichen Zahlen, die sich als Produkt von Primzahlpotenzen der Elemente aus $\mathcal{P}_n$ darstellen lassen. Durch vollständige Induktion nach n zeigen wir, dass

$$(*) \qquad \prod_{k=1}^{n} \frac{1}{1-p_k^{-s}} = \sum_{N\in\mathbb{N}(\mathcal{P}_n)} \frac{1}{N^s}.$$

Der Induktionsanfang liegt bei $n = 1$. Wegen

$$\frac{1}{1-p_1^{-s}} = \sum_{\alpha_1=0}^{\infty} \frac{1}{p_1^{\alpha_1 s}} = \sum_{N\in\mathbb{N}(\mathcal{P}_1)} \frac{1}{N^s}$$

ist $(*)$ für $n = 1$ richtig. Angenommen, $(*)$ sei für $n \in \mathbb{N}$ richtig. Dann ist

$$\begin{aligned}
\prod_{k=1}^{n+1} \frac{1}{1-p_k^{-s}} &= \left(\prod_{k=1}^{n} \frac{1}{1-p_k^{-s}}\right) \cdot \frac{1}{1-p_{n+1}^{-s}} \\
&= \left(\sum_{N\in\mathbb{N}(\mathcal{P}_n)} \frac{1}{N^s}\right) \cdot \frac{1}{1-p_{n+1}^{-s}} \\
&\qquad \text{(Induktionsannahme)} \\
&= \left(\sum_{N\in\mathbb{N}(\mathcal{P}_n)} \frac{1}{N^s}\right) \cdot \sum_{\alpha_{n+1}=0}^{\infty} \frac{1}{p_{n+1}^{\alpha_{n+1}s}} \\
&= \sum_{N\in\mathbb{N}(\mathcal{P}_{n+1})} \frac{1}{N^s},
\end{aligned}$$

da wegen absoluter Konvergenz gliedweise Multiplikation und Umordnung erlaubt sind. Damit ist der Induktionsbeweis abgeschlossen und $(*)$ für alle $n \in \mathbb{N}$ bewiesen. Daher ist

$$\begin{aligned}
\left|\zeta(s) - \prod_{k=1}^{n} \frac{1}{1-p_k^{-s}}\right| &= \left|\sum_{N\in\mathbb{N}\setminus\mathbb{N}(\mathcal{P}_n)} \frac{1}{N^s}\right| \\
&\leq \sum_{N\in\mathbb{N}\setminus\mathbb{N}(\mathcal{P}_n)} \frac{1}{N^\sigma} \\
&\leq \sum_{N=n}^{\infty} \frac{1}{N^\sigma},
\end{aligned}$$

denn wegen des Fundamentalsatzes der Arithmetik lässt sich jede natürliche Zahl (bis auf die Reihenfolge) eindeutig als Produkt von Primzahlpotenzen darstellen. Jedes $N \in \mathbb{N}\setminus\mathbb{N}(\mathcal{P}_n)$ besitzt daher einen Faktor aus $\{p_{n+1}, p_{n+2}, \ldots\}$, insbesondere ist $N \geq p_{n+1} > n$ für alle $N \in \mathbb{N} \setminus \mathbb{N}(\mathcal{P}_n)$. Damit ist gezeigt, dass für jedes $s \in \mathbb{C}$ mit $\mathrm{Re}(s) > 1$ die Folge $\{\prod_{k=1}^{n} 1/(1-p_k^{-s})\}_{n\in\mathbb{N}}$ von Partialprodukten mit $n \to \infty$ gegen $\zeta(s)$ konvergiert. Insgesamt ist der Satz damit bewiesen. □

Bemerkung: Mit Hilfe von Satz 9.3 kann man zeigen, dass es unendlich viele Primzahlen gibt. Denn angenommen, es gäbe nur endlich viele Primzahlen $\{p_1, \ldots, p_N\}$. Dann wäre

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{k=1}^{N} \frac{1}{1 - p_k^{-s}} \quad \text{für alle } s > 1.$$

Mit $s \searrow 1$ erhält man einen Widerspruch dazu, dass die harmonische Reihe divergiert. Man könnte aber auch so argumentieren: Gäbe es nur endlich viele Primzahlen $\{p_1, \ldots, p_N\}$, so wäre inbesondere

$$\zeta(2) = \sum_{n=1}^{\infty} \frac{1}{n^2} = \prod_{k=1}^{n} \frac{1}{1 - p_k^{-2}}$$

eine rationale Zahl. Andererseits weiß man (siehe z. B. M. Aigner. G. M. Ziegler (2018, S. 3 ff.)), dass $\zeta(2) = \pi^2/6$ irrational bzw. sogar transzendent ist. Beide Beweise sollten aber wohl nicht zu ernst genommen werden, da man durchaus auf Kanonenschüsse auf (stattliche) Spatzen erinnert wird. Mehr hierzu und den Beitrag von Euler findet man z. B. bei E. Sandifer (2006). □

Durch die Dirichlet-Reihe ist die Riemannsche ζ-Funktion lediglich auf der Halbebene $H(1)$ definiert. Der folgende Satz zeigt, dass man die ζ-Funktion meromorph auf die Halbebene

$$H(0) := \{s \in \mathbb{C} : \mathrm{Re}(s) > 0\}$$

fortsetzen kann, wobei die Fortsetzung holomorph bis auf einen Pol erster Ordnung an der Stelle $s = 1$ mit dem Residuum 1 ist.

Satz 9.4 *Die Funktion*

$$f(s) := \zeta(s) - \frac{1}{s-1}$$

kann holomoph von $H(1)$ auf $H(0)$ fortgesetzt werden.

Beweis: Für $\mathrm{Re}(s) > 1$ bzw. $s \in H(1)$ ist

$$f(s) = \zeta(s) - \frac{1}{s-1} = \sum_{n=1}^{\infty} \frac{1}{n^s} - \int_1^{\infty} \frac{1}{x^s}\, dx = \sum_{n=1}^{\infty} \underbrace{\int_n^{n+1} \left(\frac{1}{n^s} - \frac{1}{x^s} \right) dx}_{=:\phi_n(s)}.$$

Die Reihe auf $H(0)$ holomorpher Funktionen auf der rechten Seite konvergiert auf $H(0)$ absolut und lokal gleichmäßig, sodass f holomorph auf $H(0)$ ist. Dies folgt aus der für $s \in H(0)$ gültigen Abschätzung

$$\begin{aligned} |\phi_n(s)| &= \left| \int_n^{n+1} \left(\frac{1}{n^s} - \frac{1}{x^s} \right) dx \right| \\ &= \left| s \int_n^{n+1} \left(\int_n^x \frac{du}{u^{s+1}} \right) dx \right| \\ &\leq \max_{u \in [n, n+1]} \left| \frac{s}{u^{s+1}} \right| \end{aligned}$$

$$= \frac{|s|}{n^{\mathrm{Re}(s)+1}}.$$

Daher ist die Reihe $\sum_{n=1}^{\infty} \phi_n(s)$ auf $H(0)$ absolut und lokal gleichmäßig konvergent und folglich f auf $H(0)$ holomorph. □

Der folgende Satz ist entscheidend beim Beweis des Primzahlsatzes.

Satz 9.5 *Es ist $\zeta(s) \neq 0$ für alle $s \in \mathrm{cl}(H(1))$, also alle $s \in \mathbb{C}$ mit $\mathrm{Re}(s) \geq 1$.*

Beweis: Wegen Satz 9.3 wissen wir, dass $\zeta(s) \neq 0$ für alle $s \in H(1)$. Es bleibt daher zu zeigen, dass die ζ-Funktion keine Nullstelle s mit $\mathrm{Re}(s) = 1$ besitzt. Wir machen einen Widerspruchsbeweis und nehmen an, $s = 1 + it$ sei eine Nullstelle von ζ. Dann ist $t \neq 0$, da ζ in $s = 1$ einen Pol erster Ordnung besitzt. Ferner ist

$$(*) \qquad \lim_{\sigma \to 1} \zeta(\sigma)^3 \zeta(\sigma + it)^4 \zeta(\sigma + 2it) = 0.$$

Denn

$$\zeta(\sigma) = \frac{1}{\sigma - 1} + f(\sigma), \qquad \zeta(\sigma + it) = (\sigma - 1)\psi(\sigma)$$

mit in $\sigma = 1$ zumindestens stetigen Funktionen f und ψ. Da ζ in $1 + 2it$ keinen Pol besitzt, ist $\lim_{\sigma \to 1} \zeta(\sigma + 2it) = \zeta(1 + 2it)$, womit $(*)$ bewiesen ist. Andererseits gilt:

- Für alle $\sigma, t \in \mathbb{R}$ mit $\sigma > 1$ ist

$$|\zeta(\sigma)^3 \zeta(\sigma + it)^4 \zeta(\sigma + 2it)| \geq 1.$$

Denn: Seien wieder $\{p_1, p_2, \ldots\}$ mit $p_1 < p_2 < \cdots$ die Folge der Primzahlen. Für $\mathrm{Re}(s) > 1$ haben wir in Satz 9.3 die Eulersche Produktdarstellung der ζ-Funktion bewiesen, also die Gültigkeit von

$$\zeta(s) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-s}}.$$

Wir erinnern daran, dass der Hauptzweig des Logarithmus

$$\mathrm{Log}\colon \mathbb{C}^- := \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\} \longrightarrow \mathbb{C}$$

durch

$$\mathrm{Log}(z) := \log(|z|) + i\arg(z) \quad \text{mit} \quad \arg(z) \in (-\pi, \pi)$$

definiert ist. Insbesondere ist

$$\mathrm{Re}(\mathrm{Log}(z)) = \log(|z|).$$

Für $z \in \mathbb{C}$ mit $|z| < 1$ ist $z \in \mathbb{C}^-$ und es gilt die Entwicklung

$$\mathrm{Log}(1 - z) = -\sum_{n=1}^{\infty} \frac{z^n}{n}.$$

Unter Berücksichtigung von

$$|p_k^{-s}| = p_k^{-\mathrm{Re}(s)}| < 1, \quad k \in \mathbb{N},$$

erhalten wir

$$\log|\zeta(s)| = -\sum_{k=1}^{\infty} \log|1 - p_k^{-s}| = -\sum_{k=1}^{\infty} \mathrm{Re}(\mathrm{Log}(1 - p_k^{-s})) = \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \frac{\mathrm{Re}(p_k^{-ns})}{n}.$$

Setzt man hier $s = \sigma + it$, so erhält man

$$\log|\zeta(\sigma + it)| = \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \frac{\mathrm{Re}(p_k^{-n(\sigma+it)})}{n} = \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \frac{\cos(nt\log p_k)}{np_k^{n\sigma}}.$$

Dann ist

$$\begin{aligned}
\log|\zeta(\sigma)^3\zeta(\sigma+it)^4\zeta(\sigma+2it)| &= 3\log|\zeta(\sigma)| + 4\log|\zeta(\sigma+it)| + \log|\zeta(\sigma+2it)| \\
&= \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \frac{3 + 4\cos(nt\log p_k) + \cos(2nt\log p_k)}{np_k^{n\sigma}} \\
&= \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \frac{2(1 + \cos(nt\log p_k))^2}{np_k^{n\sigma}} \\
&\geq 0.
\end{aligned}$$

Hieraus folgt die in • behauptete Ungleichung und wir haben den gewünschten Widerspruch. Der Satz ist bewiesen. □

Insbesondere gilt die Aussage, die D. J. Newman (1980) an den Anfang seines Beweises des Primzahlsatzes stellt:

- So let us begin with the well-known fact about the ζ-function:

  $$(z-1)\zeta(z) \text{ is analytic and zero-free throughout } \mathrm{Re}(z) \geq 1.$$

  This will be assumed throughout and will allow us to give our proof of the prime number theorem.

Bemerkung: Die Riemannsche ζ-Funktion kann analytisch (von $H(1)$ oder $H(0)$) auf $\mathbb{C} \setminus \{1\}$ fortgesetzt werden. Diese analytische Fortsetzung besitzt sogenannte triviale Nullstellen in -2, -4, -6 usw. Die berühmte Riemannsche Vermutung besagt, dass alle nichttrivialen Nullstellen auf der Geraden $\{s \in \mathbb{C} : \mathrm{Re}(s) = \frac{1}{2}\}$ liegen. □

Neben der Riemannschen ζ-Funktion spielt eine weitere Funktion eine wichtige Rolle bei der folgenden Untersuchung, nämlich die von Tschebyscheff eingeführte Funktion

$$\Phi(s) := \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s}.$$

Hierbei seien wieder $\{p_k\}_{k\in\mathbb{N}}$ die der Größe nach angeordneten Primzahlen, also $p_1 = 2$, $p_2 = 3$ usw. Wir wollen uns davon überzeugen, dass Φ auf

$$H(1) := \{s \in \mathbb{C} : \mathrm{Re}(s) > 1\}$$

absolut und lokal gleichmäßig konvergiert und damit Φ auf $H(1)$ eine holomorphe Funktion ist. Dies geschieht im Prinzip genauso wie der entsprechende Beweis für die Riemannsche ζ-Funktion. Mit $s = \sigma + it$ ist

$$(*) \qquad \left|\frac{\log p_k}{p_k^s}\right| = \frac{\log p_k}{p_k^\sigma} = (\log p_k)\exp(-\sigma \log p_k) \le (\log(k+1))\exp(-\sigma\log(k+1))$$

für alle $k \in \mathbb{N}$. Hierbei sieht man die letzte Ungleichung folgendermaßen ein. Für $k = 1$ und $k = 2$ besteht wegen $p_1 = 2$ und $p_2 = 3$ sogar Gleichheit. Auf $[3, \infty)$ ist

$$f(x) := (\log x)\exp(-\sigma \log x)$$

monoton fallend, wie man

$$f'(x) = (\underbrace{1 - \sigma\log x}_{<0})\,\frac{\exp(-\sigma\log x)}{x}$$

entnimmt. Wegen $3 \le k + 1 \le p_k$, $k = 3, 4, \ldots$ folgt daher $(*)$ für alle $k \in \mathbb{N}$. Die absolute Konvergenz von $\Phi(s)$ auf $H(1)$ ist daher bewiesen, wenn

$$\sum_{k=1}^{\infty}(\log(k+1))\exp(1 - \sigma\log(k+1)) < \infty$$

bzw.

$$\int_1^\infty (\log(x+1))\exp(-\sigma\log(x+1))\,dx < \infty$$

für $\sigma > 1$. Dass der auf $[1, \infty)$ positive Integrand auf einem kleinen Anfangsintervall $[1, \exp(1/\sigma) - 1] \subset [1, e - 1]$ nicht positiv fallend ist, spielt für die Anwendung des Integralkriteriums für Reihen offenbar keine Rolle. Nun ist

$$\begin{aligned}
\int_1^N \log(x+1)\exp(-\sigma\log(x+1))\,dx &= \int_{\log 2}^{\log(N+1)} y\exp(-\sigma y)\exp(y)\,dy \\
&\qquad (y = \log(x+1),\ \exp(y)dy = dx) \\
&= \int_{\log 2}^{\log(N+1)} y\exp((1-\sigma)y)\,dy \\
&= \frac{y(1-\sigma)-1}{(1-\sigma)^2}\exp((1-\sigma)y)\,\Big|_{y=\log(2)}^{y=\log(N+1)} \\
&= \frac{(1-\sigma)\log(N+1)-1}{(1-\sigma)^2(N+1)^{\sigma-1}} \\
&\qquad -\frac{(1-\sigma)\log 2 - 1}{(1-\sigma)^2 2^{\sigma-1}}
\end{aligned}$$

$$\overset{N\to\infty}{\to} \quad -\frac{(1-\sigma)\log 2 - 1}{(1-\sigma)^2 2^{\sigma-1}},$$

womit für $\mathrm{Re}(s) > 1$ die absolute Konvergenz von

$$\Phi(s) = \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s}$$

bewiesen ist. Die lokal gleichmäßige Konvergenz der Reihe Φ auf $H(1)$ kann im Prinzip genauso wie die entsprechende Aussage für die ζ-Funktion bewiesen werden. Denn ist $s_0 \in H(1)$, also $\sigma_0 := \mathrm{Re}(s_0) > 1$, so wähle man ein $\delta_0 > 0$ mit $\sigma_0 - \delta_0 > 1$ und setze

$$U_0 := \{s \in \mathbb{C} : \mathrm{Re}(s) \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]\}.$$

Dann ist U_0 eine Umgebung von s_0 und für jedes $s \in U_0$ ist

$$\left|\frac{\log p_k}{p_k^s}\right| = \frac{\log p_k}{p_k^\sigma} \le \frac{\log p_k}{p_k^{\sigma_0-\delta_0}}, \quad k \in \mathbb{N},$$

mit

$$\sum_{k=1}^{\infty} \frac{\log p_k}{p_k^{\sigma_0-\delta_0}} < \infty.$$

Wegen des Majorantenkriteriums von Weierstraß folgt die auf $H(1)$ lokal gleichmäßige Konvergenz der Reihe Φ. Also ist Φ auf $H(1)$ definiert und holomorph.

Satz 9.6 *Die Funktion*

$$\Phi(s) - \frac{1}{s-1} = \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s} - \frac{1}{s-1},$$

die auf $H(1)$ definiert und holomorph ist, kann zu einer auf

$$H(\tfrac{1}{2}) = \{s \in \mathbb{C} : \mathrm{Re}(s) > \tfrac{1}{2}\}$$

meromorphen Funktion mit einem einzigen Pol in $s = 1$ fortgesetzt werden, welche auf

$$\mathrm{cl}(H(1)) = \{s \in \mathbb{C} : \mathrm{Re}(s) \ge 1\}$$

holomorph ist.

Beweis: Für $\mathrm{Re}(s) > 1$ bzw. $s \in H(1)$ erhalten wir aus dem Euler-Produkt (siehe Satz 9.3)

$$\zeta(s) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-s}}$$

mit Hilfe der Produktregel für Ableitungen die für $\mathrm{Re}(s) > 1$ gültige Darstellung

$$\zeta'(s) \quad = \quad \sum_{j=1}^{\infty} \frac{d}{ds}\left(\frac{1}{1-p_j^{-s}}\right) \prod_{\substack{k=1 \\ k\neq j}}^{\infty} \frac{1}{1-p_k^{-s}}$$

$$
\begin{aligned}
&= \sum_{j=1}^{\infty} \frac{d}{ds}\left(\frac{1}{1-p_j^{-s}}\right)(1-p_j^{-s}) \prod_{k=1}^{\infty} \frac{1}{1-p_k^{-s}} \\
&= \zeta(s) \sum_{k=1}^{\infty} \frac{d}{ds}\left(\frac{1}{1-p_k^{-s}}\right)(1-p_k^{-s}) \\
&= -\zeta(s) \sum_{k=1}^{\infty} \frac{(\log p_k) p_k^{-s}}{(1-p_k^{-s})^2}(1-p_k^{-s}) \\
&= -\zeta(s) \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s - 1}.
\end{aligned}
$$

Für $\mathrm{Re}(s) > 1$ ist also

$$
-\frac{\zeta'(s)}{\zeta(s)} = \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s - 1} = \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s} + \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s(p_k^s - 1)} = \Phi(s) + \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s(p_k^s - 1)}.
$$

Die linke Seite $-\zeta'(s)/\zeta(s)$ (das Negative der logarithmischen Ableitung von $\zeta(s)$) in dieser Gleichungskette ist auf $H(1)$ holomorph, da die Riemannsche ζ-Funktion dort holomorph ist und keine Nullstelle besitzt, siehe Lemma 9.1 und Satz 9.3. Wegen Satz 9.4 kann die ζ-Funktion so von $H(1)$ nach $H(0)$ fortgesetzt werden, dass

$$
f(s) := \zeta(s) - \frac{1}{s-1}
$$

holomorph auf $H(0)$ ist. Da $\zeta(s)$ nur einen einfachen Pol mit Residuum 1 in $s = 1$ und keine Nullstellen für $\mathrm{Re}(s) = 1$ besitzt (siehe Satz 9.5), gilt dasselbe für die linke Seite $-\zeta'(s)/\zeta(s)$. Die Reihe auf der rechten Seite ist für $\mathrm{Re}(s) > \frac{1}{2}$ absolut konvergent und auf $H(\frac{1}{2})$ lokal gleichmäßig konvergent, definiert also auf $H(\frac{1}{2})$ eine holomorphe Funktion. Ersteres sieht man folgendermaßen ein. Für $\sigma := \mathrm{Re}(s) > \frac{1}{2}$ und alle $k \in \mathbb{N}$ ist

$$
\left|\frac{\log p_k}{p_k^s(p_k^s - 1)}\right| = \frac{\log p_k}{|p_k^s|\,|p_k^s - 1|} \le \left(\frac{1}{1 - 1/\sqrt{2}}\right) \frac{\log p_k}{p_k^{2\sigma}}.
$$

Hierbei haben wir ausgenutzt, dass

$$
p_k^\sigma = |p_k^s| \le |p_k^s - 1| + 1 \le |p_k^s - 1| + \frac{p_k^\sigma}{\sqrt{2}}.
$$

Insgesamt folgt die Behauptung. □

9.2 Der Beweis des Primzahlsatzes von D. J. Newman

Ein relativ einfacher Beweis des Primzahlsatzes ist von D. J. Newman (1980) angegeben worden, siehe auch D. J. Newman (1998, S. 67 ff.). Wir folgen im wesentlichen der Darstellung dieses Beweises bei D. Zagier (1997) und J. Korevaar (1982), siehe auch J. Bak, D. J. Newman (2010, S. 285 ff.). Im Internet findet man einige weitere Ausarbeitungen des Beweises von D. J. Newman, z. B. von M. Baker, D. Clark (2002), A. Sutherland (2015), A. Zeilmann (2013), C. O'Rourke (2013).

Mit $\mathcal{P} = \{p_k\}_{k\in\mathbb{N}}$ bezeichnen wir die Folge der Primzahlen, wobei wir annehmen, dass diese der Größe nach geordnet sind, dass also $p_1 < p_2 < \cdots$. Für $x \in \mathbb{R}$ sei $\pi(x)$ die Anzahl der Primzahlen, die kleiner oder gleich x sind. Wenn wir $\sum_{p\le x}$ bzw. $\prod_{p\le x}$ schreiben, so heißt dies, dass wir die Summe bzw. das Produkt über alle Primzahlen p bilden, die kleiner oder gleich x sind. Drei Funktionen spielen für den Beweis des Primzahlsatzes eine herausragende Rolle. Dies sind neben den im vorigen Abschnitt untersuchten Funktionen

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}, \qquad \Phi(s) := \sum_{k=1}^{\infty} \frac{\log p_k}{p_k^s}$$

die sogenannte *erste Tschebyscheff-Funktion*, nämlich die durch

$$\vartheta(x) := \sum_{p\le x} \log p$$

definierte Funktion $\vartheta\colon \mathbb{R} \longrightarrow \mathbb{R}$. Während $\pi(x)$ die Anzahl der Primzahlen $\le x$ angibt, ist $\vartheta(x)$ sozusagen eine logarithmisch gewichtete Anzahlfunktion der Primzahlen $\le x$. Wie wir am Anfang dieses Abschnittes schon bemerkten, sagt der *Primzahlsatz* (prime number theorem) aus, dass

$$\pi(x) \sim \frac{x}{\log x},$$

wobei die Bezeichnung $f(x) \sim g(x)$ bedeutet, dass $\lim_{x\to\infty} f(x)/g(x) = 1$. Durch den folgenden Satz wird der Beweis des Primzahlsatzes auf den Nachweis von $\vartheta(x) \sim x$ reduziert.

Satz 9.7 *Ist $\vartheta(x) \sim x$, so ist $\pi(x) \sim x/\log x$.*

Beweis: Für $x > 1$ ist

$$\vartheta(x) = \sum_{p\le x} \log p \le \pi(x) \log x$$

und daher

$$\frac{\vartheta(x)}{x} \le \pi(x)\frac{\log x}{x}.$$

Für jedes $\epsilon \in (0,1)$ ist

$$\begin{aligned}
\vartheta(x) &\ge \sum_{x^{1-\epsilon}<p\le x} \log p \\
&\ge (\log x^{1-\epsilon})(\pi(x) - \pi(x^{1-\epsilon}) \\
&= (1-\epsilon)(\log x)(\pi(x) - \pi(x^{1-\epsilon})) \\
&\ge (1-\epsilon)(\log x)(\pi(x) - x^{1-\epsilon})
\end{aligned}$$

und folglich

$$\pi(x) \le \frac{1}{1-\epsilon}\frac{\vartheta(x)}{\log x} + x^{1-\epsilon}.$$

Für alle $x > 1$ und alle $\epsilon \in (0,1)$ ist dann

$$\frac{\vartheta(x)}{x} \leq \frac{\pi(x)\log x}{x} \leq \frac{1}{1-\epsilon}\frac{\vartheta(x)}{x} + \frac{\log x}{x^\epsilon}.$$

Nun gilt

$$\lim_{x\to\infty} \frac{\log x}{x^\epsilon} = 0.$$

Aus

$$\frac{\vartheta(x)}{x} - 1 \leq \frac{\pi(x)\log x}{x} - 1 \leq \frac{\epsilon}{1-\epsilon}\left(\frac{\vartheta(x)}{x} - 1\right) + \frac{\log x}{x^\epsilon}$$

erkennen wir: Ist $\vartheta(x) \sim x$, so ist auch $\pi(x) \sim x/\log x$. Der Satz ist bewiesen. □

Im folgenden Lemma erreichen wir unser eigentliches Ziel zwar noch nicht, nämlich zu zeigen, dass $\vartheta(x) \sim x$ bzw. $\lim_{x\to\infty} \vartheta(x)/x = 1$. Wir können aber zeigen, dass $\vartheta(x)/x$ auf $[1,\infty)$ beschränkt ist bzw. $\vartheta(x) = O(x)$ gilt.

Lemma 9.8 *Für alle $x \geq 1$ ist $\vartheta(x) \leq (4\log 2)x$. Daher ist $\vartheta(x) = O(x)$.*

Beweis: Für jedes $n \in \mathbb{N}$ folgt aus dem binomischen Lehrsatz, dass

$$2^{2n} = (1+1)^{2n} = \sum_{m=0}^{2n} \binom{2n}{m} \geq \binom{2n}{n} = \frac{(2n)!}{n!\,n!} \geq \prod_{n<p\leq 2n} p = \exp(\vartheta(2n) - \vartheta(n)).$$

Hierbei haben wir bei der letzten Ungleichung ausgenutzt, dass in der Primzahlzerlegung von $\binom{2n}{n}$ alle Primzahlen p aus $(n, 2n]$ vorkommen[40], denn diese p teilen $(2n)!$, aber nicht $n!$. Durch Logarithmieren erhält man

$$\vartheta(2n) - \vartheta(n) \leq 2n\log 2$$

für alle $n \in \mathbb{N}$. Für alle $m \in \mathbb{N}$ ist

$$\vartheta(2^m) = \sum_{n=1}^{m} [\vartheta(2^n) - \vartheta(2^{n-1})] \leq \sum_{n=1}^{m} 2^n \log 2 \leq 2^{m+1}\log 2.$$

Nun sei $x \geq 1$ beliebig. Man bestimme $m \in \mathbb{N}$ mit $2^{m-1} \leq x < 2^m$. Dann ist

$$\vartheta(x) \leq \vartheta(2^m) \leq 2^{m+1}\log 2 = (4\log 2)2^{m-1} \leq (4\log 2)x,$$

wie behaupet. □

Entscheidend für den Beweis des Primzahlsatzes durch D. J. Newman in der Darstellung von J. Korevaar (1982) und D. Zagier (1998) ist der folgende Satz, der von Korevaar *Auxiliary Tauberian theorem* und von Zagier *Analytic Theorem* genannt wird. Diesen Satz geben wir zunächst ohne Beweis an.

[40] Z. B. ist

$$\binom{24}{12} = 2\,704\,156 = \underbrace{13\cdot 17\cdot 19\cdot 23}_{=96\,577}\cdot 28$$

Satz 9.9 *Sei $f\colon [0,\infty) \longrightarrow \mathbb{R}$ eine beschränkte und über jedes endliche Teilintervall integrierbare, z. B. stückweise stetige Funktion. Die auf*

$$H(0) := \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$$

definierte und dort holomorphe Funktion[41]

$$g(s) := \int_0^\infty f(t)e^{-st}\,dt$$

sei holomorph auf $\mathrm{cl}(H(0))$ *fortsetzbar. Dann existiert* $\int_0^\infty f(t)\,dt$ *(und ist gleich* $g(0)$*).*

Aus (dem *noch nicht bewiesenen*) Satz 9.9 erhalten wir einen Beweis des Primzahlsatzes.

Satz 9.10 (Primzahlsatz) *Bezeichnet $\pi(x)$ die Anzahl der Primzahlen kleiner oder gleich x, so ist $\pi(x) \sim x/\log x$ bzw. $\lim_{x\to\infty} \pi(x)\log(x)/x = 1$.*

Beweis: Wegen Satz 9.7 genügt es zu zeigen, dass $\vartheta(x) \sim x$ bzw. $\lim_{x\to\infty} \vartheta(x)/x = 1$, wobei $\vartheta\colon \mathbb{R} \longrightarrow \mathbb{R}$ durch

$$\vartheta(x) := \sum_{p\le x} \log p$$

gegeben ist. Formal setzen wir $p_0 := 1$, sodass $\vartheta(p_0) = 0$. Die Funktion ϑ ist stückweise konstant mit

$$\vartheta(x) = \vartheta(p_k) \quad \text{für } x \in [p_k, p_{k+1}), \quad k = 0, 1, \ldots$$

und

$$\vartheta(p_k) = \vartheta(p_{k-1}) + \log p_k, \quad k = 1, 2 \ldots.$$

Der Beweis besteht aus zwei Teilen, wobei im ersten Teil Satz 9.9 angewandt wird, während der zweite Teil elementar ist.

- Das Integral

$$\int_1^\infty \frac{\vartheta(x) - x}{x^2}\,dx$$

  existiert.

Denn: Wir wollen (den noch nicht bewiesenen) Satz 9.9 anwenden und definieren hierzu $f\colon [0,\infty) \longrightarrow \mathbb{R}$ durch $f(t) := \theta(e^t)e^{-t} - 1$. Dann ist f beschränkt, denn wegen Lemma 9.8 ist

$$-1 \le f(t) = \vartheta(e^t)e^{-t} - 1 \le (4\log 2)e^t \cdot e^{-t} - 1 = 4\log 2 - 1$$

für alle $t \ge 0$. Weiter ist f offensichtlich f stückweise stetig. Um Satz 9.9 anwenden zu können, müssen wir uns überlegen, dass die auf $H(0)$ definierte und dort holomorphe Funktion

$$g(s) := \int_0^\infty f(t)e^{-st}\,dt$$

[41] Dies ist gerade die Laplace-Transformierte von f.

auf $\mathrm{cl}(H(0))$ holomorph fortsetzbar ist. Für $s \in H(0)$ bzw. $\mathrm{Re}(s) > 0$ ist

$$\begin{aligned}
g(s) &= \int_0^\infty [\theta(e^t)e^{-t} - 1]e^{-st}\,dt \\
&= \int_1^\infty \left(\frac{\vartheta(x)}{x} - 1\right)\frac{1}{x^{s+1}}\,dx \\
&\qquad (t = \log x,\ dt = dx/x) \\
&= \int_1^\infty \frac{\vartheta(x)}{x^{s+2}}\,dx + \frac{1}{s}x^{-s}\,\Big|_{x=1}^{x=\infty} \\
&= \sum_{k=1}^\infty \vartheta(p_k)\int_{p_k}^{p_{k+1}} \frac{dx}{x^{s+2}} - \frac{1}{s} \\
&\qquad (\vartheta(x) = 0 \text{ für } x \in [1, p_1),\ \vartheta(x) = \vartheta(p_k) \text{ für } x \in [p_k, p_{k+1})) \\
&= -\frac{1}{s+1}\sum_{k=1}^\infty \vartheta(p_k)(p_{k+1}^{-(s+1)} - p_k^{-(s+1)}) - \frac{1}{s} \\
&= -\frac{1}{s+1}\sum_{k=1}^\infty [\vartheta(p_{k-1}) - \vartheta(p_k)]p_k^{-(s+1)} - \frac{1}{s} \\
&= \frac{1}{s+1}\sum_{k=1}^\infty \frac{\log p_k}{p_k^{s+1}} - \frac{1}{s} \\
&= \frac{1}{s+1}\Phi(s+1) - \frac{1}{s} \\
&= \frac{1}{s+1}\left(\Phi(s+1) - \frac{1}{s}\right) - \frac{1}{s+1}.
\end{aligned}$$

Wegen Satz 9.6 kann die auf $H(1)$ definierte und dort holomorphe Funktion $\Phi(s) - 1/(s-1)$ zu einer auf $H(\frac{1}{2})$ meromorphen Funktion mit dem einzigen Pol in $s = 1$ fortgesetzt werden, welche auf $\mathrm{cl}(H(1))$ holomorph ist. Hieraus folgt, dass $\Phi(s+1) - 1/s$ und dann auch $g(s)$ auf $\mathrm{cl}(H(0))$ holomorph fortsetzbar ist. Wegen (des noch nicht bewiesenen) Satz 9.9 existiert

$$\int_0^\infty f(t)\,dt = \int_0^\infty [\vartheta(e^t)e^{-t} - 1]\,dt = \int_1^\infty \frac{\vartheta(x) - x}{x^2}\,dx$$

und der erste Teil ist bewiesen. Im zweiten Teil zeigen wir:

- Es ist $\vartheta(x) \sim x$.

Denn: Wir definieren

$$\Theta(x) := \int_1^x \frac{\vartheta(t) - t}{t^2}\,dt$$

und werden benutzen, dass $\lim_{x\to\infty} \Theta(x)$ existiert. Angenommen, es ist $\vartheta(x) \not\sim x$. Zwei Fälle werden unterschieden.

(a) Es existiert ein $\lambda > 1$ und eine Folge $\{x_k\} \subset [1, \infty)$ mit $x_k \to \infty$ und $\vartheta(x_k) \geq \lambda x_k$ für alle $k \in \mathbb{N}$. Da ϑ monoton nicht fallend ist, ist

$$\begin{aligned}
\Theta(\lambda x_k) - \Theta(x_k) &= \int_{x_k}^{\lambda x_k} \frac{\vartheta(t) - t}{t^2}\, dt \\
&\geq \int_{x_k}^{\lambda x_k} \frac{\lambda x_k - t}{t^2}\, dt \\
&= \left(-\frac{\lambda x_k}{t} - \log t \right) \Bigg|_{t=x_k}^{t=\lambda x_k} \\
&= \lambda - 1 - \log \lambda \\
&> 0 \quad .
\end{aligned}$$

Mit $k \to \infty$ erhalten wir einen Widerspruch zur Existenz von $\lim_{x\to\infty} \Theta(x)$.

- Es existiert ein $\lambda < 1$ und eine Folge $\{x_k\} \subset [1, \infty)$ mit $x_k \to \infty$ und $\vartheta(x_k) \leq \lambda x_k$ für alle $k \in \mathbb{N}$. Da ϑ monoton nicht fallend ist, ist

$$\begin{aligned}
\Theta(x_k) - \Theta(\lambda x_k) &= \int_{\lambda x_k}^{x_k} \frac{\vartheta(t) - t}{t^2}\, dt \\
&\leq \int_{\lambda x_k}^{x_k} \frac{\lambda x_k - t}{t^2}\, dt \\
&= \left(-\frac{\lambda x_k}{t} - \log t \right) \Bigg|_{t=\lambda x_k}^{t=x_k} \\
&= -\lambda + 1 + \log \lambda \\
&< 0 \quad .
\end{aligned}$$

Mit $k \to \infty$ erhalten wir einen Widerspruch zur Existenz von $\lim_{x\to\infty} \Theta(x)$.

Damit ist $\vartheta(x) \sim x$ bewiesen. Wegen Satz 9.7 folgt $\pi(x) \sim x/\log x$ und der Primzahlsatz ist (mit einer Lücke) bewiesen. □

Jetzt kommt es darauf an, die noch bestehende Lücke zu schließen.

Beweis von Satz 9.9: Für $T > 0$ definiere man $g_T \colon \mathbb{C} \longrightarrow \mathbb{C}$ durch

$$g_T(s) := \int_0^T f(t) e^{-st}\, dt.$$

Dann ist g_T eine ganze Funktion, also auf $\mathbb{C}$ holomorph. Es ist zu zeigen, dass

$$\lim_{T\to\infty} [g(0) - g_T(0)] = 0.$$

Wir wählen ein (großes, später genauer zu bestimmendes) $R > 0$ und hierzu ein $\delta = \delta(R) \in (0, R/2)$ so klein, dass g holomorph auf

$$D := \{s \in \mathbb{C} : |s| \leq R,\ \mathrm{Re}(s) \geq -\delta\}$$

ist. Weshalb kann ein solches $\delta > 0$ gefunden werden? Nach Voraussetzung gibt es eine auf $\mathrm{cl}(H(0))$ holomorphe Fortsetzung von g, die natürlich wieder mit g bezeichnet wird. Wir definieren

$$E := \{s \in \mathbb{C} : |s| \leq R,\ \mathrm{Re}(s) = 0\}.$$

Da g auf $\mathrm{cl}(H(0))$ holomorph ist, gibt es zu jedem $s \in E$ (insbesondere ist $\mathrm{Re}(s) = 0$ und $s = \mathrm{Im}(s)$) ein $\delta_s > 0$ mit der Eigenschaft, dass g auf

$$U_s := \{z \in \mathbb{C} : \max(|\mathrm{Re}(z)|, |\mathrm{Im}(z) - s)|) \leq \delta_s\}$$

holomorph ist. Aus der Überdeckung $\bigcup_{s\in E} U_s \supset E$ der kompakten Menge E kann eine endliche Teilüberdeckung ausgewählt werden. Es existieren also endlich viele $s_k \in E$ und zugehörige positive Zahlen $\delta_k = \delta_{s_k}$, $k = 1, \ldots, n$, mit der Eigenschaft, dass g auf $\bigcup_{k=1}^{n} U_{s_k} \supset E$ holomorph ist. Sei

$$\delta := \min_{k=1,\ldots,n} \delta_k.$$

Wir wollen zeigen, dass dann g auf

$$D := \{s \in \mathbb{C} : |s| \leq R,\ \mathrm{Re}(s) \geq -\delta\}$$

holomorph ist. Da g auf $\mathrm{cl}(H(0))$ holomorph ist, genügt es die Holomorphie von g in jedem $s \in \mathbb{C}$ mit $|s| \leq R$ und $-\delta \leq \mathrm{Re}(s) < 0$ nachzuweisen. Sei also ein solches $s = \sigma + it$ vorgegeben. Dann ist $it \in E$ und folglich gibt es mit einem gewissen $k \in \{1, \ldots, n\}$ ein $s_k \in E$ und eine zugehörige Umgebung U_{s_k}, die it enthält und auf der g holomorph ist. Die Situation machen wir uns in Abbildung 63 klar. Dann ist

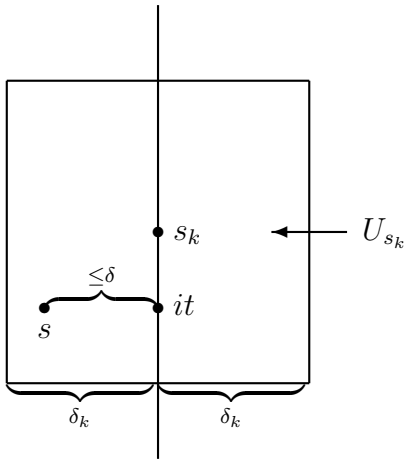


Abbildung 63: g ist auf $D = \{s \in \mathbb{C} : |s| \leq R,\ \mathrm{Re}(s) \geq -\delta\}$ holomorph

$s \in U_{s_k}$, da

$$\max(|\mathrm{Re}(s)|, |\mathrm{Im}(s) - s_k|) = \max(\underbrace{-\mathrm{Re}(s)}_{\leq\delta}, \underbrace{|it - s_k|}_{\leq\delta_k}) \leq \delta_k.$$

Folglich ist

$$\{s \in \mathbb{C} : |s| \leq R,\ -\delta \leq \mathrm{Re}(s) < 0\} \subset \bigcup_{k=1}^{n} U_{s_k}.$$

Damit ist gezeigt, dass g auf

$$D := \{s \in \mathbb{C} : |s| \leq R,\ \mathrm{Re}(s) \geq -\delta\}$$

holomorph ist. Den Rand von D nennen wir $\Gamma := \partial D$. Dieser setzt sich zusammen aus dem Halbkreis

$$\Gamma_+ := \Gamma \cap \{s \in \mathbb{C} : \mathrm{Re}(s) \geq 0\}$$

sowie aus

$$\Gamma_- := \Gamma \cap \{s \in \mathbb{C} : \mathrm{Re}(s) < 0\}.$$

In Abbildung 64 veranschaulichen wir uns die Menge D und ihren Rand $\Gamma = \Gamma_+ \cup \Gamma_-$. Nun wenden wir die Cauchysche Integralformel auf die Funktion

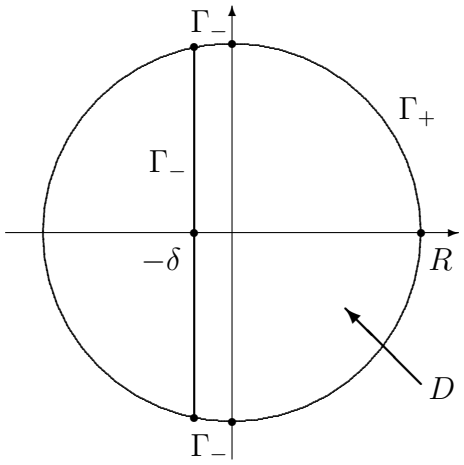


Abbildung 64: D und der Rand $\Gamma = \Gamma_+ \cup \Gamma_-$

$$G_T(s) := (g(s) - g_T(s))e^{sT}\left(1 + \frac{s^2}{R^2}\right),$$

welche im Innern von D und auf dem Rand Γ holomorph ist, an und erhalten

$$G_T(0) = \frac{1}{2\pi i}\int_\Gamma G_T(s)\,\frac{ds}{s}$$

bzw.

$$(*) \qquad g(0) - g_T(0) = \frac{1}{2\pi i}\int_\Gamma (g(s) - g_T(s))e^{sT}\left(1 + \frac{s^2}{R^2}\right)\frac{ds}{s}.$$

Da $f\colon [0,\infty) \longrightarrow \mathbb{R}$ nach Voraussetzung beschränkt ist, existiert eine Konstante $M > 0$ mit $|f(t)| \leq M$ für alle $t \geq 0$. Kurvenintegrale werden jeweils mit Hilfe der sogenannten

M-L-Formel abgeschätzt: Der Betrag eines Kurvenintegrals ist kleiner oder gleich dem Maximum des Integranden auf der Kurve multipliziert mit der Länge L der Kurve.

Wir erinnern daran, dass es unser Ziel ist, $\lim_{T\to\infty}[g(0) - g_T(0)] = 0$ nachzuweisen. Hierzu geben wir uns ein $\epsilon > 0$ vor und zeigen die Existenz eines $T_0 > 0$ mit

$$|g(0) - g_T(0)| \leq \epsilon \quad \text{für alle } T \geq T_0.$$

Zunächst schätzen wir in der Darstellung von $g(0) - g_T(0)$ in $(*)$ das Integral über den Halbkreis Γ_+ dem Betrag nach oben ab und zeigen, dass dieses durch M/R nach oben beschränkt werden kann. Danach schätzen wir das entsprechende Integral über Γ_- ab, und zwar mit Hilfe der Dreiecksungleichung getrennt für den Anteil von $g(s)$ und dem von $g_T(s)$. Der letztere Anteil kann ebenfalls durch M/R nach oben abgeschätzt werden. Wählt man daher $R > 0$ so groß, dass $R \geq 3M/\epsilon$, so ist

$$|g(0) - g_T(0)| \leq \frac{2\epsilon}{3} + \left|\frac{1}{2\pi i}\int_{\Gamma_-} g(s)e^{sT}\left(1 + \frac{s^2}{R^2}\right)\frac{ds}{s}\right|.$$

Kann man dann zeigen, dass der letzte Term mit $T \to \infty$ gegen Null konvergiert, so ist das Ziel erreicht. Dies ist der Fahrplan für den restlichen Beweis.

Nun aber zu den Einzelschritten. Für $s = \sigma + it \in \Gamma_+$ ist (zunächst für $\sigma > 0$, dann für $\sigma \geq 0$)

$$\begin{aligned}
\left|(g(s) - g_T(s))e^{sT}\left(1 + \frac{s^2}{R^2}\right)\frac{1}{s}\right| &= |g(s) - g_T(s)|e^{\sigma T}\left|1 + \frac{s^2}{R^2}\right|\frac{1}{R} \\
&= |g(s) - g_T(s)|\frac{e^{\sigma T}}{R^3}|\underbrace{\sigma^2 + t^2}_{=R^2} + (\underbrace{\sigma^2 - t^2 + 2i\sigma t}_{=s^2})| \\
&= |g(s) - g_T(s)|\frac{e^{\sigma T}}{R^3}2\sigma\underbrace{|\sigma + it|}_{=R} \\
&= |g(s) - g_T(s)|e^{\sigma T}\frac{2\sigma}{R^2} \\
&= e^{\sigma T}\frac{2\sigma}{R^2}\left|\int_T^\infty f(t)e^{-st}\,dt\right| \\
&\leq e^{\sigma T}\frac{2\sigma}{R^2}\int_T^\infty |f(t)|e^{-\sigma t}\,dt \\
&\leq e^{\sigma T}\frac{2\sigma}{R^2}\,M\frac{e^{-\sigma T}}{\sigma} \\
&= \frac{2M}{R^2}.
\end{aligned}$$

Daher ist

$$\left|\frac{1}{2\pi i}\int_{\Gamma_+}(g(s) - g_T(s))e^{sT}\left(1 + \frac{s^2}{R^2}\right)\frac{ds}{s}\right| \leq \frac{1}{2\pi}\cdot\pi R\cdot\frac{2M}{R^2} = \frac{M}{R}.$$

Jetzt muss das entsprechende Integral über Γ_- abgeschätzt werden. Hierzu benutzen wir, dass

$$\left|\frac{1}{2\pi i}\int_{\Gamma_-}(g(s)-g_T(s))e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right| \;\le\; \left|\frac{1}{2\pi i}\int_{\Gamma_-}g(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right| + \left|\frac{1}{2\pi i}\int_{\Gamma_-}g_T(s)s^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right|.$$

Beim zweiten Summanden nutzen wir aus, dass $g_T(s)e^{sT}(1+s^2/R^2)$ eine ganze Funktion, also auf ganz $\mathbb{C}$ holomorph ist. Daher kann hier Γ_- durch den Halbkreis

$$\Gamma'_- := \{s\in\mathbb{C} : |s| = R,\ \mathrm{Re}(s) < 0\}$$

ersetzt werden. Folglich ist

$$\left|\frac{1}{2\pi i}\int_{\Gamma_-}g_T(s)s^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right| = \left|\frac{1}{2\pi i}\int_{\Gamma'_-}g_T(s)s^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right|.$$

Für $s=\sigma+it\in\Gamma'_-$ ist

$$|g_T(s)| = \left|\int_0^T f(t)e^{-st}\,dt\right| \le M\int_0^T e^{-\sigma t}\,dt \le \frac{M}{|\sigma|}e^{-\sigma T}$$

und

$$\left|e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{1}{s}\right| = e^{\sigma T}\cdot\frac{2|\sigma|}{R^2}.$$

Daher ist

$$\left|\frac{1}{2\pi i}\int_{\Gamma'_-}g_T(s)s^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right| \le \frac{1}{2\pi}\cdot\pi R\cdot\frac{M}{|\sigma|}e^{-\sigma T}\cdot e^{\sigma T}\cdot\frac{2|\sigma|}{R^2} = \frac{M}{R}.$$

Bei festem $R>0$ ist

$$(**)\qquad \lim_{T\to\infty}\left|\frac{1}{2\pi i}\int_{\Gamma_-}g(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right| = 0,$$

denn der Integrand ist das Produkt aus der Funktion $g(s)(1+s^2/R^2)$, die von T unabhängig ist, und der Funktion e^{sT}, welche auf kompakten Teilmengen der Halbebene $\{s\in\mathbb{C} : \mathrm{Re}(s)<0\}$ mit $T\to\infty$ schnell gegen Null konvergiert. Von J. Bak, D. J. Newman (2010, S. 289) wird die Aussage $(**)$ mit Hilfe der M-L-Formel elementar bewiesen. Hierzu wird $T>0$ zunächst so groß gewählt, dass $1/\sqrt{T}<\delta$, danach wird Γ_- zerlegt in

$$\Gamma^1_- := \{s\in\Gamma_- : \mathrm{Re}(s)\ge -1/\sqrt{T}\},\qquad \Gamma^2_- := \{s\in\Gamma_- : \mathrm{Re}(s) < -1/\sqrt{T}\}$$

und

$$M := \max_{s\in\Gamma_-}\left|g(s)\left(1+\frac{s^2}{R^2}\right)\frac{1}{s}\right|$$

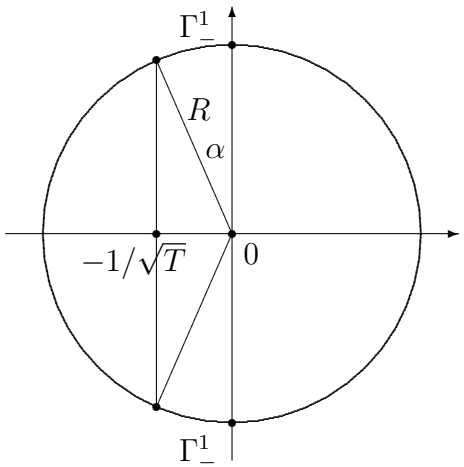


Abbildung 65: Die Kurve Γ_-^1 und ihre Länge

gesetzt. Zunächst schätzen wir das Integral über Γ_-^1 ab. Die Länge $L(\Gamma_-^1)$ von Γ_-^1 ist nach oben durch $4/\sqrt{T}$ beschränkt. Denn es ist $L(\Gamma_-^1) = 2R\alpha$ mit $\sin\alpha = 1/(R\sqrt{T})$ bzw. $L(\Gamma_-^1) = 2R \arcsin 1/(R\sqrt{T})$, siehe Abbildung 65. Nun ist $1/\sqrt{T} < \delta < R/2$ und daher $1/(R\sqrt{T}) \in (0, 1/2)$. Daher ist

$$\alpha = \arcsin \frac{1}{R\sqrt{T}} \leq \frac{2}{R\sqrt{T}}$$

und folglich

$$L(\Gamma_-^1) = 2R\alpha \leq \frac{4}{\sqrt{T}}.$$

Die Länge $L(\Gamma_-^2)$ von Γ_-^2 wird (ganz grob) durch πR, die Länge eines Halbkreises mit Radius R, nach oben abgeschätzt. Nun ist der Nachweis für die Gültigkeit der Aussage $(**)$ einfach. Es ist nämlich

$$\begin{aligned}
\left| \int_{\Gamma_-} g(s) e^{sT} \left(1 + \frac{s^2}{R^2}\right) \frac{ds}{s} \right| &\leq \int_{\Gamma_-} e^{\mathrm{Re}(s)T} \underbrace{\left| g(s) \left(1 + \frac{s^2}{R^2}\right) \frac{1}{s} \right|}_{\leq M} ds \\
&\leq M \left(\int_{\Gamma_-^1} \underbrace{e^{\mathrm{Re}(s)T}}_{\leq 1} ds + \int_{\Gamma_-^2} \underbrace{e^{\mathrm{Re}(s)T}}_{\leq e^{-\sqrt{T}}} ds \right) \\
&\leq M \left(\frac{4}{\sqrt{T}} + \pi R e^{-\sqrt{T}} \right).
\end{aligned}$$

Damit ist auch $(**)$ und insgesamt der Satz 9.9 bewiesen. □

10 Konforme Abbildungen und der Riemannsche Abbildungssatz

10.1 Einleitung, Grundlagen

Der *Riemannsche Abbildungssatz*[42] sagt aus:

Satz 10.1 (Riemannscher Abbildungssatz) *Sei $G \subset \mathbb{C}$ eine einfach zusammenhängende offene Menge mit $G \neq \mathbb{C}$ und $a \in G$ beliebig. Dann existiert genau eine konforme Abbildung $f\colon G \longrightarrow B(0;1)$, wobei $B(0;1) := \{z \in \mathbb{C} : |z| < 1\}$ die offene Einheitskreisscheibe ist, mit $f(a) = 0$ und $f'(a) > 0$.*

Unser Ziel besteht darin, diesen Satz zu beweisen. Zunächst müssen aber einige in dem Riemannschen Abbildungssatz auftretende Vokabeln erläutert werden.

1. Eine Menge $G \subset \mathbb{C}$ heißt *zusammenhängend*, wenn es es zu je zwei Punkten $u, v \in G$ eine u und v verbindende *Kurve* bzw. eine stetige Abbildung $p\colon [0,1] \longrightarrow G$ mit $p(0) = u$ und $p(1) = v$ gibt.

2. Eine offene, zusammenhängende Menge $G \subset \mathbb{C}$ heißt ein *Gebiet*.

3. Eine zusammenhängende Menge $G \subset \mathbb{C}$ heißt *einfach zusammenhängend*, wenn sich jede in G verlaufende geschlossene einfache Kurve γ stetig auf einen Punkt $z_0 \in G$ zusammenziehen lässt. Hierbei heißt eine Kurve $\gamma = p([0,1])$ *geschlossen*, wenn $p(0) = p(1)$. Sie heißt *einfach*, wenn sie keine mehrfachen Punkte besitzt, also $p(s) \neq p(t)$ für $s, t \in [0,1)$ mit $s \neq t$ gilt. Ferner sagt man, die geschlossene einfache Kurve $\gamma = p([0,1]) \subset G$ lasse sich auf *einen Punkt $z_0 \in G$ zusammenziehen*, wenn es eine stetige Abbildung $P\colon [0,1] \times [0,1] \longrightarrow G$ gibt mit

   (a) $P(0,\alpha) = P(1,\alpha)$ für alle $\alpha \in [0,1]$, d. h. $P([0,1],\alpha)$ ist für alle $\alpha \in [0,1]$ eine geschlossene Kurve,

   (b) $P(t,0) = p(t)$ für alle $t \in [0,1]$,

   (c) $P(t,0) = z_0$ für alle $t \in [0,1]$.

4. Ist $G \subset \mathbb{C}$ ein Gebiet, so heißt eine Abbildung $f\colon G \longrightarrow \mathbb{C}$ eine *auf G konforme Abbildung*, wenn f auf G holomorph und *injektiv* (d. h. $f(u) \neq f(v)$ für $u, v \in G$ mit $u \neq v$) ist. Wie wir sehen werden, ist dann $f(G) \subset \mathbb{C}$ ebenfalls offen, ferner $f'(z) \neq 0$ für alle $z \in G$ und $f^{-1}\colon f(G) \longrightarrow G \subset \mathbb{C}$ auch holomorph (d. h. komplex differenzierbar) bzw. f *biholomorph*. Die Mengen G und $f(G)$ heißen in diesem Fall *konform äquivalent*. Aus dem (noch nicht bewiesenen) Riemannschen Abbildungssatz folgt, dass je zwei einfach zusammenhängende, von $\mathbb{C}$ verschiedene, offene Teilmengen von $\mathbb{C}$ konform äquivalent sind, da sie jeweils zu $B(0;1)$ konform äquivalent sind.

[42] Beim Riemannschen Abbildungssatz muss ich an die folgende Begebenheit denken. Vor vielen Jahren war ich als Assistent Beisitzer bei einer Prüfung, die mein Doktorvater Lothar Collatz in Hamburg abnahm. Er fragte den Prüfling: Lässt sich ein Quadrat konform auf eine Kreisscheibe abbilden? Der Prüfling antwortete: Nein, denn sonst wäre ja die Quadratur des Kreises möglich!

5. Sei $z_0 \in \mathbb{C}$ und $r > 0$. Eine Funktion $f\colon B(z_0;r) \setminus \{z_0\} \longrightarrow \mathbb{C}$ hat einen *Pol* in z_0, wenn die Funktion $1/f$ mit $(1/f)(z_0) := 0$ auf einer Umgebung von z_0 holomorph ist. Ein Pol z_0 von f hat die *Ordnung* $n \in \mathbb{N}$, wenn es eine Umgebung U_0 von z_0 und eine auf U_0 nicht verschwindende holomorphe Funktion h mit $f(z) = (z - z_0)^{-n} h(z)$ für alle $z \in U_0 \setminus \{z_0\}$ gibt. Das *Residuum* einer Funktion f in einem Pol der Ordnung n ist definiert durch

$$\operatorname{res}_{z_0}(f) := \lim_{z \to z_0} \frac{1}{(n-1)!} \left(\frac{d}{dz} \right)^{n-1} (z - z_0)^n f(z).$$

Bemerkungen: Wir wollen elementare Bemerkungen[43] zu den eingeführten Begriffen machen und einige klassische Sätze der Funktionentheorie ins Gedächtnis zurückrufen. Hierzu gehören die folgenden Aussagen:

- *Satz von Cauchy (für Kreisscheibe)*: Sei f auf der offenen Kreisscheibe $B(a;r)$ (mit Mittelpunkt $a \in \mathbb{C}$ und Radius $r > 0$) holomorph. Dann ist

  $$\int_\gamma f(z)\, dz = 0$$

  für jede geschlossene, in $B(a;r)$ enthaltene Kurve γ.

  Einen Beweis findet man bei E. M. Stein, R. Shakarchi (2003, Theorem 2.2 auf S. 39). Der Satz von Goursat ist ein Spezialfall (γ besteht aus den drei Seiten eines Dreiecks T), ihn beweisen wir später unter 10. Dann hat man leicht auch die Aussage des Satzes von Cauchy mit den Seiten eines Rechtecks als Kurve γ, siehe Abbildung 66. Danach kann man zeigen, dass eine auf einer offenen

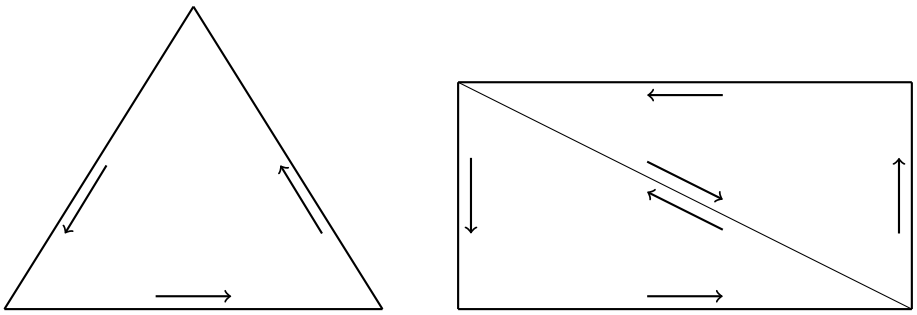


Abbildung 66: Orientiertes Dreieck und Rechteck

  Kreisscheibe B holomorphe Funktion f dort eine Stammfunktion besitzt, also eine auf B holomorphe Funktion F mit $F'(z) = f(z)$ für alle $z \in B$ existiert. Der letzte Schritt zum Beweis des Satzes von Cauchy für eine Kreisscheibe ist dann einfach.

  Allgemeiner gilt (siehe E. M. Stein, R. Shakarchi (2003, S. 97):

[43] Als Literatur ist fast jedes Lehrbuch über Funktionentheorie geeignet, z. B. E. M. Stein, R. Shakarchi (2003) oder J. B. Conway (1973).

Ist $G \subset \mathbb{C}$ ein einfach zusammenhängendes Gebiet und $f\colon G \longrightarrow \mathbb{C}$ holomorph, so ist

$$\int_\gamma f(z)\,dz = 0$$

für jede geschlossene Kurve γ in G.

- *Integralformel von Cauchy*: Sei f holomorph auf einer offenen Menge, welche den Abschluss $B[a;r] := \operatorname{cl} B(a;r)$ der offenen Kreisscheibe $B(a;r)$ um a mit Radius $r > 0$ enthält. Sei $C := \partial B(a;r)$ der positiv orientierte Rand von $B(a;r)$. Dann ist

$$f(z) = \frac{1}{2\pi i}\int_C \frac{f(\zeta)}{\zeta - z}\,d\zeta \quad \text{für alle } z \in B(a;r).$$

  Ferner ist

$$f^{(k)}(z) = \frac{1}{2\pi i}\int_C \frac{f(\zeta)}{(\zeta - z)^{k+1}}\,d\zeta \quad \text{für alle } z \in B(a;r),\ k \in \mathbb{N}.$$

  Einen Beweis findet man bei E. M. Stein, R. Shakarchi (2003, Theorem 4.1 auf S. 45, Corollary 4.2 auf S. 47).

- *Identitätssatz*: Seien f, $g\colon G \longrightarrow \mathbb{C}$ holomorphe Funktionen auf dem Gebiet $G \subset \mathbb{C}$ und $\{z_k\} \subset G$ eine Folge, die einen Häufungspunkt $z_0 \in G$ besitzt und für die $f(z_k) = g(z_k)$ für alle $k \in \mathbb{N}$. Dann ist $f(z) = g(z)$ für alle $z \in G$.

  Einen Beweis findet man z. B. bei E. M. Stein, R. Shakarchi (2003, Theorem 4.8 auf S. 52). Aus dem Identitätssatz folgt, dass die Nullstellen einer auf einem Gebiet holomorphen, nichtkonstanten Funktion isoliert sind, dass es also zu jeder Nullstelle p von f eine Umgebung von p gibt, auf der f keine weitere Nullstelle besitzt.

- *Satz von Rouché*: Seien f, $g\colon G \longrightarrow \mathbb{C}$ holomorph auf dem Gebiet $G \subset \mathbb{C}$ und $\operatorname{cl}(B(a;r)) \subset G$ mit einem $r > 0$. Haben f und g keine Nullstelle auf dem Kreis $\gamma := \partial B(a;r)$ und ist $|f(z) - g(z)| < |g(z)|$ für alle $z \in \gamma$, so haben f und g gleich viele Nullstellen (entsprechend ihrer Vielfachheit gezählt) in $B(a;r)$.

  Einen Beweis findet man z. B. bei E. M. Stein, R. Shakarchi (2003, S. 91).

- *Riemannscher Hebbarkeitssatz*: Sei $f\colon B(z_0;r) \setminus \{z_0\} \longrightarrow \mathbb{C}$ holomorph und f auf $B(z_0;r) \setminus \{z_0\}$ beschränkt. Dann existiert eine holomorphe Abbildung $\hat{f}\colon B(z_0;r) \longrightarrow \mathbb{C}$ mit $\hat{f}(z) = f(z)$ für alle $z \in B(z_0;r) \setminus \{z_0\}$.

  Einen Beweis findet man z. B. bei E. M. Stein, R. Shakarchi (2003, Theorem 3.1 auf S. 84).

- *Residuensatz*: Sei $G \subset \mathbb{C}$ eine offene Menge, f auf G außer in $z_1, \ldots, z_N \in G$ holomorph. Die Punkte $z_1, \ldots, z_N$ seien Pole von f. Ist dann $\gamma \subset G$ ein Kreis, welcher $z_1, \ldots, z_N$ im Innern enthält, so ist

$$\int_\gamma f(z)\,dz = 2\pi i \sum_{k=1}^{N} \operatorname{res}_{z_k}(f).$$

Allgemeiner kann man hier den Kreis durch sogenannte *toy contour* ersetzen, worunter z. B. auch Dreiecke und Rechtecke fallen.

Einen Beweis findet man bei E. M. Stein, R. Shakarchi (2003, S. 76).

Dann gilt:

1. *Satz von Liouville*: Ist $f\colon \mathbb{C} \longrightarrow \mathbb{C}$ holomorph (eine auf ganz $\mathbb{C}$ holomorphe Funktion nennt man eine *ganze* Funktion) und beschränkt, so ist f auf $\mathbb{C}$ konstant.

   Denn: Wegen der Integralformel von Cauchy ist

   $$f'(z) = \frac{1}{2\pi i} \int_{\partial B(z;r)} \frac{f(\zeta)}{\zeta - z(^2}\, d\zeta$$

   für alle $z \in \mathbb{C}$ und alle $r > 0$. Da f beschränkt ist, existiert eine Konstante $c > 0$ mit $|f(\zeta)| \leq c$ für alle $\zeta \in \mathbb{C}$. Daher ist

   $$|f'(z)| \leq \frac{1}{2\pi} \int_{\partial B(z;r)} \frac{|f'(\zeta)|}{|\zeta - z|^2}\, d\zeta \leq \frac{c}{r}$$

   für alle $z \in \mathbb{C}$. Mit $r \to \infty$ folgt $f'(z) = 0$ für alle $z \in \mathbb{C}$ und hieraus, dass f konstant ist.

   Der Satz von Liouville zeigt, dass man im Riemannschen Abbildungssatz $G \neq \mathbb{C}$ voraussetzen muss, da es keine holomorphe Abbildung von $\mathbb{C}$ nach $B(0;1)$ geben kann, die injektiv ist.

2. *Open Mapping Theorem*: Ist $G \subset \mathbb{C}$ ein Gebiet und ist $f\colon G \longrightarrow \mathbb{C}$ nichtkonstant und holomorph auf G, so ist $f(U) \subset \mathbb{C}$ offen für jede offene Menge $U \subset G$.

   Denn: Mit $B(p;r) := \{z \in \mathbb{C} : |z - p| < r\}$ bezeichnen wir die offene Kreisscheibe mit dem Mittelpunkt $p \in \mathbb{C}$ und dem Radius $r > 0$. Sei $U \subset G$ offen. Wir haben zu zeigen, dass es zu jedem $q \in f(U)$ ein $\epsilon > 0$ mit $B(q;\epsilon) \subset f(U)$ gibt. Zu $q \in f(U)$ wähle man ein $p \in U$ mit $f(p) = q$ und definiere $g\colon G \longrightarrow \mathbb{C}$ durch $g(\zeta) := f(\zeta) - q$. Dann ist g eine nichtkonstante holomorphe Funktion auf dem Gebiet G. Die Nullstellen einer nichtkonstanten holomorphen Funktion sind isoliert, insbesondere ist die Nullstelle p von g isoliert. Dies ist eine Folge des *Identitätssatzes für holomorphe Funktionen*. Daher existiert ein $r > 0$ mit $\mathrm{cl}(B(p;r)) \subset U$ und der Eigenschaft, dass g auf $\mathrm{cl}\, B(p;r)\backslash\{p\}$ nicht verschwindet. Da $\partial B(p;r)$ kompakt ist und keine Nullstelle von g enthält, existiert ein $\epsilon > 0$ mit $|g(\zeta)| > \epsilon$ für alle $\zeta \in \partial B(p;r)$. Wir wollen zeigen, dass $B(q;\epsilon) \subset f(U)$. Um dies nachzuweisen, wähle man ein beliebiges $w \in B(q;\epsilon)$ und definiere $h(\zeta) := f(\zeta) - w$. Für $\zeta \in \partial B(p;r)$ ist

   $$|h(\zeta)| = |g(\zeta) + q - w| \geq \underbrace{|g(\zeta)|}_{>\epsilon} - \underbrace{|q - w|}_{<\epsilon} > 0.$$

   Also haben g und h auf $\partial B(p;r)$ keine Nullstelle. Da außerdem

   $$|g(\zeta) - h(\zeta)| = |w - q| < \epsilon < |g(\zeta)| \quad \text{für alle } \zeta \in \partial B(p;r),$$

haben wegen des Satzes von Rouché die Funktionen g und h gleich viele Nullstellen in $B(p;r)$. Daher hat h (mindestens) eine Nullstelle in $B(p;r)$, womit $w \in f(B(p;r)) \subset f(U)$ bewiesen ist. Da $w \in B(q;\epsilon)$ beliebig war, ist $B(q;\epsilon) \subset f(U)$ und damit $f(U)$ offen.

3. *Satz von Casorati-Weierstraß*: Sei $f\colon B(z_0;r) \setminus \{z_0\} \longrightarrow \mathbb{C}$ holomorph auf der punktierten Kreisscheibe

$$B(z_0;r) \setminus \{z_0\} = \{z \in \mathbb{C} : 0 < |z - z_0| < r\}.$$

Ferner sei z_0 eine *wesentliche Singularität* von f, d. h. es ist weder

- f beschränkt auf einer Umgebung von z_0 (f hat eine *hebbare Singularität* in z_0, d. h. es gibt wegen des Riemannschen Hebbarkeitssatzes eine holomorphe Funktion $\hat{f}\colon B(z_0;r) \longrightarrow \mathbb{C}$ mit $\hat{f}(z) = f(z)$ für alle $z \in B(z_0;r) \setminus \{z_0\}$)

noch gilt

- $\lim_{z\to z_0} |f(z)| = \infty$ (f besitzt einen *Pol* in z_0).

Dann ist $f(B(z_0;r) \setminus \{z_0\})$ *dicht* in $\mathbb{C}$, d. h. zu jedem $w \in \mathbb{C}$ existiert eine Folge $\{z_k\} \subset B(z_0;r) \setminus \{z_0\}$ mit $\lim_{k\to\infty} f(z_k) = w$. Gleichbedeutend hiermit ist, dass jede Umgebung eines beliebigen Punktes aus $\mathbb{C}$ ein Element aus $f(B(z_0;r)\setminus\{z_0\})$ enthält.

Denn: Angenommen die Aussage sei nicht richtig, also $f(B(z_0;r) \setminus \{z_0\})$ nicht dicht in $\mathbb{C}$. Dann existieren $w \in \mathbb{C}$ und $\delta > 0$ mit

$$|f(z) - w| > \delta \quad \text{für alle } z \in B(z_0;r) \setminus \{z_0\}.$$

Wir definieren $g\colon B(z_0;r) \setminus \{z_0\} \longrightarrow \mathbb{C}$ durch

$$g(z) := \frac{1}{f(z) - w}.$$

Dann ist

$$f(z) = \frac{1}{g(z)} + w \quad \text{für alle } z \in B(z_0;r) \setminus \{z_0,$$

ferner g auf $B(z_0;r) \setminus \{z_0\}$ holomorph und durch $1/\delta$ beschränkt. Der Riemannschen Hebbarkeitssatz impliziert die Existenz einer holomorphen Funktion $\hat{g}\colon B(z_0;r) \longrightarrow \mathbb{C}$ mit $\hat{g}(z) = g(z)$ für alle $z \in B(z_0;r) \setminus \{z_0\}$. Insbesondere ist $\hat{g}(z_0) = \lim_{z\to z_0} g(z)$. Jetzt unterscheiden wir zwei Fälle. Ist $\hat{g}(z_0) \neq 0$, so ist f auf einer Umgebung von z_0 beschränkt. Ist dagegen $\hat{g}(z_0) = 0$, so ist $\lim_{z\to z_0} |f(z)| = \infty$. Beide Möglichkeiten sind ausgeschlossen, da f in z_0 eine wesentliche Singularität besitzt. Damit ist der Satz von Casorati-Weierstraß bewiesen.

4. *Maximum modulus principle*: Sei $G \subset \mathbb{C}$ ein Gebiet und $f\colon G \longrightarrow \mathbb{C}$ holomorph. Es existiere ein $p \in G$ mit $|f(z)| \leq |f(p)|$ für alle $z \in G$. Dann ist f konstant auf G.

   Denn: Angenommen, f sei nichtkonstant. Wegen des Open Mapping Theorems ist $f(G)$ offen. Daher existiert ein $\epsilon > 0$ mit $B(f(p);\epsilon) \subset f(G)$. Das ist aber ein Widerspruch dazu, dass $|f|$ auf G in $p \in G$ ein Maximum annimmt.

5. *Maximum modulus theorem*: Sei $G \subset \mathbb{C}$ ein beschränktes Gebiet, $f\colon \operatorname{cl} G \longrightarrow \mathbb{C}$ stetig und f auf G holomorph. Dann nimmt $|f|$ sein Maximum auf $\operatorname{cl} G$ in einem Punkt aus ∂G, also einem Randpunkt von G, an.

   Denn: Zunächst beachte man, dass die stetige Funktion $|f|$ ihr Maximum auf der kompakten Menge $\operatorname{cl} G$ annimmt. Ist f nichtkonstant, so sagt das maximum modulus principle aus, dass $|f|$ auf G kein Maximum haben kann. Ein Maximum muss also auf dem Rand ∂G auftreten.

6. *Schwarzsches Lemma*: Sei $f\colon B(0;1) \longrightarrow B(0;1)$ holomorph und $f(0) = 0$. Dann gilt:

   (a) Es ist $|f(z)| \leq |z|$ für alle $z \in B(0;1)$.

   (b) Ist $|f(z_0)| = |z_0|$ für ein $z_0 \in B(0;1) \setminus \{0\}$, so ist f eine Rotation bzw. Drehung, es existiert also ein $\theta \in [0, 2\pi)$ mit $f(z) = e^{i\theta} \cdot z$ für alle $z \in B(0;1)$.

   (c) Es ist $|f'(0)| \leq 1$. Gilt hier Gleichheit, so ist f eine Rotation.

   Denn: Wir definieren $g\colon B(0;1) \longrightarrow \mathbb{C}$ durch

$$g(z) := \begin{cases} \dfrac{f(z)}{z}, & z \neq 0, \\ f'(0), & z = 0. \end{cases}$$

   Dann ist g holomorph auf $B(0;1)$. Wir geben uns ein $r \in (0,1)$ vor. Für $|z| = r$ bzw. $z \in \partial B(0;r)$ ist

$$|g(z)| = \left|\frac{f(z)}{z}\right| \leq \frac{1}{r}.$$

   Wegen des maximum modulus theorem nimmt g sein Maximum auf $\operatorname{cl} B(0;r)$ in einem Randpunkt an, d. h. es ist

$$|g(z)| \leq \frac{1}{r} \quad \text{bzw.} \quad |f(z)| \leq \frac{|z|}{r} \quad \text{für alle } z \in \operatorname{cl} B(0;r).$$

   Mit $r \to 1$ folgt

$$|g(z)| \leq 1 \quad \text{bzw.} \quad |f(z)| \leq |z| \quad \text{für alle } z \in B(0;1).$$

   Damit ist (a) bewiesen.

   Zum Beweis von (b) beachten wir, dass

$$|g(z)| \leq 1 = |g(z_0)| \quad \text{für alle } z \in B(0;1).$$

Die auf $B(0;1)$ holomorphe Funktion g nimmt also in $z_0 \in B(0;1)$ ihr Maximum an. Wegen des maximum modulus principle ist g auf $B(0;1)$ konstant, es existiert also $c \in \mathbb{C}$ mit $f(z) = c \cdot z$ für alle $z \in B(0;1)$. Setzt man hier $z = z_0$, so folgt wegen $|f(z_0)| = |z_0| \neq 0$, dass $|c| = 1$ und damit die Existenz eines $\theta \in [0, 2\pi)$ mit $f(z) = e^{i\theta} \cdot z$ für alle $z \in B(0;1)$. Damit ist auch (b) bewiesen.

Jetzt kommt der Beweis von (c). Wegen $|g(z)| \leq 1$ für alle $z \in B(0;1)$ und

$$g(0) = \lim_{z \to 0} \frac{f(z) - f(0)}{z} = f'(0)$$

ist $|f'(0)| \leq 1$. Gilt hier Gleichheit, so ist $|g(0)| = 1$ und wegen des maximum modulus principle ist g konstant auf $B(0;1)$, es existiert also ein $c \in \mathbb{C}$ mit $g(z) = c$ für alle $z \in B(0;1)$. Setzt man hier $z = 0$, so erhält man $|c| = 1$ und damit die Behauptung.

7. Sei $G \subset \mathbb{C}$ ein Gebiet und $f\colon G \longrightarrow f(G) \subset \mathbb{C}$ holomorph und injektiv. Dann gilt:

   (a) Es ist $f'(z) \neq 0$ für alle $z \in G$.

   (b) Die Umkehrabbildung $f^{-1}\colon f(G) \longrightarrow G \subset \mathbb{C}$ ist ebenfalls holomorph und $(f^{-1})'(f(z_0)) = [f'(z_0)]^{-1}$ für alle $z_0 \in G$.

   Denn[44]: Angenommen, es existiert ein $z_0 \in G$ mit $f'(z_0) = 0$. Wäre die holomorphe Funktion f' konstant, so wäre $f(z) = f(z_0)$ ebenfalls konstant, ein Widerspruch zur Injektivität von f. Die Nullstellen einer nichtkonstanten holomorphen Funktion sind isoliert. Daher existiert ein $r_0 > 0$ mit $\operatorname{cl} B(z_0; r_0) \subset G$ und $f(z) - f(z_0) \neq 0$ und $f'(z) \neq 0$ für alle $z \in \operatorname{cl} B(z_0; r_0) \setminus \{z_0\}$. Es ist

$$f(z) - f(z_0) = a(z - z_0)^k + (z - z_0)^{k+1} h(z)$$

   mit $a \neq 0$, $k \geq 2$ und auf $\operatorname{cl} B(z_0; \epsilon_0)$ holomorpher Funktion h. Mit noch zu bestimmenden $w \neq 0$ setzen wir

$$\phi(z) := f(z) - f(z_0) - w, \qquad \gamma(z) := a(z - z_0)^k - w$$

   sodass

$$\phi(z) - \gamma(z) = (z - z_0)^{k+1} h(z).$$

   Wir wollen den Satz von Rouché auf einer geeigneten Kugel $B(z_0; r)$ mit $r \in (0, r_0]$ und ϕ, γ (statt f, g) anwenden. Hierzu wählen wir ein $r \in (0, r_0]$ und anschließend ein $w \neq 0$ so, dass ϕ und γ keine Nullstelle auf $\partial B(z_0; r)$ besitzen und $|\phi(z) - \gamma(z)| < |\gamma(z)|$ für alle $z \in \partial B(z_0; r)$ gilt. Wenn wir gezeigt haben, dass dies möglich ist, sagt uns der Satz von Rouché, dass die Funktionen ϕ und γ gleich viele Nullstellen (entsprechend ihrer Vielfachheit gezählt) in $B(z_0; r)$ besitzen. Da

[44] Der folgende Beweis ist eine ausführliche Ausarbeitung der Darstellung bei E. M. Stein, R. Shakarchi (2003, p. 206).

aber die Funktion bzw. das Polynom γ für $w \neq 0$ und $w/a = |w/a|e^{i\theta}$ die $k \geq 2$ Nullstellen

$$z_j := |w/a|^{1/k} e^{i(\theta + j2\pi)/k}, \quad j = 0, \ldots, k-1,$$

die für $|w| < |a|r^k$ in $B(z_0; r)$ liegen, folgt aus dem Satz von Rouché, dass auch $\phi(z) = f(z) - f(z_0) - w$ mindestens zwei Nullstellen $u_1, u_2 \in B(z_0; r)$ besitzt. Wegen $w \neq 0$ sind u_1, u_2 von z_0 verschieden. Ist $u_1 \neq u_2$, so hat man einen Widerspruch zur Injektivität von f. Ist dagegen $u_1 = u_2$, so ist u_1 eine Nullstelle von ϕ der Vielfachheit ≥ 2, d. h. es ist $\phi'(u_1) = f'(u_1) = 0$. Wegen $u_1 \neq z_0$ ist das aber ein Widerspruch dazu, dass f' in $\operatorname{cl} B(z_0; r_0) \setminus \{z_0\}$ keine Nullstelle besitzt. Zur Bestimmung von $r \in (0, r_0]$ und $w \neq 0$ definieren wir

$$\epsilon(r) := \min_{z \in \partial B(z_0; r)} |f(z) - f(z_0)| > 0, \qquad M := \max_{z \in \operatorname{cl} B(z_0; r_0)} |h(z)|.$$

Nun wählen wir $r \in (0, r_0]$ so klein, dass $r < |a|/M$. Dann ist

$$r^k|a| - r^{k+1}M = r^k(\underbrace{|a| - rM}_{>0}) > 0.$$

Anschließend wähle man ein $w \neq 0$ mit

$$|w| < \epsilon(r), \qquad |w| < |a|r^k, \qquad |w| < r^k|a| - r^{k+1}M.$$

Für $z \in \partial B(z_0; r)$ bzw. $|z - z_0| = r$ ist dann

$$|\phi(z)| = |f(z) - f(z_0) - w| \geq |f(z) - f(z_0)| - |w| \geq \epsilon(r) - |w| > 0,$$

und

$$|\gamma(z)| = |a(z - z_0)^k - w| \geq |a|r^k - |w| > 0,$$

sodass ϕ und γ auf $\partial B(z_0; r)$ keine Nullstelle besitzen. Für $z \in \partial B(z_0; r)$ ist weiter

$$\begin{aligned} |\phi(z) - \gamma(z)| &= |(z - z_0)^{k+1} h(z)| \\ &\leq r^{k+1} M \\ &< r^k|a| - |w| \\ &\leq |a(z - z_0)^k - w| \\ &= |\gamma(z)| \end{aligned}$$

bei der angegebenen Wahl von r und w. Damit ist die Aussage (a) bewiesen.

Nun zum Beweis von (b). Als auf G injektive Abbildung ist f auf G nichtkonstant. Wegen des Open Mapping Theorems ist f eine offene Abbildung, d. h. das Bild $f(U)$ einer offenen Menge $U \subset G$ unter f ist offen. Dies bedeutet aber, dass $f^{-1}\colon f(G) \longrightarrow \mathbb{C}$ stetig ist. Um dies einzusehen, seien $x = f(\xi) \in f(G)$ und $\epsilon > 0$ mit $B(\xi; \epsilon) \subset G$ vorgegeben. Da $f(B(\xi; \epsilon))$ wegen des Open Mapping Theorems offen ist, existiert ein $\delta > 0$ mit $B(f(\xi); \delta) \subset f(B(\xi; \epsilon))$. Zu jedem $y \in B(x; \delta) = B(f(\xi); \delta)$ existiert genau ein $\eta \in B(\xi; \epsilon)$ mit $y = f(\eta)$. Daher gilt die Implikation

$$y \in f(G),\ |x - y| < \delta \Longrightarrow |f^{-1}(x) - f^{-1}(y)| = |\xi - \eta| < \epsilon$$

und das beweist die Stetigkeit von $f^{-1}\colon f(G) \longrightarrow \mathbb{C}$ in $x = f(\xi)$. Wir wissen, dass $f(G)$ offen und $f^{-1}\colon f(G) \longrightarrow G \subset \mathbb{C}$ stetig ist und $f^{-1}(f(z)) = z$ sowie $f'(z) \neq 0$ für alle $z \in G$ gilt. Um die Differenzierbarkeit von f^{-1} auf $f(G)$ nachzuweisen, wählen wir ein beliebiges $w_0 = f(z_0) \in f(G)$ sowie $\epsilon > 0$. Da f in z_0 differenzierbar ist und $f'(z_0) \neq 0$ gilt, existiert ein $\delta_0 > 0$ mit

$$|z - z_0| < \delta_0 \Longrightarrow \left| \frac{z - z_0}{f(z) - f(z_0)} - \frac{1}{f'(z_0)} \right| < \epsilon.$$

Da weiter $f^{-1}\colon f(G) \longrightarrow G \subset \mathbb{C}$ stetig ist, wie wir uns gerade eben überlegt haben, existiert ein $\delta > 0$ mit $B(w_0; \delta) \subset f(G)$ und

$$|w - w_0| < \delta \Longrightarrow |f^{-1}(w) - f^{-1}(w_0)| < \delta_0.$$

Für $w = f(z) \in B(w_0; \delta)$ ist dann $z \in B(z_0; \delta_0)$ und folglich

$$\left| \frac{f^{-1}(w) - f^{-1}(w_0)}{w - w_0} - \frac{1}{f'(z_0)} \right| = \left| \frac{z - z_0}{f(z) - f(z_0)} - \frac{1}{f'(z_0)} \right| < \epsilon,$$

womit die Differenzierbarkeit und damit Holomorphie von f^{-1} auf $f(G)$ und

$$(f^{-1})'(f(z_0)) = \frac{1}{f'(z_0)}$$

bewiesen ist.

8. Eine konforme Abbildung ist *winkeltreu*. Das soll heißen: Schneiden sich zwei Kurven unter einem bestimmten Winkel, so schneiden sich die Bilder der beiden Kurven unter einer konformen Abbildung unter demselben Winkel. Um dies zu präzisieren, erinnern wir daran, dass die eindeutige *Polardarstellung* einer von Null verschiedenen komplexen Zahl z, also von $z \in \mathbb{C} \setminus \{0\}$, durch

$$z = |z| e^{i \arg(z)}$$

mit $\arg(z) \in (-\pi, \pi]$ gegeben ist. Die Abbildung $\arg\colon \mathbb{C} \setminus \{0\} \longrightarrow (-\pi, \pi]$ nennen wir den *Hauptzweig* der *Argumentfunktion*. Weiter ist der *Winkel* $\theta = \sphericalangle(z, w)$ zweier komplexer Zahlen $z, w \in \mathbb{C} \setminus \{0\}$ durch

$$e^{i\theta} = \frac{w}{|w|} \Big/ \frac{z}{|z|} = \frac{w\overline{z}}{|w|\,|z|}$$

(natürlich nur modulo 2π) bestimmt. Dann ist also

$$\frac{w}{|w|} = e^{i\theta} \frac{z}{|z|}.$$

Dies veranschaulichen wir uns in Abbildung 67. Hier ist $z = 3 + 2i$, $w = -4 + i$ und daher (Rechnung mit Octave und `format long`).

$$\frac{w\overline{z}}{|w|\,|z|} = -6.726727939963125e-01 + 7.399400733959438e-01i.$$

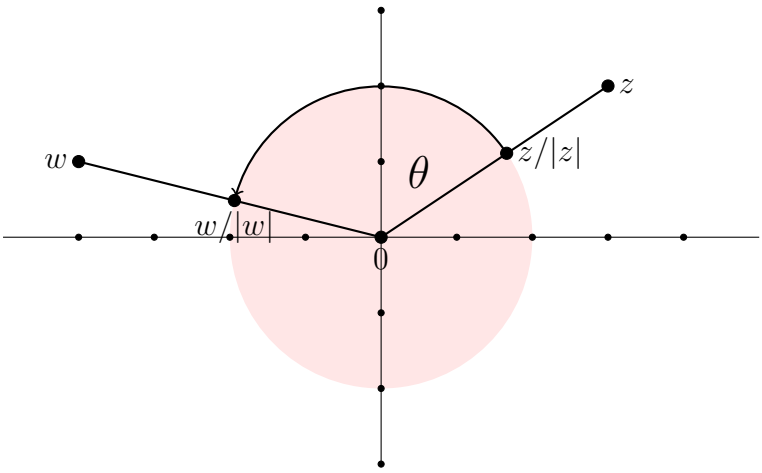


Abbildung 67: Der Winkel $\theta = \sphericalangle(z, w)$ zwischen $z, w \in \mathbb{C} \setminus \{0\}$

Dann ist (Berechnung durch `theta=arg(w/z)` oder `theta=angle(w/z)`)

$$\theta = \sphericalangle(z, w) = 2.308611386915361$$

bzw. (Gradangabe)

$$\theta = 132.2736890060937°.$$

Sind $\gamma_1, \gamma_2 \colon [a, b] \longrightarrow \mathbb{C}$ zwei sich schneidende C^1-Kurven, also etwa $\gamma_1(t_0) = \gamma_2(t_0)$ und $\gamma_1'(t_0), \gamma_2'(t_0) \in \mathbb{C} \setminus \{0\}$. Dann schneiden sich γ_1 und γ_2 unter dem Winkel $\theta = \sphericalangle(\gamma_1'(t_0), \gamma_2'(t_0))$, da $\gamma_1(t_0) + \gamma_1'(t_0)(t - t_0)$ bzw. $\gamma_2(t_0) + \gamma_2'(t_0)(t - t_0)$ die Tangenten an γ_1 bzw. γ_2 in $\gamma_1(t_0) = \gamma_2(t_0)$ sind. Es ist also

$$e^{i\theta} = \frac{\gamma_2'(t_0)}{|\gamma_2'(t_0)|} \Big/ \frac{\gamma_1'(t_0)}{|\gamma_1'(t_0)|}.$$

Ist nun f eine auf einem Gebiet $G \subset \mathbb{C}$ konforme Abbildung, so sind $f \circ \gamma_1$ und $f \circ \gamma_2$ zwei C^1-Kurven, die sich in $(f \circ \gamma_1)(t_0) = (f \circ \gamma_2)(t_0)$ schneiden. Dies geschieht unter einem Winkel $f(\theta) = \sphericalangle((f \circ \gamma_1)'(t_0), (f \circ \gamma_2)'(t_0))$, wobei

$$\begin{aligned} e^{if(\theta)} &= \frac{(f \circ \gamma_2)'(t_0)}{|(f \circ \gamma_2)'(t_0)|} \Big/ \frac{(f \circ \gamma_1)'(t_0)}{|(f \circ \gamma_1)'(t_0)|} \\ &= \frac{f'(\gamma_2(t_0))\gamma_2'(t_0)}{|f'(\gamma_2(t_0))\gamma_2'(t_0)|} \Big/ \frac{f'(\gamma_1(t_0))\gamma_1'(t_0)}{|f'(\gamma_1(t_0))\gamma_1'(t_0)|} \\ &= \frac{\gamma_2'(t_0)}{|\gamma_2'(t_0)|} \Big/ \frac{\gamma_1'(t_0)}{|\gamma_1'(t_0)|} \\ &= e^{i\theta} \end{aligned}$$

und folglich $\theta = f(\theta)$ modulo 2π, d. h. eine konforme Abbildung ist winkeltreu.

9. *Existenz einer holomorphen Quadratwurzel*: Sei $G \subset \mathbb{C}$ offen und einfach zusammenhängend, $f \colon G \longrightarrow \mathbb{C}$ holomorph und $f(z) \neq 0$ für alle $z \in G$. Dann existiert eine holomorphe Funktion $g \colon G \longrightarrow \mathbb{C}$ mit $g(z)^2 = f(z)$ für alle $z \in G$.

Denn: Da f auf G nicht verschwindet und holomorph ist, ist f'/f eine auf G holomorphe Funktion. Da G einfach zusammenhängend ist, existiert eine holomorphe Funktion $h\colon G \longrightarrow \mathbb{C}$ mit $h'(z) = f'(z)/f(z)$ für alle $z \in G$. Um dies einzusehen wähle man $z_0 \in G$ beliebig und definiere

$$h(z) := \int_\gamma \frac{f'(w)}{f(w)}\, dw,$$

wobei γ eine z_0 und z verbindende Kurve in G ist. Da G *einfach* zusammenhängend ist, ist h wohldefiniert und man kann leicht zeigen, dass h eine gesuchte Funktion (also auf G holomorph mit $h' = f'/f$) ist, siehe E. M. Stein, R. Shakarchi (2003, Theorem 5.2 auf S. 96). Mit h ist auch $h + c$ mit $c \in \mathbb{C}$ auf G holomorph mit Ableitung gleich f'/f. Daher können wir annehmen, dass $e^{h(z_0)} = f(z_0)$. Für $z \in G$ ist dann

$$\begin{aligned}
\frac{d}{dz}(f(z)e^{-h(z)}) &= f'(z)e^{-h(z)} - f(z)h'(z)e^{-h(z)} \\
&= \left(f'(z) - f(z)\frac{f'(z)}{f(z)}\right)e^{-h(z)} \\
&= 0.
\end{aligned}$$

Daher ist $f(z)e^{-h(z)}$ auf G konstant und folglich $f(z) = e^{h(z)}$ für alle $z \in G$ nach Wahl von h. Mit $g(z) := e^{h(z)/2}$ hat man die gesuchte holomorphe Quadratwurzel zu f gefunden.

Die Voraussetzung im Riemannschen Abbildungssatz, dass die Menge G *einfach zusammenhängend* ist, wird "nur" dazu benutzt, die Existenz einer holomorphen Quadratwurzel einer nicht verschwindende, holomorphen Funktion zu sichern.

10. *Satz von Goursat*: Sei $G \subset \mathbb{C}$ offen und einfach zusammenhängend, $T \subset G$ ein Dreieck[45]. Sei $f\colon G \longrightarrow \mathbb{C}$ holomorph. Dann ist

$$\int_T f(z)\, dz = 0.$$

Denn[46]: Sei $T^{(0)} := T$ das Ausgangsdreieck, es sei etwa positiv orientiert. Man halbiere die drei Seiten und verbinde die Mittelpunkte. Hierdurch gewinnt man vier kleinere Dreiecke $T_1^{(1)}, T_2^{(1)}, T_3^{(1)}, T_4^{(1)}$, siehe Abbildung 68. Für ein Dreieck T bezeichne man mit $l(T)$ die *Länge*, d. h. die Summe der drei Seitenlängen, und mit $\mathrm{diam}(\Delta)$ den *Durchmesser* der konvexen Hülle Δ von T, was mit der Länge der längsten Seite von T übereinstimmt. Daher ist $l(T_j^{(1)}) = \frac{1}{2}l(T^{(0)})$, $\mathrm{diam}(T_j^{(1)}) = \frac{1}{2}\mathrm{diam}(T^{(0)})$, $j = 1, 2, 3, 4$. Die Orientierung der kleineren Dreiecke

[45] Unter dem *Dreieck* T verstehen wir nur die Vereinigung der drei Seiten und nicht etwa die konvexe Hülle Δ von T, so wie man zwischen Kreis und Kreisscheibe unterscheidet. Die konvexe Hülle von T ist eine kompakte Menge Δ mit $T = \partial\Delta$.

[46] Siehe auch E. M. Stein, R. Shakarchi (2003, Theorem 1.1 auf S. 34)

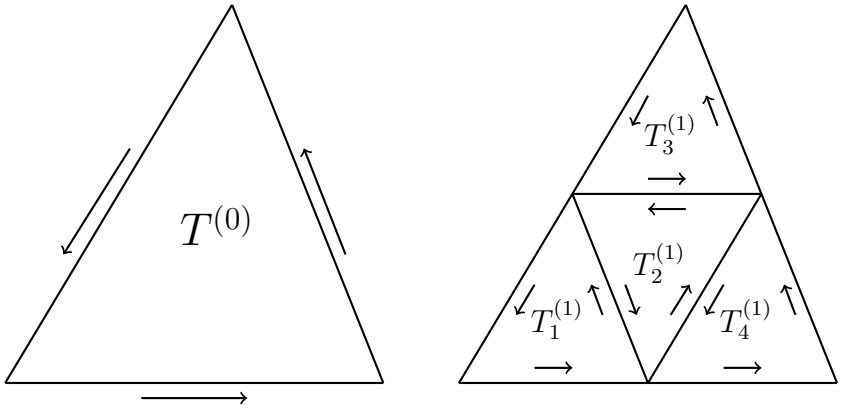


Abbildung 68: Beweis des Satzes von Goursat

wird so gewählt, dass sie konsistent mit der Orientierung des gegebenen Dreiecks ist. Dann ist

$$\int_{T^{(0)}} f(z)\,dz = \sum_{j=1}^{4} \int_{T_j^{(1)}} f(z)\,dz,$$

da die Integrale über die inneren Seiten der kleineren Dreiecke sich gegenseitig wegheben. Mit $T^{(1)}$ bezeichnen wir das (bzw. ein: es muss nicht eindeutig sein) Dreieck $T_j^{(1)}$, für das $|\int_{T_j(1)} f(z)\,dz|$, $j = 1, 2, 3, 4$, maximal ist. Dann ist also

$$\left|\int_{T^{(1)}} f(z)\,dz\right| \geq \left|\int_{T_j^{(1)}} f(z)\,dz\right|, \quad j = 1, 2, 3, 4,$$

und folglich

$$\left|\int_{T^{(0)}} f(z)\,dz\right| \leq 4 \left|\int_{T^{1)}} f(z)\,dz\right|.$$

Diesen Prozess kann man nun auf $T^{(1)}$ anwenden. Hierdurch erhält man eine Folge $\{T^{(n)}\}$ von Dreiecken mit

$$\left|\int_{T^{(n)}} f(z)\,dz\right| \leq 4 \left|\int_{T^{(n+1)}} f(z)\,dz\right|$$

und

$$l(T^{(n+1)}) = \frac{1}{2} l(T^{(n)}), \quad d(T^{(n+1)}) = \frac{1}{2} d(T^{(n)}).$$

Mit $l := l(T)$ und $d := \operatorname{diam}(\Delta)$ ist daher

$$\left|\int_{T} f(z)\,dz\right| \leq 4^n \left|\int_{T^{(n)}} f(z)\,dz\right|$$

und

$$l(T^{(n)}) = \left(\frac{1}{2}\right)^n l, \qquad d(T^{(n)}) = \left(\frac{1}{2}\right)^n d.$$

Bezeichnet man mit $\Delta^{(n)}$ die konvexe Hülle des Dreiecks $T^{(n)}$, so erhalten wir eine Folge $\{\Delta^{(n)}\}$ kompakter Mengen mit

$$\Delta := \Delta^{(0)} \supset \Delta^{(1)} \supset \Delta^{(2)} \supset \cdots$$

und der Eigenschaft, dass $\operatorname{diam}(\Delta^{(n)}) = (\frac{1}{2})^n d$ mit $n \to \infty$ gegen 0 geht. Dann existiert genau ein z_0, der in allen $\Delta^{(n)}$ enthalten ist, d. h. $\bigcap_{n=1}^{\infty} \Delta^{(n)} = \{z_0\}$ ist eine einpunktige Menge[47]. Da f in z_0 holomorph ist bzw. die Ableitung $f'(z_0)$ besitzt, ist

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \phi(z)(z - z_0)$$

mit $\lim_{z \to z_0} \psi(z) = 0$. Integriert man hier über das (geschlossene) Dreieck $T^{(n)}$, so fallen die ersten beiden Terme fort, da es hierzu eine Stammfunktion (engl: primitive) gibt:

$$\int_{T^{(n)}} [f(z_0) + f'(z_0)(z - z_0)]\, dz = \int_{T^{(n)}} \frac{d}{dz} \left[f(z_0)(z - z_0) + \frac{1}{2} f'(z_0)(z - z_0)^2 \right] dz = 0,$$

siehe auch E. M. Steni, R. Shakarchi (2003, Corollary 3.3 auf S. 23). Mit

$$\epsilon_n := \sup_{z \in T^{(n)}} |\psi(z)|$$

erhalten wir die folgenden Abschätzungen:

$$\begin{aligned}
\left| \int_T f(z)\, dz \right| &\leq 4^n \left| \int_{T^{(n)}} f(z)\, dz \right| \\
&= 4^n \left| \int_{T^{(n)}} \psi(z)(z - z_0)\, dz \right| \\
&\leq 4^n \int_{T^{(n)}} \underbrace{|\psi(z)|}_{\leq \epsilon_n} \underbrace{|z - z_0|}_{\leq \operatorname{diam}(\Delta^{(n)})}\, dz \\
&\leq 4^n\, \epsilon_n\, l(T^{(n)}) \operatorname{diam}(\Delta^{(n)}) \\
&= \epsilon_n l d,
\end{aligned}$$

woraus mit $n \to \infty$ die Behauptung folgt.

11. *Satz von Morera*: Sei $G \subset \mathbb{C}$ und sei $f\colon G \longrightarrow \mathbb{C}$ eine stetige Funktion mit der Eigenschaft, dass $\int_T f(z)\, dz = 0$ für jedes Dreieck $T \subset G$ (genauer: (geschlossener) Pfad oder Weg auf einem Dreieck). Dann ist f holomorph auf G.

    Denn[48]: O. B. d. A. ist $= B(a; r)$ eine offene Kugel um einen Punkt $a \in G$. Wir definieren $F\colon G \longrightarrow \mathbb{C}$ durch

$$F(z) := \int_{[a,z]} f(w)\, dw$$

[47] Der Beweis für diese Aussage (siehe z. B. E. M. Stein, R. Shakarchi (2003, Proposition 1.4 auf S. 7)) ist einfach. Man wähle ein $z_n \in \Delta^{(n)}$. Wegen $\lim_{n \to \infty} \operatorname{diam}(\Delta^{(n)}) = 0$ ist $\{z_n\}$ eine Cauchy-Folge. Daher existiert $z_0 := \lim_{n \to \infty} z_n$. Weiter ist $z_0 \in \Delta^{(n)}$ für alle n und z_0 ist wegen $\operatorname{diam}(\Delta^{(n)}) \to 0$ der einzige Punkt mit dieser Eigenschaft.

[48] Siehe auch E. M. Stein, R. Shakarchi (2003, Theorem 5.1 auf S. 53)

und zeigen, dass F in z_0 differenzierbar ist mit $F'(z_0) = f(z_0)$. Als Ableitung einer holomorphen Funktion ist mit F dann auch f holomorph. Betrachtet man das Dreieck $T = [a, z_0, z, a]$ und wendet die Voraussetzung

$$\int_T f(w)\,dw = 0$$

an, so erhält man

$$0 = \int_{[a,z_0]} f(w)\,dw + \int_{[z_0,z]} f(w)\,dw + \underbrace{\int_{[z,a]} f(w)\,dw}_{=-F(z)} = 0,$$

damit

$$F(z) = \int_{[a,z_0]} f(w)\,dw + \int_{[z_0,z]} f(w)\,dw = F(z_0) + \int_{[z_0,z]} f(w)\,dw$$

und

$$\frac{F(z) - F(z_0)}{z - z_0} = \frac{1}{z - z_0} \int_{z_0,z]} f(w)\,dw.$$

Hieraus folgt

$$\frac{F(z) - F(z_0)}{z - z_0} - f(z_0) = \frac{1}{z - z_0} \int_{[z,z_0]} [f(w) - f(z_0)]\,dw.$$

Da f als auf G und insbesondere in z_0 als stetig vorausgesetzt ist, existiert zu vorgegebenem $\epsilon > 0$ ein $\delta > 0$ mit

$$|w - z_0| < \delta \Longrightarrow |f(w) - f(z_0)|.$$

Für $|z - z_0| < \delta$ ist dann

$$\left| \frac{F(z) - F(z_0)}{z - z_0} - f(z_0) \right| \leq \max_{t \in [0,1]} |f(z_0 + t(z - z_0)) - f(z_0)| < \epsilon,$$

womit dann die Holomorphie von F und dann auch von f bewiesen ist.

12. *Schwarzsches Spiegelungsprinzip*: Sei $G \subset \mathbb{C}$ eine offene Menge, welche *symmetrisch bezüglich der reellen Achse* ist, d. h. es sei $z \in G$ genau dann, wenn $\overline{z} \in G$. Man definiere

$$G^+ := \{z \in G : \mathrm{Im}(z) > 0\}, \qquad G^- := \{z \in G : \mathrm{Im}(z) < 0\}$$

und

$$I := \{z \in G : \mathrm{Im}(z) = 0\}.$$

Wir veranschaulichen uns die Situation in Abbildung 69. Sei $f\colon G^+ \cup I \longrightarrow \mathbb{C}$ auf $G^+ \cup I$ stetig, auf I reell und auf G^+ holomorph. Dann existiert eine auf G holomorphe Funktion $F\colon G \longrightarrow \mathbb{C}$ mit $F(z) = f(z)$ für alle $z \in G^+ \cup I$.

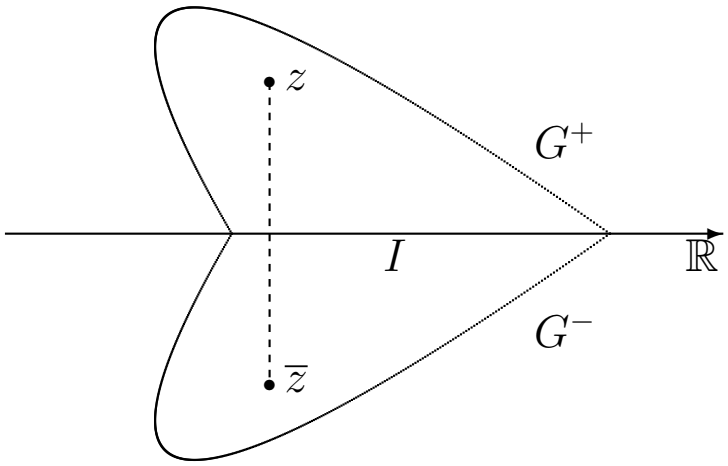


Abbildung 69: Das Schwarzsche Spiegelungsprinzip

Denn: Man definiere $F\colon G \longrightarrow \mathbb{C}$ durch

$$F(z) := \begin{cases} f(z), & z \in G^+ \cup I, \\ \overline{f(\overline{z})}, & z \in G^-. \end{cases}$$

Nach Voraussetzung ist F auf G^+ und dann auch auf G^- holomorph. Denn mit $z \in G^-$ und $|h| = |\overline{h}| \to 0$ ist $z + h \in G^-$ und es gilt

$$\frac{F(z+h) - F(z)}{h} = \frac{\overline{f(\overline{z+h})} - \overline{f(\overline{z})}}{h} = \overline{\left(\frac{f(\overline{z}+\overline{h}) - f(\overline{z})}{\overline{h}}\right)} \to \overline{f'(\overline{z})},$$

also ist F auf G^- komplex differenzierbar bzw. holomorph. Wir überlegen uns, dass F auf G stetig ist. Da F auf $G^+ \cup I$ mit f übereinstimmt und dort nach Voraussetzung stetig, genügt es

$$z_0 \in I,\ \{z_k\} \subset G^-,\ \lim_{k\to\infty} z_k = z_0 \Longrightarrow \lim_{k\to\infty} F(z_k) = F(z_0)$$

nachzuweisen. Ist aber $z_0 \in I$, $\{z_k\} \subset G^-$ und $\lim_{k\to\infty} z_k = z_0$, so ist

$$\lim_{k\to\infty} F(z_k) = \lim_{k\to\infty} \overline{f(\overline{z}_k)} = \overline{f(\overline{z}_0)} = \overline{f(z_0)} = f(z_0) = F(z_0),$$

wobei wir ausgenutzt haben, dass f auf $G^+ \cup I$ stetig und auf I reell ist. Die Holomorphie von F auf G zeigen wir mit Hilfe des Satzes von Morera. Hierzu haben wir zu zeigen, dass $\int_T F(z)\,dz = 0$ für jedes (positiv orientierte) Dreieck $T \subset G$ ist. Ist T in G^+ oder in G^- enthalten, so ist wegen des Satzes von Goursat natürlich $\int_T F(z)\,dz = 0$, da F auf G^+ bzw. G^- holomorph ist. Sei nun T in $G^+ \cup I$ bzw. in $G^- \cup I$, nicht aber in G^+ bzw. G^- enthalten. O. B. d. A. sei $T \subset G^+ \cup I$. Dann kann man diesen Fall durch eine Störung $T_\epsilon := T + \epsilon i$ von T, die ganz in G^+ liegt, mit einem Stetigkeitsargument auf den ersten Fall zurückführen. Denn wegen des Satzes von Goursat ist $\int_{T_\epsilon} F(z)\,dz = 0$. In den Abbildungen haben wir unterschieden, ob T nur eine Ecke (Abbildung 70) oder

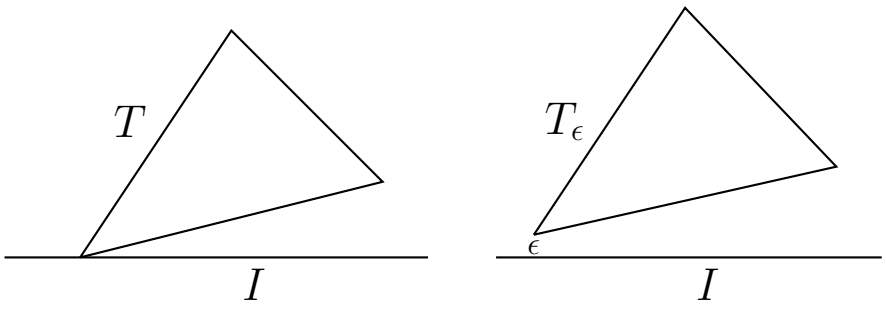


Abbildung 70: Satz von Morera und Schwarzsches Spiegelungsprinzip I

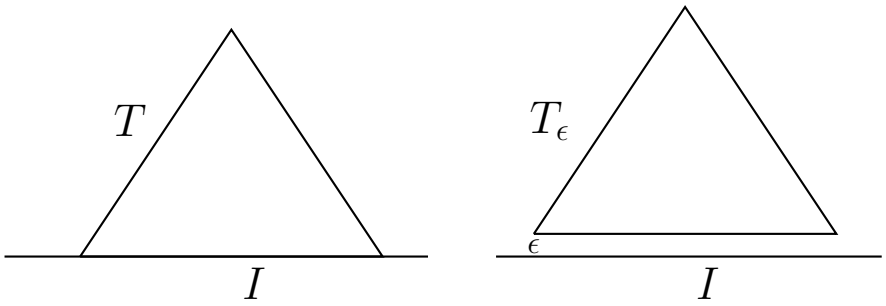


Abbildung 71: Satz von Morera und Schwarzsches Spiegelungsprinzip II

eine ganze Seite (Abbildung 71) mit I teilt. Wegen $\int_{T_\epsilon} F(z)\,dz = 0$ folgt mit $\epsilon \to 0+$ auch $\int_T F(z)\,dz = 0$. Schließlich habe T eine Ecke in G^+ und eine Ecke in G^-. Es sind zwei Fälle zu unterscheiden, nämlich ob die dritte Ecke in I oder in G^+ bzw. G^- liegt. Beide Fälle kann man auf den gerade eben untersuchten Fall zurückführen, indem man das gegebene Dreieck T als Vereinigung von Dreiecken darstellt, die eine Seite oder zwei Seiten mit I teilen. Im ersten Fall (Abbildung 72) ist T die Vereinigung von zwei, im zweiten Fall (Abbildung 73) von drei Dreiecken. Hierbei heben sich Seiten der Teildreiecke, die nicht zum gegebenen Dreieck gehören, gegenseitig weg. Also ist hier

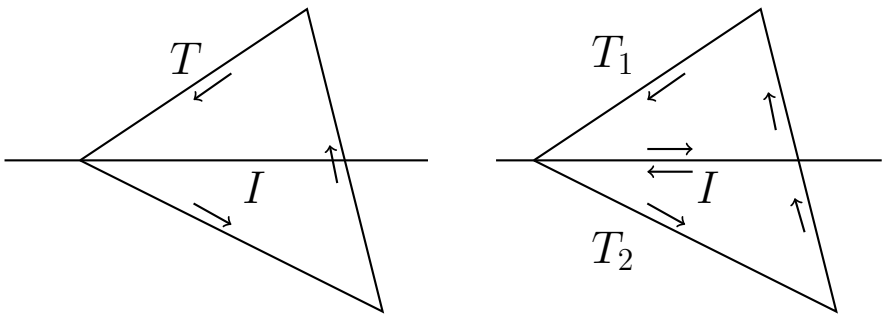


Abbildung 72: Satz von Morera und Schwarzsches Spiegelungsprinzip III

$$\int_T F(z)\,dz = \underbrace{\int_{T_1} F(z)\,dz}_{=0} + \underbrace{\int_{T_2} F(z)\,dz}_{=0} = 0.$$

Den zweiten Fall führt man dadurch auf einen schon behandelten Fall zurück,

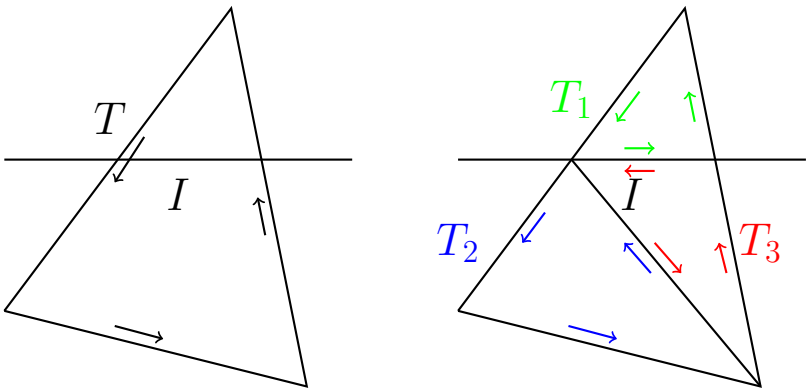


Abbildung 73: Satz von Morera und Schwarzsches Spiegelungsprinzip IV

indem man das gegebene Dreieck T als Vereinigung von drei Dreiecken T_1, T_2, T_3 aus $G^+ \cup I$ bzw. $G^- \cup I$ darstellt. Wegen

$$\int_T F(z)\,dz = \underbrace{\int_{T_1} F(z)\,dz}_{=0} + \underbrace{\int_{T_2} F(z)\,dz}_{=0} + \underbrace{\int_{T_3} F(z)\,dz}_{=0} = 0$$

ist wegen des Satzes von Morera das Schwarzsche Spiegelungsprinzip bewiesen.

13. *Der komplexe Logarithmus*: Der *Hauptzweig des komplexen Logarithmus* ist auf der geschlitzten Ebene

$$\mathbb{C}^- := \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\}$$

definiert durch

$$\operatorname{Log} z := \log r + i\theta,$$

wobei $z = re^{i\theta}$ mit $|\theta| < \pi$ bzw.

$$\operatorname{Log} z := \log |z| + i\arg(z), \quad \arg(z) \in (-\pi, \pi).$$

In Abbildung 74 geben wir links in der z-Ebene einen Kreis sowie einen Strahl

$$K := \{z \in \mathbb{C} : |z| = r\}, \qquad R := \{z \in \mathbb{C} : \arg(z) = \theta\}$$

mit $r > 0$ und $\theta \in (-\pi, \pi)$ an. Rechts findet man in der w-Ebene das Bild unter Log. Es sind ein vertikales Segment bzw. eine horizontale Gerade. Man erkennt sehr schön die Winkeltreue des Logarithmus: Kreis und Strahl sowie Segment und Gerade schneiden sich im rechten Winkel. Der Hauptzweig des Logarithmus ist auf der geschlitzten Ebene $\mathbb{C}^-$ holomorph. Dies kann verallgemeinert werden. Denn es gilt (siehe E. M. Stein, R. Shakarchi (2003, S. 98):

Sei $G \subset \mathbb{C}$ offen und einfachzusammenhängend und es gelte $1 \in G$ und $0 \notin G$. Dann existiert eine Abbildung $F\colon G \longrightarrow \mathbb{C}$ mit

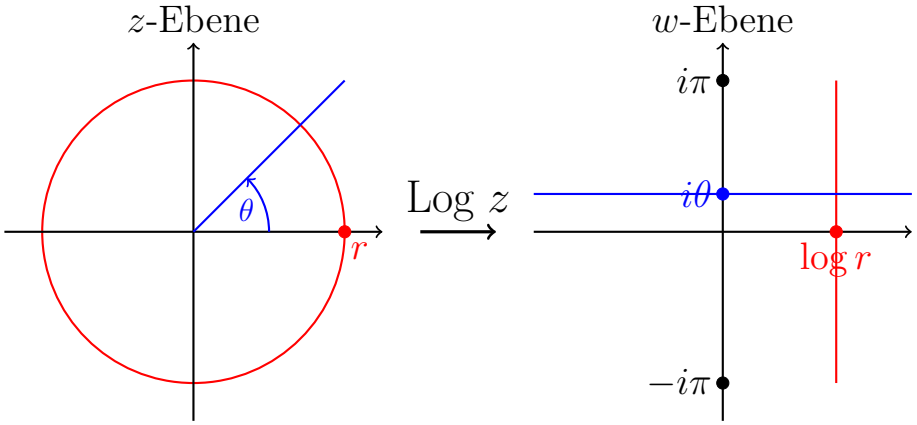


Abbildung 74: Bild von Kreis und Strahl unter Log

(a) F ist holomorph auf G,

(b) $e^{F(z)} = z$ für alle $z \in G$,

(c) Für alle reellen r, für die das Segment[49] $[1, r]$ auf der reellen Achse in G enthalten ist, ist $F(r) = \log r$. Dies ist zumindestens für alle reellen r einer Umgebung von 1 der Fall.

Statt F schreiben wir auch $\log_G$ für den so definierten Zweig des Logarithmus. Für jedes $\alpha \in \mathbb{C}$ ist durch $z^\alpha := e^{\alpha \log_G z}$ eine auf G holomorphe Funktion gegeben.

Denn: Da $0 \notin G$ ist $f(z) := 1/z$ auf G holomorph. Man definiere F als Stammfunktion von f, setze also

$$F(z) := \int_\gamma f(w)\, dw,$$

wobei γ eine Kurve in G ist, die 1 und z miteinander verbindet. Da G einfach zusammenhängend ist, ist F wohldefiniert bzw. von der 1 und z verbindenden Kurve unabhängig. Die Holomorphie von F und $F'(z) = f(z)$ kann ähnlich wie im Beweis des Satzes von Morera gezeigt werden. Damit ist die erste Aussage bewiesen. Zum Beweis der zweiten beachten wir, dass

$$\frac{d}{dz}(ze^{-F(z)}) = e^{-F(z)} - zF'(z)e^{-F(z)} = \underbrace{(1 - zF'(z))}_{=0}e^{-F(z)} = 0.$$

Hieraus folgt leicht, dass

$$0 = \int_\gamma \frac{d}{dz}(ze^{-F(z)})\, dz = ze^{-F(z)} - 1 \cdot e^{-0} = ze^{-F(z)} - 1$$

und damit $e^{F(z)} = z$ für alle $z \in G$. Ist schließlich das Segment $[1, r]$ in G enthalten, so ist

$$F(r) = \int_1^r \frac{dx}{x} = \log r.$$

[49] Das Wort Segment soll ausdrücken, dass auch $r < 1$ sein kann. Es ist einfach die "Kurve" von 1 nach r auf der reellen Achse.

□

10.2 Elementare Beispiele

Beispiel: Sei

$$\mathbb{H} := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$$

die (offene) obere Halbebene in $\mathbb{C}$. Die durch

$$f(z) := \frac{i-z}{i+z}$$

definierte Abbildung $f\colon \mathbb{H} \longrightarrow \mathbb{C}$ ist eine auf $\mathbb{H}$ holomorphe Abbildung (klar, denn natürlich ist f auf $\mathbb{H}$ komplex differenzierbar), f ist auf $\mathbb{H}$ injektiv und es ist $f(\mathbb{H}) = B(0;1)$, d. h. die (offene) obere Halbebene und die offene Einheitskreisscheibe sind konform äquivalent. Zum Beweis der Injektivität nehmen wir an, $z_1, z_2 \in \mathbb{H}$ mit $f(z_1) = f(z_2)$ seien gegeben. Dann ist offenbar der $z_1 = z_2$ bzw. f auf $\mathbb{H}$ injektiv. Für $z = x + iy \in \mathbb{H}$ ist

$$|f(z)| = \left|\frac{-x + i(1-y)}{x + i(1+y)}\right| = \sqrt{\frac{x^2 + (1-y)^2}{x^2 + (1+y)^2}} < 1$$

wegen $y > 0$. Also ist $f(\mathbb{H}) \subset B(0;1)$. Es ist aber sogar $f(\mathbb{H}) = B(0;1)$. Denn ist $w = u + iv \in B(0;1)$ vorgegeben, so setze man

$$z := i\frac{1-w}{1+w} = i\frac{1-u-iv}{1+u+iv} = \frac{2v}{(1+u)^2+v^2} + i\frac{1-(u^2+v^2)}{(1+u)^2+v^2},$$

woraus man $z \in \mathbb{H}$ abliest. Weiter ist

$$f(z) = \frac{i-z}{i+z} = \frac{1-(1-w)/(1+w)}{1+(1-w)/(1+w)} = w$$

und folglich $f(\mathbb{H}) = B(0;1)$, siehe Abbildung 75. Ist $z = x \in \mathbb{R}$ ein Randpunkt von $\mathbb{H}$,

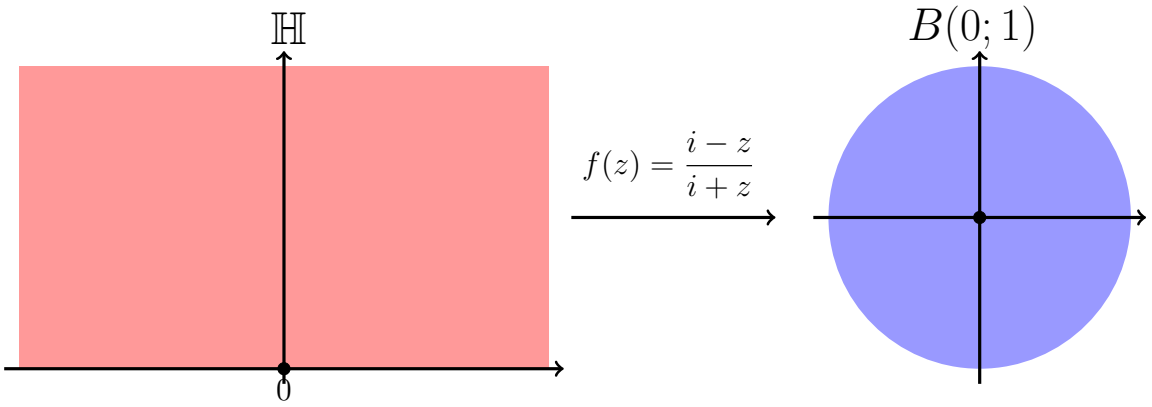


Abbildung 75: $\mathbb{H}$ und $B(0,1)$ sind konform äquivalent

so ist

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

Mit $x = \tan t$, $t \in (-\pi/2, \pi/2)$, ist $f(x) = e^{i2t}$. Der Rand von $\mathbb{H}$, also die reelle Achse $\mathbb{R}$, wird eineindeutig durch f auf den Rand $\partial B(0;1)$ der Einheitskreisscheibe mit Ausnahme des Punktes -1 abgebildet. Dieser Punkt -1 auf dem Einheitskreis entspricht einem Punkt ∞ der oberen Halbebene. □

Beispiel: Sei

$$G := \{z \in \mathbb{C} : |z| < 1,\ \mathrm{Im}(z) > 0\}$$

die (offene) obere Einheitskreishalbscheibe und $f\colon G \longrightarrow \mathbb{C}$ durch

$$f(z) := \frac{1+z}{1-z}$$

definiert. Offensichtlich ist f auf G holomorph und injektiv. Weiter ist

$$f(G) = \{w = u + iv \in \mathbb{C} : u > 0,\ v > 0\},$$

das Bild von G unter f also der (offene) erste Quadrant in $\mathbb{C}$. Denn für $z = x + iy \in G$ ist

$$f(z) = \frac{1+x+iy}{1-x-iy} = \frac{(1-(x^2+y^2))+2iy}{(1-x)^2+y^2}$$

ein Element des ersten Quadranten von $\mathbb{C}$. Ist umgekehrt $w = u + iv \in \mathbb{C}$ mit $u > 0$, $v > 0$ ein Element des ersten Quadranten, so ist

$$z := \frac{w-1}{w+1}$$

ein Element der oberen Einheitshalbkreisscheibe und $f(z) = w$. Damit ist gezeigt, dass f eine konforme Abbildung der oberen Einheitskreisscheibe auf den ersten Quadranten ist, siehe Abbildung 76. □

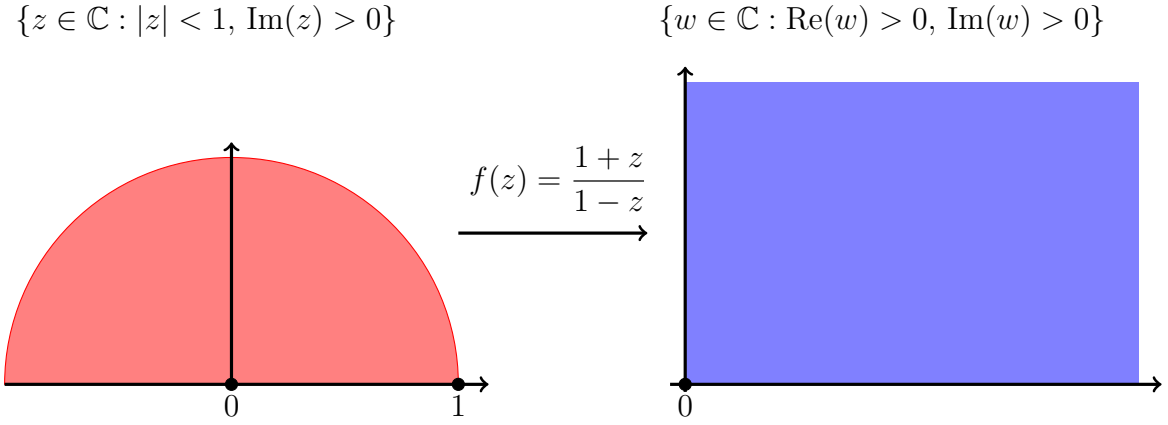


Abbildung 76: Konform äquivalente Mengen I

Beispiel: Wie man leicht nachweist, wird der Halbstreifen

$$G := \{z \in \mathbb{C} : -\pi/2 < \mathrm{Re}(z) < \pi/2,\ \mathrm{Im}(z) > 0\}$$

durch $f(z) := e^{iz}$ konform auf die rechte Einheitskreishalbscheibe

$$\{w \in \mathbb{C} : |w| < 1,\ \mathrm{Re}(w) > 0\}$$

abgebildet, siehe Abbildung 77. □

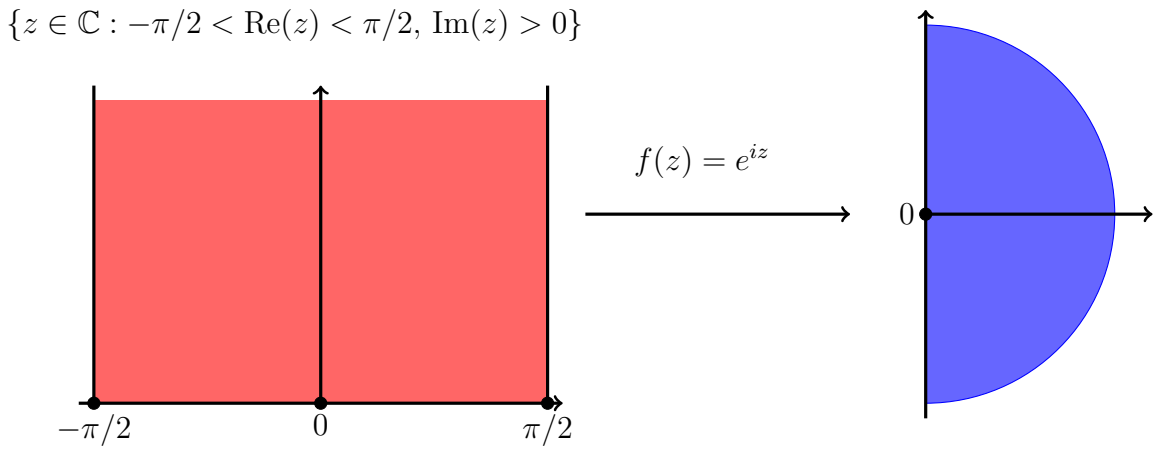


Abbildung 77: Konform äquivalente Mengen II

Beispiel: Die obere Einheitskreishalbscheibe

$$G := \{z \in \mathbb{C} : |z| < 1,\ \mathrm{Im}(z) > 0\}$$

wird durch

$$f(z) := -\frac{1}{2}\left(z + \frac{1}{z}\right)$$

konform auf die obere Halbebene $\mathbb{H} := \{w \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ abgebildet, siehe Abbildung 78. Denn f ist auf G holomorph und injektiv, da aus $z_1, z_2 \in G$, $f(z_1) = f(z_2)$ sowie $z_1 \neq z_2$ folgt, dass $z_1 z_2 = 1$, ein Widerspruch dazu, dass z_1, z_2 in der (offenen) Einheitskreisscheibe liegen. Mit $z = x + iy \in G$ ist

$$f(z) = -\frac{1}{2}\left(z + \frac{1}{z}\right) = -\frac{1}{2}\left(x + \frac{x}{x^2+y^2}\right) + i \cdot \underbrace{\frac{1}{2} y \left(\frac{1}{x^2+y^2} - 1\right)}_{>0}$$

und daher $f(G) \subset \mathbb{H}$. Um zu zeigen, dass G durch f *auf* $\mathbb{H}$ abgebildet wird, geben wir uns ein $w \in \mathbb{H}$ vor und zeigen die Existenz eines $z \in G$ mit $f(z) = w$ bzw. $z^2 + 2zw + 1 = 0$. Wegen $w \in \mathbb{H}$ ist $w \neq \pm 1$ und daher hat die quadratische Gleichung $z^2 + 2zw + 1 = 0$ zwei *verschiedene* Lösungen z_1, z_2. Dann ist $z_1 + z_2 = -2w$ und

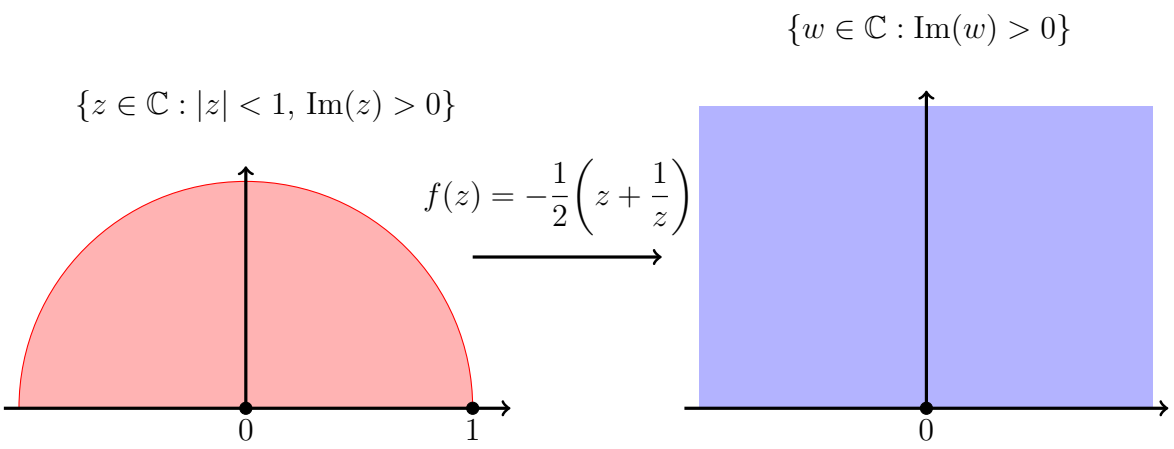


Abbildung 78: Konform äquivalente Mengen III

$z_1 z_2 = 1$. Wäre $|z_1| = 1$, so wäre $z_1 = e^{i\phi}$ mit $\phi \in [0, 2\pi)$, dann $z_2 = e^{-i\phi}$ und $z_1 + z_2 = 2\cos\phi$ reell, ein Widerspruch zu $z_1 + z_2 = -2w$ mit $\mathrm{Im}(w) > 0$. Also ist etwa $|z_1| < 1$ (und $|z_2| > 1$). Sei $z_1 = x_1 + iy_1$, wegen $|z_1| < 1$ ist also $x_1^2 + y_1^2 < 1$. Dann ist

$$0 > -2\mathrm{Im}(w) = \mathrm{Im}(z_1 + z_2) = \mathrm{Im}(z_1) + \mathrm{Im}(z_2) = y_1 \underbrace{\left(1 - \frac{1}{x_1^2 + y_1^2}\right)}_{<0}$$

und folglich $y_1 > 0$. Insgesamt ist $z_1 \in G$ und $f(z_1) = w$, womit $f(G) = \mathbb{H}$ und die konforme Äquivalenz von G und $\mathbb{H}$ bewiesen ist. □

10.3 Automorphismengruppen

Sei $\Omega \subset \mathbb{C}$ offen. Eine bijektive konforme Abbildung $f\colon \Omega \longrightarrow \Omega$ heißt ein *Automorphismus* von Ω. Die Menge der Automorphismen von Ω bezeichnen wir mit $\mathrm{Aut}(\Omega)$. Dann ist $\mathrm{Aut}(\Omega)$ in natürlicher Weise eine Gruppe. Die Verknüpfung in der Gruppe ist die Komposition der Abbildungen, das neutrale Element ist die Abbildung $f(z) = z$, die Inverse die inverse Abbildung. Sind f und g Automorphismen von Ω, so natürlich auch $f \circ g$ und das Inverse dieser Abbildung ist durch

$$(f \circ g)^{-1} = g^{-1} \circ f^{-1}$$

gegeben. Für $\Omega = B(0;1)$ und $\Omega = \mathbb{H}$ (obere Halbebene) wollen wir die zugehörige Automorphismengruppe bestimmen.

10.3.1 Die Automorphismengruppe der Einheitskreisscheibe

Wir wollen alle Automorphismen der Einheitskreisscheibe bestimmen. Im folgenden Lemma wird eine Menge von Automorphismen der Einheitskreisscheibe angegeben und in dem darauf folgenden Satz bewiesen, dass hierdurch bis auf Rotationen dieser Automorphismen *alle* Automorphismen der Einheitskreisscheibe gegeben sind.

Lemma 10.2 *Für $\alpha \in \mathbb{C}$ mit $|\alpha| < 1$ ist die durch*

$$\psi_\alpha(z) := \frac{\alpha - z}{1 - \overline{\alpha} z}$$

definierte Abbildung ψ_α ein Automorphismus der Einheitskreisscheibe $B(0;1)$.

Beweis: Zunächst ist ψ_α offenbar holomorph und injektiv auf $B(0;|\alpha|^{-1})$, einer offenen, $\operatorname{cl} B(0;1)$ enthaltenden Kreisscheibe. Wegen des maximum modulus theorem nimmt die Abbildung ψ_α ihr Maximum auf $\operatorname{cl} B(0;1)$ in einem Randpunkt an. Für einen Randpunkt $z \in \partial B(0;1)$ ist $z = e^{i\theta}$ mit $\theta \in [0, 2\pi)$, daher

$$\psi_\alpha(z) = \frac{\alpha - e^{i\theta}}{1 - \overline{\alpha} e^{i\theta}} = e^{-i\theta} \frac{\alpha - e^{i\theta}}{\overline{\alpha - e^{i\theta}}}$$

und folglich $|\psi_\alpha(z)| = 1$. Es ist also $\psi_\alpha(\partial B(0;1)) = \partial B(0;1)$. Da ψ_α nichtkonstant ist, ist wegen des maximum modulus principle $|\psi_\alpha(z)| < 1$ für alle $z \in B(0;1)$ bzw. $\psi_\alpha(B(0;1)) \subset B(0;1)$. Weiter ist $(\psi_\alpha \circ \psi_\alpha)(z) = z$ für alle $z \in \operatorname{cl} B(0;1)$ und damit $\psi_\alpha^{-1} = \psi_\alpha$. Denn für $z \in \operatorname{cl} B(0;1)$ ist

$$\begin{aligned}
(\psi_\alpha \circ \psi_\alpha)(z) &= \psi_\alpha(\psi_\alpha(z)) \\
&= \frac{\alpha - \psi_\alpha(z)}{1 - \overline{\alpha}\psi_\alpha(z)} \\
&= \frac{\alpha - (\alpha - z)/(1 - \overline{\alpha} z)}{1 - \overline{\alpha}(\alpha - z)/(1 - \overline{\alpha} z)} \\
&= \frac{\alpha - |\alpha|^2 z - \alpha + z}{1 - \overline{\alpha} z - |\alpha|^2 + \overline{\alpha} z} \\
&= \frac{(1 - |\alpha|^2) z}{1 - |\alpha|^2} \\
&= z.
\end{aligned}$$

Insbesondere ist $\psi_\alpha(B(0;1)) = B(0;1)$, womit insgesamt gezeigt ist, dass ψ_α für $\alpha \in B(0;1)$ ein Automorphismus von $B(0;1)$ ist. □

Satz 10.3 *Sei f ein Automorphismus von $B(0;1)$. Dann existieren $\theta \in [0, 2\pi)$ und $\alpha \in B(0;1)$ mit*

$$f(z) = e^{i\theta} \frac{\alpha - z}{1 - \overline{\alpha} z}.$$

Beweis: Da f ein Automorphismus von $B(0;1)$ ist, existiert genau ein $\alpha \in B(0;1)$ mit $f(\alpha) = 0$. Mit dem in Lemma 10.2 definierten Automorphismus ψ_α definiere man den Automorphismus $g := f \circ \psi_\alpha$ von $B(0;1)$. Dann ist $g(0) = 0$ und das Schwarz Lemma liefert, dass

$$|g(z)| \leq |z| \quad \text{für alle } z \in B(0;1).$$

Da auch $g^{-1}(0) = 0$, liefert eine Anwendung des Schwarz Lemma auf g^{-1}, dass

$$|g^{-1}(w)| \leq |w| \quad \text{für alle } w \in B(0;1).$$

Setzt man hier $w = g(z)$ mit $z \in B(0;1)$, so erhält man

$$|z| \leq |g(z)| \quad \text{für alle } z \in B(0;1).$$

Insgesamt ist $|g(z)| = |z|$ für alle $z \in B(0;1)$. Aus dem Schwarz Lemma folgt, dass g eine Rotation ist, also ein $\theta \in [0, 2\pi)$ mit $g(z) = e^{i\theta}z$ für alle $z \in B(0;1)$ existiert. Ersetzt man hier z durch $\psi_\alpha(z)$ und nutzt $(\psi_\alpha \circ \psi_\alpha)(z) = z$ aus, so erhält man $f(z) = e^{i\theta}\psi_\alpha(z)$ und das war zu zeigen. □

Bemerkung: Eine etwas andere Darstellung eines Automorphismus von $B(0;1)$ erhalten wir durch

- Ist f ein Automorphismus von $B(0;1)$, so existieren $a, b \in \mathbb{C}$ mit $|a|^2 - |b|^2 = 1$ derart, dass
$$f(z) = \frac{az+b}{\overline{b}z + \overline{a}} \quad \text{für alle } z \in B(0;1).$$

Denn: Wegen Satz 10.3 existieren $\theta \in [0, 2\pi)$ sowie $\alpha \in \mathbb{C}$ mit $|\alpha| < 1$ derart, dass

$$f(z) = e^{i\theta}\frac{\alpha - z}{1 - \overline{\alpha}z} \quad \text{für alle } x \in B(0;1).$$

Man setze

$$a := i\frac{e^{i\theta/2}}{\sqrt{1-|\alpha|^2}}, \qquad b := -i\alpha\frac{e^{i\theta/2}}{\sqrt{1-|\alpha|^2}}.$$

Dann ist $|a|^2 - |b|^2 = 1$ und

$$\frac{az+b}{\overline{b}z+\overline{a}} = \frac{ie^{i\theta/2}z - i\alpha e^{i\theta/2}}{i\overline{\alpha}e^{-i\theta/2}z - ie^{-i\theta/2}} = e^{i\theta}\frac{z-\alpha}{\overline{\alpha}z - 1} = f(z).$$

Umgekehrt gilt:

- Sind $a, b \in \mathbb{C}$ gegeben mit $|a|^2 - |b|^2 = 1$, so ist die durch
$$f(z) := \frac{az+b}{\overline{b}z+\overline{a}}$$
auf $B(0;1)$ definierte Abbildung f ein Automorphismus von $B(0;1)$.

Denn: Die Abbildung f ist auf $B(0;1)$ holomorph, denn wegen

$$1 = |a|^2 - |b|^2 = (|a| - |b|)(|a| + |b|)$$

ist

$$|\overline{b}z + \overline{a}| \geq |a| - |b||z| \geq |a| - |b| > 0$$

für beliebiges $z \in B(0;1)$. Eine leichte Rechnung zeigt außerdem, dass f auf $B(0;1)$ injektiv ist. Für $z \in B(0;1)$ ist

$$|\overline{b}z + \overline{a}|^2 - |az+b|^2 = (\underbrace{|a|^2 - |b|^2}_{=1})(\underbrace{1 - |z|^2}_{>0}) > 0$$

und daher

$$|f(z)| = \frac{|az+b|}{|\overline{b}z+\overline{a}|} < 1$$

bzw. $f(B(0;1)) \subset B(0;1)$. Es ist sogar $f(B(0;1)) = B(0;1)$, denn für $w \in B(0;1)$ ist

$$z := \frac{-\overline{a}w+b}{\overline{b}w-a} \in B(0;1)$$

und $f(z) = w$. Damit ist gezeigt, dass f ein Automorphismus von $B(0;1)$ ist. □

Als unmittelbare Anwendung aus dem letzten Satz erhalten wir

Korollar 10.4 *Ist f ein Automorphismus von $B(0;1)$ mit $f(0) = 0$, so ist f eine Rotation, es existiert also $\theta \in [0, 2\pi)$ mit $f(z) = e^{i\theta}z$ für alle $z \in B(0;1)$.*

10.3.2 Die Automorphismengruppe der oberen Halbebene

Nun wollen wir die Automorphismengruppe der oberen Halbebene

$$\mathbb{H} := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$$

bestimmen. In einem Beispiel im vorigen Abschnitt haben wir nachgewiesen, dass die Abbildung $F\colon \mathbb{H} \longrightarrow \mathbb{C}$, definiert durch

$$F(z) := \frac{i-z}{i+z},$$

eine auf $\mathbb{H}$ konforme Abbildung mit $F(\mathbb{H}) = B(0;1)$ ist. Die Umkehrabbildung

$$F^{-1}\colon B(0;1) \longrightarrow \mathbb{H}$$

zu F ist gegeben durch

$$F^{-1}(w) := i\frac{1-w}{1+w}.$$

Da wir die Menge der Automorphismen von $B(0;1)$ kennen, können wir auch die Darstellung eines Automorphismus von $\mathbb{H}$ bestimmen. Genauer gilt

Satz 10.5 *Ist f ein Automorphismus der oberen Halbebene $\mathbb{H}$, so ist f gegeben durch*

$$f(z) = \frac{Az+B}{Cz+D},$$

wobei $A, B, C, D \in \mathbb{R}$ und $AD - BC > 0$. Sind umgekehrt $A, B, C, D \in \mathbb{R}$ reelle Zahlen mit $AD - BC > 0$, so ist durch

$$f(z) := \frac{Az+B}{Cz+D}$$

ein Automorphismus von $\mathbb{H}$ gegeben.

Beweis: Sei $f \in \mathrm{Aut}(\mathbb{H})$ ein Automorphismus der oberen Halbebene $\mathbb{H}$. Mit der Abbildung $F\colon \mathbb{H} \longrightarrow B(0;1)$, definiert durch

$$F(z) := \frac{i-z}{i+z},$$

ist $g := F \circ f \circ F^{-1} \in \mathrm{Aut}(B(0;1))$. Wegen der Bemerkung im Anschluss an Satz 10.3 existieren $a, b \in \mathbb{C}$ mit $|a|^2 - |b|^2 = 1$ und

$$g(z) = F \circ f \circ F^{-1}(z) = \frac{az+b}{\overline{b}z + \overline{a}} \quad \text{für alle } z \in B(0;1).$$

Für $z \in B(0;1)$ ist dann

$$\begin{aligned}
f(z) &= F^{-1} \circ g \circ F(z) \\
&= F^{-1} \circ g\left(\frac{i-z}{i+z}\right) \\
&= F^{-1}\left(\frac{a(i-z)/(i+z)+b}{\overline{b}(i-z)/(i+z)+\overline{a}}\right) \\
&= F^{-1}\left(\frac{a(i-z)+b(i+z)}{\overline{b}(i-z)+\overline{a}(i+z)}\right) \\
&= i\frac{1-[a(i-z)+b(i+z)]/[\overline{b}(i-z)+\overline{a}(i+z)]}{1+[a(i-z)+b(i+z)]/[\overline{b}(i-z)+\overline{a}(i+z)]} \\
&= i\frac{[\overline{b}(i-z)+\overline{a}(i+z)]-[a(i-z)+b(i+z)]}{[\overline{b}(i-z)+\overline{a}(i+z)]+[a(i-z)+b(i+z)]} \\
&= i\frac{(a+\overline{a}-b-\overline{b})z+i(\overline{a}-a-b+\overline{b})}{(\overline{a}-a+b-\overline{b})z+i(a+\overline{a}+b+\overline{b})} \\
&= \frac{(\mathrm{Re}(a)-\mathrm{Re}(b))z+(\mathrm{Im}(a)+\mathrm{Im}(b))}{(-\mathrm{Im}(a)+\mathrm{Im}(b))z+(\mathrm{Re}(a)+\mathrm{Re}(b))} \\
&= \frac{Az+B}{Cz+D}
\end{aligned}$$

mit den reellen Zahlen

$$A := \mathrm{Re}(a) - \mathrm{Re}(b), \quad B := \mathrm{Im}(a) + \mathrm{Im}(b)$$

und

$$C := -\mathrm{Im}(a) + \mathrm{Im}(b), \quad D := \mathrm{Re}(a) + \mathrm{Re}(b).$$

Es ist

$$AD - BC = (\mathrm{Re}(a)^2 - \mathrm{Re}(b)^2) - (\mathrm{Im}(b)^2 - \mathrm{Im}(a)^2) = |a|^2 - |b|^2 = 1.$$

Damit ist der erste Teil des Satzes bewiesen. Umgekehrt sei die Abbildung $f\colon \mathbb{H} \longrightarrow \mathbb{C}$ durch

$$f(z) := \frac{Az+B}{Cz+D}$$

definiert, wobei $A, B, C, D \in \mathbb{R}$ mit $AD - BC > 0$ vorgegeben sind. Dann ist f auf $\mathbb{H}$ holomorph, da $Cz + D \neq 0$ für alle $z \in \mathbb{H}$ (denn C und D sind nicht beide gleich 0 und reell, während $z \in \mathbb{H}$ nicht reell ist). Ebenso einfach erkennt man wegen $AD - BC > 0$, dass f auf $\mathbb{H}$ injektiv ist. Ferner ist $f(\mathbb{H}) \subset \mathbb{H}$. Denn ist $z \in \mathbb{H}$ bzw. $\mathrm{Im}(z) > 0$, so ist

$$f(z) = \frac{Az + B}{Cz + D} = \frac{(Az + B)(C\overline{z} + D)}{|Cz + D|^2} = \frac{AC|z|^2 + ADz + BC\overline{z} + BD}{|Cz + D|^2}.$$

Hieraus liest man ab, dass

$$\mathrm{Im}(f(z)) = \frac{(AD - BC)\mathrm{Im}(z)}{|Cz + D|^2} > 0$$

bzw. $f(z) \in \mathbb{H}$. Weiter ist sogar $f(\mathbb{H}) = \mathbb{H}$, denn bei gegebenem $w \in \mathbb{H}$ ist $f(z) = w$ mit

$$z := \frac{Dw - B}{-Cw + A} \in \mathbb{H}.$$

Damit ist der Satz bewiesen. □

Bemerkung: Sei $f \in \mathrm{Aut}(\mathbb{H})$. Wegen Satz 10.5 existieren $A, B, C, D \in \mathbb{R}$ mit $AD - BC > 0$ derart, dass f auf $\mathbb{H}$ durch

$$f(z) = \frac{Az + B}{Cz + D}$$

gegeben ist. Dann kann f auf naheliegende weise auf $\mathbb{H} \cup (\mathbb{R} \cup \{\infty\})$ fortgesetzt werden. Für $C \neq 0$ setze man nämlich

$$f(z) := \begin{cases} \dfrac{Az + B}{Cz + D}, & z \in \mathbb{H} \cup (\mathbb{R} \setminus \{-D/C\}), \\ \infty, & z = -D/C, \\ A/C, & z = \infty, \end{cases}$$

für $C = 0$ (wegen $AD - BC > 0$ ist dann $D \neq 0$) sei

$$f(z) := \begin{cases} \dfrac{Az + B}{D}, & z \in \mathbb{H} \cup \mathbb{R}, \\ \infty, & z = \infty. \end{cases}$$

Man überlegt sich leicht, dass diese auf $\mathbb{H} \cup (\mathbb{R} \cup \{\infty\})$ fortgesetzte Abbildung f eine stetige, bijektive Abbildung von $\mathbb{H} \cup (\mathbb{R} \cup \infty\})$ auf sich ist. Daher ist auch klar, was wir meinen, wenn wir im folgenden ein $f \in \mathrm{Aut}(\mathbb{H})$ auf ein Argument aus $\mathbb{R} \cup \{\infty\}$ anwenden. □

Jetzt wollen wir noch den folgenden Satz beweisen.

Satz 10.6 *Es gelten die folgenden beiden Aussagen über Automorphismen der oberen Halbebene* $\mathbb{H}$.

1. *Ein Automorphismus von* $\mathbb{H}$, *der drei verschiedene Punkte aus* $\mathbb{R} \cup \{\infty\}$ *festlässt, ist die Identität.*

2. *Sind $u_1 < u_2 < u_3$ und $v_1 < v_2 < v_3$ Punkte aus $\mathbb{R} \cup \{\infty\}$, so existiert genau ein $f \in \mathrm{Aut}(\mathbb{H})$ mit $f(u_j) = v_j$, $j = 1, 2, 3$.*

Beweis: Ein $f \in \mathrm{Aut}(H)$ ist gegeben durch

$$f(z) = \frac{Az + B}{Cz + D}$$

mit reellen A, B, C, D und $AD - BC > 0$. Ist einer der drei Punkte ∞ und $f(\infty) = \infty$, so ist $C = 0$ und daher $AD > 0$. Also hat

$$f(z) = \frac{A}{D}z + \frac{B}{D}$$

zwei verschiedene Fixpunkte. Hieraus folgt $A = D$ und $B = 0$, d. h. f ist die Identität. Sind dagegen die drei verschiedenen Fixpunkte von f aus $\mathbb{R}$, so sind diese drei Punkte Lösungen der quadratischen Gleichung

$$Cx^2 + (D - A)x - B = 0.$$

Daher ist $C = B = 0$ und $A = D$, also f die Identität. Damit ist die erste Aussage bewiesen. Zum Beweis der zweiten Aussage definieren wir

$$g(z) := \begin{cases} \dfrac{(z - u_1)(u_2 - u_3)}{(z - u_3)(u_2 - u_1)}, & u_3 \neq \infty, \\ \dfrac{z - u_1}{u_2 - u_1}, & u_3 = \infty, \end{cases}$$

und

$$h(w) := \begin{cases} \dfrac{(w - v_1)(v_2 - v_3)}{(w - v_3)(v_2 - v_1)}, & v_3 \neq \infty, \\ \dfrac{w - v_1}{v_2 - v_1}, & v_3 = \infty. \end{cases}$$

Dann sind $g, h \in \mathrm{Aut}(\mathbb{H})$, da

$$U := (u_3 - u_1)(u_3 - u_2)(u_2 - u_1) > 0, \qquad V := (v_3 - v_1)(v_3 - v_2)(v_2 - v_1) > 0,$$

und $g(u_1) = h(v_1) = 0$, $g(u_2) = h(v_2) = 1$ und $g(u_3) = h(v_3) = \infty$. Daher ist $f := h^{-1} \circ g \in \mathrm{Aut}(\mathbb{H})$ mit $f(u_j) = v_j$, $j = 1, 2, 3$. Die Eindeutigkeit von f folgt aus dem ersten Teil des Satzes. □

10.3.3 Die Automorphismengruppe der komplexen Ebene

Wir zeigen:

Satz 10.7 *Sei $f \in \mathrm{Aut}(\mathbb{C})$ ein Automorphismus der komplexen Ebene $\mathbb{C}$. Dann ist f gegeben durch $f(z) = az + b$, wobei $a, b \in \mathbb{C}$ mit $a \neq 0$. Sind umgekehrt $a, b \in \mathbb{C}$ mit $a \neq 0$ gegeben, so ist durch $f(z) := az + b$ ein Automorphismus von $\mathbb{C}$ gegeben.*

Beweis: Wir beginnen mit der einfachen Richtung. Sind $a, b \in \mathbb{C}$ mit $a \neq 0$, so ist die durch $f(z) := az + b$ definierte Abbildung f ganz offensichtlich ein Automorphismus von $\mathbb{C}$. Nun zum schwierigeren Teil. Hier zeigen wir zunächst eine Hilfsaussage:

- Sei g holomorph und injektiv auf $B(0;1)\setminus\{0\}$. Dann ist 0 keine wesentliche Singularität von g, d. h. g ist auf $B(0;1)\setminus\{0\}$ beschränkt oder es gilt $\lim_{z\to 0} |g(z)| = \infty$.

Denn: Angenommen, 0 sei eine wesentliche Singularität von g. Wegen des Satzes von Casorati-Weierstraß ist $g(B(0;\epsilon) \setminus \{0\})$ für jedes $\epsilon \in (0,1)$ dicht in $\mathbb{C}$. Wir setzen $B_1 := B(0;\epsilon_1)$ mit einem $\epsilon_1 \in (0,1)$, wählen ein $z \in B(0;1) \setminus B_1$ und anschließend ein $\epsilon_2 > 0$ mit $B_2 := B(z;\epsilon_2) \subset B(0;1)$ und $B_1 \cap B_2 = \emptyset$. In Abbildung 79 verdeutlichen wir uns die Situation. Wegen des Open Mapping Theorems ist $g(B_2)$ offen. Da $g(B_1 \setminus \{0\})$

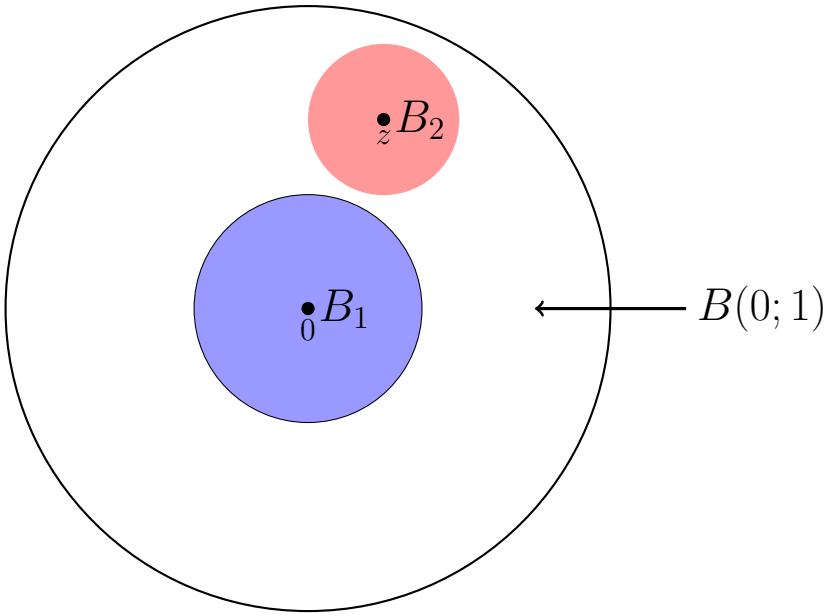


Abbildung 79: Beweis der Hilfsaussage im Beweis von Satz 10.7

dicht in $\mathbb{C}$ ist, ist $g(B_1 \setminus \{0\}) \cap g(B_2) \neq \emptyset$. Daher existieren $z_0 \in B_1 \setminus \{0\}$ und $w_0 \in B_2$ mit $g(z_0) = g(w_0)$. Da g auf $B(0;1) \setminus \{0\}$ injektiv ist, ist $z_0 = w_0 \in B_1 \cap B_2$, ein Widerspruch zu $B_1 \cap B_2 = \emptyset$. Damit ist die Hilfsaussage bewiesen.

Sei nun $f \in \mathrm{Aut}(\mathbb{C})$ ein Automorphismus von $\mathbb{C}$. Die Potenzreihenentwicklung von f sei

$$f(z) = \sum_{n=0}^{\infty} a_n z^n.$$

Wir definieren $g\colon B(0;1) \setminus \{0\} \longrightarrow \mathbb{C}$ durch

$$g(z) := f(1/z) = a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots.$$

Natürlich ist dann g auf $B(0;1) \setminus \{0\}$ holomorph und injektiv. Aus der gerade bewiesenen Hilfssaussage folgt, dass 0 keine wesentliche Singularität von g ist. Daher gibt es eine nichtnegative ganze Zahl m mit $a_n = 0$ für alle $n > m$, d. h. f ist ein Polynom m-ten Grades. Da f injektiv ist, ist f wegen des Fundamentalsatzes der Algebra ein Polynom vom Grad 1, also $f(z) = a_0 + a_1 z$ mit $a_1 \neq 0$. Das war zu zeigen. □

10.4 Der Beweis des Riemannschen Abbildungssatzes

Unser Ziel besteht darin, Satz 10.1, den Riemannschen Abbildungssatz, zu beweisen. Hierbei ist die Eindeutigkeitsaussage einfach einzusehen. Denn seien $f_1, f_2 \colon G \longrightarrow B(0;1)$ zwei Abbildungen, die die offene, einfach zusammenhängende Menge $G \subset \mathbb{C}$ mit $G \neq \mathbb{C}$ konform auf die offene Einheitskreisscheibe $B(0;1)$ abbilden, wobei mit einem vorgegebenen $a \in G$ die Bedingungen $f_1(a) = f_2(a) = 0$ und $f_1'(a) > 0$, $f_2'(a) > 0$ erfüllt sind. Dann ist $f := f_2 \circ f_1^{-1} \in \mathrm{Aut}(B(0;1))$ ein Automorphismus der Einheitskreisscheibe mit $f(0) = 0$. Wegen Korollar 10.4 ist f eine Drehung, es existiert also ein $\theta \in [0, 2\pi)$ mit $f(z) = e^{i\theta} z$ für alle $z \in B(0;1)$. Nun ist aber

$$f'(0) = f_2'(f_1^{-1}(0))(f_1^{-1})'(0) = f_2'(a)(f_1^{-1})'(f_1(a)) = f_2'(a)\frac{1}{f_2'(a)} > 0$$

und andererseits $f'(0) = e^{i\theta}$. Folglich ist notwendig $\theta = 0$, daher f die Identität bzw. $f_1 = f_2$. Damit ist die Eindeutigkeitsaussage im Riemannschen Abbildungssatz bewiesen, sodass wir uns im weiteren auf die Existenzaussage konzentrieren können.

Hier wollen wir zunächst die *Struktur* des Beweises andeuten. Wir gehen im folgenden davon aus, dass G eine offene, einfach zusammenhängende Teilmenge von $\mathbb{C}$ mit $G \neq \mathbb{C}$ ist und $a \in G$ vorgegeben ist. Der Beweis besteht im wesentlichen aus zwei Schritten.

1. Es existiert eine auf G holomorphe, injektive Abbildung f mit $f(G) \subset B(0;1)$ und $f(a) = 0$, d. h. es ist

   $$\mathcal{F} := \{f \in H(G) : f \text{ injektiv}, f(G) \subset B(0;1), f(a) = 0\} \neq \emptyset.$$

   In Unterabschnitt 10.1 haben wir gezeigt: Ist $f \colon G \longrightarrow \mathbb{C}$ holomorph und injektiv, so ist $f'(z) \neq 0$ für alle $z \in G$. Ist also $g \in \mathcal{F}$ und $\theta \in [0, 2\pi)$ geeignet gewählt, so ist $f := e^{i\theta} g \in \mathcal{F}$ und $f'(a) > 0$. I. Allg. ist natürlich $f(G)$ eine echte Teilmenge von $B(0;1)$, denn andernfalls wären wir fertig.

2. Mit $H(G)$ bezeichnen wir die Menge der auf G holomorphen Funktionen. Weiter sei

   $$\mathcal{F} := \{f \in H(G) : f \text{ injektiv}, f(G) \subset B(0;1), f(a) = 0\}.$$

   Wegen des ersten Schrittes ist $\mathcal{F} \neq \emptyset$. Unser Ziel besteht natürlich darin, die Existenz eines $f \in \mathcal{F}$ mit $f(G) = B(0;1)$ zu zeigen. Hierzu definieren wir

   $$s := \sup_{f \in \mathcal{F}} |f'(a)|.$$

   Da die Ableitung einer auf einem Gebiet G holomorphen, injektiven Funktion nicht verschwindet, ist $s > 0$. Mit Hilfe der Integralformel von Cauchy werden wir zeigen können, dass $s < \infty$. Folglich existiert eine Folge $\{f_n\} \subset \mathcal{F}$ mit $\lim_{n\to\infty} |f_n'(a)| = s$. Mit Hilfe des Satzes von Montel (Satz 10.11) werden wir zeigen, dass $\{f_n\}$ eine in einem geeigneten Sinne konvergente Teilfolge $\{g_n\}$ besitzt. Natürlich muss genauer geklärt werden, welcher Konvergenzbegriff in $H(G)$ zugrunde gelegt wird. Um es vorweg zu nehmen: Der wichtigste Konvergenzbegriff auf $H(G)$ ist die *gleichmäßige Konvergenz auf kompakten Teilmengen* von

G. Dann werden wir mit Hilfe eines Satzes von Hurwitz zeigen, dass der Grenzwert g der Teilfolge $\{g_n\}$ ein Element von $\mathcal{F}$ ist. In einem letzten Schritt wird $g(G) = B(0;1)$ nachgewiesen. Damit wird der Riemannsche Abbildungssatz bewiesen sein.

10.4.1 Nachweis von $\mathcal{F} \neq \emptyset$

Unser Ziel hier ist es, den folgenden Satz zu beweisen.

Satz 10.8 *Sei $G \subset \mathbb{C}$ offen und einfach zusammenhängend, $G \neq \mathbb{C}$ und $a \in G$. Sei $H(G)$ die Menge der auf G holomorphen Funktionen und*

$$\mathcal{F} := \{f \in H(G) : f \text{ injektiv}, f(G) \subset B(0;1), f(a) = 0\}.$$

Dann ist $\mathcal{F} \neq \emptyset$.

Beweis: Da G eine echte Teilmenge von $\mathbb{C}$ ist, kann ein $b \in \mathbb{C} \setminus G$ gewählt werden. Im einleitenden Unterabschnitt haben wir begründet, dass eine auf einer offenen, einfach zusammenhängende holomorphe und nicht verschwindende Funktion dort eine holomorphe Quadratwurzel besitzt. Daher existiert eine auf G holomorphe Abbildung g mit $g(z)^2 = z - b$ für alle $z \in B$. Offensichtlich ist g auf G injektiv, denn

$$g(z_1) = g(z_2) \Longrightarrow z_1 - b = g(z_1)^2 = g(z_2)^2 = z_2 - b \Longrightarrow z_1 = z_2.$$

Wegen des Open Mapping Theorems ist $g(G)$ offen. Für ein beliebiges $w_0 \in g(G)$ existiert daher ein $r > 0$ mit $B(w_0; r) \subset g(G)$. Als zwischenbehauptung zeigen wir:

- Es ist $B(-w_0; r) \subset \mathbb{C} \setminus g(G)$ und daher $|g(z) + w_0| \geq r$ für alle $z \in G$.

Denn: Angenommen, • sei nicht richtig. Dann existiert ein $w \in B(-w_0; r) \cap g(G)$, folglich ist $w = g(z_1)$ mit einem $z_1 \in G$ und außerdem $-w \in B(-w_0; r) \subset g(G)$ und daher $-w = g(z_2)$ mit einem $z_2 \in G$. Es ist $w = g(z_1) = -g(z_2)$, folglich

$$z_1 - b = g(z_1)^2 = g(z_2)^2 = z_2 - b,$$

also $z_1 = z_2$. Wegen $g(z_2) = -g(z_1) = -g(z_2)$ ist $g(z_2) = 0$ bzw. $z_2 = b$, was ein Widerspruch zu $b \notin G$ ist. Damit ist die Zwischenbehauptung bewiesen.
Nun definiere man

$$f_0(z) := \frac{r}{2(g(z) + w_0)}.$$

Wegen der gerade eben bewiesenen Zwischenbehauptung ist $|g(z) + w_0| \geq r$ für alle $z \in G$ und daher f_0 auf G holomorph und $f_0(G) \subset B(0; \frac{1}{2}) \subset B(0;1)$. Da g injektiv ist, ist es auch f_0. Durch Vorschalten eines geeigneten Automorphismus $\phi \in \mathrm{Aut}(B(0;1))$ können wir erreichen, dass $f(z) = (\phi \circ f_0)(z)$ der Bedingung $f(a) = 0$ genügt. Man setze also

$$f(z) := \frac{f_0(a) - f_0(z)}{1 - \overline{f_0(a)} f_0(z)}.$$

Der Satz ist bewiesen. □

10.4.2 Gleichmäßige Konvergenz auf kompakten Teilmengen von G in $H(G)$

Mit $H(G)$ bezeichnen wir wieder die Menge der holomorphen Abbildungen der offenen Teilmenge $G \subset \mathbb{C}$ nach $\mathbb{C}$. Den für die weiteren Untersuchungen geeigneten Konvergenzbegriff auf $H(G)$ findet man in der folgenden Definition.

Definition 10.9 Sei $G \subset \mathbb{C}$ offen. Wir sagen, die Folge $\{f_n\} \subset H(G)$ *konvergiere gleichmäßig auf kompakten Teilmengen von* G gegen eine Funktion $f\colon G \longrightarrow \mathbb{C}$, wenn für jede kompakte Teilmenge $K \subset G$ gilt, dass

$$\lim_{n\to\infty} \sup_{z\in K} |f_n(z) - f(z)| = 0.$$

Bemerkung: Wir wollen uns klarmachen, dass eine Folge $\{f_n\} \subset H(G)$ genau dann gleichmäßig auf kompakten Teilmengen von G gegen eine Funktion $f\colon G \longrightarrow \mathbb{C}$ konvergiert (gelegentlich spricht man auch von *normaler* oder *kompakter Konvergenz*), wenn sie *lokal gleichmäßig auf* G gegen $f\colon G \longrightarrow \mathbb{C}$ konvergiert. Lokal gleichmäßige Konvergenz auf G bedeutet hierbei, dass es zu jedem $z \in G$ eine offene Umgebung U von z gibt mit der Eigenschaft, dass $\{f_n\}$ auf U gleichmäßig gegen f konvergiert, dass also

$$\lim_{n\to\infty} \sup_{w\in U} |f_n(w) - f(w)| = 0.$$

Dass kompakte Konvergenz lokal gleichmäßige Konvergenz impliziert, ist hierbei ziemlich offensichtlich. Daher nehmen wir umgekehrt an, $\{f_n\}$ würde lokal gleichmäßig gegen f konvergieren, ferner sei $K \subset G$ kompakt. Bei vorgegebenem $\epsilon > 0$ gibt es zu jedem $z \in K$ eine offene Umgebung U_z von z und $n_z \in \mathbb{N}$ mit

$$\sup_{w\in U_z} |f_n(w) - f(w)| < \epsilon \quad \text{für alle } n \geq n_z.$$

Da K kompakt ist, kann aus der Überdeckung $K \subset \bigcup_{z\in K} U_z$ eine endliche Teilüberdeckung ausgewählt werden. Es existieren also $z_1, \ldots, z_r \in K$ mit $K \subset \bigcup_{i=1}^{r} U_{z_i}$. Man setze $n_0 := \max_{i=1,\ldots,r} n_{z_i}$. Ist $w \in K$, so ist $w \in U_{z_i}$ für ein $i \in \{1, \ldots, r\}$. Für $n \geq n_0$ ist daher $|f_n(w) - f(w)| < \epsilon$ und folglich $\sup_{w\in K} |f_n(w) - f(w)| < \epsilon$. Damit ist gezeigt, dass $\{f_n\}$ gleichmäßig auf kompakten Teilmengen von G gegen f konvergiert. □

Durch den nächsten Satz zeigen wir, dass im Sinne der gerade eben definierten gleichmäßigen Konvergenz auf kompakten Teilmengen auch die Grenzfunktion einer Folge $\{f_n\} \subset H(G)$ in $H(G)$ liegt. Einen Beweis findet man auch bei E. M. Stein, R. Shakarchi (2003, Theorem 5.2, 5.3 auf S. 53—55).

Satz 10.10 *Sei $G \subset \mathbb{C}$ offen. Ist $\{f_n\} \subset H(G)$ eine Folge, die auf kompakten Teilmengen K von G gleichmäßig gegen eine Funktion $f\colon G \longrightarrow \mathbb{C}$ konvergiert, so ist $f \in H(G)$. Für jedes $k \in \mathbb{N}$ konvergiert ferner die Folge $\{f_n^{(k)}\}$ der k-ten Ableitungen auf jeder kompakten Teilmenge von G gleichmäßig gegen die k-te Ableitung $f^{(k)}$ von f.*

Beweis: Wir haben zu zeigen, dass in jedem $a \in G$ die komplexe Ableitung $f'(a)$ existiert. Man wähle also ein $a \in G$. Da G offen ist, existiert ein $r > 0$ mit $B(a; r) \subset G$.

Anschließend wähle man $\delta \in (0, r)$. Dann ist $K := \operatorname{cl} B(a;\delta) \subset G$ kompakt und nach Voraussetzung die Folge $\{f_n\}$ auf K gleichmäßig konvergent gegen f, insbesondere ist f auf K stetig. Sei T ein in $B(a;\delta)$ enthaltenes Dreieck. Da f_n auf $B(a;\delta)$ holomorph ist, folgt aus dem Satz von Goursat, dass

$$\int_T f_n(z)\,dz = 0, \qquad n \in \mathbb{N}.$$

Andererseits folgt aus der gleichmäßigen Konvergenz von $\{f_n\}$ gegen f auf K, dass

$$\lim_{n\to\infty} \int_T f_n(z)\,dz = \int_T f(z)\,dz.$$

Folglich ist $\int_T f(z)\,dz = 0$, aus dem Satz von Morera folgt, dass f holomorph auf G ist. Dann sind aber auch sämtliche Ableitungen $f^{(k)}$ holomorph auf G, also Elemente von $H(G)$ für $k = 0$ und alle $k \in \mathbb{N}$.

Wir zeigen nun, dass bei gegebenem $k \in \mathbb{N}$ die Folge $\{f_n^{(k)}\}$ der k-ten Ableitungen gleichmäßig auf kompakten Teilmengen von G gegen die k-te Ableitung $f^{(k)}$ von f konvergiert. In der Bemerkung im Anschluss an Definition 10.9 hatten wir gezeigt, dass eine Folge in $H(G)$ genau dann auf kompakten Teilmengen von G gleichmäßig konvergiert, wenn sie lokal gleichmäßig konvergiert. Daher zeigen wir jetzt, dass $\{f^{(k)}\}$ lokal gleichmäßig gegen $f^{(k)}$ konvergiert. Hierzu sei $z \in G$ vorgegeben. Da G offen ist, existiert ein $s > 0$ mit $B(z;s) \subset G$. Anschließend wähle man $r \in (0, s)$, bezeichne mit $S := \partial B(z;r)$ den Kreis um z mit dem Radius r und setze $B := \operatorname{cl} B(z;r/2)$. Dann ist $B(z;r/2)$ eine offene Umgebung von z. Für alle $w \in B$ und $\zeta \in S$ ist dann

$$|\zeta - w| \geq \underbrace{|\zeta - z|}_{=r} - \underbrace{|w - z|}_{\leq r/2} \geq \frac{r}{2}.$$

Für alle $w \in B$ und alle $n \in \mathbb{N}$ ist wegen der Integralformel von Cauchy

$$\begin{aligned} |f_n^{(k)}(w) - f^{(k)}(w)| &= \left| \frac{k!}{2\pi i} \int_S \frac{[f_n(\zeta) - f(\zeta)]}{(\zeta - w)^{k+1}}\,d\zeta \right| \\ &\leq \frac{2^{k+1}k!}{r^k} \sup_{\zeta\in S} |f_n(\zeta) - f(\zeta)|. \end{aligned}$$

Folglich ist

$$\sup_{w\in B} |f_n^{(k)}(w) - f^{(k)}(w)| \leq \frac{2^{k+1}k!}{r^k} \sup_{\zeta\in S} |f_n(\zeta) - f(\zeta)|.$$

Da $\{f_n\}$ gleichmäßig auf der kompakten Menge S gegen f konvergiert, konvergiert $\{f_n^{(k)}\}$ auf B und dann auch auf der offenen Umgebung $B(z;r/2)$ gleichmäßig gegen f. Der Satz ist damit vollständig bewiesen. □

10.4.3 Die Sätze von Montel und Hurwitz

Das entscheidende Hilfsmittel beim Beweis des Riemannschen Abbildungssatzes ist der folgende Satz, in seinem Beweis steckt die Hauptarbeit.

Satz 10.11 (Montel) *Sei $G \subset \mathbb{C}$ offen und $\mathcal{F} \subset H(G)$ gleichmäßig beschränkt auf kompakten Teilmengen von G, d. h. zu jeder kompakten Menge $K \subset G$ existiert ein $M_K > 0$ mit*

$$\sup_{z \in K} |f(z)| \leq M_K \quad \text{für alle } f \in \mathcal{F}.$$

Dann besitzt jede Folge $\{f_n\} \subset \mathcal{F}$ eine auf (allen) kompakten Teilmengen von G gleichmäßig konvergente Teilfolge (deren Grenzwert in $H(G)$, aber nicht notwendig in $\mathcal{F}$ liegt).

Beweis: Im ersten Schritt, dem im Gegensatz zu den beiden folgenden Schritten *funktionentheoretischen*[50] Teil des Beweises zeigen wir:

- Die auf kompakten Teilmengen von G gleichmäßig beschränkte Menge $\mathcal{F} \subset H(G)$ ist auf jeder kompakten Menge $K \subset G$ *gleichgradig stetig*, d. h. zu jedem $\epsilon > 0$ und jeder kompakten Menge $K \subset G$ existiert ein $\delta = \delta_{\epsilon,K} > 0$ mit

$$w, z \in K,\ |w - z| < \delta \Longrightarrow |f(w) - f(z)| < \epsilon \quad \text{für alle } f \in \mathcal{F}.$$

Denn: Sei eine kompakte Menge $K \subset G$ gegeben. Der Rand ∂G von G ist eine abgeschlossene Menge und es ist $K \cap \partial G = \emptyset$. Dann ist der *Abstand* der kompakten Menge K zu der abgeschlossenen Menge ∂G positiv, d. h. es ist

$$\operatorname{dist}(K, \partial G) := \inf_{w \in K, z \in \partial G} |w - z| > 0.$$

Nun wähle man ein $r \in (0, \frac{1}{3}\operatorname{dist}(K, \partial G))$. Dann ist $B(3r; z) \subset G$ für alle $z \in K$. Wir definieren

$$K_{2r} := \{w \in G : \inf_{z \in K} |w - z| \leq 2r\},$$

dies ist also die Menge aller Punkte aus G, die einen Abstand $\leq 2r$ zu K haben. Dann ist K_{2r} eine kompakte Teilmenge von G. Für beliebige $w, z \in K$ mit $|w - z| < r$, $\gamma := \partial B(2r; w)$ (Kreis um w mit dem Radius $2r$) und $\zeta \in \gamma$ ist dann

$$|\zeta - z| \geq \underbrace{|\zeta - w|}_{=2r} - \underbrace{|w - z|}_{<r} > r$$

und daher gilt für alle $f \in \mathcal{F}$ wegen der Integralformel von Cauchy

$$\begin{aligned} |f(w) - f(z)| &= \left| \frac{1}{2\pi i} \int_\gamma f(\zeta) \left[\frac{1}{\zeta - w} - \frac{1}{\zeta - z} \right] d\zeta \right. \\ &\leq \frac{1}{2\pi} \int_\gamma f(\zeta) | \frac{|w - z|}{|\zeta - w|\,|\zeta - z|}\, d\zeta \\ &\leq \frac{1}{2\pi} 2\pi r \frac{|w - z|}{2r^2} \sup_{\zeta \in \gamma} |f(\zeta)| \end{aligned}$$

[50] Hiermit meinen wir, dass die *Holomorphie* auftretender Funktionen ausgenutzt wird, was danach nicht mehr der Fall ist.

$$
\begin{aligned}
&= \left(\frac{1}{2r}\sup_{\zeta\in\gamma}|f(\zeta)|\right)|w-z| \\
&\leq \frac{M_{K_{2r}}}{2r}|w-z|
\end{aligned}
$$

mit einer wegen der Beschränktheitsvoraussetzung existierenden Konstanten $M_{K_{2r}} > 0$ mit

$$\sup_{\zeta\in K_{2r}}|f(\zeta)| \leq M_{K_{2r}} \quad \text{für alle } f \in \mathcal{F}.$$

Setzt man also

$$\delta := \frac{2r\epsilon}{M_{K_{2r}}},$$

so erkennt man, dass der erste Teil des Beweises erbracht ist.

Jetzt kommen die beiden funktionalanalytischen Beweisteile. Im zweiten Schritt zeigen wir:

- Sei $K \subset G$ kompakt. Dann besitzt jede Folge $\{f_n\} \subset \mathcal{F}$ eine auf K gleichmäßig konvergente Teilfolge.

Denn: Sei $K \subset G$ kompakt und $\{f_n\} \subset H(G)$ eine Folge in $H(G)$. Nach Voraussetzung existiert eine Konstante $M_K > 0$ mit

$$\sup_{w\in K}|f_n(w)| \leq M_K \quad \text{für alle } n \in \mathbb{N}.$$

Sei $\{w_j\}$ eine abzählbare, dichte Teilmenge von K. Als eine solche kann man z. B. $K \cap (\mathbb{Q}+i\mathbb{Q})$ wählen. Die Folge $\{f_n(w_1)\}$ ist wegen $|f_n(w_1)| \leq M_K$, $n \in \mathbb{N}$, beschränkt. Daher existiert eine Teilfolge $\{f_{n,1}\} \subset \{f_n\}$ mit der Eigenschaft, dass $\{f_{n,1}(w_1)\}$ konvergiert. Auch die Folge $\{f_{n,1)}(w_2)\}$ ist beschränkt. Daher existiert eine Teilfolge $\{f_{n,2}\} \subset \{f_{n,1}\}$, für welche $\{f_{n,2}(w_2)\}$ (und sozusagen automatisch auch $\{f_{n,2}(w_1)\}$) konvergiert. In dieser Weise kann man fortfahren und kann für jedes $j \in \mathbb{N}$ aus $\{f_{n,j-1}\}$ eine Teilfolge $\{f_{n,j}\}$ auswählen, für die $\{f_{n,j}(w_k)\}$ für $k = 1, \ldots, j$ konvergiert. Nun setze man $g_n := f_{n,n}$, $n \in \mathbb{N}$, bilde also die Diagonalfolge. Nach Konstruktion ist $\{g_n\}$ eine Teilfolge von $\{f_{n,j}\}$ für alle $j \in \mathbb{N}$, daher konvergiert $\{g_n(w_k)\}$ für alle $k \in \mathbb{N}$. Wir zeigen nun, dass $\{g_n\}$ eine auf K gleichmäßige Cauchy-Folge und damit auf K gleichmäßig konvergent ist. Da die Folge $\{g_n\} \subset \mathcal{F}$ wegen des ersten Teiles des Beweises auf K gleichgradig stetig ist, existiert zu jedem $\epsilon > 0$ ein $\delta > 0$ mit

$$w, z \in K,\ |w-z| < \delta \Longrightarrow |g_n(w) - g_n(z)| < \epsilon \quad \text{für alle } n \in \mathbb{N}.$$

Da die Folge $\{w_j\} \subset K$ dicht in K ist, ist jedes Element in einer offenen δ-Kugel eines der Folgenglieder enthalten, d./,h. es ist $K \subset \bigcup_{j=1}^{\infty} B(w_j;\delta)$. Wegen der Kompaktheit von K kann aus dieser Überdeckung eine endliche Teilüberdeckung ausgewählt werden. Daher existiert ein $J \in \mathbb{N}$ mit $K \subset \bigcup_{j=1}^{J} B(w_j;\delta)$. Die J Folgen $\{g_n(w_j)\}_{n\in\mathbb{N}}$, $j = 1, \ldots, J$, sind als konvergente Folgen auch Cauchy-Folgen. Daher existiert $N = N_{\epsilon,K}$ derart, dass

$$|g_n(w_j) - g_m(w_j)| < \epsilon \quad \text{für } j = 1, \ldots, J \text{ und alle } n, m \geq N.$$

Nun geben wir uns $z \in K$ beliebig vor. Wegen $K \subset \bigcup_{j=1}^{J} B(w_j;\delta)$ existiert ein $j \in \{1,\dots,J\}$ mit $z \in B(w_j;\delta)$ bzw. $|w_j - z| < \delta$. Für alle $n, m \geq N$ ist daher

$$|g_n(z) - g_m(z)| \leq |g_n(z) - g_n(w_j)| + |g_n(w_j) - g_m(w_j)| + |g_m(w_j) - g_m(z)| < 3\epsilon.$$

Damit ist nachgewiesen, dass $\{g_n\}$ eine auf K gleichmäßige Cauchy-Folge ist, also auf K gleichmäßig konvergiert.

Im dritten Teil des Beweises zeigen wir die eigentliche Behauptung:

- Jede Folge $\{f_n\} \subset \mathcal{F}$ besitzt eine auf allen kompakten Teilmengen $K \subset G$ gleichmäßig konvergente Teilfolge.

Zum Beweis führen wir zunächst eine Bezeichnung bzw. eine Definition ein. Wir nennen eine Folge $\{K_n\}$ kompakter Teilmengen von G eine *Aussscḧopfung* von G, wenn

(a) $K_n \subset \text{int}\,(K_{n+1})$, $n \in \mathbb{N}$,

(b) $\bigcup_{n=1}^{\infty} K_n = G$.

(c) Zu jeder kompakten Teilmenge $K \subset G$ existiert ein $n \in \mathbb{N}$ mit $K \subset K_n$,

Für den Nachweis dafür, dass jede offene Menge $G \subset \mathbb{C}$ eine Ausschöpfung besitzt, machen wir eine Fallunterscheidung. Ist $G = \mathbb{C}$, so setzen wir $K_n := \text{cl}\,B(0;n)$, $n \in \mathbb{N}$. Ganz offensichtlich ist hierdurch eine Ausschöpfung von G gegeben. Ist dagegen $G \neq \mathbb{C}$ eine echte Teilmenge von $\mathbb{C}$, so setzen wir

$$K_n := \left\{ z \in G : \text{dist}(z, \mathbb{C} \setminus G) \geq \frac{1}{n} \right\} \cap \text{cl}\,B(0;n), \quad n \in \mathbb{N},$$

wobei $\text{dist}(z, \mathbb{C} \setminus G) := \inf_{w \in \mathbb{C} \setminus G} |z - w|$ den Abstand zwischen z und $\mathbb{C} \setminus G$ bezeichnet. Als Durchschnitt einer abgeschlossenen und einer kompakten Menge ist K_n kompakt. Ist $z \in K_n$, so ist

$$B\left(z; \frac{1}{n} - \frac{1}{n+1}\right) \subset K_{n+1}$$

und damit $K_n \subset \text{int}\,(K_{n+1})$. Denn ist $y \in B(z; 1/n - 1/(n+1))$ und $w \in \mathbb{C} \setminus G$ beliebig, so ist

$$|y - w| \geq |z - w| - |y| \geq \text{dist}(z, \mathbb{C} \setminus G) - |y| > \frac{1}{n} - \left(\frac{1}{n} - \frac{1}{n+1}\right) = \frac{1}{n+1}$$

und daher $\text{dist}(z, \mathbb{C} \setminus G) > 1/(n+1)$ sowie

$$|y| \leq |y - z| + |z| < \frac{1}{n} - \frac{1}{n+1} + n \leq n + 1.$$

Damit ist $K_n \subset \text{int}\,(K_{n+1})$, $n \in \mathbb{N}$, bzw. die Eigenschaft (a) einer Ausschöpfung von G bewiesen. Zum Beweis von (b) geben wir uns $z \in G$ vor. Da G offen ist, existiert ein $\delta > 0$ mit $B(z;\delta) \subset G$. Dann ist $\text{dist}(z, \mathbb{C} \setminus G) \geq \delta$. Ist dann $n \in \mathbb{N}$ so groß, dass $\delta \geq 1/n$ und $|z| \leq n$, so ist $z \in K_n$. Damit ist $G \subset \bigcup_{n=1}^{\infty} K_n$. Da jede der Mengen

K_n in G enthalten ist, gilt hier Gleichheit. Damit ist auch die Eigenschaft (b) einer Ausschöpfung bewiesen. Nun sei $K \subset G$ kompakt. Dann ist

$$\delta := \operatorname{dist}(K, \mathbb{C} \setminus G) = \inf_{z \in K, w \in \mathbb{C} \setminus G} |z - w| > 0.$$

Man wähle $n \in \mathbb{N}$ so groß, dass $\delta \geq 1/n$ und $K \subset \operatorname{cl} B(0; n)$. Offenbar ist dann $K \subset K_n$ und auch die Eigenschaft (c) einer Ausschöpfung von G erfüllt.

Gegeben sei nun eine Ausschöpfung $\{K_n\}$ eine Ausschöpfung der offenen Menge G. Die *Existenz* einer solchen Ausschöpfung haben wir gerade eben bewiesen. Sei ferner $\{f_n\} \subset \mathcal{F}$ gegeben. Die gesuchte, auf allen kompakten Teilmengen von G gleichmäßig konvergente Teilfolge von $\{f_n\}$ gewinnen wir wieder durch einen Diagonalisierungsprozess. Wegen des zweiten Beweisschrittes, den wir wiederholt anwenden, existiert eine Teilfolge $\{f_{n,1}\} \subset \{f_n\}$, die auf K_1 gleichmäßig konvergiert. Entsprechend existiert eine Teilfolge $\{f_{n,2}\}$ von $\{f_{n,1}\}$ (und von $\{f_n\}$), welche auf K_2 gleichmäßig konvergiert. So kann man fortfahren und erhält Teilfolgen $\{f_{n,j}\}_{n \in \mathbb{N}}$ von $\{f_n\}$, welche auf $K_1, \ldots, K_j$ gleichmäßig konvergieren. Bildet man die Diagonalfolge $\{g_n\}$ mit $g_n := f_{n,n}$, so hat man eine Teilfolge von $\{g_n\}$ gewonnen, welche auf allen K_j, $j \in \mathbb{N}$, gleichmäßig konvergiert. Da es nach Eigenschaft (c) einer Ausschöpfung zu jeder kompakten Menge $K \subset G$ ein $j \in \mathbb{N}$ mit $K \subset K_j$ gibt, ist die Existenz einer Teilfolge von $\{f_n\}$, welche auf allen kompakten Teilmengen von G gleichmäßig konvergiert, bewiesen.

Der Satz von Montel ist vollständig bewiesen. □

Der folgende Satz ist eigentlich ein Korollar zu einem allgemeineren Satz von Hurwitz.

Satz 10.12 (Hurwitz) *Sei $G \subset \mathbb{C}$ ein Gebiet und $\{f_n\}$ eine Folge auf G holomorpher und injektiver Funktionen, welche auf kompakten Teilmengen von G gleichmäßig gegen eine Funktion $f\colon G \longrightarrow \mathbb{C}$ konvergiert. Dann ist f auf G injektiv oder konstant.*

Beweis: Angenommen, f sei nicht injektiv auf G. Dann existieren $a, b \in G$ mit $a \neq b$ und $f(a) = f(b)$. Man definiere $g_n\colon G \longrightarrow \mathbb{C}$ durch $g_n(z) := f_n(z) - f_n(a)$. Dann konvergiert $\{g_n\}$ auf kompakten Teilmengen von G gleichmäßig gegen $g\colon G \longrightarrow \mathbb{C}$ mit $g(z) := f(z) - f(a)$. Wegen Satz 10.10 sind $f, g \in H(G)$. Wir nehmen an, es sei f nicht konstant bzw. $g \not\equiv 0$. Da g die beiden isolierten Nullstellen a und b besitzt, existiert ein $\delta \in (0, \frac{1}{2}|a - b|)$ mit $\operatorname{cl} B(a; \delta) \subset G$, $\operatorname{cl} B(b; \delta) \subset G$ und der Eigenschaft, dass g auf $\partial B(a; \delta)$ und auf $\partial B(b; \delta)$ nicht verschwindet. Mit Hilfe des Satzes von Rouché werden wir zeigen, dass g_n und g für alle hinreichend großen n dieselbe Anzahl an Nullstellen (unter Berücksichtigung der Vielfachheiten) in $B(a; \delta)$ bzw. $B(b; \delta)$ besitzen. Da g eine Nullstelle in $B(a; \delta)$ und in $B(b; \delta)$ besitzt, nämlich a bzw. b, besitzt g_n für alle hinreichend großen n (mindestens) eine Nullstelle a_n bzw. b_n in $B(a; \delta)$ bzw. $B(b; \delta)$. Da wir $\delta \in (0, \frac{1}{2}|a - b|)$ gewählt haben, ist $B(a; \delta) \cap B(b; \delta) = \emptyset$ und damit $a_n \neq b_n$. Da mit f_n auch g_n auf G injektiv ist, hätten wir den gewünschten Widerspruch zu der Annahme, f sei weder injektiv noch konstant auf G. Es bleibt, die folgende Aussage mit Hilfe des Satzes von Rouché zu zeigen:

- Sei $G \subset \mathbb{C}$ ein Gebiet und $\{g_n\} \subset H(G)$ eine Folge, die auf kompakten Teilmengen von G gleichmäßig gegen $g \in H(G)$ konvergiert. Sei $g \not\equiv 0$, $\operatorname{cl} B(a; \delta) \subset G$ und $g(z) \neq 0$ für alle $z \in \partial B(a; \delta)$. Dann existiert ein $n_0 \in \mathbb{N}$ mit der Eigenschaft, dass

g_n und g für alle $n \geq n_0$ dieselbe Anzahl von Nullstellen (unter Berücksichtigung der Vielfachheit) in $B(a;\delta)$ besitzen.

Denn: Da g auf $\partial B(a;\delta)$ nicht verschwindet, ist

$$\epsilon := \inf_{z \in \partial B(a;\delta)} |g(z)| > 0.$$

Da $\{g_n\}$ auf der kompakten Menge $\partial B(a;\delta)$ gleichmäßig gegen g konvergiert, existiert $n_0 \in \mathbb{N}$ mit

$$\sup_{z \in \partial B(a;\delta)} |g_n(z) - g(z)| < \frac{\epsilon}{2} \quad \text{für alle } n \geq n_0.$$

Für alle $z \in \partial B(a;\delta)$ und alle $n \geq n_0$ ist

$$|g_n(z)| = |g(z) + g_n(z) - g(z)| \geq \underbrace{|g(z)|}_{\geq \epsilon} - \underbrace{|g_n(z) - g(z)|}_{< \epsilon/2} > \frac{\epsilon}{2}.$$

Also hat auch g_n für alle $n \geq n_0$ auf $\partial B(a;r)$ keine Nullstelle. Für alle $z \in \partial B(a;r)$ ist ferner

$$|g_n(z) - g(z)| < \frac{\epsilon}{2} < \epsilon \leq |g(z)|.$$

Aus dem Satz von Rouché folgt die Behauptung • und damit ist der ganze Satz bewiesen. □

10.4.4 Zusammensetzen der Beweisteile

Jetzt werden die bisher entwickelten Beweisteile zusammengesetzt. Hier folgen wir neben der Darstellung bei E. M. Stein, R. Shakarchi auch den Ausführungen von C. J. Bishop (2010) und F. Monard (2017).

Beweis von Satz 10.1, dem Riemannschen Abbildungssatz: Mit $H(G)$ bezeichnen wir die Menge der auf der offenen, einfach zusammenhängenden Menge $G \subset \mathbb{C}$ holomorphen Funktionen. Mit einem gegebenen $a \in G$ definieren wir

$$\mathcal{F} := \{f \in H(G) : f \text{ injektiv}, f(G) \subset B(0;1), f(a) = 0\}.$$

Nach Satz 10.8 ist $\mathcal{F} \neq \emptyset$. Anschließend definiere man

$$s := \sup_{f \in \mathcal{F}} |f'(a)|.$$

Die Ableitung einer auf einem Gebiet holomorphen, injektiven Abbildung ist von Null verschieden. Daher (und wegen $\mathcal{F} \neq \emptyset$) ist $s > 0$. Ist $r > 0$ so klein, dass $\operatorname{cl} B(a;r) \subset G$, so ergibt die Integralformel von Cauchy

$$f'(a) = \int_{\partial B(a;r)} \frac{f(\zeta)}{(\zeta - a)^2}\, d\zeta.$$

Für $f \in \mathcal{F}$ ist $f(G) \subset B(0;1)$ und insbesondere $|f(\zeta)| \leq 1$ für alle $\zeta \in \partial B(a;r)$. Für $f \in \mathcal{F}$ ist daher

$$|f'(a)| \leq \frac{1}{2\pi} \cdot \frac{2\pi r}{r^2} = \frac{1}{r}.$$

Daher ist $s < \infty$. Nach Definition von s existiert eine Folge $\{f_n\} \subset \mathcal{F}$ mit

$$\lim_{n\to\infty} |f_n'(a)| = s.$$

Wegen Satz 10.11, dem Satz von Montel, gibt es eine Teilfolge $\{g_n\}$ von $\{f_n\}$, welche auf allen kompakten Teilmengen von G gleichmäßig gegen eine Funktion g konvergiert. Diese Funktion g ist wegen Satz 10.10 aus $H(G)$. In diesem Satz wird außerdem ausgesagt, dass für jedes $k \in \mathbb{N}$ auch die Folge $\{g_n^{(k)}\}$ der k-ten Ableitungen auf kompakten Teilmengen von G gleichmäßig gegen die k-te Ableitung $g^{(k)}$ von g konvergiert.Daher ist einerseits $\lim_{n\to\infty} |g_n'(a)| = s$ (da $\{g_n\}$ eine Teilfolge von $\{f_n\}$ ist), andererseits $\lim_{n\to\infty} |g_n'(a)| = |g'(a)|$, also $|g'(a)| = s$. Wegen $g_n(a) = 0$ ist $g(a) = 0$. Wegen $s > 0$ ist g nicht konstant. Satz 10.12 liefert, dass g auf G injektiv ist. Insgesamt ist also $g \in \mathcal{F}$ eine Lösung der Aufgabe

$$\text{(P)} \qquad \text{Maximiere} \quad |f'(a)|, \quad f \in \mathcal{F}.$$

Es bleibt zu zeigen, dass $g(G) = B(0;1)$. Angenommen, dies wäre nicht der Fall. Dann existiert ein $w \in B(0;1) \setminus f(G)$. Wir zeigen die Existenz eines $f \in \mathcal{F}$ mit $|f'(a)| > |g'(a)|$, was dann einen Widerspruch dazu ergibt, dass $g \in \mathcal{F}$ eine Lösung von (P) ist. Hierzu betrachte man den durch

$$\psi_w(z) := \frac{w - z}{1 - \overline{w}z}$$

definierten Automorphismus $\psi_w \in \mathrm{Aut}(B(0;1))$ von $B(0;1)$. Dieser Automorphismus der offenen Einheitskreisscheibe $B(0;1)$ verschwindet genau in $w \notin g(G)$. Daher ist $\psi_w \circ g \colon G \longrightarrow B(0;1)$ eine auf der offenen, einfach zusammenhängenden Menge G holomorphe, nicht verschwindende Funktion und besitzt daher eine holomorphe Quadratwurzel. Es existiert also eine holomorphe Funktion $q \colon G \longrightarrow \mathbb{C}$ mit $q(z)^2 = (\psi_w \circ g)(z)$ für alle $z \in G$. Offenbar ist mit g auch q injektiv. Da q auf G nicht verschwindet, ist $\lambda := q(a) \neq 0$. Ferner ist

$$\lambda^2 = q(a)^2 = (\psi_w \circ g)(a) = \psi_w(g(a)) = \psi_w(0) = w.$$

Wegen $w \in B(0;1)$ ist $|\lambda| < 1$. Wir zeigen, dass $f := \psi_\lambda \circ q$ ein Element aus $\mathcal{F}$ mit $|f'(a)| > |g'(a)|$ ist, womit wir dann den gewünschten Widerspruch zu der Annahme $B(0;1) \setminus g(G) \neq \emptyset$ erhalten haben. Offensichtlich ist f holomorph und injektiv auf G, $f(G) \subset B(0;1)$ (man beachte hierzu, dass mit $\psi_w \circ g$ auch die Quadratwurzel q eine Abbildung von G nach $B(0;1)$ ist), $f(a) = 0$ wegen $q(a) = \lambda$ und daher insgesamt $f \in \mathcal{F}$. Jetzt gilt es, $f'(a)$ zu berechnen. Wegen $q(a) = \lambda$ und $f = \psi_\lambda \circ q$ erhält man mit Hilfe der Kettenregel

$$(*) \qquad f'(a) = \psi_\lambda'(q(a))q'(a) = \psi_\lambda'(\lambda(q'(a) = -\frac{q'(a)}{1 - |\lambda|^2}.$$

Jetzt müssen wir $q'(a)$ berechnen. Dies geschieht durch Differenzieren von $q^2 = \psi_w \circ g$ und Auswertung an der Stelle a und ergibt unter erneuter Berücksichtigung von $q(a) = \lambda$, dass

$$(**) \quad 2\lambda q'(a) = \psi_w'(g(a))g'(a) = \psi_w'(0)g'(a) = -(1 - |w|^2)g'(a) = -(1 - |\lambda|^4)g'(a).$$

Aus $(*)$ und $(**)$ erhalten wir

$$f'(a) = \frac{|1+|\lambda|^2}{2\lambda} g'(a)$$

und damit

$$|f'(a)| = \frac{1+|\lambda|^2}{2|\lambda|} |g'(a)| > |g'(a)|,$$

wobei wir die wegen $|\lambda| < 1$ gültige Ungleichung

$$\frac{1+|\lambda|^2}{2|\lambda|} > 1$$

ausgenutzt haben. Der Riemannsche Abbildungssatz ist damit bewiesen. □

10.5 Die Schwarz-Christoffel Abbildung

10.5.1 Einführung

In diesem Unterabschnitt beschäftigen wir uns mit der konformen Abbildung der oberen Halbebene[51] auf einen *Polygonbereich*. Hierunter verstehen wir eine offene, einfach zusammenhängende Menge P in $\mathbb{C}$, die von einem Polygon berandet ist. Unter einem *Polygon* oder einem (geschlossenen) *Polygonzug* in $\mathbb{C}$ verstehen wir hierbei eine Folge von endlich vielen Ecken $w_1, \ldots, w_n$, wobei w_k, w_{k+1}, $k = 1, \ldots, n-1$, und w_n, w_1 durch eine Strecke oder Kante miteinander verbunden sind. Hierbei entstehen *Innenwinkel* $\alpha_1\pi, \ldots, \alpha_n\pi$ mit $\alpha_1, \ldots, \alpha_n \in (0, 2)$. Die Anordnung der Ecken auf dem Polygon sei im mathematisch positiven Sinne vorgenommen, sodass der Polygonbereich beim Übergang von w_k nach w_{k+1} links von der entsprechenden Kante liegt. Ist der Innenwinkel $\alpha_k\pi$ an der k-ten Ecke gegeben, so ist $(1-\alpha_k)\pi \in (-\pi, \pi)$ die Richtungsänderung beim Durchlaufen der k-ten Ecke, siehe Abbildung 80. Hierbei spielt es keine Rolle, ob der Innenwinkel ein spitzer oder stumpfer Winkel ($0 < \alpha_k \leq 1$, Richtungsänderung nach links) oder ein überstumpfer Winkel ($1 < \alpha_k < 2$, Richtungsänderung nach rechts) ist. Weil das Gebiet genau einmal umlaufen wird, ist die Summe der Richtungsänderungen 2π, d. h.

$$2\pi = \sum_{k=1}^{n} (1-\alpha_k)\pi \quad \text{bzw.} \quad \sum_{k=1}^{n} \alpha_k = n-2.$$

Die einfachsten beschränkten Polygonbereiche sind natürlich Dreiecke und Rechtecke, siehe Abbildung 81. In Abbildung 82 geben wir einen beschränkten Polygonbereich

[51] Die offene Einheitskreisscheibe und die obere Halbebene sind konform äquivalent. Denn $G\colon B(0;1) \longrightarrow \mathbb{H}$, definiert durch

$$G(w) := i\frac{1-w}{1+w},$$

ist eine konforme Abbildung von $B(0;1)$ auf $\mathbb{H}$. Offenbar kann G auf triviale Art zu einer stetigen, bijektiven Abbildung von $\operatorname{cl} B(0;1)$ auf $\operatorname{cl} H$ fortgesetzt werden, wobei dem Punkt $w = -1$ auf dem Rande von $B(0;1)$ der Punkt ∞ auf $\mathbb{R} \cup \{\infty\}$ entspricht und jeder andere Punkt des Randes von $B(0;1)$ den Punkten der reellen Achse $\mathbb{R}$ entsprechen. Es ist von der Notation her etwas einfacher, von der oberen Halbebene statt der Einheitskreisscheibe auszugehen.

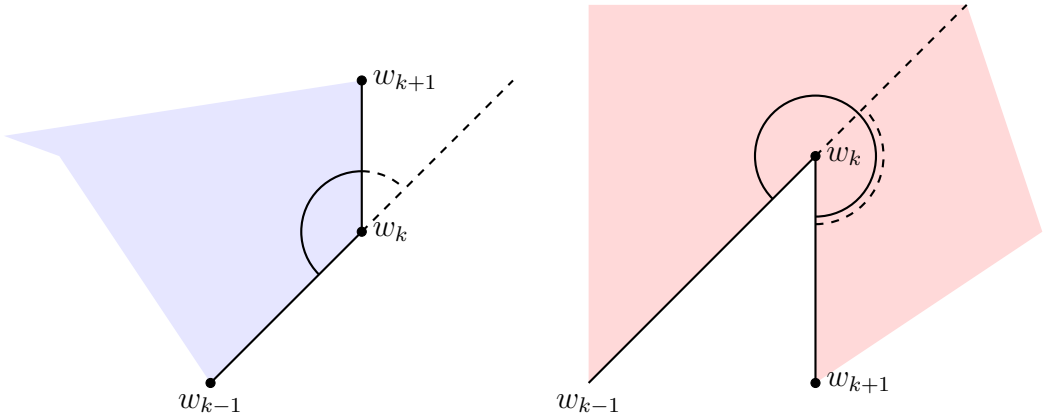


Abbildung 80: Richtungsänderung in der k-ten Ecke nach links und nach rechts

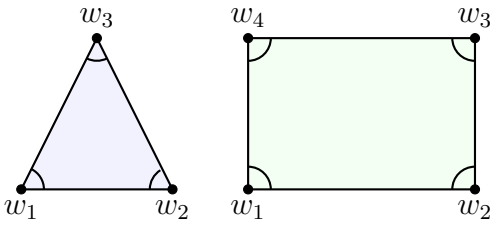


Abbildung 81: Dreieck und Rechteck

sowie seine Ecken und Innenwinkel an. Wir erlauben auch Ecken $w_n = \infty$, wodurch unbeschränkte Polygonbereiche zugelassen sind. Bisher traten keine unendlichen Ecken auf. Das ist in der Abbildung 83 anders. Schließlich geben wir in Abbildung 84 noch

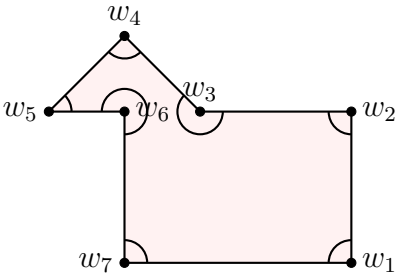


Abbildung 82: Ein beschänkter Polygonbereiche

einen unbeschränkten Polygonbereich mit drei endlichen und einer unendlichen Ecke an. Die Berechnung des Innenwinkels $\alpha_4\pi$ in der Ecke $w_4 = \infty$ aus Abbildung 84 erfolgt über die Gleichung $\sum_{k=1}^{4} \alpha_k = 2$, d. h. es ist $\alpha_4 = 2 - (\alpha_1 + \alpha_2 + \alpha_3)$. Entsprechend ist in Abbildung 83 links $\alpha_1 = -1$.

Wegen des Riemannschen Abbildungssatzes existiert eine konforme Abbildung von $\mathbb{H}$, der oberen Zahlenebene in $\mathbb{C}$, auf den offenen und einfach zusammenhängenden Polygonbereich P. Unser Ziel in diesem Unterabschnitt ist es, die beiden folgenden Sätze 10.13 und 10.14 zu beweisen und an einigen Beispielen zu erläutern. In diesen Sätzen

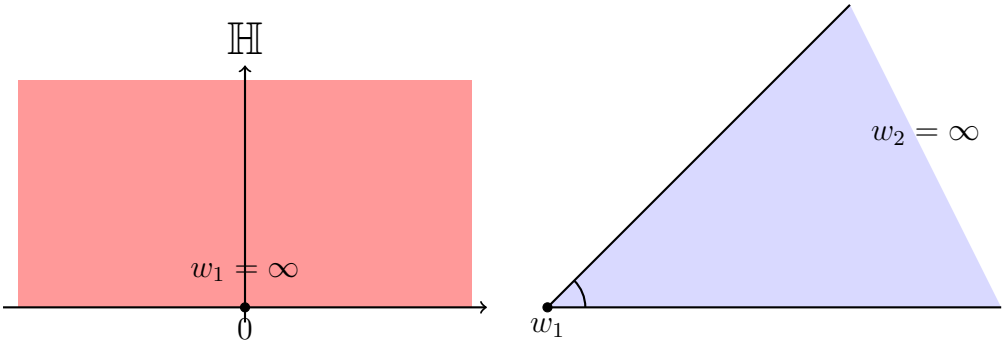


Abbildung 83: Polygonbereiche mit einer unendlichen Ecke

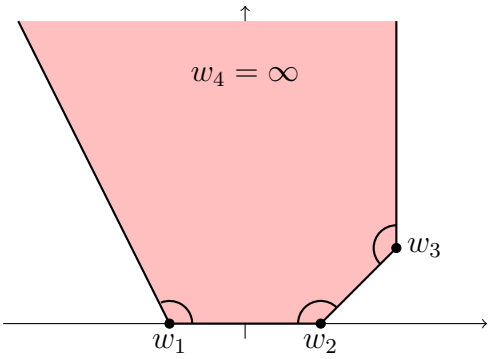


Abbildung 84: Polygonbereich mit drei endlichen und einer unendlichen Ecke

wird eine Darstellung einer konformen Abbildung von $\mathbb{H}$ auf einen Polygonbereich P angegeben. Schon bei der Formulierung der Sätze wird (ohne Beweis!) benutzt, dass sich eine konforme Abbildung f von $\mathbb{H}$ auf den offenen, einfach zusammenhängenden Polygonbereich P mit den n Ecken $w_1, \ldots, w_n$ zu einer stetigen Bijektion von $\operatorname{cl} H$ auf $\operatorname{cl} P$ fortsetzen lässt. Diese Fortsetzung wird wieder mit f bezeichnet. Hierbei ist $\operatorname{cl} \mathbb{H}$ als $\mathbb{R} \cup \{\infty\}$ zu interpretieren. Die konforme Abbildung f hat also die Eigenschaft, dass sie die obere Halbebene $\mathbb{H}$ auf den offenen Polygonbereich P und den Rand von $\mathbb{H}$ (einschließlich des Punktes ∞) auf den Rand von P abbildet. Ist also eine konforme Abbildung f von $\mathbb{H}$ auf P gegeben, so existiert zu jeder Ecke (vertex) w_k von P genau eine *prevertex* $z_k \in \mathbb{R} \cup \{\infty\}$ mit $f(z_k) = w_k$, $k = 1, \ldots, n$. Es ist $z_1 < z_2 < \cdots < z_{n-1} < z_n$, da wir annehmen, dass die Ecken w_k von P entgegen dem Uhrzeigersinn angeordnet sind. Im ersten Satz 10.13 wird angenommen, dass die Urbilder aller Ecken des Polygonbereichs endlich sind, dass also $z_k := f^{-1}(w_k) \in \mathbb{R}$, $k = 1, \ldots, n$, während $f(\infty) = w_n$ im zweiten Satz 10.14 ist. Dieser zweite Satz folgt relativ einfach aus dem ersten, wie wir im Anschluss an ihre Formulierung zeigen werden.

Wir überlegen uns *notwendige* Bedingungen, die die konforme Abbildung f mit den angegebenen Eigenschaften zu erfüllen hat. Hierbei gehen wir vom ersten Fall aus, in dem alle prevertices z_k endlich sind.

Für ein $k \in \{1, \ldots, n-1\}$ definieren wir $z\colon [0,1] \longrightarrow \mathbb{R}$ durch $z(t) := z_k + t(z_{k+1} - z_k)$

und anschließend $w\colon [0,1] \longrightarrow \mathbb{C}$ durch $w(t) := f(z(t))$. Offenbar ist

$$w([0,1]) = f \circ z([0,1])$$

das w_k und w_{k+1} verbindende (zum Rand von P gehörende) Geradenstück $[w_k, w_{k+1}]$. Folglich ist $\arg(w'(t)) = \arg(w_{k+1} - w_k)$ für $t \in (0,1)$. Da

$$w'(t) = f'(z(t))z'(t) = f'(z(t))(\underbrace{z_{k+1} - z_k}_{>0}),$$

ist $\arg(f'(z)) = \arg(w_{k+1} - w_k)$ für $z \in (z_k, z_{k+1})$, insbesondere gilt[52]:

(1) Es ist $\arg(f'(z))$ konstant auf (z_k, z_{k+1}).

Weiter gilt:

(2) Ist $z \in (z_{k-1}, z_k)$ und $\tilde{z} \in (z_k, z_{k+1})$, so ist $\arg(f'(\tilde{z})) \equiv \arg(f'(z)) + (1-\alpha_k)\pi$ modulo 2π, wobei $\alpha_k\pi$ der Innenwinkel in der Ecke w_k bezüglich des Polygonbereichs P ist.

Für einen Beweis dieser Aussage verweisen wir auf Abbildung 85. Mit $z \in (z_{k-1}, z_k)$,

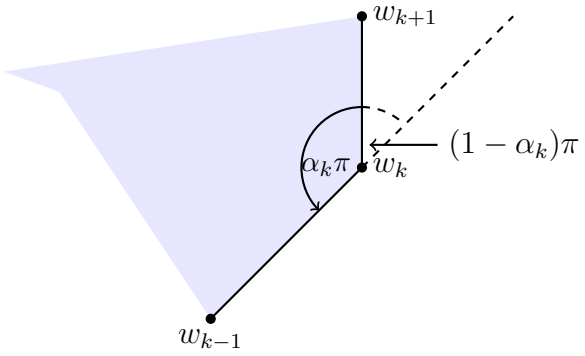


Abbildung 85: Veranschaulichung von Aussage (2)

$\tilde{z} \in (z_k, z_{k+1})$ ist genauer

$$\begin{aligned}
\exp(i\alpha_k\pi) &= \left(\frac{w_{k-1} - w_k}{|w_{k-1} - w_k|}\right) \Big/ \left(\frac{w_{k+1} - w_k}{|w_{k+1} - w_k}\right) \\
&= -\left(\frac{w_k - w_{k-1}}{|w_k - w_{k-1}|}\right) \Big/ \left(\frac{w_{k+1} - w_k}{|w_{k+1} - w_k}\right) \\
&= \exp(i(\pi + \arg(w_k - w_{k-1}) - \arg(w_{k+1} - w_k))) \\
&= \exp(i(\pi + \arg(f'(z)) - \arg(f'(\tilde{z}))))
\end{aligned}$$

und daher

$$\alpha_k\pi \equiv \pi + \arg(f'(z)) - \arg(f'(\tilde{z})) \quad \text{modulo } 2\pi$$

[52]Bemerkungen zur Argumentfunktion arg finden sich auf Seite 264

bzw.

$$\arg(f'(\tilde{z})) \equiv \arg(f'(z)) + (1-\alpha_k)\pi \quad \text{modulo } 2\pi,$$

womit die Aussage (2) bewiesen ist. Eine Funktion f', welche den Bedingungen (1) und (2) genügt, ist

$$f'(z) = c\prod_{k=1}^{n}(z-z_k)^{\alpha_k-1} \tag{3}$$

mit beliebigem $c \in \mathbb{C}$. Hierbei ist f' auf der n-fach geschlitzten komplexen Ebene $\mathbb{C}\setminus\{\bigcup_{k=1}^{n}\{z_k+iy : y \le 0\}$, einer einfach zusammenhängenden Menge, definiert. Hierbei sind die Faktoren in der Definition von f' durch

$$\begin{aligned}(z-z_k)^{\alpha_k-1} &= \exp[(\alpha_k-1)\log(z-z_k)] \\ &= \exp[(\alpha_k-1)(\log|z-z_k| + i\arg(z-z_k))]\end{aligned}$$

mit

$$-\pi < \arg(z-z_k) \le \pi$$

gegeben. Für *reelles* z ist daher

$$\arg(z-z_k) = \begin{cases} 0, & z > z_k, \\ \pi, & z < z_k, \end{cases} \qquad k = 1,\ldots,n,$$

und daher

$$(z-z_k)^{\alpha_k-1} = \begin{cases} (z-z_k)^{\alpha_k-1}, & z > z_k, \\ |z-z_k|^{\alpha_k-1}e^{i(\alpha_k-1)\pi}, & z < z_k, \end{cases} \qquad k = 1,\ldots,n.$$

Mit $f'(z) = c\prod_{k=1}^{n}(z-z_k)^{\alpha_k-1}$ wie in (3) und $z \in (z_{i-1}, z_i)$ ist daher

$$\arg(f'(z)) \equiv \arg(c) + \sum_{k=1}^{n}(\alpha_k-1)\arg(z-z_k) = \arg(c) + \pi\sum_{k=i}^{n}(\alpha_k-1)$$

modulo 2π. Daher ist $\arg(f'(z))$ konstant auf jedem der Intervalle, die durch aufeinanderfolgende prevertices gebildet werden und damit die Bedingung (1) erfüllt. Zum Nachweis der Bedingung (2) sei $\tilde{z} \in (z_i, z_{i+1})$. Dann ist

$$\begin{aligned}\arg(f'(\tilde{z})) &\equiv \arg(c) + \pi\sum_{k=i+1}^{n}(\alpha_k-1) \\ &= \arg(c) + \pi\sum_{k=i}^{n}(\alpha_k-1) + (1-\alpha_i)\pi \\ &\equiv \arg(f'(z)) + (1-\alpha_i)\pi,\end{aligned}$$

womit auch die Gültigkeit der Bedingung (2) nachgewiesen ist. Eine Integration von f' in der Darstellung (3) liefert

$$f(z) = a + c\int^{z}\prod_{k=1}^{n}(\zeta-z_k)^{\alpha_k-1}\,d\zeta$$

und das ist genau die Darstellung, wie sie in den beiden folgenden Sätzen von Schwarz-Christoffel behauptet wird. Man beachte, dass der Integrand $f'(\zeta) = c\prod_{k=1}^{n}(\zeta - z_k)^{\alpha_k - 1}$ in der geschlitzten Ebene holomorph ist und daher das Integral von einem Anfangspunkt z_0 bis zum Endpunkt z wegunabhängig ist, wenn der Weg nur in der geschlitzten Ebene verläuft.

Satz 10.13 (Schwarz-Christoffel 1) *Sei $P \subset \mathbb{C}$ ein offener, einfach zusammenhängender Polygonbereich, dessen Rand ein Polygon mit Ecken $w_1, \ldots, w_n$, die entgegen dem Uhrzeigersinn angeordnet sind, und zugehörigen Innenwinkeln $\alpha_1\pi, \ldots, \alpha_n\pi$ ist. Sei f eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf P. Seien $z_k := f^{-1}(w_k) \in \mathbb{R}$, $k = 1, \ldots, n$. Dann existieren komplexe Konstanten a, c derart, dass sich f in der Form*

$$(*) \qquad f(z) = a + c\int^{z} \prod_{k=1}^{n}(\zeta - z_k)^{\alpha_k - 1}\, d\zeta$$

darstellen lässt.

Satz 10.14 (Schwarz-Christoffel 2) *Sei $P \subset \mathbb{C}$ ein offener, einfach zusammenhängender Polygonbereich, dessen Rand ein Polygon mit Ecken $w_1, \ldots, w_n$, die entgegen dem Uhrzeigersinn angeordnet sind und zugehörigen Innenwinkeln $\alpha_1\pi, \ldots, \alpha_n\pi$ ist. Sei f eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf P mit $z_k := f^{-1}(w_k)$, $k = 1, \ldots, n$, und $z_n = \infty$. Dann existieren komplexe Konstanten a, c derart, dass sich f in der Form*

$$(**) \qquad f(z) = a + c\int^{z} \prod_{k=1}^{n-1}(\zeta - z_k)^{\alpha_k - 1}\, d\zeta$$

darstellen lässt.

In beiden Sätzen ist die untere Integrationsgrenze unbestimmt geblieben, weil sie lediglich die Konstante a beeinflusst. Durch $(*)$ bzw. $(**)$ ist eine sogenannte *Schwarz-Christoffel-Darstellung* der konformen Abbildung f von $\mathbb{H}$ auf den Polygonbereich P gegeben.

Wir *nehmen nun an*, Satz 10.13 sei schon bewiesen und folgern hieraus die Gültigkeit von Satz 10.14, siehe E. M. Stein, R. Shakarchi (2003, S. 244 ff.).

Beweis von Satz 10.14: Wir können o. B. d. A. annehmen, dass $z_k \neq 0$, $k = 1, \ldots, n-1$. Man wähle ein (reelles) $z_n^* > 0$ und definiere $\Phi\colon \mathbb{H} \longrightarrow \mathbb{C}$ durch

$$\Phi(z) := z_n^* - \frac{1}{z}.$$

Wegen Satz 10.5 ist Φ ein Automorphismus der oberen Halbebene $\mathbb{H}$, weiter ist $\Phi(\infty) = z_n^*$. Man definiere $z_k^* := \Phi(z_k)$, $k = 1, \ldots, n-1$. Dann ist $f \circ \Phi^{-1}$ eine konforme Abbildung von $\mathbb{H}$ auf den Polygonbereich P und

$$(f \circ \Phi^{-1})(z_k^*) = f(z_k) = w_k, \quad k = 1, \ldots, n.$$

Da wir annehmen, Satz 10.13 sei schon bewiesen, existieren komplexe Konstanten a, c derart, dass sich $f \circ \Phi^{-1}$ in der Form

$$\begin{aligned}
(f \circ \Phi^{-1})(z') &= a + c \int^{z'} \prod_{k=1}^{n} (\zeta - z_k^*)^{\alpha_k - 1}\, d\zeta \\
&= a + c \int^{\Phi^{-1}(z')} \left(\prod_{k=1}^{n} (\Phi(\eta) - z_k^*)^{\alpha_k - 1} \right) \Phi'(\eta)\, d\eta \\
&\qquad \text{(Substitution } \zeta = \Phi(\eta)\text{)} \\
&= a + c \int^{\Phi^{-1}(z')} \left(\prod_{k=1}^{n-1} (\Phi(\eta) - \Phi(z_k))^{\alpha_k - 1} \right) \left(-\frac{1}{\eta} \right)^{\alpha_n - 1} \frac{1}{\eta^2}\, d\eta \\
&= a + c' \int^{\Phi^{-1}(z')} \prod_{k=1}^{n-1} (\eta - z_k)^{\alpha_k - 1} \eta^{\sum_{k=1}^{n} (1 - \alpha_k) - 2}\, d\eta \\
&= a + c' \int^{\Phi^{-1}(z')} \prod_{k=1}^{n-1} (\eta - z_k)^{\alpha_k - 1}\, d\eta \\
&\qquad \text{(wegen } \textstyle\sum_{k=1}^{n} \alpha_k = n - 2\text{)}
\end{aligned}$$

darstellen lässt. Hierbei ist

$$c' := c(-1)^{\alpha_n - 1} \prod_{k=1}^{n-1} z_k^{1 - \alpha_k}.$$

Mit $z = \Phi^{-1}(z')$ erhalten wir

$$f(z) = a + c' \int^{z} \prod_{k=1}^{n-1} (\zeta - z_k)^{\alpha_k - 1}\, d\zeta.$$

Damit ist nachgewiesen, dass aus Satz 10.13 die Gültigkeit von Satz 10.14 folgt. □

Bemerkung: Sei $P \subset \mathbb{C}$ ein offener, einfach zusammenhängender Polygonbereich, dessen Rand ein Polygon mit Ecken $w_1, \ldots, w_n$, die entgegen dem Uhrzeigersinn angeordnet sind, und zugehörigen Innenwinkeln $\alpha_1\pi, \ldots, \alpha_n\pi$ ist. Sei f eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf P. Wir unterscheiden zwei Fälle.

Im ersten Fall sei $z_k := f^{-1}(w_k) \in \mathbb{R}$, $k = 1, \ldots, n$, alle n prevertices seien also endlich. Wir können annehmen, dass $z_1 < \cdots < z_n$. Seien $y_p < y_q < y_r$ mit $1 \le p < q < r \le n$ drei vorgegebene Punkte auf der reellen Achse $\mathbb{R}$. Wegen des zweiten Teils von Satz 10.6 existiert (genau ein) Automorphismus $F \in \mathrm{Aut}(\mathbb{H})$ mit[53] $F(y_p) = z_p$, $F(y_q) = z_q$ und $F(y_r) = z_r$. Nun definiere man

$$y_k := F^{-1}(z_k), \quad k \in \{1, \ldots, n\} \setminus \{p, q, r\}.$$

[53] Wir erinnern daran, dass man einen Automorphismus $F \in \mathrm{Aut}(\mathbb{H})$ als stetige, bijektive Abbildung auf $\mathbb{R} \cup \{\infty\}$ fortsetzen kann, siehe die Bemerkung im Anschluss an Satz 10.5. Insbesondere ist diese Fortsetzung von F eine bijektive Abbildung von $\mathbb{R} \cup \{\infty\}$ auf sich.

Dann ist $f \circ F^{-1}$ eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf den Polygonbereich P mit

$$(f \circ F^{-1})(y_k) = f(z_k) = w_k, \quad k = 1, \ldots, n.$$

Wegen Satz 10.13 existieren daher komplexe Parameter a, c mit

$$(f \circ F^{-1})(z) = a + c \int^z \prod_{k=1}^{n} (\zeta - y_k)^{\alpha_k - 1} \, d\zeta.$$

Mit anderen Worten: Es existiert eine konforme Abbildung von $\mathbb{H}$ auf P, bei deren Schwarz-Christoffel-Darstellung *drei* der n reellen prevertices beliebig vorgegeben werden können.

Im zweiten Fall ist $z_n = \infty$ bzw. $f(\infty) = w_n$ und nach wie vor $z_k := f^{-1}(w_k)$, $k = 1, \ldots, n-1$. Dann gilt eine zum ersten Fall ganz analoge Aussage. Denn gibt man sich $y_p < y_q$ mit $1 \leq p < q \leq n-1$ beliebig vor und definiert man $y_n := \infty$, so existiert wiederum wegen des zweiten Teiles von Satz 10.6 genau ein $F \in \mathrm{Aut}(\mathbb{H})$ mit $F(y_p) = z_p$, $F(y_q) = z_q$ und $F(\infty) = \infty$. Definiert man dann

$$y_k := F^{-1}(z_k), \quad k \in \{1, \ldots, n-1\} \setminus \{p, q\},$$

so ist

$$(f \circ F^{-1})(y_k) = f(z_k) = w_k, \quad k = 1, \ldots, n-1, \qquad (f \circ F^{-1})(\infty) = w_n.$$

Wegen Satz 10.14 existieren daher komplexe Parameter a, c mit

$$(f \circ F^{-1})(z) = a + c \int^z \prod_{k=1}^{n-1} (\zeta - y_k)^{\alpha_k - 1} \, d\zeta.$$

□

Wir werden noch klären müssen, weshalb sich eine konforme Abbildung von $\mathbb{H}$ auf einen Polygonbereich zu einer stetigen bijektiven Abbildung von $\mathrm{cl}\,\mathbb{H}$ auf $\mathrm{cl}\,P$ fortsetzen lässt (dies geschieht durch Satz 10.17 und eine anschließende Bemerkung). Bevor wir dies tun, wollen wir durch einige Beispiele ein Gefühl für die auftretenden Probleme gewinnen.

10.5.2 Spezialfälle, Beispiele

Beispiel: Wir wollen die konforme Abbildung der oberen Halbebene

$$\mathbb{H} := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$$

auf den in Abbildung 86 dargestellten Sektor

$$P := \{w \in \mathbb{C} : 0 < \arg(w) < \alpha\pi\},$$

wobei $\alpha \in (0, 2)$, mit Hilfe der Schwarz-Christoffel Darstellung bestimmen. Wegen Satz 10.14 und der anschließenden Bemerkung existieren zu einer konformen Abbildung f

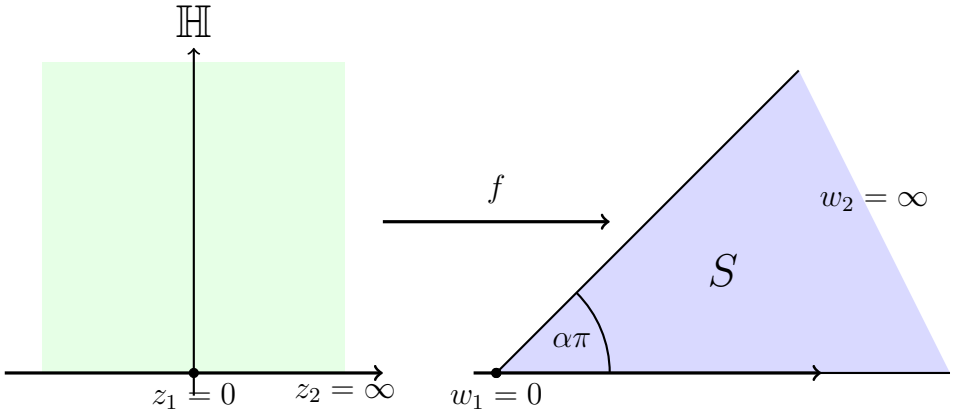


Abbildung 86: Ein Sektor mit Innenwinkel $\alpha\pi$

von $\mathbb{H}$ auf den Sektor S mit $f(0) = 0$ und $f(\infty) = \infty$ komplexe Konstanten a, a_0 und c, c_0 mit

$$f(z) = a + c \int^z \zeta^{\alpha-1}\, d\zeta = a_0 + c_0 z^\alpha.$$

Wegen $f(0) = 0$ ist $a_0 = 0$ und damit $f(z) = c_0 z^\alpha$ mit $c_0 \in \mathbb{C}$. Es ist sogar c_0 reell und positiv, wenn das Segment $[0, \infty)$ im Abschluss von $\mathbb{H}$ auf $[0, \infty)$ im Abschluss von S abgebildet wird. Ist dagegen $x \in (-\infty, 0]$, so ist

$$f(x) = c_0 x^\alpha = c_0(-1)^\alpha(-x)^\alpha = c_0(-x)^\alpha e^{i\pi\alpha}$$

und daher

$$f((-\infty, 0]) = \{\lambda e^{i\pi\alpha} : \lambda > 0\}.$$

Die Situation verdeutlichen wir uns in Abbildung 87. In Abbildung 88 geben wir Bilder

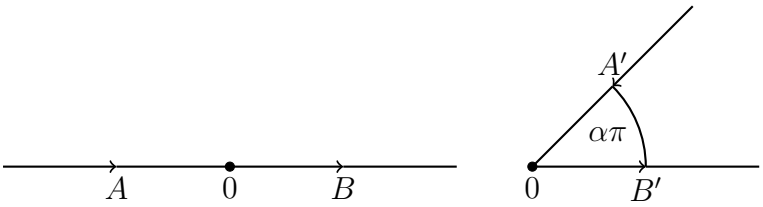


Abbildung 87: Die konforme Abbildung z^α

der Geraden $\{z \in \mathbb{C} : \mathrm{Im}(z) = 0.005k\}$, $k = 0, \ldots, 20$, unter der Abbildung $z^{1/4}$ an. □

Beispiel: Wir wollen die konforme Abbildung der oberen Halbebene

$$\mathbb{H} := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$$

auf den in Abbildung 89 dargestellten Halbstreifen

$$P := \{w \in \mathbb{C} : \mathrm{Re}(w) > 0,\ 0 < \mathrm{Im}(w) < \pi\}$$

mit Hilfe der Schwarz-Christoffel Darstellung bestimmen. In P sind die Innenwinkel zu

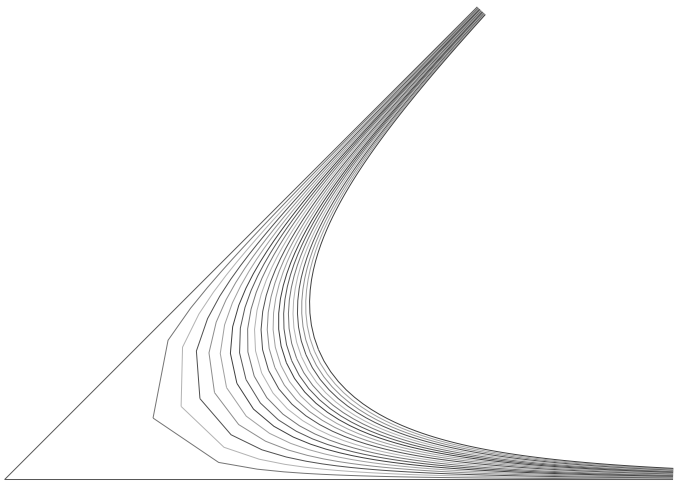


Abbildung 88: Bilder von zu $\mathbb{R}$ parallelen Linien in $\mathbb{H}$ unter $z^{1/4}$

$w_1 = \pi i$ bzw. $w_2 = 0$ jeweils durch $\frac{1}{2}\pi$ gegeben. Wegen der Bemerkung im Anschluss an die Sätze 10.13 und 10.14 existieren zu einer konformen Abbildung f von $\mathbb{H}$ auf P mit $f(-1) = w_1$, $f(1) = w_2$ und $f(\infty) = \infty$ komplexe Konstanten a, a_0 und c derart, dass

$$f(z) = a + c\int^z \frac{d\zeta}{\sqrt{(\zeta+1)(\zeta-1)}} = a + c\int^z \frac{d\zeta}{\sqrt{\zeta^2-1}} = a_0 + c\,\operatorname{arcosh} z.$$

Aus

$$w_1 = \pi i = f(-1) = a_0 + c\pi i, \qquad w_2 = 0 = f(1) = a_0$$

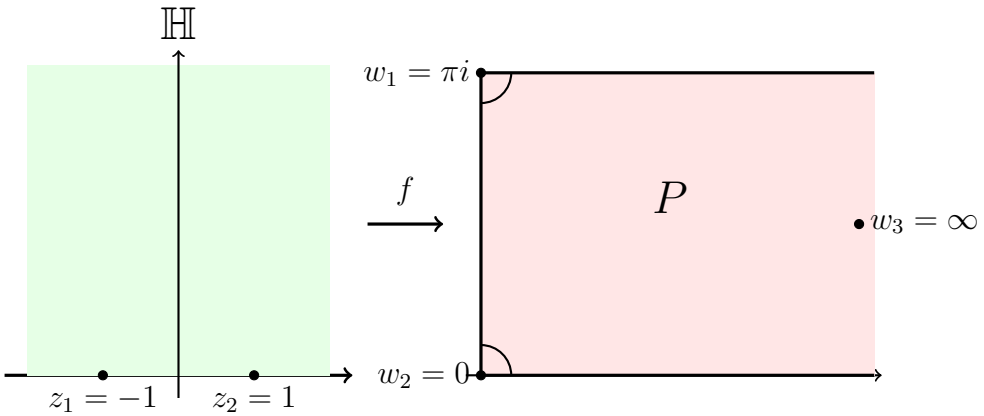


Abbildung 89: Ein Halbstreifen als Polygonbereich

erhalten wir $a_0 = 0$ und $c = 1$ und damit

$$f(z) = \operatorname{arcosh} z,$$

wobei $f(z) = \operatorname{arcosh} z$ die inverse hyperbolische Kosinusfunktion ist, also eine Lösung der Gleichung

$$\cosh f(z) = \frac{e^{f(z)} + e^{-f(z)}}{2} = z.$$

In Abbildung 90 links geben wir in der oberen Halbebene $\mathbb{H}$ einige Parallelen zur $\mathbb{R}$-Achse an, in derselben Abbildung sieht man rechts deren Bilder unter der Funktion arcosh.

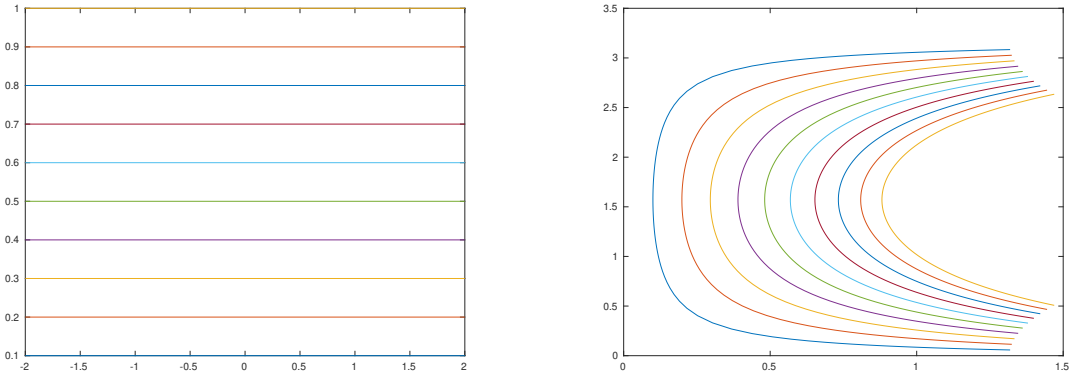


Abbildung 90: Zu $\mathbb{R}$ parallele Linien in $\mathbb{H}$ und deren Bilder unter arcosh

Ein nur unwesentlich anderes Ergebnis erhält man, wenn man die obere Halbebene $\mathbb{H}$ auf den in Abbildung 91 dargestellten vertikalen Halbstreifen

$$Q := \left\{ w \in \mathbb{C} : -\frac{\pi}{2} < \operatorname{Re}(w) < \frac{\pi}{2},\ \operatorname{Im}(w) > 0 \right\}$$

mit den Ecken $w_1 = -\frac{\pi}{2}$, $w_2 = \frac{\pi}{2}$ sowie $w_3 = \infty$ konform abbilden will. Wegen $Q = \frac{\pi}{2} + iP$ ist eine konforme Abbildung f von $\mathbb{H}$ auf Q mit $f(-1) = w_1$, $f(1) = w_2$ und $f(\infty) = \infty$ durch

$$f(z) = \frac{\pi}{2} + i \operatorname{arcosh} z$$

gegeben. Dann ist

$$\begin{aligned} z &= \cosh\left(-i\left(f(z) - \frac{\pi}{2}\right)\right) \\ &= \frac{e^{-i(f(z)-\pi/2)} + e^{i(f(z)-\pi/2)}}{2} \\ &= \frac{ie^{-if(z)} - ie^{if(z)}}{2} \\ &= \frac{e^{if(z)} - e^{-if(z)}}{2i} \end{aligned}$$

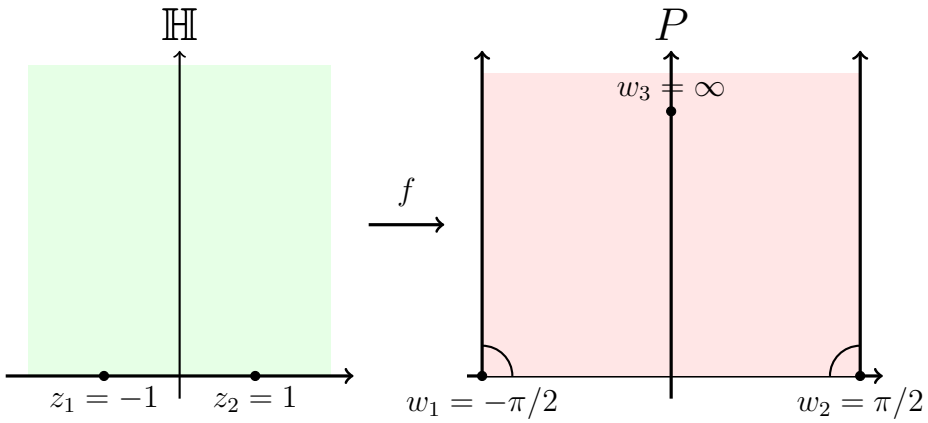


Abbildung 91: Ein vertikaler Halbstreifen als Polygonbereich

$$= \sin f(z)$$

und folglich

$$f(z) = \arcsin z.$$

In Abbildung 92 geben wir die Bilder zu $\mathbb{R}$ paralleler Linien in $\mathbb{H}$ unter der Abbildung

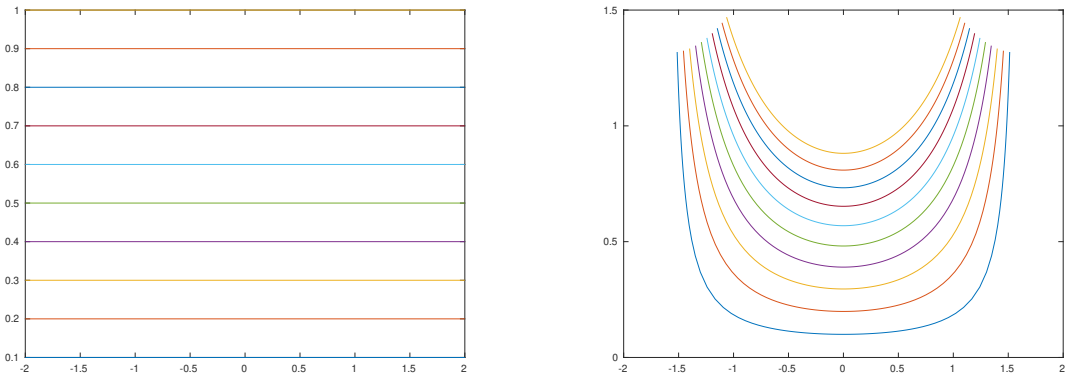


Abbildung 92: Zu $\mathbb{R}$ parallele Linien in $\mathbb{H}$ und deren Bilder unter arcsin

arcsin wieder. □

Bevor wir auf die konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf ein Dreieck eingehen, ist es zweckmäßig, einiges zur (vollständigen) Beta- und der Gamma-Funktion auszusagen bzw. ins Gedächtnis zu rufen. Für positive p, q definieren wir die *vollständige Beta-Funktion* durch

$$B(p,q) := \int_0^1 \zeta^{p-1}(1-\zeta)^{q-1}\, d\zeta.$$

Interessant ist der Zusammenhang zwischen der Beta- und der Gamma-Funktion. Diese ist bekanntlich für $x > 0$ (oder: komplexe x mit positivem Realteil) durch

$$\Gamma(x) := \int_0^\infty e^{-t} t^{x-1}\, dt$$

definiert.

Lemma 10.15 *Für positive p, q ist*

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Beweis: Seien p, q positiv. Dann ist

$$\Gamma(p)\Gamma(q) = \int_0^\infty e^{-t}t^{p-1}\,dt \int_0^\infty e^{-s}s^{q-1}\,ds = \int_0^\infty \int_0^\infty e^{-(t+s)}t^{p-1}s^{q-1}\,dt\,ds.$$

Nun definieren wir $F\colon (0,\infty)\times(0,1) \longrightarrow (0,\infty)\times(0,\infty)$ durch

$$F(x,y) := \begin{pmatrix} xy \\ x(1-y) \end{pmatrix}.$$

Dann ist F ein Diffeomorphismus von $(0,\infty)\times(0,1)$ auf $(0,\infty)\times(0,\infty)$, wobei die Umkehrabbildung $F^{-1}\colon (0,\infty)\times(0,\infty) \longrightarrow (0,\infty)\times(0,1)$ durch

$$F^{-1}(t,s) = \begin{pmatrix} t+s \\ \dfrac{t}{t+s} \end{pmatrix}$$

gegeben ist. Die Funktionalmatrix von F ist

$$F'(x,y) = \begin{pmatrix} y & x \\ 1-y & -x \end{pmatrix}.$$

Bei der Substitution $(t,s) = F(x,y)$ liefert die Transformationsformel für Integrale

$$\begin{aligned}
\Gamma(p)\Gamma(q) &= \int_0^\infty \int_0^\infty e^{-(t+s)}t^{p-1}s^{q-1}\,dt\,ds \\
&= \int_0^1 \int_0^\infty e^{-x}(xy)^{p-1}x^{q-1}(1-y)^{q-1} \underbrace{|\det F'(x,y)|}_{=x}\,dx\,dy \\
&= \int_0^1 \int_0^\infty e^{-x}x^{p+q-1}y^{p-1}(1-y)^{q-1}\,dx\,dy \\
&= \int_0^\infty e^{-x}x^{p+q-1}\,dx \int_0^1 y^{p-1}(1-y)^{q-1}\,dy \\
&= \Gamma(p+q)\,B(p,q),
\end{aligned}$$

womit das Lemma bewiesen ist. □

Es folgt die *Eulersche Reflektionsformel*, siehe z. B. E. M. Stein, R. Shakarchi (2003, S. 164):

Lemma 10.16 *Für $a \in (0,1)$ ist*

$$\Gamma(a)\Gamma(1-a) = \frac{\pi}{\sin \pi a}.$$

Beweis: Nach Definition der Gamma-Funktion ist

$$\Gamma(a)\Gamma(1-a) = \int_0^\infty \int_0^\infty e^{-(t+s)} t^{-a} s^{a-1}\, dt\, ds.$$

Wir definieren $F\colon (0,\infty)\times(0,\infty) \longrightarrow (0,\infty)\times(0,\infty)$ durch

$$F(x,y) := \begin{pmatrix} \dfrac{x}{1+y} \\ \dfrac{xy}{1+y} \end{pmatrix}.$$

Dann ist F ein Diffeomorphismus von $(0,\infty)\times(0,\infty)$ auf $(0,\infty)\times(0,\infty)$, wobei die Umkehrabbildung $F^{-1}\colon (0,\infty)\times(0,\infty) \longrightarrow (0,\infty)\times(0,\infty)$ durch

$$F^{-1}(t,s) = \begin{pmatrix} t+s \\ \dfrac{s}{t} \end{pmatrix}$$

gegeben ist. Die Funktionalmatrix von F ist

$$F'(x,y) = \begin{pmatrix} \dfrac{1}{1+y} & -\dfrac{x}{(1+y)^2} \\ \dfrac{y}{1+y} & \dfrac{x}{(1+y)^2} \end{pmatrix},$$

daher ist

$$\det F'(x,y) = \frac{x}{(1+y)^2}.$$

Bei der Substitution $(t,s) = F(x,y)$ liefert die Transformationsformel für Integrale

$$\begin{aligned}
\Gamma(a)\Gamma(1-a) &= \int_0^\infty \int_0^\infty e^{-(t+s)} t^{-a} s^{a-1}\, dt\, ds \\
&= \int_0^\infty \int_0^\infty s^{-1} \left(\frac{s}{t}\right)^a e^{-(t+s)}\, dt\, ds \\
&= \int_0^\infty \int_0^\infty \left(\frac{1+y}{xy}\right) y^a e^{-x} \underbrace{\left|\det F'(x,y)\right|}_{=x/(1+y)^2}\, dx\, dy \\
&= \int_0^\infty \int_0^\infty \left(\frac{y^{a-1}}{1+y}\right) e^{-x}\, dx\, dy \\
&= \int_0^\infty \frac{y^{a-1}}{1+y}\, dy \\
&= \int_{-\infty}^\infty \frac{e^{ax}}{1+e^x}\, dx \\
&\qquad \text{(Substitution } y = e^x\text{)} \\
&= \frac{\pi}{\sin \pi a}.
\end{aligned}$$

Einen Beweis der letzten Gleichung mit Hilfe des *Residuensatzes* findet man bei E. M. Stein, R. Shakarchi (2003, S. 79–81). Hierzu definiert man

$$g(z) := \frac{e^{az}}{1+e^z}$$

und integriert g bei gegebenem $R > 0$ über einen Weg γ_R, der in Abbildung 93 angegeben ist. In dem von γ_R umrandeten Gebiet hat g in πi die einzige Polstelle erster

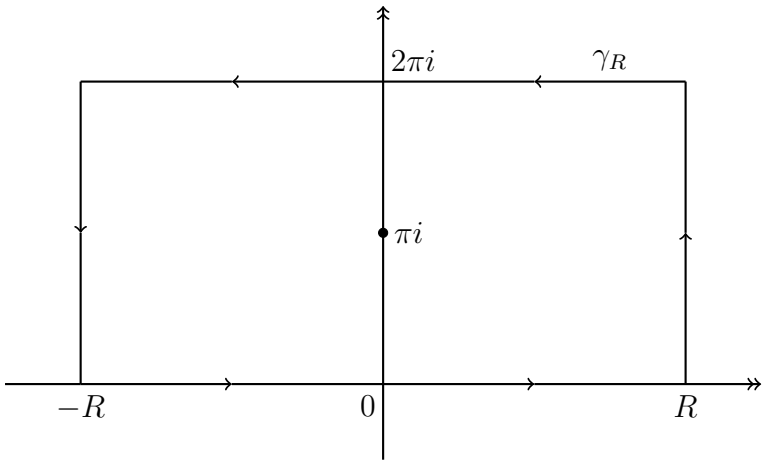


Abbildung 93: Der Weg γ_R

Ordnung und das Residuum ist gegeben durch

$$\mathrm{Res}_{\pi i}(g) = \lim_{z\to\pi i}(z-\pi i)g(z) = \lim_{z\to\pi i} e^{az}\frac{z-\pi i}{e^z - e^{\pi i}} = -e^{a\pi i}.$$

Der Residuensatz liefert

$$\int_{\gamma_R} g(z)\,dz = -2\pi i e^{-a\pi i}.$$

Daher ist

$$\begin{aligned}
-2\pi i e^{-a\pi i} &= \int_{\gamma_R} g(z)\,dz \\
&= \int_{-R}^{R} g(x)\,dx + \int_{R}^{R+2\pi i} g(z)\,dz + \int_{R+2\pi i}^{-R+2\pi i} g(z)\,dz + \int_{-R+2\pi i}^{-R} g(z)\,dz \\
&= \int_{-R}^{R} g(x)\,dx + i\int_0^{2\pi} \frac{e^{a(R+it)}}{1+e^{R+it}}\,dt - \int_{-R}^{R}\frac{e^{a(s+2\pi i)}}{1+e^{s+2\pi i}}\,ds \\
&\qquad -i\int_0^{2\pi}\frac{e^{a(-R+it)}}{1+e^{-R+it}}\,dt \\
&= (1-e^{2a\pi i})\int_{-R}^{R} f(x)\,dx + i\int_0^{2\pi}\frac{e^{a(R+it)}}{1+e^{R+it}}\,dt - i\int_0^{2\pi}\frac{e^{a(-R+it)}}{1+e^{-R+it}}\,dt.
\end{aligned}$$

Die letzten beiden Terme konvergieren wegen $a \in (0,1)$ mit $R \to \infty$ gegen Null, denn für $t \in [0, 2\pi]$ ist

$$\left|\frac{e^{a(R+it)}}{1+e^{R+it}}\right| \le \frac{e^{aR}}{e^R - 1}, \qquad \left|\frac{e^{a(-R+it)}}{1+e^{-R+it}}\right| \le \frac{e^{-aR}}{1-e^{-R}}.$$

Also ist

$$-2\pi i e^{-a\pi i} = (1 - e^{2a\pi i}) \int_{-\infty}^{\infty} \frac{e^{ax}}{1+e^x}\, dx$$

und folglich

$$\int_{-\infty}^{\infty} \frac{e^{ax}}{1+e^x}\, dx = -\frac{2\pi i e^{-a\pi i}}{1 - e^{2a\pi i}} = \frac{2\pi i}{e^{a\pi i} - e^{-a\pi i}} = \frac{\pi}{\sin \pi a}.$$

Das war zu zeigen. □

Beispiel: Gesucht sei eine konforme Abbildung f der oberen Halbebene $\mathbb{H}$ auf ein Dreieck $\mathbb{T}$ mit den Ecken w_1, w_2 und w_3 sowie den Innenwinkeln $\alpha_1\pi$, $\alpha_2\pi$ und $\alpha_3\pi$ mit $\alpha_3 := 1 - (\alpha_1 + \alpha_2)$, siehe Abbildung 94. Wir setzen voraus, dass $\alpha_1, \alpha_2 \in (0,1)$ und

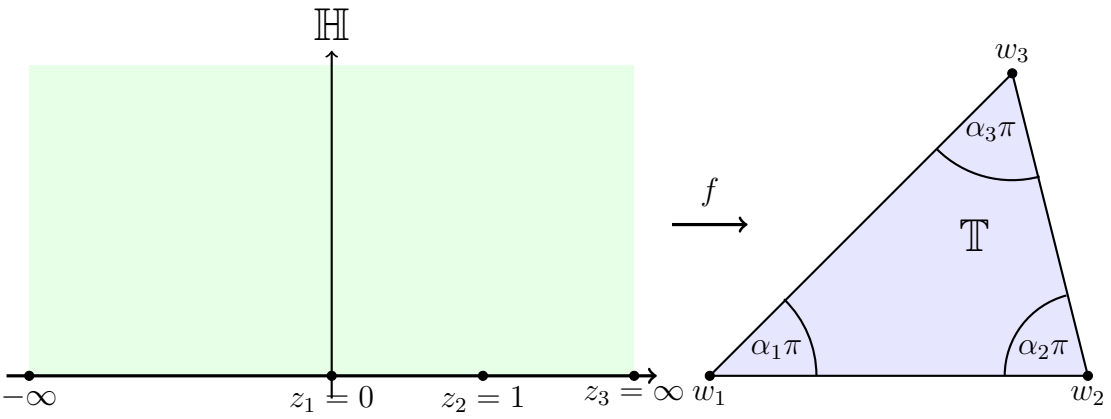


Abbildung 94: Konforme Abbildung von $\mathbb{H}$ auf ein Dreieck

auch $\alpha_1 + \alpha_2 \in (0,1)$. Durch w_1 und w_2 sowie die zugehörigen Innenwinkel $\alpha_1\pi$ und $\alpha_2\pi$ ist natürlich der Innenwinkel $\alpha_3\pi$ der dritten Ecke w_3 (wegen $\alpha_3 = 1-(\alpha_1+\alpha_2)$), aber auch die Ecke w_3 selbst bestimmt. Nach leichter Rechnung (siehe auch die Abbildung 94 rechts) erhält man

$$w_3 = w_1 + e^{i\alpha_1\pi} \frac{\sin \alpha_2\pi}{\sin(\alpha_1+\alpha_2)\pi}(w_2 - w_1).$$

Denn mit dem Ansatz

$$w_3 - w_1 = \lambda(w_2 - w_1)$$

mit $\lambda \in \mathbb{C}$ erhalten wir aus dem Sinussatz, dass

$$\frac{|\lambda|\,|w_2 - w_1|}{\sin \alpha_2\pi} = \frac{|w_3 - w_1|}{\sin \alpha_2\pi} = \frac{|w_2 - w_1|}{\sin \alpha_3\pi}$$

und damit

$$|\lambda| = \frac{\sin \alpha_2\pi}{\sin \alpha_3\pi} = \frac{\sin \alpha_1\pi}{\sin(\pi - (\alpha_1+\alpha_2)\pi)} = \frac{\sin \alpha_1\pi}{\sin(\alpha_1+\alpha_2)\pi}.$$

Weiter ist

$$\cos \alpha_1\pi \;=\; \cos \sphericalangle(w_2 - w_1, w_3 - w_1)$$

$$
\begin{aligned}
&= \frac{\frac{1}{2}[(\overline{w_2 - w_1})(w_3 - w_1) + (w_2 - w_1)(\overline{w_3 - w_1})]}{|w_1 - w_2|\,|w_3 - w_2|} \\
&= \frac{\mathrm{Re}(\lambda)}{|\lambda|},
\end{aligned}
$$

woraus insgesamt die behauptete Darstellung für die dritte Eckew_3 folgt.

Wegen der Bemerkung im Anschluss an die Sätze 10.13 und 10.14 existieren zu einer konformen Abbildung f von $\mathbb{H}$ auf $\mathbb{T}$ mit $f(0) = w_1$, $f(1) = w_2$ und $f(\infty) = w_3$ komplexe Konstanten a, c derart, dass

$$
f(z) = a + c \int_0^z \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta.
$$

Aus $f(0) = w_1$ folgt $a = w_1$, während c aus

$$
w_2 = f(1) = w_1 + c \int_0^1 \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta = w_1 - c e^{i\alpha_2 \pi} \underbrace{\int_0^1 \zeta^{\alpha_1 - 1}(1 - \zeta)^{\alpha_2 - 1}\, d\zeta}_{= B(\alpha_1, \alpha_2)}
$$

berechnet werden kann. Dann ist also

$$
c = e^{-i\alpha_2 \pi} \left(\frac{w_1 - w_2}{B(\alpha_1, \alpha_2} \right)
$$

und folglich

$$
f(z) = w_1 + e^{-i\alpha_2 \pi} \left(\frac{w_1 - w_2}{B(\alpha_1, \alpha_2)} \right) \int_0^z \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta.
$$

Wir wollen uns überlegen, dass das Bild der reellen Achse $\mathbb{R}$ bzw. des Randes der oberen Halbebene $\mathbb{H}$ unter f genau der Rand des Dreiecks $\mathbb{T}$ ist. Bei der Berechnung von $f(x)$ für $x \in \mathbb{R}$ unterscheiden wir drei Fälle, je nachdem ob $x \in [0, 1]$, $x \in (1, \infty)$ oder $x \in (-\infty, 0)$ ist.

- Für $x \in [0, 1]$ ist

$$
\begin{aligned}
f(x) &= w_1 + e^{-i\alpha_2 \pi} \left(\frac{w_1 - w_2}{B(\alpha_1, \alpha_2)} \right) \int_0^x \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta \\
&= w_1 + \left(\frac{w_2 - w_1}{B(\alpha_1, \alpha_2)} \right) \int_0^x \zeta^{\alpha_1 - 1}(1 - \zeta)^{\alpha_2 - 1}\, d\zeta \\
&= w_1 + \lambda(x)(w_2 - w_1)
\end{aligned}
$$

  mit

$$
\lambda(x) := \frac{1}{B(\alpha_1, \alpha_2)} \int_0^x \zeta^{\alpha_1 - 1}(1 - \zeta)^{\alpha_2 - 1}\, d\zeta.
$$

  Offenbar ist $\lambda(\cdot)$ auf $(0, 1]$ positiv, monoton wachsend und $\lambda(0) = 0$, $\lambda(1) = 1$. Insbesondere ist $f(x)$ für $x \in [0, 1]$ eine Konvexkombination von $w_1 = f(0)$ und $w_2 = f(1)$ und damit ein Punkt auf der Strecke von w_1 nach w_2.

- Für $x \in (1,\infty)$ und $\alpha_3 := 1-(\alpha_1+\alpha_2)$ ist

$$\begin{aligned}
f(x) &= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\int_0^x \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta \\
&= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\left[\int_0^1 \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta\right. \\
&\qquad \left. + \int_1^x \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta\right] \\
&= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\left[e^{i\pi(\alpha_2-1)}B(\alpha_1,\alpha_2) + \int_1^x \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta\right] \\
&= w_2 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\int_1^x \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta \\
&= w_2 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\int_{1/x}^1 \left(\frac{1}{\eta}\right)^{\alpha_1-1}\left(\frac{1}{\eta}-1\right)^{\alpha_2-1}\frac{1}{\eta^2}\,d\eta \\
&\qquad \text{(Substitution } \zeta = 1/\eta) \\
&= w_2 + e^{-i\alpha_2\pi}\left(\frac{w_1-w_2}{B(\alpha_1,\alpha_2)}\right)\int_{1/x}^1 \eta^{\alpha_3-1}(1-\eta)^{\alpha_2-1}\,d\eta \\
&= w_2 + e^{-i\alpha_2\pi}\mu(x)(w_1-w_2)
\end{aligned}$$

  mit

$$\mu(x) := \frac{1}{B(\alpha_1,\alpha_2)}\int_{1/x}^1 \eta^{\alpha_3-1}(1-\eta)^{\alpha_2-1}\,d\eta.$$

  Offenbar ist $\mu(\cdot)$ auf $(1,\infty)$ positiv und monoton wachsend, $\mu(1)=0$ und

$$\begin{aligned}
\mu(\infty) &= \lim_{x\to\infty}\mu(x) \\
&= \frac{B(\alpha_3,\alpha_2)}{B(\alpha_1,\alpha_2)} \\
&= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3+\alpha_2)} \\
&\qquad \text{(Lemma 10.15)} \\
&= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(1-(\alpha_1+\alpha_2))}{\Gamma(\alpha_1)\Gamma(1-\alpha_1)} \\
&= \frac{\sin\alpha_1\pi}{\sin(\alpha_1+\alpha_2)\pi} \\
&\qquad \text{(Lemma 10.16).}
\end{aligned}$$

  Daher ist

$$\begin{aligned}
f(\infty) &= w_2 + e^{-i\alpha_2\pi}\frac{\sin\alpha_1\pi}{\sin(\alpha_1+\alpha_2)\pi}(w_1-w_2) \\
&= w_1 + e^{i\alpha_1\pi}\frac{\sin\alpha_2\pi}{\sin(\alpha_1+\alpha_2)\pi} \\
&= w_3.
\end{aligned}$$

- Für $x \in (-\infty, 0)$ ist

$$\begin{aligned}
f(x) &= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1 - w_2}{B(\alpha_1,\alpha_2)}\right)\int_0^x \zeta^{\alpha_1-1}(\zeta-1)^{\alpha_2-1}\,d\zeta \\
&= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1 - w_2}{B(\alpha_1,\alpha_2)}\right)e^{i\pi(\alpha_1+\alpha_2-2)}\int_0^x (-\zeta)^{\alpha_1-1}(1-\zeta)^{\alpha_2-1}\,d\zeta \\
&= w_1 + e^{i\alpha_1\pi}\left(\frac{w_1 - w_2}{B(\alpha_1,\alpha_2)}\right)\int_0^x (-\zeta)^{\alpha_1-1}(1-\zeta)^{\alpha_2-1}\,d\zeta \\
&= w_1 + e^{i\alpha_1\pi}\left(\frac{w_1 - w_2}{B(\alpha_1,\alpha_2)}\right)\int_1^{1/(1-x)} \left(\frac{1}{\eta}-1\right)^{\alpha_1-1}\left(\frac{1}{\eta}\right)^{\alpha_2-1}\frac{1}{\eta^2}\,d\eta \\
&\qquad \text{(Substitution } \zeta = 1 - 1/\eta) \\
&= w_1 - e^{i\alpha_1\pi}\left(\frac{w_1 - w_2}{B(\alpha_1,\alpha_2)}\right)\int_{1/(1-x)}^1 \eta^{\alpha_3-1}(1-\eta)^{\alpha_1-1}\,d\eta \\
&\qquad \text{(wegen } \alpha_3 = 1-(\alpha_1+\alpha_2)) \\
&= w_1 + e^{i\alpha_1\pi}\nu(x)(w_2 - w_1)
\end{aligned}$$

  mit

$$\nu(x) := \frac{1}{B(\alpha_1,\alpha_2)}\int_{1/(1-x)}^1 \eta^{\alpha_3-1}(1-\eta)^{\alpha_1-1}\,d\eta.$$

  Offenbar ist $\nu(\cdot)$ auf $(-\infty, 0)$ positiv, monoton fallend, $\nu(0) = 0$ und

$$\begin{aligned}
\nu(-\infty) &= \lim_{x\to-\infty}\nu(x) \\
&= \frac{B(\alpha_3,\alpha_1)}{B(\alpha_1,\alpha_2)} \\
&= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_1)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3+\alpha_1)} \\
&\qquad \text{(Lemma 10.15)} \\
&= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(1-(\alpha_1+\alpha_2))}{\Gamma(\alpha_2)\Gamma(1-\alpha_2)} \\
&= \frac{\sin\alpha_2\pi}{\sin(\alpha_1+\alpha_2)\pi} \\
&\qquad \text{(Lemma 10.16).}
\end{aligned}$$

  Daher ist

$$f(-\infty) = w_1 + e^{i\alpha_1\pi}\frac{\sin\alpha_2\pi}{\sin(\alpha_1+\alpha_2)\pi}(w_2 - w_1) = w_3 = f(\infty).$$

Wandert also ein Punkt z von $-\infty$ nach 0, so läuft $f(z)$ von w_3 nach w_1. Läuft z von 0 nach 1, so wandert $f(z)$ von w_1 nach w_2. Schließlich ist das Bild von $(0,\infty)$ genau die Dreiecksseite von w_2 nach w_3. Dies veranschaulichen wir uns in Abbildung 95. Wir fassen das erhaltene Ergebnis zusammen:

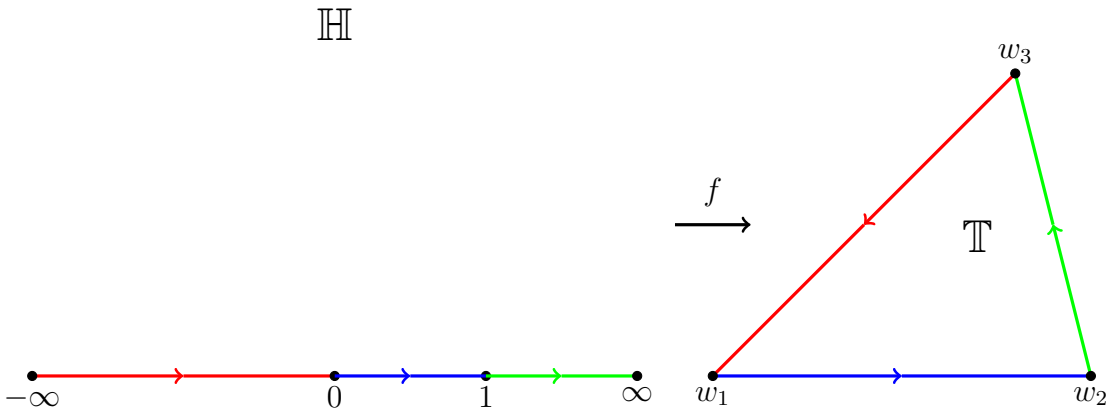


Abbildung 95: Die Abbildung des Randes von $\mathbb{H}$ auf den Rand des Dreiecks $\mathbb{T}$

- Gegeben sei ein Dreieck $\mathbb{T} \subset \mathbb{C}$ mit den Ecken w_1, w_2, w_3 (im mathematisch positiven Sinne angeordnet) mit zugehörigen Innenwinkeln $\alpha_1\pi, \alpha_2\pi, \alpha_3\pi$, wobei $\alpha_3 := 1 - (\alpha_1 + \alpha_2)$. Durch

$$f(z) := w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1 - w_2}{B(\alpha_1, \alpha_2)}\right)\int_0^z \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta$$

ist eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf das Dreieck $\mathbb{T}$ mit $f(0) = w_1$, $f(1) = w_2$ und $f(\infty) = w_3$ gegeben. Hierbei ist

$$B(\alpha_1, \alpha_2) = \int_0^1 \zeta^{\alpha_1 - 1}(1 - \zeta)^{\alpha_2 - 1}\, d\zeta$$

die vollständige Beta-Funktion. Ferner ist

$$\begin{aligned} f(\infty) &= w_1 + e^{i\alpha_1\pi}\frac{\sin\alpha_2\pi}{\sin(\alpha_1 + \alpha_2)\pi}(w_2 - w_1) \\ &= w_3 \\ &= w_2 + e^{-i\alpha_2\pi}\frac{\sin\alpha_1\pi}{\sin(\alpha_1 + \alpha_2)\pi}(w_1 - w_2) \\ &= f(-\infty). \end{aligned}$$

Um dieses Ergebnis zu überprüfen, betrachten wir das in Abbildung 96 angegebene Dreieck mit Ecken $w_1 = 0$, $w_2 = 1$, $w_3 = i$ und zugehörigen Innenwinkeln $\alpha_1\pi$, $\alpha_2\pi$ und $\alpha_3\pi$ mit $(\alpha_1, \alpha_2, \alpha_3) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. Als konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf das Dreieck mit den angegebenen Ecken und Innenwinkeln erhalten wir

$$\begin{aligned} f(z) &= w_1 + e^{-i\alpha_2\pi}\left(\frac{w_1 - w_2}{B(\alpha_1, \alpha_2)}\right)\int_0^z \zeta^{\alpha_1 - 1}(\zeta - 1)^{\alpha_2 - 1}\, d\zeta \\ &= -\frac{e^{-i\pi/4}}{B(\frac{1}{2}, \frac{1}{4})}\int_0^z \zeta^{-1/2}(\zeta - 1)^{-3/4}\, d\zeta \\ &= -\frac{1 - i}{\sqrt{2}B(\frac{1}{2}, \frac{1}{4})}\int_0^z \zeta^{-1/2}(\zeta - 1)^{-3/4}\, d\zeta. \end{aligned}$$

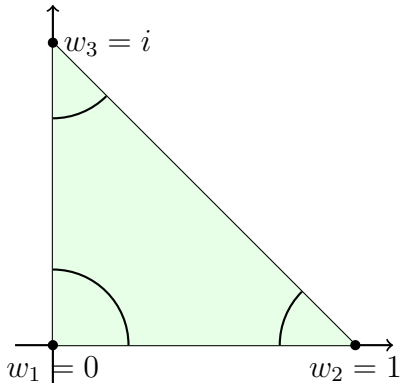


Abbildung 96: Ein Testdreieck

Mit MATLAB oder GNU Octave erhält man, dass $B(\frac{1}{2}, \frac{1}{4}) \approx 5.244115108584238$, wobei die Funktion `beta` benutzt wurde. □

Beispiel: Gesucht sei (siehe M. Stein, R. Shakarchi (2003, S. 233 ff.)) eine konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf ein Rechteck mit den Ecken $w_1 := -K + iK'$, $w_2 := -K$, $w_3 := K$ und $w_4 := K + iK'$, wobei $K > 0$ und $K' > 0$ vorgegeben sind, siehe Abbildung 97. Wir benutzen die Darstellung bei E. M. Stein, R. Shakarchi

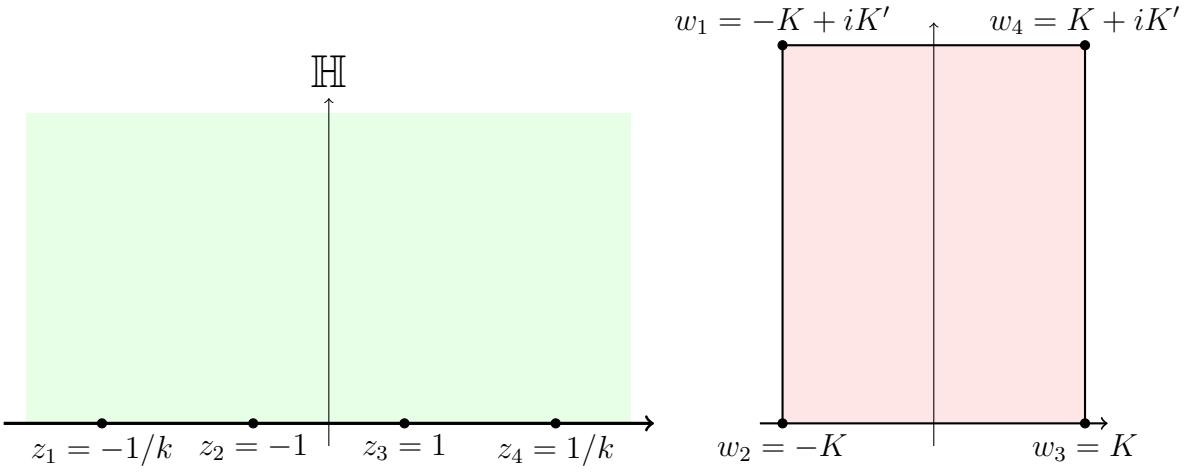


Abbildung 97: Konforme Abbildung von $\mathbb{H}$ auf ein Rechteck

(2003, S. 233 ff.). Das gegebene Rechteck hat also die Seitenlängen $2K$ und K'. Die Innenwinkel sind $\alpha_i \pi$ mit $\alpha_i = \frac{1}{2}$, $i = 1, \ldots, 4$. Sei f eine konforme Abbildung von $\mathbb{H}$ auf dieses Rechteck, wobei wir annehmen, dass f schiefsymmetrisch zur imaginären Achse ist bzw. $f(x + iy) = -f(-x + iy)$ für alle $z = x + iy \in \mathbb{H}$ und damit $f(0) = 0$ gilt. Als prevertices wählen wir aus Symmetriegründen $z_1 = -1/k$, $z_2 = -1$, $z_3 = 1$ und $z_4 = 1/k$ mit (noch unbekanntem) $k \in (0, 1)$. Dann hat f wegen Satz 10.13 eine

Darstellung

$$f(z) = c\int_0^z \frac{d\zeta}{\sqrt{(\zeta^2 - 1/k^2)(\zeta^2 - 1)}} = ck\int_0^z \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}$$

mit einer Konstanten $c \in \mathbb{C}$. Die unbekannten Parameter c und k sind zu bestimmen aus den Gleichungen

$$(*) \qquad f(1) = ck\int_0^1 \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} = K = w_3$$

und

$$(**) \qquad f(1/k) = ck\int_0^{1/k} \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} = K + iK' = w_4.$$

Wegen

$$\begin{aligned} ck\int_0^{1/k} \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} &= \underbrace{ck\int_0^1 \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}}_{= K \text{ wegen } (*)} \\ &\quad + ck\int_1^{1/k} \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} \\ &= K + ick\int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}} \end{aligned}$$

ist $(**)$ gleichbedeutend mit

$$ck\int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}} = K'.$$

Der Parameter k wird bestimmt aus dem "Seitenverhältnis" K/K' mittels der Gleichung

$$(***) \qquad \int_0^1 \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} \Big/ \int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}} = \frac{K}{K'},$$

anschließend kann die Konstante c aus $(*)$ oder $(**)$ berechnet werden. Das im Zähler von $(***)$ auftretende Integral

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

(bei der Umrechnung wird die Substitution $\zeta = \sin\theta$ benutzt) nennt man *elliptisches Integral erster Art zum Modul k*. Das im Nenner von $(***)$ auftretende Integral

$$K'(k) := \int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}}$$

kann als ein elliptisches Integral erster Art zum Modul $\tilde{k} := \sqrt{1-k^2}$ entlarvt werden. Denn mittels der Variablentransformation

$$\zeta = (1 - \tilde{k}^2\eta^2)^{-1/2},$$

was

$$\sqrt{\zeta^2 - 1} = \frac{\tilde{k}\eta}{\sqrt{1-\tilde{k}^2\eta^2}}, \qquad \sqrt{1-k^2\zeta^2} = \frac{\tilde{k}\sqrt{1-\eta^2}}{\sqrt{1-\tilde{k}^2\eta^2}}$$

sowie

$$d\zeta = (1 - \tilde{k}^2\eta^2)^{-3/2}\tilde{k}^2\eta d\eta$$

impliziert, erhält man

$$K'(k) = \int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}} = \int_0^1 \frac{d\eta}{\sqrt{(1-\eta^2)(1-\tilde{k}^2\eta^2)}} = K(\underbrace{\sqrt{1-k^2}}_{=\tilde{k}}).$$

Die Gleichung $(***)$ kann also umgeschrieben werden zu

$$(***) \qquad g(k) := \frac{K(k)}{K(\sqrt{1-k^2})} - \frac{K}{K'} = 0.$$

In Abbildung 98 geben wir plots von $K(\cdot)$, $K'(\cdot)$ und $K(\cdot)/K'(\cdot)$ über dem Intervall $[0,1]$

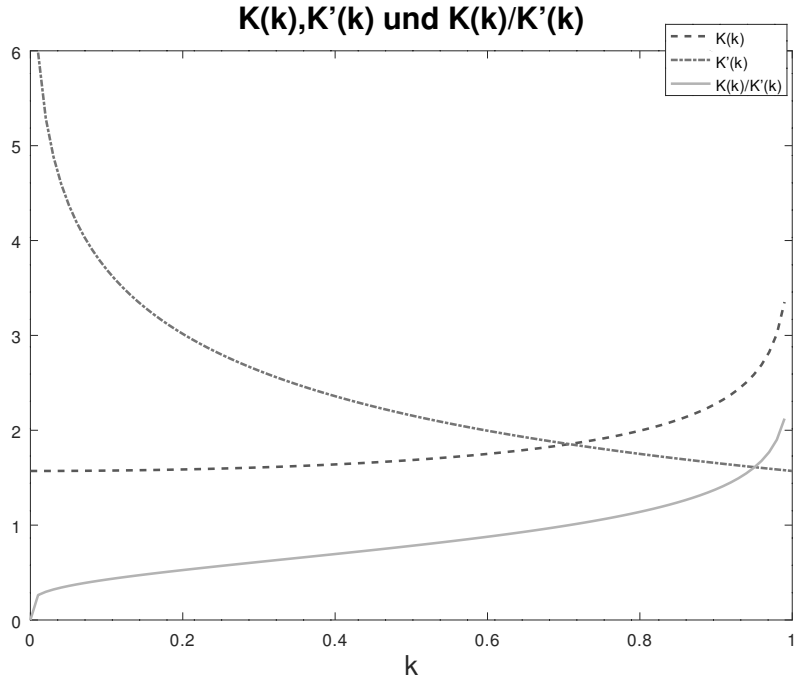


Abbildung 98: $K(k)$, $K'(k)$ und $K(k)/K'(k)$ auf dem Intervall $[0,1]$

an. Man erkennt, dass $K(\cdot)$ monoton wachsend, $K'(\cdot)$ monoton fallend und $K(\cdot)/K'(\cdot)$

monoton wachsend sind. Diese Aussagen können natürlich auch leicht bewiesen werden. Die Gleichung $(***)$ besitzt bei gegebenen positiven K, K' offenbar genau eine Lösung $k \in (0,1)$. Diese kann z. B. mit dem Sekantenverfahren gewonnen werden. Bei diesem werden zwei Startwerte k_0, k_1 vorgegeben und nach der Vorschrift

$$k_{n+1} := k_n - \left(\frac{k_n - k_{n-1}}{g(k_n) - g(k_{n-1})}\right) g(k_n)$$

iteriert. Als Beispiel nehmen wir $K/K' = 0.8$. Mit den Startwerten $k_0 = 0.5$ und $k_1 = 0.8$ erhalten wir die Werte

n	k_n	$g(k_n)$
0	0.5	$-1.829903865194438e-02$
1	0.8	$3.396821039838378e-01$
2	0.5153351977010943	$-4.377265555675303e-03$
3	0.5189568210147795	$-1.056122879871002e-03$
4	0.5201084966930702	$2.791342399999230e-06$
5	0.5201054608275730	$-1.789606463020732e-09$
6	0.5201054627727029	$-2.997602166487923e-15$

Bei der Rechnung haben wir die MATLAB- bzw. GNU Octave-Funktion `ellipke` zur Berechnung vollständiger elliptischer Integrale benutzt. Hierbei ist zu beachten, dass $K(k)$ durch `ellipke(m)` mit $m := k^2$ berechnet wird.

Wir wollen uns überlegen, dass die durch

$$f(z) := ck \int_0^z \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}$$

definierte Abbildung f, wobei $k \in (0,1)$ aus der obigen Gleichung $(***)$ und $c \in \mathbb{R}$ durch

$$c = \frac{K}{kK(k)} = \frac{K'}{kK'(k)}$$

bestimmt sind, den Rand der oberen Halbebene $\mathbb{H}$ (einschließlich des Punktes ∞) auf den Rand des Rechtecks mit den Ecken $w_1 := -K + iK'$, $w_2 := -K$, $w_3 := K$ und $w_4 := K + iK'$ abbildet. Wir definieren

$$A := (-\infty, -1/k], \quad B := [-1/k, -1], \quad C := [-1, 1]$$

sowie

$$D := [1, 1/k], \quad E := [1/k, \infty),$$

siehe Abbildung 99 links. Es ist

$$f(-1) = ck \int_0^{-1} \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} = -ckK(k) = -K,$$

wie man nach der Variablentransformation $\zeta = -\eta$ nach Wahl von c und k erhält und natürlich $f(1) = K$. Da $f'(x) > 0$ für $x \in (0,1)$, wird das Segment $C = [-1,1]$ des

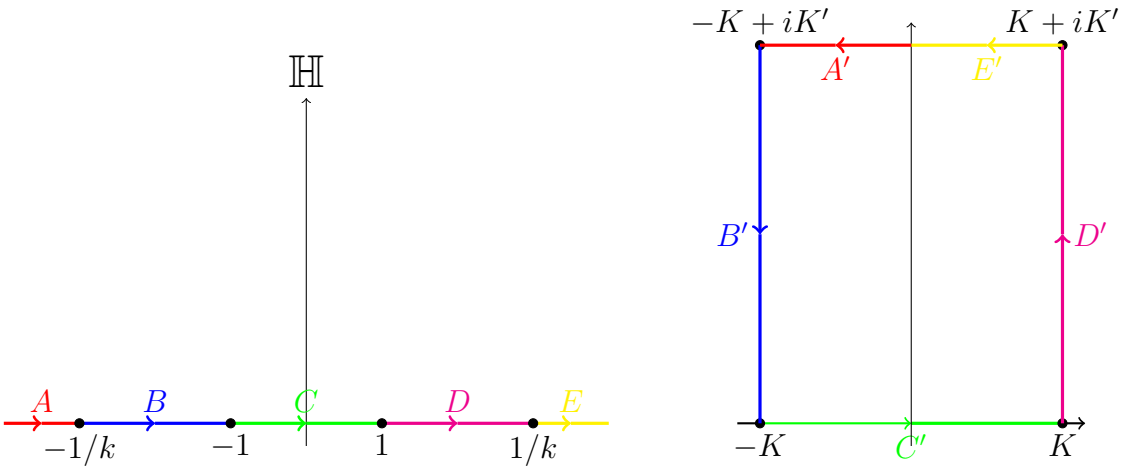


Abbildung 99: Abbildung des Randes von $\mathbb{H}$ auf den Rand des Rechtecks

Randes von $\mathbb{H}$ auf die Seite $C' = [-K, K]$ des Rechtecks abgebildet. Für $x \in (1, 1/k)$ ist

$$\begin{aligned} f(x) &= \underbrace{ck \int_0^1 \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}}_{=K} + ck \int_1^x \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} \\ &= K + ick \int_1^x \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}}. \end{aligned}$$

Da außerdem

$$ck \int_1^{1/k} \frac{d\zeta}{\sqrt{(\zeta^2-1)(1-k^2\zeta^2)}} = K',$$

bildet f das Segment $D = [1, 1/k]$ auf die Seite $D' = [K, K+iK']$ des Rechtecks ab. Entsprechend wird $B = [-1/k, -1]$ auf die Seite $B' = [-K+iK', -K]$ des Rechtecks abgebildet. Für $x \in [1/k, \infty)$ ist

$$\begin{aligned} f(x) &= ck \int_0^x \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} \\ &= \underbrace{ck \int_0^1 \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}}_{=K} + \underbrace{ck \int_1^{1/k} \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}}_{=iK'} \\ &\qquad + ck \int_{1/k}^x \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}} \\ &= K + iK' - ck \int_{1/k}^x \frac{d\zeta}{\sqrt{(\zeta^2-1)(k^2\zeta^2-1)}}, \end{aligned}$$

wobei wir ausgenutzt haben, dass

$$\sqrt{(1-\zeta^2)(1-k^2\zeta^2)} = i\sqrt{(\zeta^2-1)}i\sqrt{(k^2\zeta^2-1)} = -\sqrt{(\zeta^2-1)(k^2\zeta^2-1)}$$

für $\zeta > 1/k$. Nach der Variablentransformation $\zeta = 1/(k\eta)$ erkennt man, dass

$$\int_{1/k}^{\infty} \frac{d\zeta}{\sqrt{(\zeta^2-1)(k^2\zeta^2-1)}} = \int_0^1 \frac{d\eta}{\sqrt{(1-\eta^2)(1-k^2\eta^2)}}.$$

Für $x \in E = [1/k, \infty)$ ist also $f(x) \in E' = [K + iK', iK')$ und $f(\infty) = iK'$. Entsprechend wird das Segment $A = (-\infty, -1/k]$ auf das Segment $A' = (iK', -K + iK']$ abgebildet. □

10.5.3 Der Satz von Osgood-Carathéodory

Unser Ziel besteht darin, den folgenden Satz zu beweisen (siehe z. B. E. M. Stein, R. Shakarchi (2003, S. 238)).

Satz 10.17 (Osgood-Carathéodory) *Sei $P \subset \mathbb{C}$ ein beschränkter Polygonbereich. Dann lässt sich eine konforme Abbildung der offenen Einheitskreisscheibe $B(0;1)$ auf P zu einer stetigen, bijektiven Abbildung von $B[0;1] := \operatorname{cl} B(0;1)$ auf $\operatorname{cl} P$ fortsetzen.*

Bemerkungen: Sei $P \subset \mathbb{C}$ ein beschränkter Polygonbereich und f eine konforme Abbildung der (offenen) oberen Halbebene $\mathbb{H}$ auf P. *Angenommen*, Satz 10.17 sei schon bewiesen. Wir wollen uns überlegen, dass f zu einer stetigen, bijektiven Abbildung von $\operatorname{cl} H$ (diese Menge enthält als Rand die reelle Achse $\mathbb{R}$ und den Punkt ∞) auf $\operatorname{cl} P$ fortgesetzt werden kann. Man definiere $G\colon B(0;1) \longrightarrow \mathbb{H}$ durch

$$G(z) := i\frac{1-z}{1+z}.$$

Dann ist G eine konforme Abbildung von $B(0;1)$ auf $\mathbb{H}$, welche sich zu einer stetigen, bijektiven Abbildung von $B[0;1]$ auf $\operatorname{cl}\mathbb{H}$ fortsetzen lässt, wobei der Punkt -1 des Randes von $B(0;1)$ dem Punkt ∞ auf dem Rand von $\mathbb{H}$ entspricht. Da $f_G := f \circ G$ eine konforme Abbildung von $B(0;1)$ auf den Polygonbereich P ist, lässt sich f_G zu einer stetigen, bijektiven Abbildung $F_G\colon B[0;1] \longrightarrow \operatorname{cl} P$ fortsetzen. Anschließend definieren wir $F := F_G \circ G^{-1}$. Dann ist F eine stetige, bijektive Abbildung von $\operatorname{cl}\mathbb{H}$ auf $\operatorname{cl} P$, welche eine Erweiterung (von $\mathbb{H}$ auf $\operatorname{cl}\mathbb{H}$) der konformen Abbildung f ist. Als Konsequenz von Satz 10.17 haben wir damit nachgewiesen:

- *Sei $P \subset \mathbb{C}$ ein beschränkter Polygonbereich. Dann lässt sich eine konforme Abbildung der (offenen) oberen Halbebene $\mathbb{H}$ auf P zu einer stetigen, bijektiven Abbildung von $\operatorname{cl}\mathbb{H} = \mathbb{H} \cup (\mathbb{R} \cup \{\infty\})$ auf $\operatorname{cl} P$ fortsetzen.*

Erwähnt sei lediglich, dass Satz 10.17 eigentlich nicht *der* Satz von Osgood-Caratheodory ist, sondern nur ein (für uns geeigneter) Spezialfall. Allgemeiner gilt nämlich: Sei f eine konforme Abbildung der offenen Einheitskreisscheibe $B(0;1)$ auf ein Gebiet R, welches von einer Jordankurve J berandet ist. Dann lässt sich f zu einer stetigen, bijektiven Abbildung von $B[0;1]$ auf $\operatorname{cl} R$ fortsetzen. Wir verweisen lediglich auf E. J. McShane (1937) und W. P. Novinger (1975). □

Beweis von Satz 10.17: Zunächst zeigen wir:

(1) Sei z_0 ein Punkt auf dem Rand von $B(0;1)$. Dann existiert $\lim_{z\to z_0} f(z) \in \operatorname{cl} P$. Also kann die Abbildung f zu einer Abbildung von $B[0;1]$ nach $\operatorname{cl} P$ erweitert werden. Diese Abbildung nennen wir ebenfalls f.

Hierzu genügt es offenbar zu zeigen: Sind $\{z_k\}, \{z_k'\} \subset B(0;1)$ zwei Folgen, die gegen z_0 konvergieren, und für die $\zeta := \lim_{k\to\infty} f(z_k)$ und $\zeta' := \lim_{k\to\infty} f(z_k')$ existieren, so ist $\zeta = \zeta'$. Da $\{f(z_k)\} \subset P$ und $\{f(z_k')\} \subset P$, sind ζ und ζ' aus $\operatorname{cl} P$. Da f eine konforme Abbildung ist, liegen ζ und ζ' auf dem Rand ∂P von P, also den P berandenden Polygonzug. Im Widerspruch zur Behauptung nehmen wir an, es sei $\zeta \neq \zeta'$. Seien B und B' zwei offene disjunkte Kreisscheiben mit dem Mittelpunkt ζ bzw. ζ', welche eine positiven Abstand $d > 0$ voneinander haben. Da die Folgen $\{f(z_k)\}$ bzw. $\{f(z_k')\}$ gegen ζ bzw. ζ' konvergieren, den Mittelpunkten der Kreisscheiben B bzw. B', ist $f(z_k) \in B$ und $f(z_k') \in B'$ und daher $|f(z_k) - f(z_k')| > d$ für alle hinreichend großen k. Bei gegebenem $r > 0$ bezeichnen wir mit $C_r(z_0) := \{z \in \mathbb{C} : |z - z_0| = r\}$ den Kreis um z_0 mit dem Radius r. Dann existieren zu jedem hinreichend kleinen $r > 0$ Punkte

$$z(r), z'(r) \in C_r(z_0) \cap B(0;1) \quad \text{mit} \quad f(z(r)) \in B,\ f(z'(r)) \in B'$$

und damit

$$(*) \qquad \rho(r) := |f(z(r)) - f(z'(r))| > d.$$

Im Widerspruch zu $(*)$ zeigen wir nun, dass eine Folge $\{r_k\}$ positiver Zahlen mit $\lim_{k\to\infty} \rho(r_k) = 0$ existiert. Die Annahme $\zeta \neq \zeta'$ ist dann zu einem Widerspruch geführt und damit bewiesen, dass $f(z)$ gegen einen Grenzwert konvergiert, wenn sich z aus der offenen Kreisscheibe $B(0;1)$ dem Randpunkt z_0 annähert. Wenn keine Folge $\{r_k\}$ positiver Zahlen mit $\lim_{k\to\infty} \rho(r_k) = 0$ existiert, so gibt es ein $c > 0$ und ein hinreichend kleines $R \in (0,1)$ derart, dass $c \leq \rho(r)$ für alle $r \in (0, R]$. Mit dem $z(r)$ und $z'(r)$ innerhalb von $B(0;1)$ verbindenden Bogen

$$\gamma(r) := \{z_0 + re^{i\theta} : \theta \in [\theta_1(r), \theta_2(r)]\}$$

mit $0 \leq \theta_1(r) < \theta_2(r)$ ist

$$\begin{aligned}
c &\leq \rho(r) \\
&= |f(z(r)) - f(z'(r))| \\
&= \left| \int_{\gamma(r)} f'(\eta) d\eta \right| \\
&\leq \int_{\theta_1(r)}^{\theta_2(r)} |f'(z_0 + re^{i\theta})| r \, d\theta \\
&\leq \left(\int_{\theta_1(r)}^{\theta_2(r)} |f'(z_0 + re^{i\theta})|^2 r \, d\theta \right)^{1/2} \left(\int_{\theta_1(r)}^{\theta_2(r)} r \, d\theta \right)^{1/2} \\
&\qquad \text{(Cauchy-Schwarz)} \\
&\leq \left(\int_{\theta_1(r)}^{\theta_2(r)} |f'(z_0 + re^{i\theta})|^2 r \, d\theta \right)^{1/2} \sqrt{2\pi}\sqrt{r}.
\end{aligned}$$

Quadrieren und Division mit $r > 0$ liefert

$$\frac{c^2}{r} \leq 2\pi \int_{\theta_1(r)}^{\theta_2(r)} |f'(z_0 + re^{i\theta})|^2 r \, d\theta.$$

Mit beliebigem $\epsilon \in (0, R)$ erhält man durch Integration über das Intervall $[\epsilon, R] \subset (0, R]$, dass

$$\begin{aligned} c^2(\log R - \log \epsilon) &= c^2 \int_\epsilon^R \frac{dr}{r} \\ &\leq 2\pi \int_0^R \int_{\theta_1(r)}^{\theta_2(r)} |f'(z_0 + re^{i\theta})|^2 r \, d\theta dr \\ &\leq 2\pi \iint_{B(0;1)} |f'(x+iy)|^2 \, dxdy \\ &= 2\pi \iint_{f(B(0;1))} dxdy \\ &= 2\pi \operatorname{area}(P). \end{aligned}$$

Hierbei erhält man die Gleichung

$$\iint_{B(0;1)} |f'(x+iy)|^2 \, dxdy = \iint_{f(B(0;1))} dxdy$$

als Konsequenz der Transformationsformel für Integrale. Ist nämlich

$$f(x+iy) = u(x,y) + iv(x,y)$$

und $\Phi\colon B(0;1) \longrightarrow \mathbb{R}^2$ durch $\Phi(x,y) := (u(x,y), v(x,y))$ erklärt, so ist

$$\operatorname{area}(\Phi(B(0;1)) = \iint_{B(0;1)} |\det \Phi'(x,y)| \, dxdy.$$

Mit Hilfe der Cauchy-Riemannschen Gleichungen

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

erhalten wir

$$\begin{aligned} \det \Phi'(x,y) &= \det \begin{pmatrix} \frac{\partial u}{\partial x}(x,y) & \frac{\partial u}{\partial y}(x,y) \\ \frac{\partial v}{\partial x}(x,y) & \frac{\partial v}{\partial y}(x,y) \end{pmatrix} \\ &= \frac{\partial u}{\partial x}(x,y)\frac{\partial v}{\partial y}(x,y) - \frac{\partial u}{\partial y}(x,y)\frac{\partial v}{\partial x}(x,y) \\ &= \left(\frac{\partial u}{\partial x}(x,y)\right)^2 + \left(\frac{\partial v}{\partial x}(x,y)\right)^2 \\ &= |f'(x+iy)|^2 \end{aligned}$$

Da P als beschränkt vorausgesetzt ist, ist $\operatorname{area}(P) < \infty$. Mit $\epsilon \to 0+$ erhalten wir einen Widerspruch. Damit ist die Zwischenbehauptung (1) bewiesen.

(2) Die nach (1) existierende Fortsetzung $f\colon B[0;1] \longrightarrow \operatorname{cl} P$ der konformen Abbildung f von $B(0;1)$ auf P ist stetig.

Denn: Nach (1) ist die Fortsetzung der Abbildung f von $B(0;1)$ auf $B[0;1]$ durch

$$f(z_0) := \lim_{z \to z_0} f(z)$$

wohldefiniert. Um die Stetigkeit von f auf $B[0;1]$ zu zeigen, geben wir uns ein z_0 aus dem Rand von $B(0;1)$ vor. Nach Definition von $f(z_0)$ als Limes von $f(z)$, wenn $z \in B(0;1)$ den Randpunkt z_0 approximiert, existiert zu vorgegebenem $\epsilon > 0$ ein $\delta > 0$ mit

$$z \in B(0;1),\ |z - z_0| \leq \delta \Longrightarrow |f(z) - f(z_0)| < \epsilon.$$

Ist dagegen $z \in \partial B(0;1)$, also z ein Randpunkt von $B(0;1)$, so können wir ein $w \in B(0;1)$ mit $|w - z_0| \leq \delta$ und $|f(z) - f(w)| < \epsilon$ wählen, wieder nach Definition von $f(z)$. Dann ist aber

$$|f(z) - f(z_0)| \leq \underbrace{|f(z) - f(w)|}_{<\epsilon} + \underbrace{|f(w) - f(z_0)|}_{<\epsilon} < 2\epsilon,$$

womit auch (2) bewiesen ist.

(3) Die nach (1) und (2) existierende stetige Fortsetzung $f\colon B[0;1] \longrightarrow \operatorname{cl} P$ der konformen Abbildung f von $B(0;1)$ auf P ist eine bijektive Abbildung von $B[0;1]$ auf $\operatorname{cl} P$.

Denn: Die Abbildung $g := f^{-1}\colon P \longrightarrow B(0;1)$ kann wie in Beweisteil (1), (2) zu einer stetigen Abbildung von $\operatorname{cl} P$ nach $B[0;1]$ fortgesetzt werden. Denn im Beweisteil (1) haben wir als geometrische Eigenschaft der Einheitskreisscheibe $B[0;1]$ nur ausgenutzt, dass mit einem Randelement z_0 und einem kleinen Kreis $C_r(z_0)$ mit einem Radius $r > 0$ um z_0 die Menge $C_r(z_0)$ aus einem Kreisbogen besteht. Diese Eigenschaft hat aber auch der Polygonbereich P. Es bleibt jetzt noch zu zeigen, dass die Fortsetzungen von f und g Inverse voneinander sind. Ist $z \in \partial B(0;1)$ und $\{z_k\} \subset B(0;1)$ eine Folge, die gegen z konvergiert, so ist $g(f(z_k)) = z_k$ für alle k. Wegen der Stetigkeit von f und g ist $g(f(z)) = z$, insgesamt also $g(f(z)) = z$ für alle $z \in B[0;1]$. Entsprechend kann $f(g(w)) = w$ für alle $w \in \operatorname{cl} P$ bewiesen werden. Insgesamt ist Satz 10.17 bewiesen. □

10.5.4 Beweis der Schwarz-Christoffel-Sätze

Unser Ziel besteht darin, die Sätze 10.13 und 10.14, in denen die Struktur einer konformen Abbildung der oberen Halbebene $\mathbb{H}$ auf einen Polygonbereich angegeben wird, zu beweisen. Da Satz 10.14 aus Satz 10.13 folgt, genügt es Satz 10.13 zu beweisen. Um Satz 10.17 anwenden zu können, setzen wir voraus, dass der Polygonbereich P beschränkt ist und formulieren Satz 10.13 neu. Beim Beweis folgen wir der sehr schönen Darstellung bei E. M. Stein, R. Shakarchi (2003, S. 241 ff.).

Satz 10.18 *Sei $P \subset \mathbb{C}$ ein beschränkter, offener, einfach zusammenhängender Polygonbereich, dessen Rand ein Polygon mit Ecken $w_1, \ldots, w_n$, die entgegen dem Uhrzeigersinn angeordnet sind, und zugehörigen Innenwinkeln $\alpha_1\pi, \ldots, \alpha_n\pi$ ist. Sei f eine*

konforme Abbildung der oberen Halbebene $\mathbb{H}$ auf P. Die stetige, bijektive Fortsetzung von f von $\operatorname{cl} H$ auf $\operatorname{cl} P$ werde ebenfalls mit f bezeichnet. Seien $z_k := f^{-1}(w_k) \in \mathbb{R}$, $k = 1, \ldots, n$. Dann existieren komplexe Konstanten a, c derart, dass sich f in der Form

$$(*) \qquad f(z) = a + c \int^z \prod_{k=1}^{n} (\zeta - z_k)^{\alpha_k - 1} \, d\zeta$$

darstellen lässt.

Beweis: Mit $1 < k < n$ betrachten wir in der oberen Halbebene $\mathbb{H}$ zunächst einen Halbstreifen

$$\mathbb{H}_k^+ := \{z \in \mathbb{H} : z_{k-1} < \operatorname{Re}(z) < z_{k+1}\}.$$

Wir wissen, dass $f([z_{k-1}, z_k]) = [w_{k-1}, w_k]$ und $f([z_k, z_{k+1}]) = [w_k, w_{k+1}]$, wobei

$$\pi\alpha_k = \sphericalangle(w_{k-1} - w_k, w_{k+1} - w_k)$$

bzw.

$$e^{i\pi\alpha_k} = \frac{(w_{k+1} - w_k)}{|w_{k+1} - w_k|} \Big/ \frac{(w_{k-1} - w_k)}{|w_{k-1} - w_k]},$$

siehe Abbildung 100. Man definiere die holomorphe Funktion $h_k \colon \mathbb{H}_k^+ \longrightarrow \mathbb{C}$ durch

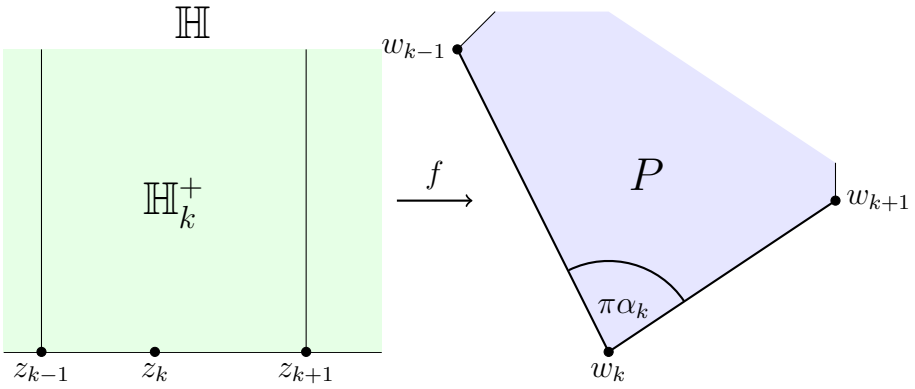


Abbildung 100: Abbildung von $[z_{k-1}, z_k]$ bzw. $[z_k, z_{k+1}]$ auf $[w_{k-1}, w_k]$ bzw. $[w_k, w_{k+1}]$

$$h_k(z) := (f(z) - w_k)^{1/\alpha_k}.$$

Da f wegen Satz 10.17 (bzw. einer darauffolgenden Bemerkung) bis zum Rand von $\mathbb{H}$ zu einer stetigen, bijektiven Abbildung fortgesetzt werden kann, kann auch h_k bis zum Rand von $\mathbb{H}_k^+$, also dem Segment $[z_{k-1}, z_{k+1}]$, stetig und bijektiv fortgesetzt werden. Wir definieren $L_k := h_k([z_{k-1}, z_{k+1}])$. Dann ist

$$\begin{aligned} L_k &= [h_k(z_{k-1}), h_k(z_k)] \cup [h_k(z_k), h_k(z_{k+1}] \\ &= [(w_{k-1} - w_k)^{1/\alpha_k}, 0] \cup [0, (w_{k+1} - w_k)^{1/\alpha_k}] \\ &= [h_k(z_{k-1}), h_k(z_{k+1})] \end{aligned}$$

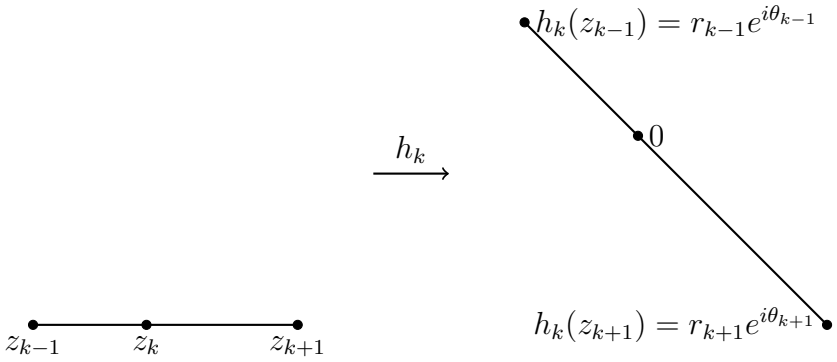


Abbildung 101: Abbildung von $[z_{k-1}, z_{k+1}]$ auf $L_k = [h_k(z_{k-1}), h_k(z_{k+1})]$

ein *Geraden*segment, da

$$\sphericalangle((w_{k-1} - w_k)^{1/\alpha_k}, (w_{k+1} - w_k)^{1/\alpha_k}) = \pi.$$

Wegen $h_k(z_k) = 0$ ist $0 \in (h_k(z_{k-1}), h(z_{k+1}))$. Wir veranschaulichen uns die Situation in Abbildung 101. Wie dort angegeben sei

$$h_k(z_{k-1}) = r_{k-1}e^{i\theta_{k-1}}, \qquad h_k(z_{k+1}) = r_{k+1}e^{i\theta_{k+1}}$$

mit $\theta_{k-1}, \theta_{k+1} \in (-\pi, \pi]$. Da $h_k(z_{k-1})$, 0 und $h_k(z_{k+1})$ auf einer Geraden liegen, ist

$$\theta_{k+1} - \theta_{k-1} \equiv \pi \pmod{2\pi}.$$

Wir wollen h_k mit Hilfe des Schwarzschen Spiegelungsprinzips von $\mathbb{H}_k^+$ auf den gesamten Halbstreifen

$$S_k := \{z \in \mathbb{C} : z_{k-1} < \mathrm{Re}(z) < z_{k+1}\}$$

holomorph fortsetzen. Hierbei haben wir die (kleine) Schwierigkeit zu überwinden, dass hierzu die fortzusetzende Funktion, in unserem Fall also h_k, auf $[z_{k-1}, z_{k+1}]$ *reell* sein müsste, was aber i. Allg. nicht der Fall ist. Daher betrachten wir die durch

$$g_k(z) := -e^{-i\theta_{k-1}}h_k(z)$$

definierte holomorphe Funktion $g_k \colon \mathbb{H}_k^+ \longrightarrow \mathbb{C}$. Diese ist natürlich ebenfalls bis zum Rand $I_k = [z_{k-1}, z_{k+1}]$ von $\mathbb{H}_k^+$ stetig fortsetzbar. Da

$$\begin{aligned} g_k(I_k) &= [g_k(z_{k-1}), g_k(z_{k+1})] \\ &= [-r_{k-1}, -e^{-i\theta_{k-1}}h_k(z_{k+1})] \\ &= [-r_{k-1}, -r_{k+1}e^{i(\theta_{k+1}-\theta_{k-1})}] \\ &= [-r_{k-1}, r_{k+1}], \end{aligned}$$

kann das Schwarzsche Spiegelungsprinzip auf g_k angewandt werden. Es existiert also eine holomorphe Fortsetzung von g_k und dann auch von h_k auf den gesamten Streifen S_k.

Wir behaupten, dass h_k' in diesem Streifen nicht verschwindet. Für $z \in \mathbb{H}_k^+$ ist

$$h_k'(z) = \frac{1}{\alpha_k}(f(z) - w_k)^{1/\alpha_k - 1} f'(z) = \frac{1}{\alpha_k} \cdot \frac{h_k(z) f'(z)}{f(z) - w_k} \neq 0,$$

da f auf $\mathbb{H}$ eine konforme Abbildung ist und folglich f' auf $\mathbb{H}$ nicht verschwindet. Für ein z aus dem unteren Halbstreifen $\{z \in \mathbb{C} : \operatorname{Im}(z) < 0,\ z_{k+1} < \operatorname{Re}(z) < z_{k+1}\}$ ist

$$g_k(z) = \overline{g_k(\overline{z})}, \qquad g_k'(z) = \overline{g_k'(\overline{z})} \neq 0$$

(siehe Beweis des Schwarzschen Spiegelungprinzips auf S. 269) und daher auch $h_k'(z) \neq 0$. Jetzt müssen noch die Punkte auf dem Segment (z_{k-1}, z_{k+1}) untersucht werden. Sei also $x \in (z_{k-1}, z_{k+1})$. Sei $r > 0$ so klein, dass $[x-r, x+r] \subset [z_{k-1}, z_{k+1}]$. Auf $B(0;r) \cap \operatorname{cl} \mathbb{H}$ ist h_k injektiv, da es f ist. Aus Symmetrie- bzw. Spiegelungsgründen ist h_k injektiv sogar in der gesamtem Kugel $B(0;r)$. Da h_k auf $B(0;r)$ holomorph ist, folgt hieraus aber (siehe Aussage auf S. 262), dass $h_k'(z) \neq 0$ für alle $z \in B(0;r)$ und insbesondere $h_k'(x) \neq 0$. Damit ist gezeigt, dass h_k' auf dem gesamten Streifen S_k nicht verschwindet. Mit $\beta_k := 1 - \alpha_k$ ist nun

$$h_k'(z) = \frac{1}{\alpha_k} h_k(z)^{1-\alpha_k} f'(z) = \frac{1}{\alpha_k} h_k(z)^{\beta_k} f'(z)$$

und daher

$$f'(z) = \alpha_k h_k(z)^{-\beta_k} h_k'(z)$$

und

$$f''(z) = -\alpha_k \beta_k h_k(z)^{-\beta_k - 1} h_k'(z)^2 + \alpha_k h_k(z)^{-\beta_k} h_k''(z).$$

Folglich ist

$$\begin{aligned} \frac{f''(z)}{f'(z)} &= -\beta_k \frac{h_k'(z)}{h_k(z)} + \frac{h_k''(z)}{h_k'(z)} \\ &= -\frac{\beta_k}{z - z_k} + E_k(z) \end{aligned}$$

mit der im Streifen S_k, wie wir gleich einsehen werden, holomorphen Funktion

$$E_k(z) := \beta_k \left(-\frac{h_k'(z)}{h_k(z)} + \frac{1}{z - z_k} \right) + \frac{h_k''(z)}{h_k'(z)}.$$

Der zweite Summand h_k''/h_k' ist selbstverständlich in dem genannten Streifen holomorph, da es h_k ist und h_k' in dem Streifen nicht verschwindet. Beim ersten Summanden beachten wir, dass $h_k(z) = (z - z_k)\tilde{h}_k(z)$ mit einer auf dem Streifen S_k holomorphen und nicht verschwindenden Funktion $\tilde{h}_k$ (denn h_k hat in dem Streifen *nur* die Nullstelle z_k), für die $\tilde{h}_k(z_k) = h_k'(z_k) \neq 0$. Daher ist $\tilde{h}_k(z) - h_k'(z) = (z - z_k) h_k^*(z)$ mit einer im Streifen holomorphen Funktion h^*. Daher ist

$$-\frac{h_k'(z)}{h_k(z)} + \frac{1}{z - z_k} = \frac{-h_k'(z)(z - z_k) + h_k(z)}{h_k(z)(z - z_k)}$$

$$
\begin{aligned}
&= \frac{(\tilde{h}_k(z) - h_k'(z))(z - z_k)}{\tilde{h}_k(z)(z - z_k)^2} \\
&= \frac{h_k^*(z)}{\tilde{h}_k(z)}
\end{aligned}
$$

als Quotient zweier holomorpher Funktionen, von denen der Nenner im Streifen S_k nicht verschwindet, eine holomorphe Funktion.
Für $k = 1$ und $k = n$ hat man entsprechende Ergebnisse. Denn im Streifen

$$S_1 := \{z \in \mathbb{C} : -\infty < \mathrm{Re}(z) < z_2\}$$

ist

$$\frac{f''(z)}{f'(z)} = -\frac{\beta_1}{z - z_1} + E_1(z)$$

mit einer in S_1 holomorphen Funktion und im Streifen

$$S_n := \{z \in \mathbb{C} : z_{n-1} < \mathrm{Re}(z) < \infty\}$$

ist

$$\frac{f''(z)}{f'(z)} = -\frac{\beta_n}{z - z_n} + E_n z)$$

mit einer in S_n holomorphen Funktion E_n. Zur Abkürzung definieren wir

$$E(z) := \frac{f''(z)}{f'(z)} + \sum_{k=1}^{n} \frac{\beta_k}{z - z_k}.$$

Da

$$E_1(z) = f''(z)/f'(z) + \beta_1/(z - z_1)$$

auf dem Streifen S_1 holomorph ist und $z_2, \ldots, z_n$ nicht in S_1 liegen, ist E auf dem Streifen S_1 holomorph. Entsprechend ist

$$E_2(z) = f''(z)/f'(z) + \beta_2/(z - z_2)$$

auf S_2 holomorph. Da $z_1, z_3, \ldots, z_n$ nicht in S_2 liegen, ist E auf $S_1 \cup S_2$ holomorph. Nach n Schritten erhalten wir, dass E auf $\bigcup_{k=1}^{n} S_k = \mathbb{C}$ holomorph bzw. eine *ganze* Funktion ist. Wir wollen zeigen, dass E identisch verschwindet. *Angenommen*, dies sei schon gelungen. Dann ist

$$\frac{f''(z)}{f'(z)} = -\sum_{k=1}^{n} \frac{\beta_k}{z - z_k}.$$

Mit

$$Q(z) := \prod_{k=1}^{n} (z - z_k)^{-\beta_k}$$

ist

$$\frac{Q'(z)}{Q(z)} = -\sum_{k=1}^{n} \frac{\beta_k}{z - z_k}$$

und daher

$$\begin{aligned}\frac{d}{dz}\left(\frac{f'(z)}{Q(z)}\right) &= \frac{f''(z)Q(z) - f'(z)Q'(z)}{Q(z)^2} \\ &= \frac{f''(z)}{Q(z)} - \frac{f'(t)}{Q(z)} \cdot \underbrace{\frac{Q'(z)}{Q(z)}}_{=f''(z)/f'(z)} \\ &= 0.\end{aligned}$$

Mit einer Konstanten c ist also $f'(z) = cQ(z)$. Durch Integration dieser Gleichung folgt die Behauptung.

Jetzt bleibt noch der Beweis dafür, dass die ganze Funktion E identisch verschwindet. Das Bild unter f von $(-\infty, z_1) \cup (z_n, \infty)$ ist das Geradensegment (w_1, w_n), siehe Abbildung 102. Wir wollen uns überlegen, dass wir f holomorph auf das Äußere bzw.

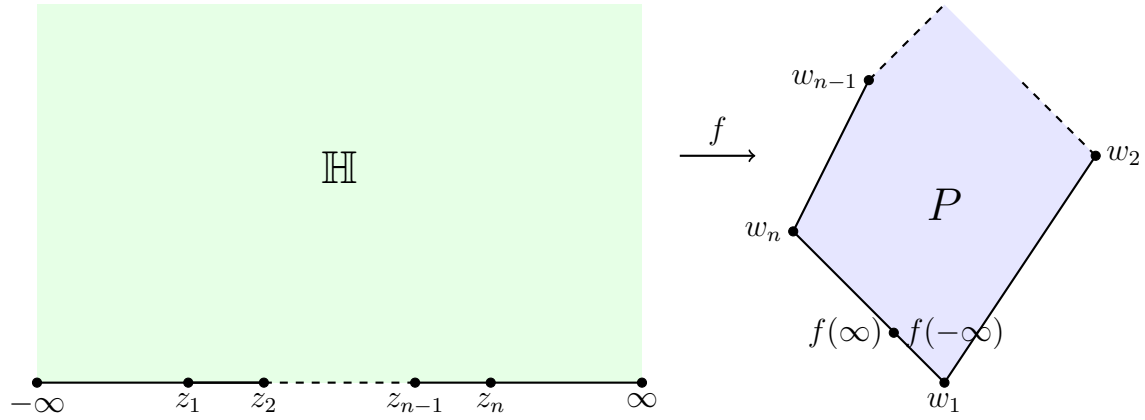


Abbildung 102: Abbildung von $(-\infty, z_1) \cup (z_n, \infty)$ auf (w_1, w_n)

Komplement einer hinreichend großen Kreisscheibe fortsetzen können. Hierzu wählen wir $R > \max_{k=1,\dots,n} |z_k|$ (dann sind $z_1, \dots, z_n \in B(0; R)$) und setzen

$$G := \{z \in \mathbb{C} : |z| > R\}, \quad G^+ := \{z \in G : \operatorname{Im}(z) > 0\}, \quad G^- := \{z \in G : \operatorname{Im}(z) < 0\}$$

sowie

$$I := \{z \in G : \operatorname{Im}(z) = 0\}.$$

Dann ist G, das Äußere der Kreisscheibe $B[0; R]$, eine offene und bezüglich der reellen Achse symmetrische Menge. Die Abbildung $f\colon \operatorname{cl}\mathbb{H} \longrightarrow \mathbb{C}$ ist auf $\mathbb{H}$ holomorph und auf $\operatorname{cl}\mathbb{H}$ stetig. Erst recht ist $f\colon G^+ \cup I \longrightarrow \mathbb{C}$ stetig und f auf G^+ holomorph. Um das Schwarzsche Spiegelungsprinzip ohne weiteres anwenden zu können, müsste f auf I reell sein, was aber i. Allg. nicht der Fall ist. Allerdings ist $f(I)$ ein *Segment* in $\mathbb{C}$, sodass man durch eine Translation und eine Drehung erreichen kann, dass die neugewonnenene Funktion I auf ein Intervall in $\mathbb{R}$ abbildet. Daher definieren wir $g\colon G^+ \cup I \longrightarrow \mathbb{C}$ durch

$$g(z) := A + Bf(z),$$

wobei $A, B \in \mathbb{C}$ mit $B \neq 0$ so bestimmt werden, dass $g(I) \subset \mathbb{R}$. Wir nehmen an, es sei

$$I = (-\infty, y_1) \cup (y_n, \infty) \quad \text{mit} \quad y_1 < z_1 < z_n < y_n.$$

Sei $v_1 := f(y_1)$ und $v_n := f(y_n)$. Wir stellen die Situation in Abbildung 103 dar. Da

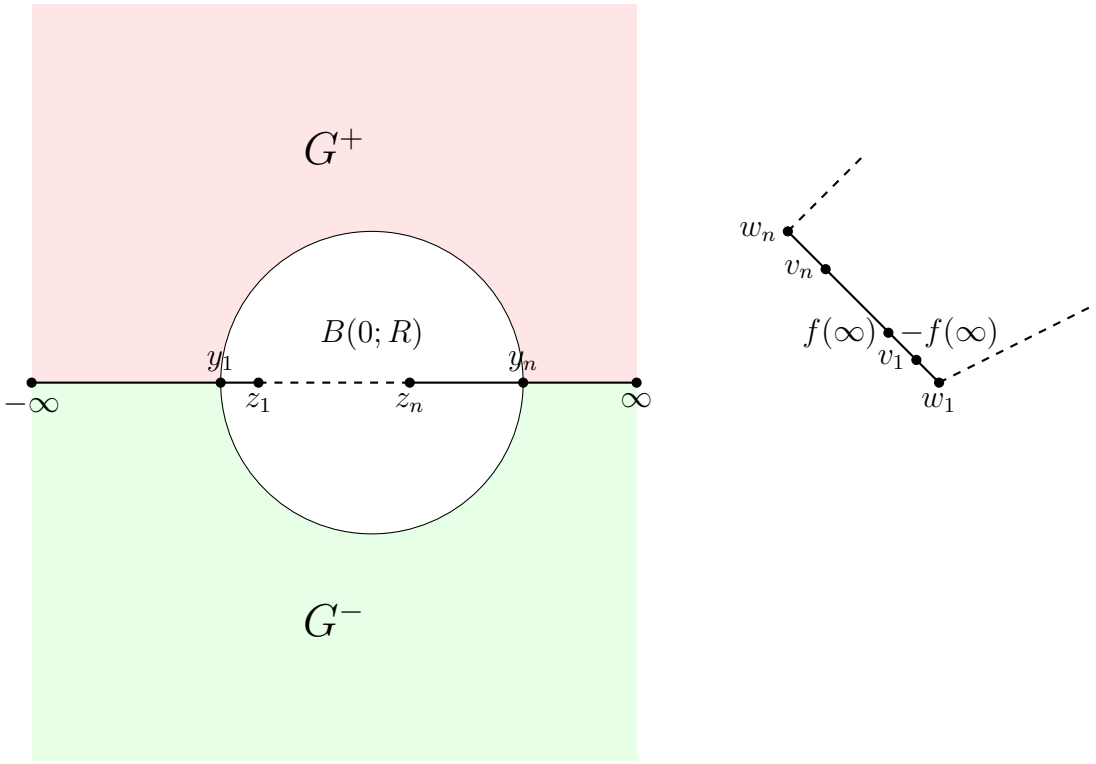


Abbildung 103: Abbildung von $(-\infty, y_1) \cup (y_n, \infty)$ auf (v_1, v_n)

$f(\infty) = f(-\infty) \in (v_1, v_n)$, existiert ein $\lambda \in (0, 1)$ mit

$$f(\infty) = f(-\infty) = (1 - \lambda)v_1 + \lambda v_n.$$

Wir bestimmen die Konstanten A und B so, dass g die Menge $I = (-\infty, y_1) \cup (y_n, \infty)$ auf das reelle Intervall $(-|v_1 - f(\infty)|, |v_n - f(\infty)|)$ abbildet. Dies erreichen wir dadurch, dass wir

$$g(-\infty) = g(\infty) = 0, \qquad g(y_1) = -|v_1 - f(\infty)|, \qquad g(y_n) = |v_n - f(\infty)|$$

fordern, was für A, B auf die Bedingungen

$$A + Bf(\infty) = 0, \qquad A + Bv_1 = -|v_1 - f(\infty)|, \qquad A + Bv_n = |v_n - f(\infty)|$$

bzw.

$$A + B((1 - \lambda)v_1 + \lambda v_n) = 0$$

und

$$A + Bv_1 = -\lambda|v_n - v_1|, \qquad A + Bv_n = (1 - \lambda)|v_n - v_1|$$

führt. Subtrahiert man von der letzten Gleichung die vorletzte, so erhält man

$$B(v_n - v_1) = |v_n - v_1|$$

und hieraus

$$B := e^{-i\arg(v_n - v_1)}, \qquad A := -Bf(\infty).$$

Mit diesen Werten ist $g\colon G^+ \cup I \longrightarrow \mathbb{C}$ mit $g(z) = A + Bf(z)$ eine auf $G^+ \cup I$ stetige und auf G^+ holomorphe Abbildung mit, wie man leicht nachweist,

$$g(I) = (-|v_1 - f(\infty)|, |v_n - f(\infty)|) \subset \mathbb{R}.$$

Wegen des Schwarzschen Spiegelungsprinzips kann g von $G^+ \cup I$ zu einer holomorphen Funktion g (wir verwenden für die Fortsetzung also denselben Namen) auf G fortgesetzt werden. Hierbei gewinnt man $g(z)$ für ein $z \in G^-$, indem man zu $\overline{z} \in G^+$ übergeht und $g(z) := \overline{g(\overline{z})}$ definiert. Das entsprechende gilt dann auch für $f = (g - A)/B$.

Nun kommen wir zum Schluss und zeigen, dass die durch

$$E(z) := \frac{f''(z)}{f'(z)} + \sum_{k=1}^{n} \frac{\beta_k}{z - z_k}$$

definierte (wie wir wissen) ganze Funktion $E\colon \mathbb{C} \longrightarrow \mathbb{C}$ identisch verschwindet, woraus, wie wir gesehen haben, die Aussage des Satzes folgt. Da $f(\mathrm{cl}\,\mathbb{H}) = \mathrm{cl}\,P$ und P beschränkt ist, ist f auf $\mathrm{cl}\,H$ und erst recht auf $G^+ \cup I$ beschränkt. Da die Werte von f in G^- im wesentlichen durch Spiegeln der Werte in G^+ (einschließlich Translation und Drehung) erhalten werden, ist f auch auf G^- beschränkt, insgesamt also auf G, dem Komplement einer hinreichend großen Kreisscheibe. Wegen des Riemannschen Hebbarkeitssatzes hat $h(z) := f(1/z)$ in $z = 0$ eine hebbare Singularität bzw. ist f in ∞ holomorph. Daher ist auch f''/f' in ∞ holomorph und wir behaupten, dass $\lim_{|z|\to\infty} f''(z)/f'(z) = 0$. Da f in ∞ holomorph ist, besitzt f in $z = \infty$ eine Entwicklung

$$f(z) = c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \cdots .$$

Hieraus folgt

$$\lim_{|z|\to\infty} \frac{f''(z)}{f'(z)} = 0.$$

Folglich ist E eine ganze, also auf ganz $\mathbb{C}$ holomorphe Funktion mit $\lim_{|z|\to\infty} E(z) = 0$. Aus der letzteren Eigenschaft folgt, dass E auf $\mathbb{C}$ beschränkt ist. Der Satz von Liouville liefert, dass E auf $\mathbb{C}$ konstant ist. Wegen $\lim_{|z|\to\infty} E(z) = 0$ verschwindet E auf $\mathbb{C}$ identisch und der Satz ist bewiesen. □

10.6 Die Integralgleichung von Theodorsen

10.6.1 Herleitung der Integralgleichung von Theodorsen

In diesem Unterabschnitt wird die Berechnung einer konformen Abbildung f der offenen Einheitskreisscheibe $B(0;1)$ (in der z-Ebene) auf ein vorgegebenes Gebiet $G \subset \mathbb{C}$

(in der w-Ebene) mit $f(0) = 0$, $f'(0) > 0$ auf die Lösung einer (singulären) Integralgleichung, nämlich der Integralgleichung von Theodorsen, zurückgeführt. Unser Ziel hier besteht darin, diese Integralgleichung herzuleiten. Es wird vorausgesetzt, dass der Rand $\Gamma := \partial G$ bezüglich des Nullpunktes $w = 0$ *sternig* ist, d. h. jeder Randpunkt $w \in \Gamma$ besitzt eine eindeutige Polardarstellung $w = \rho(\theta)e^{i\theta}$, $\theta \in [0, 2\pi]$, mit einer positiven, stetigen (und 2π-periodischen) Funktion ρ. In Abbildung 104 geben wir ein Beispiel für ein Gebiet an, das bezüglich des Nullpunktes sternig ist. In Abbildung

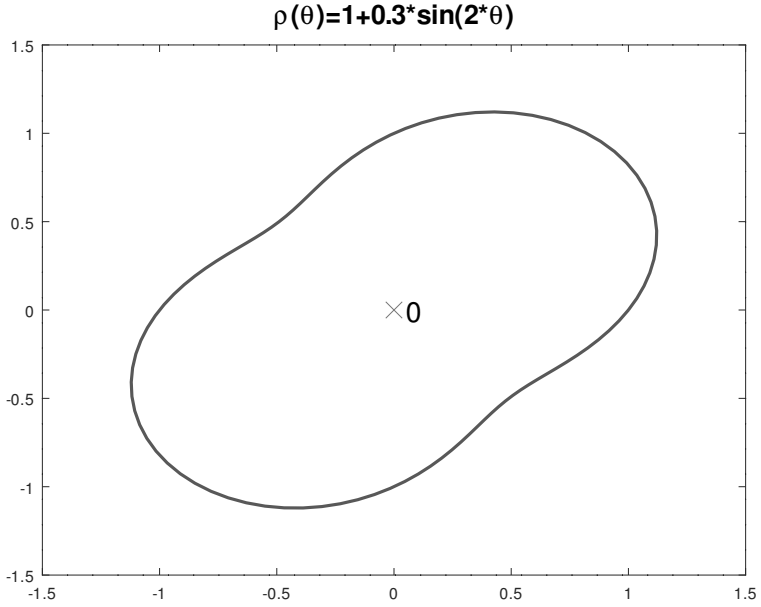


Abbildung 104: Ein sterniges Gebiet

105 zeigen wir eine Ellipse und geben ihre Polardarstellung an. Wegen des Satzes von Osgood-Carathéodory (Satz 10.17 und anschließende Bemerkung) lässt sich die konforme Abbildung von $B(0;1)$ auf G zu einer stetigen und bijektiven Abbildung von $B[0;1]$ auf $\operatorname{cl} G$ fortsetzen. Hierbei wird jedem Randpunkt $e^{i\phi}$ von $B(0;1)$ genau ein Randpunkt $f(e^{i\phi}) = \rho(\theta(\phi))e^{i\theta(\phi)} \in \Gamma$ von G zugeordnet. Diesen Zusammenhang stellen wir in Abbildung 106 dar. Kennt man die Abbildung $\theta\colon [0, 2\pi] \longrightarrow [0, 2\pi]$, so sind die Funktionswerte von f auf dem Rand von $B(0;1)$ bekannt. Sind die Funktionswerte von f auf dem Rand von $B(0;1)$ bekannt, so sind diese wegen der Integralformel von Cauchy

$$f(z) = \frac{1}{2\pi i} \int_{\partial B(0;1)} \frac{f(\zeta)}{\zeta - z}\, d\zeta \quad \text{für alle } z \in B(0;1)$$

auch in $B(0;1)$ bekannt. Unser Ziel in diesem Unterabschnitt besteht darin, für die gesuchte, die Randzuordnung vermittelnde Funktion $\theta = \theta(\phi)$ eine Integralgleichung, die *Integralgleichung von Theodorsen* aufzustellen und diese zu untersuchen. Das zweidimensionale Problem wird dann auf ein eindimensionales reduziert. Wir halten uns bei unserer Darstellung zum Teil eng an D. Gaier (1964).

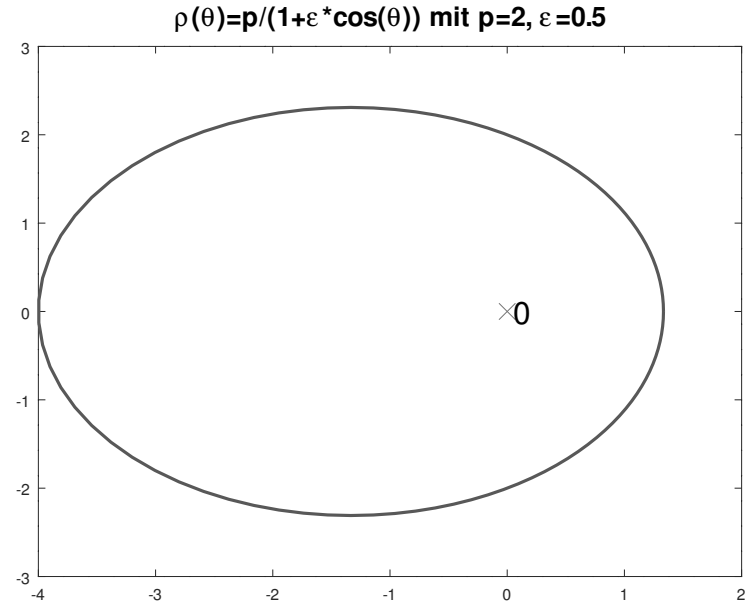


Abbildung 105: Eine Ellipse und ihre Polardarstellung

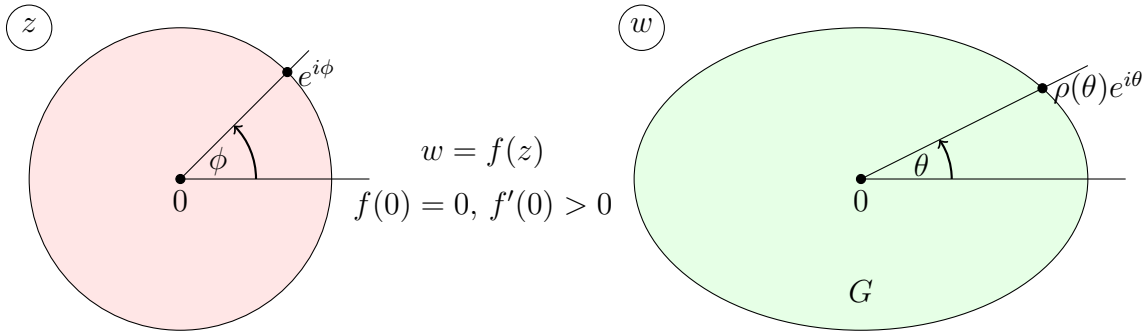


Abbildung 106: Ränderzuordnung bei einer Abbildung von $B(0;1)$ auf G

In dem folgenden Satz wird gezeigt, wie Real- und Imaginärteil einer auf $B(0;1)$ holomorphen und auf $B[0;1]$ stetigen Funktion auf dem Rand von $B(0;1)$ zusammenhängen, siehe D. Gaier (1964, S. 62).

Satz 10.19 *Sei $f\colon B[0;1] \longrightarrow \mathbb{C}$ stetig und f auf $B(0;1)$ holomorph.*

1. *Für jedes $\phi \in [0, 2\pi)$ ist*

$$f(e^{i\phi}) = f(0) + \frac{i}{2\pi} \int_{0(H)}^{2\pi} f(e^{i\vartheta}) \cot \frac{\phi - \vartheta}{2}\, d\vartheta.$$

2. *Sei $u(z) := \mathrm{Re}(f(z))$ und $v(z) := \mathrm{Im}(f(z))$. Für jedes $\phi \in [0, 2\pi)$ ist*

$$v(e^{i\phi}) = v(0) + \frac{1}{2\pi} \int_{0(H)}^{2\pi} u(e^{i\vartheta}) \cot \frac{\phi - \vartheta}{2}\, d\vartheta.$$

Hierbei sind die Integrale bezüglich $\vartheta = \phi$ als Cauchyscher Hauptwert[54] *zu nehmen.*

Beweis: Sei $\phi \in [0, 2\pi)$ vorgegeben und $\{\epsilon_k\} \subset \mathbb{R}_+$ eine Nullfolge positiver Zahlen. In Abbildung 107 geben wir einen geschlossenen Weg C_k an, der den Teilbogen

$$\gamma_k := \{z \in \mathbb{C} : |z - e^{i\phi}| = \epsilon_k\} \cap B(0;1)$$

enthält. Die durch

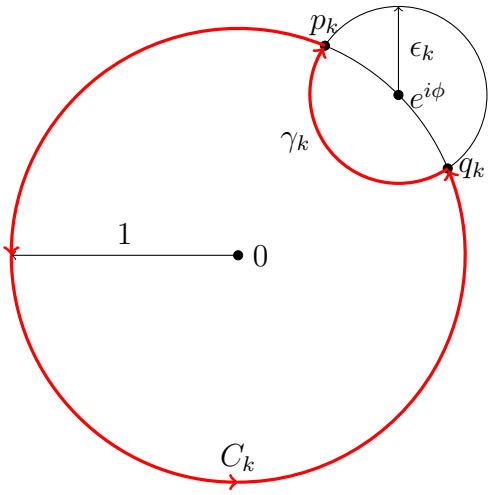


Abbildung 107: Der Weg C_k mit dem Teilbogen γ_k

$$g(z) := f(z)\frac{z + e^{i\phi}}{z(z - e^{i\phi})}$$

definierte Funktion hat in dem vom geschlossenen Weg C_k berandeten Gebiet genau einen Pol in $z = 0$ mit dem Residuum

$$\mathrm{Res}_{z=0}(g) = f(0)\frac{0 + e^{i\phi}}{0 - e^{i\phi}} = -f(0).$$

Wegen des Residuensatzes ist daher

$$\frac{1}{2\pi i}\int_{C_k} f(z)\frac{z + e^{i\phi}}{z - e^{i\phi}}\,\frac{dz}{z} = -f(0).$$

Der geschlossene Weg C_k besteht aus dem Weg von $p_k = e^{i(\phi+\delta_k)}$ nach $q_k = e^{i(2\pi+\phi-\delta_k)}$ auf dem Einheitskreis, wobei $\{\delta_k\} \subset \mathbb{R}_+$ eine Nullfolge ist, sowie dem Kreisbogen γ_k. Daher ist

$$(*) \qquad -f(0) = \frac{1}{2\pi i}\int_{p_k}^{q_k} f(z)\frac{z + e^{i\phi}}{z - e^{i\phi}}\,\frac{dz}{z} + \frac{1}{2\pi i}\int_{\gamma_k} f(z)\frac{z + e^{i\phi}}{z - e^{i\phi}}\,\frac{dz}{z}.$$

[54] Hierbei sagt man, das Integral $\int_{a(H)}^{b} f(x)\,dx$ sei *Cauchyscher Hauptwert* bezüglich $x = c \in (a, b)$, wenn $f\colon [a, c) \longrightarrow \mathbb{R}$ und $f\colon (c, b] \longrightarrow \mathbb{R}$ Riemann-integrierbar sind und

$$\int_{a(H)}^{b} f(x)\,dx := \lim_{\epsilon \to 0^+}\left(\int_a^{c-\epsilon} f(x)\,dx + \int_{c-\epsilon}^{b} f(x)\,dx\right)$$

existiert.

Im ersten Integral in $(*)$ setzen wir $z = e^{i\vartheta}$ und erhalten

$$\begin{aligned}
\frac{1}{2\pi i}\int_{p_k} f(z)\frac{z+e^{i\phi}}{z-e^{i\phi}}\,\frac{dz}{z} &= \frac{1}{2\pi}\int_{\phi+\delta_k}^{2\pi+\phi-\delta_k} f(e^{i\vartheta})\frac{e^{i\vartheta}+e^{i\phi}}{e^{i\vartheta}-e^{i\phi}}\,d\vartheta \\
&= \frac{i}{2\pi}\left(\int_0^{\phi-\delta_k} f(e^{i\vartheta})\cot\frac{\phi-\vartheta}{2}\,d\vartheta\right. \\
&\qquad \left. + \int_{\phi+\delta_k}^{2\pi} f(e^{i\vartheta})\cot\frac{\phi-\vartheta}{2}\,d\vartheta\right) \\
&\overset{\to}{_{k\to\infty}} \frac{i}{2\pi}\int_{0(H)}^{2\pi} f(e^{i\vartheta})\cot\frac{\phi-\vartheta}{2}\,d\vartheta,
\end{aligned}$$

wobei benutzt wurde, dass

$$\frac{e^{i\vartheta}+e^{i\phi}}{e^{i\vartheta}-e^{i\phi}} = \frac{1+e^{i(\phi-\vartheta)}}{1-e^{i(\phi-\vartheta)}} = \frac{e^{-i(\phi-\vartheta)/2}+e^{i(\phi-\vartheta)/2}}{e^{-i(\phi-\vartheta)/2}-e^{i(\phi-\theta)/2}} = i\cot\frac{\phi-\vartheta}{2}.$$

Jetzt zeigen wir, dass das zweite Integral in $(*)$ mit $k\to\infty$ gegen $-f(e^{i\phi})$ konvergiert. Ist das gelungen, so haben wir

$$-f(0) = \frac{i}{2\pi}\int_{0(H)}^{2\pi} f(e^{i\vartheta})\cot\frac{\phi-\vartheta}{2}\,d\vartheta - f(e^{i\phi})$$

bzw.

$$f(e^{i\phi}) = f(0) + \frac{i}{2\pi}\int_{0(H)}^{2\pi} f(e^{i\vartheta})\cot\frac{\phi-\vartheta}{2}\,d\vartheta,$$

also die erste Aussage bewiesen. Nimmt man hier auf beiden Seiten den Imaginärteil, so erhält man die zweite Behauptung. Nun zeigen wir, dass das zweite Integral in $(*)$ mit $k\to\infty$ gegen $-f(e^{i\phi})$ konvergiert. Für $z = e^{i\phi}+\epsilon_k e^{i\theta_k}\in\gamma_k$ ist

$$\frac{f(z)(z+e^{i\phi})}{z} = 2f(e^{i\phi}) + h_k(z),$$

wobei mit einer Konstanten $C>0$ die Abschätzung $|h_k(z)|\le C\epsilon_k$ gilt. Dann ist

$$\frac{1}{2\pi i}\int_{\gamma_k} f(z)\frac{z+e^{i\phi}}{z-e^{i\phi}}\,\frac{dz}{z} = \frac{1}{2\pi i}\left(2f(e^{i\phi})\underbrace{\int_{\gamma_k}\frac{dz}{z-e^{i\phi}}}_{\to -\pi i} + \underbrace{\int_{\gamma_k}\frac{h_k(z)\,dz}{z-e^{i\phi}}}_{\to 0}\right)\to -f(e^{i\phi}),$$

wobei die zweite Konvergenzaussage offensichtlich ist und die erste jetzt begründet wird. Der Weg γ_k führt von q_k nach p_k (siehe Abbildung 107) auf einem Kreis um $e^{i\phi}$ mit dem Radius ϵ_k. Ein Schnittpunkt

$$e^{i\phi}+\epsilon_k e^{i\theta}\in\partial B(0;1)\cap\partial B(e^{i\phi};\epsilon_k)$$

des Einheitskreises um den Nullpunkt und des Kreises um $e^{i\phi}$ mit dem Radius ϵ_k berechnet sich aus

$$1 = |e^{i\phi}+\epsilon_k e^{i\theta}|^2 = 1+\epsilon_k(e^{i(\phi-\theta)}+e^{-i(\phi-\theta)})+\epsilon_k^2 = 1+2\epsilon_k\cos(\phi-\theta)+\epsilon_k^2,$$

was auf

$$\theta_k^+ = \phi + \arccos(-\epsilon_k/2), \qquad \theta_k^- = \phi - \arccos(-\epsilon_k/2)$$

führt. Nach der Variablentransformation $z = e^{i\phi} + \epsilon_k e^{i\theta}$ erhalten wir

$$\int_{\gamma_k} \frac{dz}{z - e^{i\phi}} = -i \int_{\theta_k^-}^{\theta_k^+} d\theta = -i(\theta_k^+ - \theta_k^-) = -2i \underbrace{\arccos(-\epsilon_k/2)}_{\to \pi/2} \to -\pi i.$$

Hierbei haben wir berücksichtigt, dass der Weg γ_k auf dem Kreis $\partial B(e^{i\phi}; \epsilon_k)$ im mathematisch *negativen* Sinne, also im Uhrzeigersinn, erfolgt. Der Satz ist bewiesen. □

Als Korollar zu Satz 10.19 notieren wir:

Korollar 10.20 *Es ist*

$$\frac{1}{2\pi} \int_{0(H)}^{2\pi} e^{in\vartheta} \cot \frac{\phi - \vartheta}{2}\, d\vartheta = \begin{cases} -ie^{in\phi}, & n = 1, 2, \ldots, \\ 0, & n = 0, \\ ie^{in\phi}, & n = -1, -2, \ldots \end{cases}$$

bzw.

$$\frac{1}{2\pi} \int_{0(H)}^{2\pi} \left\{ \begin{array}{c} \sin n\vartheta \\ \cos n\vartheta \end{array} \right\} \cot \frac{\phi - \vartheta}{2}\, d\vartheta = \left\{ \begin{array}{c} -\cos n\phi \\ \sin n\phi \end{array} \right\}, \quad n = 1, 2, \ldots.$$

Beweis: Wir wenden den ersten Teil von Satz 10.19 mit $f(z) := z^n$ an, $n = 0, 1, 2, \ldots$. Die konjugierten Beziehungen liefern die Aussage für $n = -1, -2, \ldots$. Die entsprechenden reellen Beziehungen folgen dann. □

Nun kommen wir zur Aufstellung der Integralgleichung von Theodorsen. Wir gehen weiter von einem Gebiet $G \subset \mathbb{C}$ aus, dessen Rand Γ bezüglich des Nullpunktes $w = 0$ *sternig* ist, d. h. jeder Randpunkt $w \in \Gamma$ besitzt eine eindeutige Polardarstellung $w = \rho(\theta)e^{i\theta}$, $\theta \in [0, 2\pi]$, mit einer positiven, stetigen (und 2π-periodischen) Funktion ρ. Sei $f\colon B[0;1] \longrightarrow \mathbb{C}$ die eindeutige Funktion, die $B(0;1)$ holomorph auf G abbildet, die auf $B[0;1]$ stetig und bijektiv ist und für die $f(0) = 0$ und $f'(0) > 0$ gilt. Wir hatten uns oben überlegt, dass es (im Prinzip) genügt, die Ränderzuordnungsfunktion $\theta = \theta(\phi)$ zu bestimmen, durch die einem Randpunkt $e^{i\phi}$ von $B(0;1)$ ein Randpunkt

$$f(e^{i\phi}) = \rho(\theta(\phi))e^{i\theta(\phi)}$$

von G zugeordnet wird. Diese Ränderzuordnungsfunktion genügt einer nichtlinearen, singulären Integralgleichung, der Integralgleichung von Theodorsen. Zu ihrer Ableitung definieren wir die Hilfsfunktion

$$F(z) := \log \frac{f(z)}{z} = U(z) + iV(z).$$

Diese ist durch die Vorschrift, dass $F(0) = \log f'(0)$ reell ist, eindeutig festgelegt, auf $B(0;1)$ holomorph und in $B[0;1]$ stetig. Ferner ist

$$U(z) = \log\left|\frac{f(z)}{z}\right|, \qquad \text{also} \quad U(e^{i\phi}) = \log \rho(\theta(\phi)),$$

$$V(z) = \arg\frac{f(z)}{z}, \qquad \text{also} \quad V(e^{i\phi}) = \theta(\phi) - \phi.$$

Aus Satz 10.19 erhalten wir

$$\theta(\phi) = \phi + \frac{1}{2\pi} \int_{0(H)}^{2\pi} \log \rho(\theta(\vartheta)) \cot \frac{\phi - \vartheta}{2} \, d\vartheta. \tag{Theo}$$

Dies ist die *Integralgleichung von Theodorsen*. Wegen des Riemannschen Abbildungssatzes hat die Integralgleichung von Theodorsen mindestens eine Lösung. Definiert man den Operator K für eine 2π-periodische Funktion f aus $L^2[0, 2\pi]$ durch

$$K(f)(\phi) := \frac{1}{2\pi} \int_{0(H)}^{2\pi} f(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta,$$

so kann die Theodorsensche Integralgleichung auch als

$$\theta(\phi) = \phi + K[\log \rho(\theta)](\phi) \tag{Theo}$$

geschrieben werden. Für $f \in L^2[0, 2\pi]$ heißt $\overline{f} = K(f)$ die zu f *konjugierte Funktion*.

10.6.2 Ein Exkurs über vollständige Orthonormalsysteme

Wir erinnern an einige Begriffe und Ergebnisse im Zusammenhang mit einem Orthonormalsystem in einem Hilbertraum.

1. Sei $(V, \langle \cdot, \cdot \rangle)$ ein Prä-Hilbertraum über dem Körper $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ und $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ die von dem inneren Produkt induzierte Norm. Eine Menge $\{v_k\}_{k \in \mathbb{N}} \subset V \setminus \{0\}$ heißt ein *Orthogonalsystem* von V, wenn $\langle v_k, v_l \rangle = 0$ für $k \neq l$, wenn also je zwei Elemente senkrecht aufeinander stehen. Haben zusätzlich sämtliche Elemente die Norm 1, so spricht man von einem Orthonormalsystem. Eine Menge $\{v_k\}_{k \in \mathbb{N}} \subset V$ heißt also ein *Orthonormalsystem* von V, wenn $\langle v_k, v_l \rangle = \delta_{kl}$, $k, l \in \mathbb{N}$, wobei

   $$\delta_{kl} := \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \qquad (k, l \in \mathbb{N})$$

   das Kronecker-Symbol ist. Ein Orthonormalsystem $\{v_k\}_{k \in \mathbb{N}}$ von V heißt *vollständig*, wenn die lineare Hülle span $(\{v_k\}_{k \in \mathbb{N}})$ von $\{v_k\}_{k \in \mathbb{N}}$, also die Menge aller endlichen Linearkombinationen von $\{v_k\}_{k \in \mathbb{N}}$, *dicht* in V liegt. Dann gilt:

   (a) Ist $\{v_k\}_{k \in \mathbb{N}}$ ein Orthonormalsystem im Prä-Hilbertraum $(V, \langle \cdot, \cdot \rangle)$, so gilt die *Besselsche Ungleichung*:

   $$\sum_{k=1}^{\infty} |\langle f, v_k \rangle|^2 \leq \|f\|^2 \qquad \text{für alle } f \in V.$$

   Denn: Für jedes $f \in V$ und alle $n \in \mathbb{N}$ ist

   $$0 \leq \left\| f - \sum_{k=1}^{n} \langle f, v_k \rangle v_k \right\|^2 = \|f\|^2 - \sum_{k=1}^{n} |\langle f, v_k \rangle|^2,$$

   woraus die Behauptung folgt.

(b) Ein Orthonormalsystem $\{v_k\}_{k\in\mathbb{N}}$ in einem Prä-Hilbertraum $(V, \langle\cdot,\cdot\rangle)$ ist genau dann vollständig, wenn die *Parsevalsche Gleichung* gilt:

$$\|f\|^2 = \sum_{k=1}^{\infty} |\langle f, v_k\rangle|^2 \qquad \text{für alle } f \in V.$$

Denn: Sei das Orthonormasystem $\{v_k\}_{k\in\mathbb{N}}$ vollständig. Angenommen, die Parsevalsche Gleichung gilt nicht für jedes $f \in V$. Wegen der Besselschen Ungleichung existiert dann ein $f \in V$ mit $\sum_{k=1}^{\infty} |\langle f, v_k\rangle|^2 < \|f\|^2$. Für beliebiges $n \in \mathbb{N}$ und beliebige $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$ ist

$$\begin{aligned}
\left\| f - \sum_{k=1}^{n} \alpha_k v_k \right\|^2 &= \underbrace{\sum_{k=1}^{n} |\alpha_k - \langle f, v_k\rangle|^2}_{\geq 0} + \|f\|^2 - \sum_{k=1}^{n} |\langle f, v_k\rangle|^2 \\
&\geq \|f\|^2 - \sum_{k=1}^{n} |\langle f, v_k\rangle|^2 \\
&\geq \underbrace{\|f\|^2 - \sum_{k=1}^{\infty} |\langle f, v_k\rangle|^2}_{>0}.
\end{aligned}$$

Dies ist ein Widerspruch dazu, dass das Orthonormalsystem $\{v_k\}_{k\in\mathbb{N}}$ vollständig ist, da sich f nicht beliebig genau durch eine endliche Linearkombination von Elementen des Orthonormalsystems approximieren lässt. Gilt umgekehrt für jedes $f \in V$ die Parsevalsche Gleichung, so ist offenbar das Orthonormalsystem vollständig, da $f = \lim_{n\to\infty} \sum_{k=1}^{n} \langle f, v_k\rangle v_k$.

(c) Ist $(V, \langle\cdot,\cdot\rangle)$ ein Hilbertraum, so ist ein Orthonormalsystem $\{v_k\}_{k\in\mathbb{N}}$ genau dann ein vollständiges Orthonormalsystem, wenn das othogonale Komplement von $\{v_k\}_{k\in\mathbb{N}}$ genau das Nullelement von V ist, also die Äquivalenz

$$\langle f, v_k\rangle = 0 \quad (k \in \mathbb{N}) \Longleftrightarrow f = 0$$

gilt.

Denn: Sei $\{v_k\}_{k\in\mathbb{N}}$ ein vollständiges Orthonormalsystem. Für jedes $f \in V$ gilt dann die Parsevalsche Gleichung $\|f\|^2 = \sum_{k=1}^{\infty} |\langle f, v_k\rangle|^2$, wie wir gerade eben bewiesen haben. Folglich ist $f = 0$ genau dann, wenn $\langle f, v_k\rangle = 0$ für alle $k \in \mathbb{N}$. Für diese Richtung wird also die Vollständigkeit von V nicht benötigt. Sei nun $\{v_k\}_{k\in\mathbb{N}} \subset V$ ein Orthonormalsystem mit der Eigenschaft, dass $\langle f, v_k\rangle = 0$ für alle $k \in \mathbb{N}$, wenn $f = 0$, wenn also das Nullelement das einzige Element von V ist, welches auf allen v_k, $k \in \mathbb{N}$, senkrecht steht. Angenommen, $\{v_k\}_{k\in\mathbb{N}}$ sei *kein* vollständiges Orthonormalsystem in V. Dann wäre die lineare Hülle von $\{v_k\}_{k\in\mathbb{N}}$ nicht dicht in V bzw.

$$L := \mathrm{cl}\,(\mathrm{span}\,(\{v_k\}_{k\in\mathbb{N}}))$$

ein *echter* linearer, abgeschlossener Teilraum des Hilbertraums V. Nun gilt aber als leichte Folgerung aus dem *Projektionssatz* für abgeschlossene, konvexe Mengen in einem Hilbertraum die folgende Aussage, aus der sofort der gewünschte Widerspruch folgt:

- Sei $(V, \langle\cdot,\cdot\rangle)$ ein Hilbertraum und $L \subset V$ ein echter, abgeschlossener linearer Teilraum von V. Dann ist das orthogonale Komplement von L nichttrivial bzw. $L^\perp \neq \{0\}$, es existiert also $f \in V \setminus \{0\}$ mit $\langle f, l\rangle = 0$ für alle $l \in L$.
  Das gesuchte f gewinnt man, indem man von einem beliebigen $x \in V \setminus L$ ausgeht, die Projektion $P_L(x) \in L$ von x auf L bestimmt, also die (eindeutige) Lösung $P_L(x) \in L$ der Aufgabe, $\|\cdot - x\|$ auf L zu minimieren und $f := x - P_L(x)$ setzt. In Abbildung 108 veranschaulichen wir uns die Aussage.

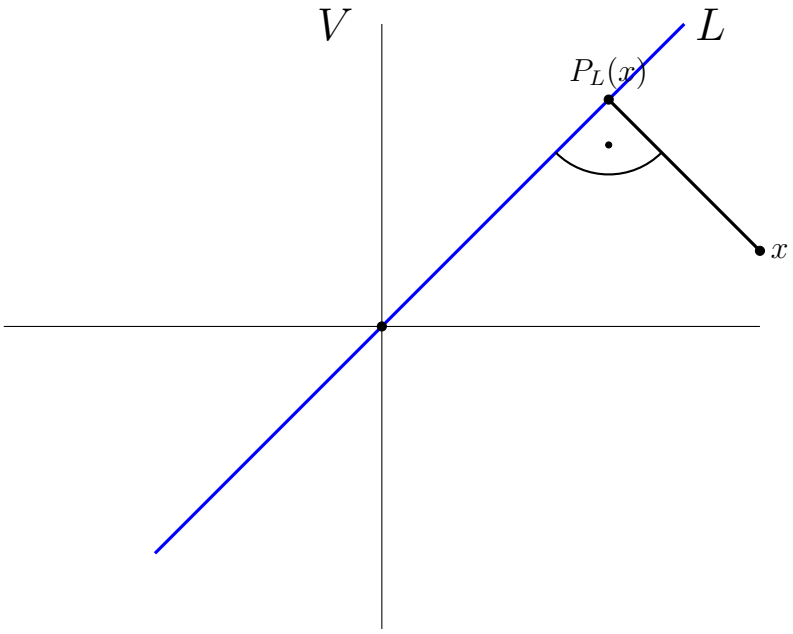


Abbildung 108: Das orthogonale Komplement eines echten, linearen, abgeschlossenen Teilraums eines Hilbertraums ist nichttrivial

(d) Ist $(V, \langle\cdot,\cdot\rangle)$ ein Prä-Hilbertraum und $\{v_k\}_{k\in\mathbb{N}}$ ein Orthonormalsystem von V, so ist

$$P_n(f) := \sum_{k=1}^{n} \langle f, v_k\rangle v_k$$

für jedes $f \in V$ die Projektion von f auf den (abgeschlossenen) endlichdimensionalen linearen Teilraum $V_n := \text{span}\,(\{v_k\}_{k=1,\ldots,n})$ von V, also

$$d(f, V_n) := \|P_n(f) - f\| \leq \|v_n - f\| \quad \text{für alle } v_n \in V_n.$$

Ist $(V, \langle\cdot,\cdot\rangle)$ sogar ein Hilbertraum, so ist

$$f = \sum_{k=1}^{\infty} \langle f, v_k\rangle v_k := \lim_{n\to\infty} \sum_{k=1}^{n} \langle f, v_k\rangle v_k \quad \text{für alle } f \in V.$$

Die Darstellung $f = \sum_{k=1}^{\infty}\langle f, v_k\rangle v_k$ nennen wir die *Fourier-Reihe* von f bezüglich des vollständigen Orthonormalsystems $\{v_k\}_{k\in\mathbb{N}}$, die Koeffizienten $\langle f, v_k\rangle$ heißen die zugehörigen *Fourier-Koeffizienten*.

2. Eine Funktion $f\colon \mathbb{R} \longrightarrow \mathbb{C}$ ist 2π-periodisch, wenn $f(x) = f(x + 2\pi)$ für alle $x \in \mathbb{R}$. Eine 2π-periodische Funktion auf $\mathbb{R}$ kann mit einer Funktion auf einem Kreis bzw. auf $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$ identifiziert werden, wobei $\mathbb{T}$ aus $\mathbb{R}$ dadurch entsteht, dass Punkte in $\mathbb{R}$, die sich um $2\pi n$ mit $n \in \mathbb{Z}$ unterscheiden als gleich angesehen werden. Der Raum $C(\mathbb{T})$ ist der Raum der stetigen Funktionen von $\mathbb{T}$ nach $\mathbb{C}$ und $L^2(\mathbb{T})$ ist die Vervollständigung von $C(\mathbb{T})$ bezüglich der L^2-Norm

$$\|f\| = \left(\int_{\mathbb{T}} |f(x)|^2\, dx\right)^{1/2}.$$

Das Integral über $\mathbb{T}$ ist als ein Integral bezüglich x über ein beliebiges Intervall der Länge 2π zu verstehen. Ein Element aus $L^2(\mathbb{T})$ kann interpretiert werden als eine Äquivalenzklasse Lebesgue-messbarer, quadrat-integierbarer Funktionen von $\mathbb{T}$ nach $\mathbb{C}$. Hierbei heißen zwei Funktionen *äquivalent*, wenn sie fast überall gleich sind. Der Raum $\mathbb{L}^2(\mathbb{T})$ ist ein Hilbertraum mit dem inneren Produkt

$$\langle f, g\rangle = \int_{\mathbb{T}} f(x)\overline{g(x)}\, dx.$$

Durch $\{e_n\}_{n\in\mathbb{Z}}$ mit

$$e_n(x) := \frac{1}{\sqrt{2\pi}} e^{inx}, \qquad n \in \mathbb{Z},$$

ist ein Orthonormalsystem in $L^2(\mathbb{T})$ gegeben, wie man aus

$$\langle e_m, e_n\rangle = \frac{1}{2\pi}\int_0^{2\pi} e^{i(m-n)x}\, dx = \begin{cases} 1, & m = n, \\ 0, & m \neq n, \end{cases}$$

erkennt. Dieses Orthonormalsystem ist ein vollständiges Orthonormalsystem in $L^2(\mathbb{T})$. Einen genauen Beweis hierfür wollen wir nicht angeben. Dieser kann in zwei Schritten jeweils durch ein $\epsilon/2$-Argument erfolgen. Vorgegeben sei ein $f \in L^2(\mathbb{T})$ und ein $\epsilon > 0$. Da $C(\mathbb{T})$ dicht in $L^2(\mathbb{T})$ ist, gibt es ein $g \in C(\mathbb{T})$ mit $\|f-g\|_2 \leq \epsilon/2$. Nun überlegt man sich, dass span $(\{e_n\}_{n\in\mathbb{Z}})$ dicht in $(C(\mathbb{T}), \|\cdot\|_\infty)$ liegt, wobei

$$\|f\|_\infty := \max_{x\in\mathbb{T}} |f(x)|.$$

Dies kann mit dem Satz von Stone-Weierstraß (siehe z. B. J. Werner (2013)) gezeigt werden. Zu $g \in C(\mathbb{T})$ existiert also ein $p \in$ span $(\{e_n\}_{n\in\mathbb{Z}})$ mit $\|g-p\|_\infty \leq$

$\epsilon/(2\sqrt{2\pi})$. Insgesamt gibt es bei vorgegebenem $\epsilon > 0$ zu jedem $f \in L^2(\mathbb{T})$ ein $p \in \text{span}\,(\{e_n\}_{n\in\mathbb{Z}})$ mit

$$\|f-p\| \le \underbrace{\|f-g\|}_{\le\epsilon/2} + \|g-p\| \le \frac{\epsilon}{2} + \underbrace{\sqrt{2\pi}\|g-p\|_\infty}_{\le\epsilon/2} \le \epsilon.$$

Damit ist wenigstens skizziert, weshalb das Orthonormalsystem $\{e_n\}_{n\in\mathbb{Z}}$ sogar ein vollständiges Orthonormalsystem in $L^2(\mathbb{T})$ ist.

Die Fourier-Reihe einer Funktion $f \in \mathbb{L}^2(\mathbb{T})$ bezüglich des vollständigen Orthonormalsystems $\{e_n\}_{n\in\mathbb{Z}}$ ist durch

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \hat{f}_n e^{inx} \quad \text{mit} \quad \hat{f}_n := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{T}} f(x) e^{-inx}\,dx, \quad n \in \mathbb{Z},$$

gegeben. Die Parsevalsche Gleichung sagt aus, dass $\|f\|^2 = \sum_{n=-\infty}^{\infty} |\hat{f}_n|^2$ für alle $f \in L^2(\mathbb{T})$.

Für *reellwertige* Funktionen aus $L^2(\mathbb{T})$, auch hier benutzen wir das innere Produkt

$$\langle f, g\rangle := \int_0^{2\pi} f(x)g(x)\,dx,$$

bilden die reellwertigen Funktionen

$$\{1, \cos kx, \sin kx : k \in \mathbb{N}\}$$

ein Orthogonalsystem. Dies erkennt man aus

$$\cos kx = \frac{e^{ikx} + e^{-ikx}}{2}, \qquad \sin kx = \frac{e^{ikx} - e^{-ikx}}{2i}$$

und

$$\int_0^{2\pi} e^{i(m-n)x}\,dx = \begin{cases} 0, & m \ne n, \\ 2\pi, & m = n \end{cases}$$

oder aus oder den Beziehungen

$$\begin{aligned}
\int_0^{2\pi} \cos(kx)\cos(lx)\,dx &= \begin{cases} 0, & k \ne l, \\ \pi, & k = l \ne 0, \\ 2\pi, & k = l = 0, \end{cases} \\
\int_0^{2\pi} \sin(kx)\sin(lx)\,dx &= \begin{cases} 0, & k \ne l, \\ \pi, & k = l \ne 0, \end{cases} \\
\int_0^{2\pi} \cos(kx)\sin(lx)\,dx &= 0.
\end{aligned}$$

Daher ist

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos kx, \frac{1}{\sqrt{\pi}}\sin kx : k \in \mathbb{N} \right\}$$

ein Orthonormalsystem in $L^2(\mathbb{T})$ (über den reellen Zahlen). Dieses ist sogar ein vollständiges Orthonormalsystem, was im Prinzip genau wie in dem oben skizzierten komplexen Fall in zwei Schritten bewiesen werden kann. Als Fourier-Reihe einer rellwertigen Funktion $f \in L^2(\mathbb{T})$ bzw. $f \in L^2[0, 2\pi]$ erhalten wir

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

mit

$$a_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad k = 0, 1, 2, \ldots$$

und

$$b_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx, \quad k = 1, 2, \quad \ldots.$$

Die Parsevalsche Gleichung besagt dann, dass

$$\int_0^{2\pi} f^2(x) \, dx = \frac{a_0^2}{2} + \sum_{k=1}^{\infty} (a_k^2 + b_k^2).$$

Hiermit beenden wir unseren Exkurs über Orthonormalsysteme und Fourier-Reihen.

10.6.3 Eine Eindeutigkeitsaussage

Wir kommen zurück auf den durch

$$K(f)(\phi) := \frac{1}{2\pi} \int_{0(H)}^{2\pi} f(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta$$

definierten Operator $K \colon L^2[0, 2\pi] \longrightarrow L^2[0, 2\pi]$. Hierbei müssen wir darauf hinweisen, dass wir *nicht* bewiesen haben, dass $K(x)$ für $x \in L^2[0, 2\pi]$ definiert und in $L^2[0, 2\pi]$ liegt. Genau wie D. Gaier (1964, S. 63) verweisen wir lediglich auf A. Zygmund (1935, S. 76). Unter diesen Voraussetzungen können wir zeigen, dass der folgende Satz (siehe D. Gaier (1964, S. 64)) gilt.

Satz 10.21 *Der durch*

$$K(f)(\phi) := \frac{1}{2\pi} \int_{0(H)}^{2\pi} f(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta$$

definierte Operator $K \colon L^2[0, 2\pi] \longrightarrow L^2[0, 2\pi]$ *ist normvermindernd, d. h. es ist*

$$\|K(f)\| \leq \|f\| \quad \textit{für alle } f \in L^2[0, 2\pi].$$

Beweis: Sei $f \in L^2[0, 2\pi]$ und

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

die zugehörige Fourier-Reihe bezüglich des vollständigen Orthonormalsystems

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos kx, \frac{1}{\sqrt{\pi}} \sin kx : k \in \mathbb{N} \right\}$$

in $L^2[0, 2\pi]$. Hierbei ist

$$a_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad k = 0, 1, 2, \ldots$$

und

$$b_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx, \quad k = 1, 2, \ldots.$$

Wir berechnen jetzt die Fourier-Koeffizienten $\overline{a}_k$, $\overline{b}_k$ von $K(f)$ bezüglich desselben Orthonormalsystems. Es ist

$$\begin{aligned}
\overline{a}_k &= \frac{1}{\pi} \int_0^{2\pi} K(f)(\phi) \cos k\phi \, d\phi \\
&= \frac{1}{\pi} \int_0^{2\pi} \left(\frac{1}{2\pi} \int_{0(H)}^{2\pi} f(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta \right) \cos k\phi \, d\phi \\
&= \frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \left(\frac{1}{2\pi} \int_{0(H)}^{2\pi} \cos k\phi \cot \frac{\phi - \vartheta}{2} \, d\phi \right) d\vartheta \\
&= -\frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \Bigg(\underbrace{\frac{1}{2\pi} \int_{0(H)}^{2\pi} \cos k\phi \cot \frac{\vartheta - \phi}{2} \, d\phi}_{= \sin k\vartheta} \Bigg) d\vartheta \\
&= -\frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \sin k\vartheta \, d\vartheta \\
&= \begin{cases} 0, & k = 0, \\ -b_k, & k = 1, 2, \ldots. \end{cases}
\end{aligned}$$

Für $k \in \mathbb{N}$ ist entsprechend

$$\begin{aligned}
\overline{b}_k &= \frac{1}{\pi} \int_0^{2\pi} K(f)(\phi) \sin k\phi \, d\phi \\
&= \frac{1}{\pi} \int_0^{2\pi} \left(\frac{1}{2\pi} \int_{0(H)}^{2\pi} f(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta \right) \sin k\phi \, d\phi \\
&= \frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \left(\frac{1}{2\pi} \int_{0(H)}^{2\pi} \sin k\phi \cot \frac{\phi - \vartheta}{2} \, d\phi \right) d\vartheta \\
&= -\frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \Bigg(\underbrace{\frac{1}{2\pi} \int_{0(H)}^{2\pi} \sin k\phi \cot \frac{\vartheta - \phi}{2} \, d\phi}_{= -\cos k\vartheta} \Bigg) d\vartheta \\
&= \frac{1}{\pi} \int_0^{2\pi} f(\vartheta) \cos k\vartheta \, d\vartheta
\end{aligned}$$

$$= \quad a_k.$$

Wegen der Parsevalschen Gleichung ist daher

$$\|K(f)\|^2 = \underbrace{\frac{\overline{a}_0^2}{2}}_{=0} + \sum_{k=1}^{\infty}(\overline{a}_k^2 + \overline{b}_k^2) = \sum_{k=1}^{\infty}(a_k^2 + b_k^2) = \|f\|^2 - \frac{a_0^2}{2} \leq \|f\|^2.$$

Damit ist die Behauptung bewiesen. □

Unter gewissen Zusatzannahmen über den Rand $\Gamma := \partial G$ des Gebietes G kann die *Eindeutigkeit* einer Lösung von (Theo) bewiesen werden. Hierzu ist die folgende Definition wichtig.

Definition 10.22 Sei $G \subset \mathbb{C}$ ein Gebiet mit der Eigenschaft, dass jeder Randpunkt $w \in \partial G$ eine eindeutige Polardarstellung $w = \rho(\theta)e^{i\theta}$, $\theta \in [0, 2\pi]$, besitzt, wobei $\rho\colon \mathbb{R} \longrightarrow \mathbb{R}_+$ eine (positive), 2π-periodische und stetige Funktion ist. Wir sagen, dass G einer *ϵ-Bedingung* genügt, wenn die folgenden beiden Bedingungen erfüllt sind:

(i) ρ ist auf $[0, 2\pi]$ stetig differenzierbar,

(ii) Es ist
$$\frac{|\rho'(\theta)|}{\rho(\theta)} \leq \epsilon \quad \text{für alle } \theta \in \mathbb{R}.$$

Beispiel: In Abbildung 104 haben wir ein Gebiet G dargestellt, dessen Rand durch

$$\partial G = \{\rho(\theta)e^{i\theta} : \theta \in [0, 2\pi]\}$$

mit

$$\rho(\theta) := 1 + 0.3\sin(2\theta)$$

gegeben ist. Die Funktion

$$h(\theta) := \frac{|\rho'(\theta)|}{\rho(\theta)|} = \frac{0.6|\cos(2\theta)|}{1 + 0.3\sin(2\theta)}$$

haben wir über dem Intervall $[0, 2\pi[$ in Abbildung 109 dargestellt. Dann ist

$$\max_{\theta\in[0,2\pi]} h(\theta) = 0.6\frac{\sqrt{0.91}}{0.91} \approx 0.62897.$$

Also genügt das Gebiet G einer 0.63-Bedingung. □

Damit ist der Beweis des folgenden Satzes einfach (siehe D. Gaier (1964, S. 66)).

Satz 10.23 *Sei $G \subset \mathbb{C}$ ein Gebiet, welches einer ϵ-Bedingung mit einem $\epsilon \in (0, 1)$ genügt. Dann besitzt die Theodorsensche Integralgleichung (Theo) genau eine stetige Lösung.*

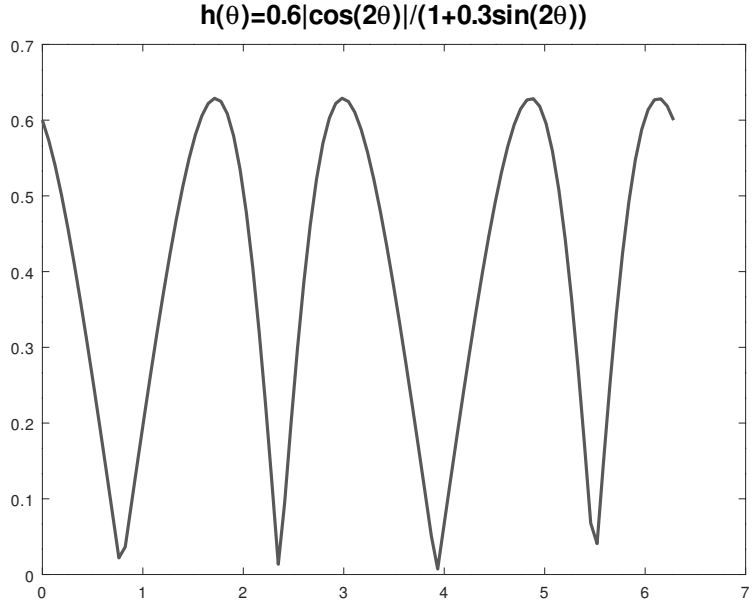


Abbildung 109: Die Funktion $h(\theta) := 0.6|\cos(2\theta)|/(1 + 0.3\sin(2\theta))$

Beweis: Die Existenz einer Lösung der Theodorsenschen Integralgleichung ist durch den Riemannschen Abbioldungssatz gesichert. Wir nehmen an, θ_1, θ_2 seien zwei auf $[0, 2\pi]$ stetige und 2π-periodische Lösungen der Theodorsenschen Integralgleichung (Theo), also

$$\theta_i(\phi) = \phi + K[\log \rho(\theta_i)](\phi) \quad \text{für alle } \phi \in [0, 2\pi], \quad i = 1, 2,$$

mit dem in Satz 10.21 angegebenen Operator K. Folglich ist

$$\begin{aligned}
\|\theta_2 - \theta_1\| &= \|K[\log \rho(\theta_2) - \log \rho(\theta_1)]\| \\
&\leq \|\log \rho(\theta_2) - \log \rho(\theta_1)\| \\
&\qquad \text{(wegen Satz 10.21)} \\
&= \left(\int_0^{2\pi} (\log \rho(\theta_2(\phi)) - \log \rho(\theta_1(\phi)))^2 \, d\phi \right)^{1/2} \\
&= \left(\int_0^{2\pi} \left(\int_{\theta_1(\phi)}^{\theta_2(\phi)} \frac{\rho'(t)}{\rho(t)} \, dt \right)^2 d\phi \right)^{1/2} \\
&\leq \epsilon \left(\int_0^{2\pi} (\theta_2(\phi) - \theta_1(\phi))^2 \, d\phi \right)^{1/2} \\
&= \epsilon \|\theta_2 - \theta_1\|.
\end{aligned}$$

Wegen $\epsilon \in (0, 1)$ folgt $\theta_1 = \theta_2$, was zu beweisen war. □

Beispiel: Die Polarkoordinatendarstellung einer Ellipse kann durch

$$\rho(\theta) := \frac{p}{1 + \epsilon \cos \theta}$$

erfolgen, wobei $p > 0$ und $\epsilon \in (0,1)$ die sogenannte *Exzentrizität* ist. Dann ist

$$h(\theta) := \frac{|\rho'(\theta)|}{\rho(\theta)} = \frac{\epsilon|\sin\theta|}{1+\epsilon\cos\theta} \leq \frac{\epsilon}{\sqrt{1-\epsilon^2}}.$$

Wegen Satz 10.23 besitzt die zu der Ellipse mit der angegebenen (Polarkoordinaten-) Darstellung gehörende Theodorsensche Integralgleichung eine eindeutige Lösung, falls $\epsilon/\sqrt{1-\epsilon^2} < 1$ bzw. $\epsilon < 1/\sqrt{2} \approx 0.70711$. □

10.6.4 Die numerische Behandlung der Theodorsenschen Integralgleichung

Jetzt wollen wir auf die numerische Behandlung der Theodorsenschen Integralgleichung eingehen, wobei wir uns im wesentlichen an D. Gaier (1964, S. 85 ff.) halten werden, die dortige oft etwas altmodische Terminologie (Vektorgleichung statt Gleichungssystem, "für die Handrechnung ist es bequem…") aber zu vermeiden suchen. Wir gehen nach wie vor von einem bezüglich des Nullpunktes sternigem Gebiet G aus, dessen Rand sich mit Hilfe der Funktion ρ beschreiben lässt. Die Theodorsensche Integralgleichung ist gegeben durch

$$\text{(Theo)} \qquad \theta(\phi) = \phi + \frac{1}{2\pi}\int_{0(H)}^{2\pi} \log\rho(\theta(\vartheta)) \cot\frac{\phi-\vartheta}{2}\, d\vartheta, \quad \phi \in [0,2\pi],$$

bzw.

$$\text{(Theo)} \qquad \theta(\phi) = \phi + K[\log\rho(\theta)](\phi), \quad \phi \in [0,2\pi],$$

wobei der Operator K durch

$$K(f)(\phi) := \frac{1}{2\pi}\int_{0(H)}^{2\pi} f(\vartheta)\cot\frac{\phi-\vartheta}{2}\, d\vartheta$$

definiert ist. *Lösung* von (Theo) ist also eine auf $[0,2\pi]$ stetige Funktion $\theta(\cdot)$ mit $\theta(0) = \theta(2\pi)$ und der Eigenschaft, dass (Theo) für alle $\phi \in [0,2\pi]$ erfüllt ist. Eine naheliegende Idee, das kontinuierliche Problem numerisch approximativ zu lösen, besteht darin, die folgenden Schritte durchzuführen.

1. Ersetze die Integralgleichung (Theo) mit der unbekannten, 2π-periodischen Lösung $\theta(\cdot)$ durch ein nichtlineares Gleichungssystem. Die Unbekannten in diesem Gleichungssystem sind $\theta_0,\ldots,\theta_{n-1}$, wobei θ_k eine Näherung für $\theta(\phi_k)$ bei vorgegebenen Stützstellen $0 \leq \phi_0 < \cdots < \phi_{n-1} < 2\pi$ ist, $k = 0,\ldots,n-1$. Wir werden uns gleich auf die Betrachtung *äquidistanter* Stützstellen beschränken, genauer sei im weiteren $\phi_k := 2\pi k/n$, $k = 0,\ldots,n-1$. Wir werden uns aus bestimmten Gründen auf *gerade* n beschränken, im weiteren sei also $n = 2m$ gerade. In diesem Schritt gehen wir also von einem kontinuierlichen zu einem diskreten Problem über, bzw. genauer zu einer Folge diskreter Probleme.

2. Löse das diskrete Problem bzw. das nichtlineare Gleichungssystem iterativ. Wir werden auf das sogenannte Gesamtschrittverfahren und auf das Newton-Verfahren eingehen.

Nun gehen wir genauer auf die angegebenen Schritte ein. Hierzu müssen wir ein wenig über trigonometrische Interpolation ins Gedächtnis zurückrufen. Der folgende Satz (siehe z. B. R. KRESS (1997, S. 165 ff.), J. WERNER (1992a, S. 148)) dient dazu, im ersten Schritt dem kontinuierlichen Problem ein diskretes bzw. endlichdimensionales Problem zuzuordnen. In dem Satz wird ausgesagt, dass bei äquidistanten Stützstellen $\phi_k := 2\pi k/n$ und beliebigen Stützwerten $f = (f_k)_{k=0,\dots,n-1}$, ein (eindeutiges) reelles trigonometrisches Polynom $T_m(f)$ vom Grad $\le m := \lfloor \frac{1}{2}n \rfloor$ mit $T_m(f)(\phi_k) = f_k$, $k = 0, \dots, n-1$, existiert. Genauer gilt:

Satz 10.24 *Sei $n \in \mathbb{N}$ und*

$$m := \begin{cases} \frac{1}{2}(n-1), & n \text{ ungerade,} \\ \frac{1}{2}n, & n \text{ gerade.} \end{cases}$$

Die äquidistanten Stützstellen $\phi_k := 2\pi k/n$ und Stützwerte $f = (f_k)_{k=0,\dots,n-1}$, seien gegeben. Sei

$$a_j := \frac{2}{n} \sum_{l=0}^{n-1} f_l \cos j\phi_l, \qquad j = 0, \dots, m,$$

$$b_j := \frac{2}{n} \sum_{l=0}^{n-1} f_l \sin j\phi_l, \qquad j = 1, \dots, m.$$

Definiert man dann das reelle trigonometrische Polynom $T_m(f)$ vom Grad $\le m$ durch

$$T_m(f)(\phi) := \begin{cases} \frac{a_0}{2} + \sum_{j=1}^{m} (a_j \cos j\phi + b_j \sin j\phi), & n = 2m+1, \\ \frac{a_0}{2} + \sum_{j=1}^{m-1} (a_j \cos j\phi + b_j \sin j\phi) + \frac{a_m}{2} \cos m\phi, & n = 2m, \end{cases}$$

so ist $T_m(f)(\phi_k) = f_k$, $k = 0, \dots, n-1$.

Bemerkung: Durch die n Interpolationsbedingungen ist das trigonometrische Interpolationspolynom $T_m(f)$ vom Grad $\le m$ *eindeutig* bestimmt, wenn man für gerades $n = 2m$ nur trigonometrische Polynome vom Grad $\le m$ mit $b_m = 0$, also verschwindendem höchsten sin-Term zulässt, weil dann die Anzahl der Unbekannten mit der Anzahl der Gleichungen übereinstimmt, und die *Existenz* eines trigonometrischen Interpolationspolynoms durch Satz 10.24 gesichert ist. □

Bei gegebenem $n \in \mathbb{N}$ definieren wir nun als Approximation der Abbildung K die Abbildung K_n, die mittels

$$K_n(f)(\phi) := \frac{1}{2\pi} \int_{0(H)}^{2\pi} T_m(f)(\vartheta) \cot \frac{\phi - \vartheta}{2} \, d\vartheta$$

einem Vektor $f = (f_k)_{k=0,\dots,n-1}$ die konjugierte Funktion des f an den Stützstellen $\phi_k := 2\pi k/n$, $k = 0, \dots, n-1$, interpolierenden trigonometrischen Polynoms $T_m(f)$ vom Grad $\le m := \lfloor \frac{1}{2}n \rfloor$ zuordnet. Es liegt nahe, bei gegebenem $n \in \mathbb{N}$ als diskretes Analogon

zur Theodorsenschen Integralgleichung die Aufgabe zu formulieren, $\theta = (\theta_k)_{k=0,\dots,n-1}$ zu bestimmen mit

$$\text{(Theo)}_n \qquad \theta_k = \phi_k + K_n[\log \rho(\theta)](\phi_k), \quad k = 0, \dots, k-1,$$

bzw.

$$\text{(Theo)}_n \qquad \theta = \phi + K_n[\log(\rho(\theta)].$$

Hierbei ist

$$\phi := (\phi_k)_{k=0,\dots,n-1}, \quad \log \rho(\theta) := (\log \rho(\theta_k))_{k=0,\dots,n-1}$$

und

$$K_n[\log \rho(\theta)] := (K_n[\log \rho(\theta)](\phi_k))_{k=0,\dots,n-1}.$$

Um das nichtlineare Gleichungssystem $(\text{Theo})_n$ mit n Unbekannten und n Gleichungen genauer analysieren zu können, geben wir bei gegebenem $f = (f_k)_{k=0,\dots,n-1}$ den Wert $K_n(f)(\phi_k)$ in Abhängigkeit von f genauer an. Wie in Satz 10.24 sei im folgenden $m := \lfloor \frac{1}{2} n \rfloor$ und

$$a_j := \frac{2}{n} \sum_{l=0}^{n-1} f_l \cos j\phi_l, \qquad b_j := \frac{2}{n} \sum_{l=0}^{n-1} f_l \sin j\phi_l \qquad (j = 0, \dots, m).$$

Wir betrachten im folgenden nur den Fall, dass $n = 2m$ *gerade* ist. Dann ist

$$\begin{aligned}
K_n(f)(\phi_k) &= \frac{1}{2\pi} \int_{0(H)}^{2\pi} T_n(f)(\vartheta) \cot \frac{\phi_k - \vartheta}{2} \, d\vartheta \\
&= \frac{1}{2\pi} \int_{0(H)}^{2\pi} \left(\frac{a_0}{2} + \sum_{j=1}^{m-1} (a_j \cos j\vartheta + b_j \sin j\vartheta) \right. \\
&\qquad \left. + \frac{a_m}{2} \cos m\vartheta \right) \cot \frac{\phi_k - \vartheta}{2} \, d\vartheta \\
&= \sum_{j=1}^{m-1} (a_j \sin j\phi_k - b_j \cos j\phi_k) + \frac{a_m}{2} \underbrace{\sin m\phi_k}_{=0} \\
&= \frac{2}{n} \sum_{j=1}^{m-1} \left[\sum_{l=0}^{n-1} (\sin j\phi_k \cos j\phi_l - \cos j\phi_k \sin j\phi_l) f_l \right] \\
&= \frac{2}{n} \sum_{l=0}^{n-1} \left[\sum_{j=1}^{m-1} (\sin j\phi_k \cos j\phi_l - \cos j\phi_k \sin j\phi_l) f_l \right] \\
&= \frac{2}{n} \sum_{\substack{l=0 \\ l \neq k}}^{n-1} \left(\sum_{j=0}^{m-1} \sin j(\phi_k - \phi_l) \right) f_l \\
&= \frac{2}{n} \sum_{\substack{l=0 \\ l \neq k}}^{n-1} \left(\sum_{j=0}^{m-1} \operatorname{Im}(e^{ij(\phi_k - \phi_l)}) \right) f_l
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} \mathrm{Im}\left(\sum_{j=0}^{m-1}(e^{i(\phi_k-\phi_l)})^j\right) f_l \\
&= \frac{2}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} \mathrm{Im}\left(\frac{1-e^{im(\phi_k-\phi_l)}}{1-e^{i(\phi_k-\phi_l)}}\right) f_l \\
&= \frac{2}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} (1-(-1)^{k-l})\mathrm{Im}\left(\frac{1}{1-e^{i(\phi_k-\phi_l)}}\right) f_l \\
&= \frac{2}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} (1-(-1)^{k-l})\mathrm{Im}\left(\frac{e^{-i(\phi_k-\phi_l)/2}}{e^{-i(\phi_k-\phi_l)/2}-e^{i(\phi_k-\phi_l)/2}}\right) f_l \\
&= \frac{1}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} (1-(-1)^{k-l})\cot\frac{\phi_k-\phi_l}{2} \\
&= \frac{1}{n}\sum_{\substack{l=0\\ l\neq k}}^{n-1} (1-(-1)^{k-l})\cot\frac{(k-l)\pi}{n}.
\end{aligned}
$$

Damit ist schließlich $K_n(f) = Af$ mit $A = (a_{kl})_{k,l=0,\ldots,n-1}$ und

$$
a_{kl} := \begin{cases} 0, & k-l \text{ gerade}, \\ \dfrac{2}{n}\cot\dfrac{(k-l)\pi}{n}, & k-l \text{ ungerade}. \end{cases}
$$

Die Aufstellung der sogenannten *Wittich-Matrix* A ist für ungerades n deutlich aufwendiger als für gerades n (siehe auch D. Gaier (1964, S. 76)). Daher werden wir im folgenden annehmen, $n = 2m$ sei gerade. Mit den oben eingeführten Bezeichnungen gilt (siehe D. Gaier (1964, S. 87) und A. Ostrowski (1952, S. 170)):

Satz 10.25 *Sei $G \subset \mathbb{C}$ ein Gebiet mit der Eigenschaft, dass jeder Randpunkt $w \in \partial G$ eine eindeutige Polardarstellung $w = \rho(\theta)e^{i\theta}$, $\theta \in [0, 2\pi]$, besitzt, wobei $\rho\colon [0, 2\pi] \longrightarrow \mathbb{R}_+$ eine (positive), 2π-periodische und stetige Funktion ist. Das Gebiet G genüge einer ϵ-Bedingung mit $\epsilon \in (0, 1)$, siehe Definition 10.22. Sei $n = 2m$ gerade. Dann besitzt das nichtlineare Gleichungssystem*

$$
\text{(Theo)}_n \qquad\qquad \theta = \phi + K_n[\log(\rho(\theta)] = \phi + A\log\rho(\theta)
$$

genau eine Lösung.

Beweis: Der Satz ist eine einfache Folgerung aus dem *Banachschen Fixpunktsatz* bzw. dem *Kontraktionssatz*. Diesen formulieren wir hier folgendermaßen (siehe z. B. J. Werner (1992a, S. 86)):

- Sei $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^n$ und $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ eine Abbildung. Es gelte:

1. $D \subset \mathbb{R}^n$ ist nichtleer und abgeschlossen.
2. F bildet D in sich ab, d. h. es ist $F(D) \subset D$.
3. F ist bezüglich der Norm $\|\cdot\|$ kontrahierend auf D mit der Lipschitzkonstanten $L \in (0,1)$, d. h. es ist

$$\|F(x) - F(y)\| \leq L\|x - y\| \quad \text{für alle } x, y \in D.$$

Dann gilt:

(a) Die Folge $\{x^{(k)}\}$ mit $x^{(k+1)} := F(x^{(k)})$, $k = 0, 1, \ldots$, konvergiert für jedes $x^{(0)} \in D$ gegen ein $x^* \in D$, den einzigen Fixpunkt von F in D.

(b) A priori Fehlerabschätzung:

$$\|x^{(k)} - x^*\| \leq \frac{L^k}{1-L}\|x^{(1)} - x^{(0)}\|, \qquad k = 1, 2, \ldots.$$

(c) A posteriori Fehlerabschätzung:

$$\|x^{(k)} - x^*\| \leq \frac{L}{1-L}\|x^{(k)} - x^{(k-1)}\|, \qquad k = 1, 2, \ldots.$$

Als Norm im $\mathbb{R}^n$ wählen wir die euklidische Norm $\|\cdot\|_2$, ferner definieren wir die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch

$$F(\theta) := \phi + K_n[\log \rho(\theta)] = \phi + A \log \rho(\theta),$$

wobei $A = (a_{kl})_{k,l=0,\ldots,n-1} \in \mathbb{R}^{n\times n}$ mit $m := n/2$ durch

$$(*) \qquad a_{kl} := \begin{cases} 0, & k-l \text{ gerade}, \\ \dfrac{1}{m} \cot \dfrac{(k-l)\pi}{n}, & k-l \text{ ungerade}. \end{cases}$$

gegeben ist. Es bleibt zu zeigen, dass die Abbildung F auf dem $\mathbb{R}^n$ kontrahiert. Für beliebige $\theta = (\theta_k)_{k=0,\ldots,n-1}, \eta = (\eta_k)_{k=0,\ldots,n-1} \in \mathbb{R}^n$ ist zunächst komponentenweise

$$|\log \rho(\theta_k) - \log \rho(\eta_k)| = \left| \int_{\eta_k}^{\theta_k} \frac{\rho'(t)}{\rho(t)}\, dt \right| \leq \epsilon |\theta_k - \eta_k|, \quad k = 0, \ldots, n-1,$$

dann

$$\|\log \rho(\theta) - \log \rho(\eta)\|_2 \leq \epsilon \|\theta - \eta\|_2$$

und damit

$$\|F(\theta) - F(\eta)\|_2 \leq \epsilon \|A\|_2 \|\theta - \eta\|_2.$$

Zu zeigen bleibt daher $\|A\|_2 \leq 1$. Hierzu überlegen wir uns, dass die folgende Aussage gilt:

- Ist $A \in \mathbb{R}^{n\times n}$ schiefsymmetrisch, d. h. ist $A = -A^T$, und haben alle Eigenwerte von A einen Betrag ≤ 1, ist also der Spektralradius von A kleiner oder gleich 1, so ist $\|A\|_2 \leq 1$.

Denn: Die Eigenwerte einer reellen, schiefsymmetrischen Matrix A sind gleich Null oder sind rein imaginär (und treten in konjugierten Paaren auf). Denn sei $\lambda \in \mathbb{C}$ ein Eigenwert von A mit zugehörigem Eigenvektor $x \in \mathbb{C}^n$. Dann ist, unter Berücksichtigung, dass eine A *reelle* Matrix ist

$$\lambda \|x\|_2^2 = \lambda \overline{x}^T x = \overline{x}^T A x = (Ax)^T \overline{x} = x^T A^T \overline{x} = -x^T A \overline{x} = -\overline{\lambda} x^T \overline{x} = -\overline{\lambda} \|x\|_2^2$$

und folglich $\lambda = -\overline{\lambda}$. Daher ist $\mathrm{Re}(\lambda) = 0$, also $\lambda = 0$ oder λ rein imaginär. Ist der Spektralradius einer schiefsymmetrischen Matrix kleiner oder gleich Eins, so gilt dies auch für den Spektralradius von $A^T A = -A^2$ und das bedeutet, dass $\|A\|_2 \leq 1$. Denn die (reellen, nichtnegativen) Eigenwerte von $A^T A = -A^2$ sind das Negative der Quadrate der Eigenwerte von A.

Nun berechnen wir die Eigenwerte der durch $(*)$ gegebenen schiefsymmetrischen Matrix A. Hierbei nutzen wir aus, dass A eine *zyklische* bzw. *zirkulante* Matrix ist, also die Form

$$A = \begin{pmatrix} a_0 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ a_2 & a_1 & a_0 & \cdots & a_3 \\ & \ddots & \ddots & \ddots & \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_0 \end{pmatrix}$$

besitzt. In unserem Falle ist

$$a_k := a_{k0} = \begin{cases} 0, & k \text{ gerade}, \\ \dfrac{1}{m} \cot \dfrac{k\pi}{2m}, & k \text{ ungerade}. \end{cases}$$

Die Eigenwerte (und die zugehörigen Eigenvektoren) einer zyklischen Matrix, insbesondere die Eigenwerte der durch $(*)$ gegebenen Matrix, können explizit angegeben werden. Denn mit

$$Z := \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

ist

$$A = a_0 I + a_1 Z + \cdots a_{n-1} Z^{n-1} = p(Z)$$

mit dem Polynom

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

vom Grad $n-1$. Die Eigenwerte (und zugehörigen Eigenvektoren) der speziellen zyklischen Matrix Z können angegeben werden. Die charakterische Gleichung ist nämlich $\det(\lambda I - A) = \lambda^n - 1$. Daher besitzt Z die n paarweise verschiedenen (auf dem Einheitskreis liegenden) Eigenwerte

$$\lambda_p := e^{2\pi i p/n} = e^{\pi i p/m}, \quad p = 0, \ldots, n-1.$$

Die Eigenwerte der zyklischen Matrix $A = p(Z)$ sind folglich

$$p(\lambda_p) = \sum_{k=0}^{n-1} a_k e^{2\pi i p k/n} = -\sum_{k=0}^{n-1} a_{0k} e^{ip\phi_k} = -K_n(f_p)(\phi_0), \quad p = 0, \ldots, n-1,$$

wobei $f_p(\phi) := e^{ip\phi}$ bzw. der Stützstellenvektor $f_p := (e^{ip\phi_k})_{k=0,\ldots,n-1}$, ist. Wegen $\phi_0 = 0$ und $T_m(f_p) = f_p$ für $p = 0, \ldots, m-1$, ist

$$p(\lambda_p) = -K_n(f_p)(0) = -K(f_p)(0) = \begin{cases} 0, & p = 0, \\ -i, & p = 1, \ldots, m-1. \end{cases}$$

Damit hat A zunächst den Eigenwert $p(\lambda_0) = 0$ und den $(m-1)$-fachen Eigenwert $-i$. Weiter ist

$$p(\lambda_m) = \sum_{k=0}^{n-1} a_k e^{\pi i k} = -\sum_{k=0}^{m-1} a_k = -p(\lambda_0) = 0.$$

Die restlichen $m-1$ Eigenwerte der reellen Matrix A sind natürlich die komplex konjugierten Werte zu dem $(m-1)$-fachen Eigenwert $-i$, also i. Insgesamt ist nachgewiesen, dass A den Spektralradius 1 besitzt und damit auch $\|A\|_2 \leq 1$ gilt. Daher ist die durch

$$F(\theta) := \phi + A \log \rho(\theta)$$

definierte Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ (auf dem $\mathbb{R}^n$) kontrahierend. Folglich besitzt F genau einen Fixpunkt bzw. das nichtlineare Gleichungssystem $(\text{Theo})_n$ genau eine Lösung. Damit ist der Satz bewiesen. □

Bemerkung: In Satz 10.25 wird die Existenz genau einer Lösung der diskreten Theodorsenschen Integralgleichung mit Hilfe des Banachschen Fixpunktsatzes bewiesen. Wegen dieses Fixpunktsatzes, den wir im Beweis von Satz 10.25 noch einmal zitiert haben, sind aber auch die folgenden Aussagen bewiesen bzw. leicht einzusehen (siehe D. Gaier (1964, S. 87)), wobei die Voraussetzungen von Satz 10.25 erfüllt seien:

- Für einen beliebigen Anfangswert $\theta^{(0)} \in \mathbb{R}^n$ konvergiert die durch

  $$\theta^{(k+1)} = \phi + A \log \rho(\theta^{(k)}), \quad k = 0, 1, \ldots,$$

  gewonnene Folge $\{\theta^{(k)}\}$ (D. Gaier (1964, S. 87) spricht vom *Gesamtschrittverfahren*) gegen die einzige Lösung θ^* von $(\text{Theo})_n$ und es gilt die Fehlerabschätzung

  $$\|\theta^{(k)} - \theta^*\|_2 \leq \frac{\epsilon^k}{1-\epsilon} \|\theta^{(1)} - \theta^{(0)}\|_2, \quad k = 1, 2, \ldots.$$

  Existiert eine Konstante $a > 0$ mit

  $$\frac{a}{1+\epsilon} \leq \rho(\theta) \leq a(1+\epsilon) \quad \text{für alle } \theta \in \mathbb{R},$$

  so gilt die Fehlerabschätzung

  $$\|\theta^{(k)} - \theta^*\|_2 \leq (\sqrt{n}\epsilon + \|\theta^{(0)} - \phi\|)\epsilon^k, \quad k = 1, 2, \ldots,$$

  für $\theta^{(0)} := \phi$ also

  $$\|\theta^{(k)} - \theta^*\|_2 \leq \sqrt{n}\epsilon^{k+1}, \quad k = 1, 2, \ldots.$$

Denn: Zu zeigen bleibt nur die zweite Aussage, da die erste eine unmittelbare Folgerung aus dem Kontraktionssatz ist. Nun ist

$$\begin{aligned}\|\theta^{(k)} - \theta^*\|_2 &= \|F(\theta^{(k-1)}) - F(\theta^*)\|_2 \\ &\leq \epsilon\|\theta^{(k-1)} - \theta^*\|_2 \\ &\vdots \\ &\leq \epsilon^k\|\theta^{(0)} - \theta^*\|_2 \\ &\leq \epsilon^k(\|\theta^{(0)} - \phi\|_2 + \|\phi - \theta^*\|_2)\end{aligned}$$

und

$$\begin{aligned}\|\theta^* - \phi\|_2 &= |A\log\rho(\theta^*)\|_2 \\ &= \left\|A\log\frac{\rho(\theta^*)}{a}\right\|_2 \\ &\qquad \text{(da 0 Eigenwert von } A \text{ mit Eigenvektor } e = (1,\ldots,1)^T) \\ &\leq \left\|\log\frac{\rho(\theta^*)}{a}\right\|_2 \\ &\leq \sqrt{n}\log(1+\epsilon) \\ &\qquad \text{(da } |\log\rho(\theta)/a| \leq \log(1+\epsilon) \text{ für alle } \theta \in \mathbb{R}) \\ &\leq \sqrt{n}\epsilon.\end{aligned}$$

Damit ist auch die zweite Aussage bewiesen. □

Bemerkung: Von M. H. Gutknecht (1977) ist mit Hilfe des Brouwerschen Fixpunktsatzes gezeigt worden, dass die diskrete Theodorsen'sche Integralgleichung bei stetigem $\rho(\cdot)$ mindestens eine Lösung besitzt. Dies ist sehr einfach einzusehen. Man setze nämlich (wir nutzen aus, dass $\rho(\cdot)$ 2π-periodisch ist)

$$C := \max_{t\in[0,2\pi]} |\log\rho(t)| = \max_{t\in\mathbb{R}} |\log\rho(t)|.$$

Die durch

$$F(\theta) := \phi + A\log\rho(\theta)$$

definierte Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ ist stetig. Für ein beliebiges $\theta \in \mathbb{R}^n$ ist ferner

$$\|F(\theta)\|_2 \leq \|\phi\|_2 + \underbrace{\|A\|_2}_{\leq 1}\|\log\rho(\theta)\|_2 \leq \|\phi\|_2 + C\sqrt{n} =: R.$$

Daher bildet F die Kugel $B[0;R] := \{\theta \in \mathbb{R}^n : \|\theta\|_2 \leq R\}$ in sich ab, besitzt also wegen des Brouwerschen Fixpunktsatzes mindestens eine Lösung. Von M. H. Gutknecht (1981) wird statt $\theta = \phi + A\log\rho(\theta)$ die äquivalente Aufgabe $y = A\log\rho(\phi+y)$ betrachtet und hierzu verschiedene Iterationsverfahren untersucht. □

Neben dem sogenannten Gesamtschrittverfahren (engl.: nonlinear Jacobi method)

$$\theta^{(k+1)} = \phi + A\log\rho(\theta^{(k)})$$

zur iterativen Lösung des nichtlinearen Gleichungssystems $(\text{Theo})_n$ wird von D. Gaier (1964, S. 89) auch noch ein Einzelschrittverfahren sowie eine Mittelung zwischen Gesamt- und Einzelschrittverfahren angegeben. Da hierüber offenbar keine Konvergenzaussagen bewiesen werden können, wollen wir darauf nicht näher eingehen. Stattdessen betrachten wir nun das Newton-Verfahren zur Lösung des nichtlinearen Gleichungssystems

$$(\text{Theo})_n \qquad \theta = \phi + A[\log(\rho(\theta)],$$

wobei nach wie vor $n = 2m$ gerade ist und die Matrix $A = (a_{kl})_{k,l=0,\dots,n-1} \in \mathbb{R}^{n\times n}$ durch

$$a_{kl} := \begin{cases} 0, & k-l \text{ gerade}, \\ \dfrac{1}{m}\cot\dfrac{(k-l)\pi}{n}, & k-l \text{ ungerade}. \end{cases}$$

gegeben ist. Das Newton-Verfahren zur iterativen Lösung einer Nullstellenaufgabe $F(\theta) = 0$, wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$, ist bekanntlich durch die Iterationsvorschrift

$$\theta^{(k+1)} := \theta^{(k)} - F'(\theta^{(k)})^{-1}F(\theta^{(k)})$$

gegeben, wobei $F'(\theta)$ die *Funktionalmatrix* von F in $\theta \in \mathbb{R}^n$ bedeutet und natürlich im k-ten Schritt davon ausgegangen wird, dass $F'(\theta^{(k)})$ nichtsingulär ist. In jedem Schritt ist also $\theta^{(k+1)}$ durch Lösen des linearen Gleichungssystems

$$F'(\theta^{(k)})\theta^{(k+1)} = F'(\theta^{(k)})\theta^{(k)} - F(\theta^{(k)})$$

zu bestimmen. In unserem speziellen Fall ist

$$F(\theta) := \theta - \phi - A\log\rho(\theta).$$

Wegen

$$\begin{aligned} F(\theta+\eta) - F(\theta) &= \eta - A[\log\rho(\theta+\eta) - \log\rho(\theta)] \\ &= \eta - A\begin{pmatrix} \dfrac{\rho'(\theta_0)}{\rho(\theta_0)}\eta_0 \\ \vdots \\ \dfrac{\rho'(\theta_{n-1})}{\rho(\theta_{n-1})}\eta_{n-1} \end{pmatrix} + o(\eta) \\ &= (I - AD(\theta))\eta + o(h), \end{aligned}$$

wobei

$$D(\theta) := \operatorname{diag}\left(\frac{\rho'(\theta_0)}{\rho(\theta_0)}, \dots, \frac{\rho'(\theta_{n-1})}{\rho(\theta_{n-1})}\right).$$

Die Funktionalmatrix $F'(\theta)$ von F an der Stelle θ ist daher durch

$$F'(\theta) = (I - AD(\theta))$$

gegeben. Beim Newton-Verfahren zur Lösung des nichtlinearen Gleichungssystems

$$\text{(Theo)}_n \qquad \theta = \phi + A \log \rho(\theta)$$

ist daher die neue Näherung $\theta^{(k+1)}$ aus der aktuellen Näherung $\theta^{(k)}$ mittels der Vorschrift

$$(I - AD(\theta^{(k)}))\theta^{(k+1)} = \phi + A(\log \rho(\theta^{(k)}) - D(\theta^{(k)})\theta^{(k)})$$

zu bestimmen. Gleichbedeutend hiermit ist, zunächst $d^{(k)}$ aus $F'(\theta^{(k)})d^{(k)} = -F(\theta^{(k)})$ bzw.

$$(I - AD(\theta^{(k)}))d^{(k)} = \phi + A \log \rho(\theta^{(k)}) - \theta^{(k)}$$

zu bestimmen und anschließend

$$\theta^{(k+1)} := \theta^{(k)} + d^{(k)}$$

zu setzen.

Bemerkung: Das Gebiet $G \subset \mathbb{C}$ genüge einer ϵ-Bedingung mit $\epsilon \in (0,1)$, siehe Definition 10.22. Dann ist die Funktionalmatrix $F'(\theta) = I - AD(\theta)$ für jedes $\theta \in \mathbb{R}^n$ nichtsingulär. Denn ist $F'(\theta)\eta = 0$, so ist $\eta = AD(\theta)\eta$ und folglich

$$\|\eta\|_2 = \|AD(\theta)\eta\|_2 \leq \underbrace{\|A\|_2}_{\leq 1} \underbrace{\|D(\theta)\|_2}_{\leq \epsilon} \|\eta\|_2 \leq \epsilon \|\eta\|_2,$$

wegen $\epsilon \in (0,1)$ folgt $\eta = 0$. □

Bekannte Konvergenzaussagen für das allgemeine Newton-Verfahren (siehe z. B. J. Werner (1992a, S. 102)) gelten entsprechend in dem hier vorliegenden speziellen Fall, darauf wollen wir nicht näher eingehen (siehe D. Gaier (1964, S. 92)). Bei D. Gaier (1964, S. 91) findet man die folgende einfache Konvergenzaussage, welche die unter einfachen Voraussetzungen vorliegende lokale Konvergenzgüte (superlineare bzw. quadratische Konvergenz bei hinreichend guter Ausgangsnäherung) nicht wiedergibt.:

Satz 10.26 *Sei $G \subset \mathbb{C}$ ein Gebiet mit der Eigenschaft, dass jeder Randpunkt $w \in \partial G$ eine eindeutige Polardarstellung $w = \rho(\theta)e^{i\theta}$, $\theta \in [0, 2\pi]$, besitzt, wobei $\rho\colon [0, 2\pi] \longrightarrow \mathbb{R}_+$ eine (positive), 2π-periodische und stetige Funktion ist. Das Gebiet G genüge einer ϵ-Bedingung mit $\epsilon \in (0,1)$, siehe Definition 10.22. Sei θ^* die nach Satz 10.25 eindeutig existierende Lösung von*

$$\text{(Theo)}_n \qquad \theta = \phi + K_n[\log(\rho(\theta)] = \phi + A \log \rho(\theta)$$

und $\{\theta^{(k)}\}$ die durch das Newton-Verfahren mit Anfangswert $\theta^{(0)}$ gewonnene Folge. Dann gilt:

1. *Es ist*

$$\|\theta^{(k)} - \theta^*\|_2 \leq \left(\frac{2\epsilon}{1-\epsilon}\right)^k \|\theta^{(0)} - \theta^*\|_2, \quad k = 0, 1, \ldots,$$

   und daher $\lim_{k\to\infty} \theta^{(k)} = \theta^*$, falls $\epsilon < \frac{1}{3}$.

2. Es ist

$$\|\theta^{(k+1)} - \theta^*\|_2 \leq \frac{2\epsilon}{1-\epsilon}\|\theta^{(k+1)} - \theta^{(k)}\|_2, \quad k = 0, 1, \ldots.$$

Beweis: Subtrahiert man von der Gleichung

$$\theta^{(k+1)} = \phi + A \log \rho(\theta^{(k)}) + AD(\theta^{(k)})(\theta^{(k+1)} - \theta^{(k)})$$

die Gleichung

$$\theta^* = \phi + A \log \rho(\theta^*),$$

so erhält man

$$\theta^{(k+1)} - \theta^* = A(\log \rho(\theta^{(k)}) - \log \rho(\theta^*)) + AD(\theta^{(k)})(\theta^{(k+1)} - \theta^{(k)})$$

und hieraus

$$\begin{aligned} \|\theta^{(k+1)} - \theta^*\|_2 &\leq \epsilon\|\theta^{(k)} - \theta^*\|_2 + \epsilon\|\theta^{(k+1)} - \theta^{(k)}\|_2 \\ &\leq \epsilon\|\theta^{(k)} - \theta^*\|_2 + \epsilon(\|\theta^{(k+1)} - \theta^*\|_2 + \|\theta^* - \theta^{(k)}\|_2) \end{aligned}$$

bzw.

$$\|\theta^{(k+1)} - \theta^*\|_2 \leq \frac{2\epsilon}{1-\epsilon}\|\theta^{(k)} - \theta^*\|_2, \quad k = 0, 1, \ldots,$$

woraus sofort die erste Aussage folgt. Die zweite Aussage erhält man aus

$$\begin{aligned} \|\theta^{(k+1)} - \theta^*\|_2 &\leq \epsilon\|\theta^{(k)} - \theta^*\|_2 + \epsilon\|\theta^{(k+1)} - \theta^{(k)}\|_2 \\ &\leq \epsilon(\|\theta^{(k)} - \theta^{(k+1)}\|_2 + \|\theta^{(k+1)} - \theta^*\|_2) + \epsilon\|\theta^{(k+1)} - \theta^{(k)}\|_2, \end{aligned}$$

woraus die zweite Aussage folgt. □

Auf die Abschätzung des Fehlers zwischen diskreter und kontinuierlicher Lösung der Theodorsenschen Integralgleichung wollen wir nicht mehr eingehen, siehe D. Gaier (1964, S. 92 ff.).

Die wesentliche Arbeit bei der Durchführung eines Iterationsschrittes beim Gesamtschrittverfahren (und anderer Iterationsverfahren, siehe M. H. Gutknecht (1981)) zur Lösung des diskreten besteht in der Berechnung des Matrix-Vektorproduktes Af, wobei $A \in \mathbb{R}^{n\times n}$ mit $n = 2m$ die Wittich-Matrix und $f \in \mathbb{R}^n$ ist. Bei naiver Herangehensweise erfordert dies $O(n^2)$ Multiplikationen. Dies kann aber wesentlich verbessert werden, wie M. H. Gutknecht (1979) zeigte. Hierauf wollen wir aber nicht mehr eingehen.

10.6.5 Ein numerisches Beispiel

Von D. Gaier (1964, S. 100) wird das nichtlineare Gleichungssystem $(\text{Theo})_n$ für ein Gebiet $G \subset \mathbb{C}$ näherungsweise gelöst, bei dem Randpunkte $w \in \partial G$ durch $w = \rho(\theta)e^{i\theta}$ gegeben sind, wobei $\rho\colon [0, 2\pi] \longrightarrow \mathbb{R}_+$ durch

$$\rho(\theta) := \sqrt{1 - (1-p^2)\cos^2\theta}$$

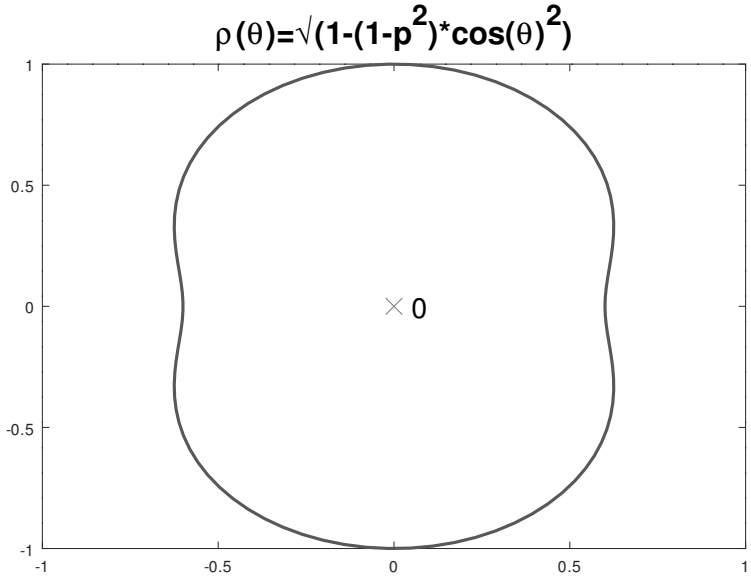


Abbildung 110: Gebiet mit $\rho(\theta) = \sqrt{1-(1-p^2)\cos^2\theta}$, $p = 0.6$

mit $p \in (0,1)$ definiert ist. In Abbildung 110 geben wir das zugehörige Gebiet G für $p = 0.6$ an. Zur Überprüfung der ϵ-Bedingung in Definition 10.22 notieren wir, dass

$$\epsilon(p) := \max_{\theta\in[0,2\pi]} \frac{|\rho'(\theta)|}{\rho(\theta)} = \max_{\theta\in[0,2\pi]} \frac{(1-p^2)|\cos\theta\sin\theta|}{1-(1-p^2)\cos^2\theta} = \frac{1-p^2}{2p}.$$

Denn

$$\begin{aligned} 2p|\cos\theta\sin\theta| - (1-(1-p^2)\cos^2\theta) &= 2p|\cos\theta\sin\theta| - \sin^2\theta - p^2\cos^2\theta \\ &= -(|\sin\theta| - p|\cos\theta|)^2 \\ &\le 0, \end{aligned}$$

wobei Gleichheit genau für $\theta = \arctan p$ eintritt. Es ist $\epsilon(p) < 1$, falls $p > \sqrt{2}-1$. Die Lösung der Theodorsenschen Integralgleichung (Theo) ist $\theta^*(\phi) = \arctan(p\tan\phi)$.

Wie bei D. Gaier (1964, S. 101 ff.) haben wir zunächst das Gesamtschrittverfahren mit $n = 36$ bzw. $m = 18$ angewandt, wobei der Startwert $\theta^{(0)} = \phi$ genommen wurde. Es wurde ein einfaches Programm mit dem MATLAB-ähnlichen Octave geschrieben, wobei die Symmetrien des Gebietes G bezüglich der reellen und imaginären Achse *nicht* ausgenutzt wurden. Zunächst wurde $p = 0.6$ gewählt, siehe Abbildung 110. Hier ist also die ϵ-Bedingung mit $\epsilon(p) \approx 0.5333$ erfüllt. Wir geben das Ergebnis der ersten 9 Iterationen an den Stellen $\phi_k := 2\pi k/36$, $k = 1,\ldots,8$, in Tabelle 1 an, wobei wir bei der Ausgabe `format short` gewählt haben. In der letzten Spalte ist die exakte Lösung θ^* der kontinuierlichen Aufgabe (Theo) angegeben.

Jetzt geben wir noch die entsprechenden Ergebnisse für das Newton-Verfahren an. Wieder wurde $n = 36$ bzw. $m = 18$ und der Startwert $\theta^{(0)} = \phi$ gewählt. In Tabelle 2 geben

k	$\theta^{(0)}$	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$\theta^{(6)}$	$\theta^{(7)}$	$\theta^{(8)}$	$\theta^{(9)}$	θ^*
1	0.17453	0.06323	0.11311	0.10658	0.10438	0.10537	0.10555	0.10540	0.10538	0.10540	0.10540
2	0.34907	0.15286	0.21881	0.22155	0.21313	0.21424	0.21543	0.21510	0.21492	0.21500	0.21501
3	0.52360	0.28103	0.32067	0.34424	0.33430	0.33117	0.33357	0.33396	0.33340	0.33337	0.33347
4	0.69813	0.44622	0.43781	0.47222	0.47213	0.46482	0.46514	0.46685	0.46671	0.46629	0.46641
5	0.87266	0.64094	0.59191	0.61548	0.62663	0.62223	0.61945	0.62035	0.62106	0.62087	0.62076
6	1.04720	0.85707	0.79101	0.79388	0.80568	0.80695	0.80468	0.80414	0.80457	0.80474	0.80463
7	1.22173	1.08767	1.02896	1.01874	1.02371	1.02634	1.02594	1.02543	1.02544	1.02554	1.02553
8	1.39626	1.32713	1.29336	1.28372	1.28394	1.28501	1.28511	1.28496	1.28494	1.28497	1.28497

Tabelle 1: Ergebnisse des Gesamtschrittverfahrens bei $p = 0.6$

k	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	θ^*
1	0.1081248245391915	0.1054062066458169	0.1054041767615610	0.1054041767577630	0.1054040985272823
2	0.2150361597515907	0.2150093179170700	0.2150064767055571	0.2150064766693733	0.2150066121104445
3	0.3258309220854920	0.3334908129394060	0.3334733392856593	0.3334733392381359	0.3334731722518320
4	0.4519685436756308	0.4663989068396909	0.4664114208674329	0.4664114209728364	0.4664115997988986
5	0.6063840963271301	0.6207270717707002	0.6207569868158371	0.6207569868630232	0.6207568104401688
6	0.7980405128297497	0.8046314577615986	0.8046335167598830	0.8046335167547549	0.8046336771011123
7	1.028435688619008	1.025517142381823	1.025525670567771	1.025525670578006	1.025525543825224
8	1.290666042336790	1.284964150470367	1.284965117283288	1.284965117274267	1.284965189025312

Tabelle 2: Ergebnisse des Newton-Verfahrens bei $p = 0.6$

wir die ersten Iterierten $\theta^{(1)}, \ldots, \theta^{(4)}$ und die exakte Lösung θ^* der kontinuierlichen Aufgabe (Theo) an den Stellen $\phi_k = 2\pi k/36$, $k = 1, \ldots, 8$, an. Diesmal haben wir bei der Ausgabe `format long` gewählt.

Jetzt wollen wir noch die entsprechenden Ergebnisse für $p = 0.3$ angeben. Hier ist $\epsilon(p) \approx 1.5167$ und daher die ϵ-Bedingung nicht mit einem $\epsilon \in (0, 1)$ erfüllt. Das entsprechende Gebiet wird in Abbildung 111 dargestellt. Das Gesamtschrittverfahren erweist sich als unbrauchbar, was mit der Beobachtung bei D. Gaier (1964, S. 102) übereinstimmt. Für das Newton-Verfahren wurde wieder $n = 36$ bzw. $m = 18$ und der Startwert $\theta^{(0)} = \phi$ gewählt. In Tabelle 3 geben wir die ersten Iterierten $\theta^{(1)}, \ldots, \theta^{(4)}$ und die exakte Lösung θ^* der kontinuierlichen Aufgabe (Theo) an den Stellen $\phi_k = 2\pi k/36$, $k = 1, \ldots, 8$, an, wobei wir bei der Ausgabe wirder `format long` gewählt haben.

k	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	θ^*
1	0.0658419099111064	0.0549129160829280	0.0531681934201397	0.0531346941963439	0.0528488369220771
2	0.0844237119372500	0.1189914060558913	0.1088423491897791	0.1084146171592609	0.1087601979784230
3	0.0781606466889441	0.2012991164565672	0.1722709540168950	0.1718377983765066	0.1715035540024133
4	0.1003664486994592	0.2631550506704262	0.2459861595615124	0.2462911995335085	0.2466061304128036
5	0.1995448383707780	0.3301328083911047	0.3434683216237466	0.3436760186254007	0.3433637690433021
6	0.4040226654098709	0.4666962687955464	0.4786657720442328	0.4789027084730283	0.4792163808447530
7	0.7174743269110464	0.6774032757845120	0.6894831174178525	0.6896618428062338	0.6893495841297208
8	1.119437950274170	1.030905952156330	1.039104041567822	1.039193274961155	1.039427968683735

Tabelle 3: Ergebnisse des Newton-Verfahrens bei $p = 0.3$

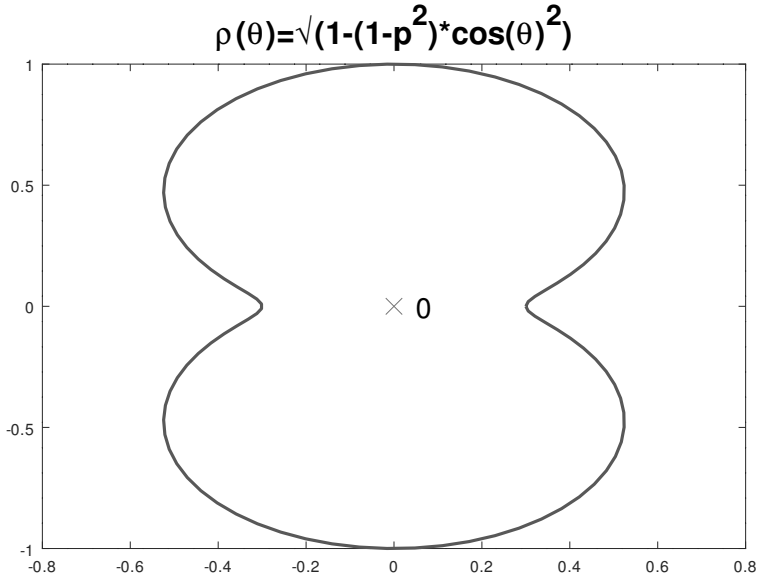


Abbildung 111: Gebiet mit $\rho(\theta) = \sqrt{1 - (1 - p^2)\cos^2\theta}$, $p = 0.3$

11 Der Satz von Poincaré-Bendixson

11.1 Beispiele, Einführung

Der Satz von Poincaré-Bendixson macht Aussagen über das qualitative Verhalten der Lösungen eines zweidimensionalen, autonomen Differentalgleichungssystems, also einer Aufgabe der Form

$$\text{(P)} \qquad \begin{cases} \dot{x}_1 &= f_1(x_1, x_2), \\ \dot{x}_2 &= f_2(x_1, x_2), \end{cases} \qquad \text{bzw.} \quad \dot{x} = f(x)$$

mit einer Abbildung $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, die wir i. Allg. als glatt, z. B. stetig differenzierbar voraussetzen, sodass die lokale Existenz und Eindeutigkeit der Lösung einer Anfangswertaufgabe gesichert ist. Besonders interessieren wir uns für *geschlossene Bahnen* (engl: *closed orbits, periodic orbits*) bzw. *geschlossene Trajektorien* in der Phasenebene $\mathbb{R}^2$ oder *periodische Lösungen*, also eine Lösung $x(\cdot)$ von (P), zu der es ein $T > 0$ mit $x(t) = x(t + T)$ für alle $t \in \mathbb{R}$ gibt. Ein Punkt $x \in \mathbb{R}^2$ mit $f(x) = 0$ heißt ein *stationärer Punkt* (*kritischer, singulärer Punkt*) oder ein *Gleichgewichtspunkt* zu (P). Dies sind sozusagen triviale geschlossene Bahnen bzw. periodische Lösungen. Die unabhängige Veränderliche wird stets mit t bezeichnet, Differentiation nach t wird durch einen Punkt $\dot{}$ gekennzeichnet. Ein Doppelpunkt $\ddot{}$ bedeutet daher eine zweimalige Differentiation nach t. Grob gesagt und nicht exakt formuliert kann eine intuitiv einfach verständliche Folgerung aus dem Satz von Poincaré-Bendixson folgendermaßen formuliert werden:

- Bleibt eine Bahn bzw. Trajektorie zu einem planaren autonomen Differentialgleichungssystems in einem beschränkten Gebiet, welches keinen kritischen Punkt enthält, so ist sie selbst eine geschlossene Bahn oder nähert sich einer solchen mit wachsender Zeit an.

Wir beginnen mit einigen Beispielen.

Beispiel: Die homogene Differentialgleichung zweiter Ordnung

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0$$

mit dem Parameter $\mu \geq 0$ heißt *van der Polsche Differentialgleichung*. Sie beschreibt eine nichtlinear gedämpfte Oszillation. Für große x ist $-\mu(1-x^2)$ positiv, d. h. es findet eine Dämpfung der durch x beschriebenen Bewegung statt. Für kleine x ist dagegen $-\mu(1 - x^2)$ negativ, d. h. es findet eine Anregung statt. Daher kann man vermuten, dass es eine periodische Lösung gibt. Schreibt man diese Differentialgleichung zweiter Ordnung in kanonischer Weise als ein System von zwei Differentialgleichungen erster Ordnung, so erhält man

$$\text{(vdPol)} \qquad \begin{cases} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \mu(1 - x_1^2)x_2 - x_1. \end{cases}$$

Für $\mu = 0$ (linearer harmonischer Oszillator) erhalten wir als Lösung von

$$\begin{cases} \dot{x}_1 &= x_2, & x_1(0) &= x_{10}, \\ \dot{x}_2 &= -x_1, & x_2(0) &= x_{20} \end{cases}$$

die Trajektorie

$$(x_1(t), x_2(t)) = (x_{10}\cos t + x_{20}\sin t, -x_{10}\sin t + x_{20}\cos t).$$

Diese stellen in der Ebene natürlich Kreise dar. In Abbildung 112 links wird für den Parameter $\mu = 2$ die von $x_0 = (x_1(0), x_2(0)) = (1, 0)$ startende Trajektorie $(x_1(t), x_2(t))$ über das Zeitintervall $[0, 30]$ dargestellt, während rechts die entsprechende Trajektorie für $\mu = 0.5$ abgebildet wird. In beiden Beispielen, die sich nur in der Wahl des Parameters μ unterscheiden ($\mu = 2$ bzw. $\mu = 0.5$), geben wir nun noch in Abbildung 113 die bei $x_0 = (5, 3)$ startenden Trajektorien an. Wir erkennen, dass diese sich in allen Abbildungen einem sogenannten *Grenzzyklus* annähern (das wird präzisiert werden müssen) und vermuten, dass es wie im Fall $\mu = 0$ (hier für *jeden* Anfangszustand) auch für $\mu > 0$ (für einen bestimmten Anfangszustand) *geschlossene* Trajektorien gibt, also eine Lösung (x_1, x_2) von (vdPol), zu der es ein $T > 0$ mit $(x_1(t), x_2(t)) = (x_1(t + T), x_2(t + T))$ für alle t gibt. Zum Schluss dieses Beispiels bemerken wir noch, dass man die van der Polsche Differentialgleichung zweiter Ordnung auch auf etwas andere Weise als ein System von zwei Differentialgleichungen erster Ordnung schreiben kann, nämlich als

$$\dot{x}_1 \;=\; x_2 - \mu\left(\frac{x_1^3}{3} - x_1\right),$$

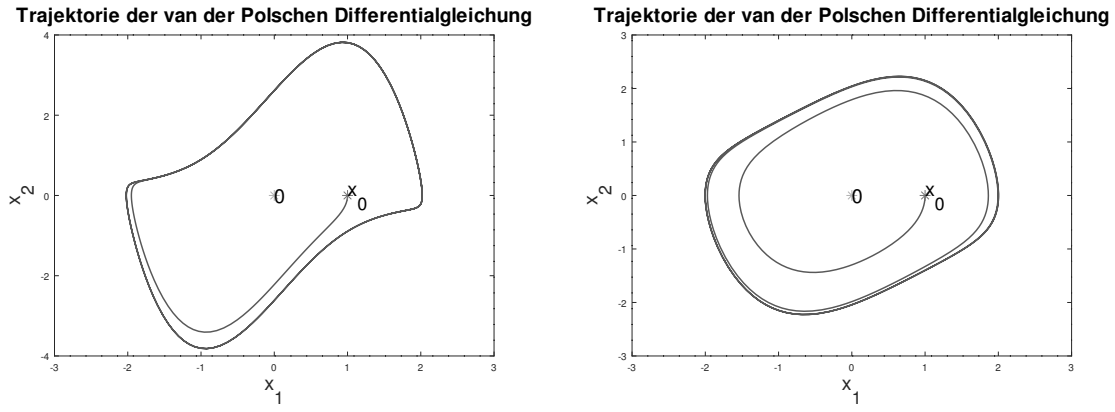


Abbildung 112: Trajektorie der van der Polschen Differentialgleichung für $\mu = 2$ bzw. $\mu = 0.5$ mit Startwert $x_0 = (1, 0)$

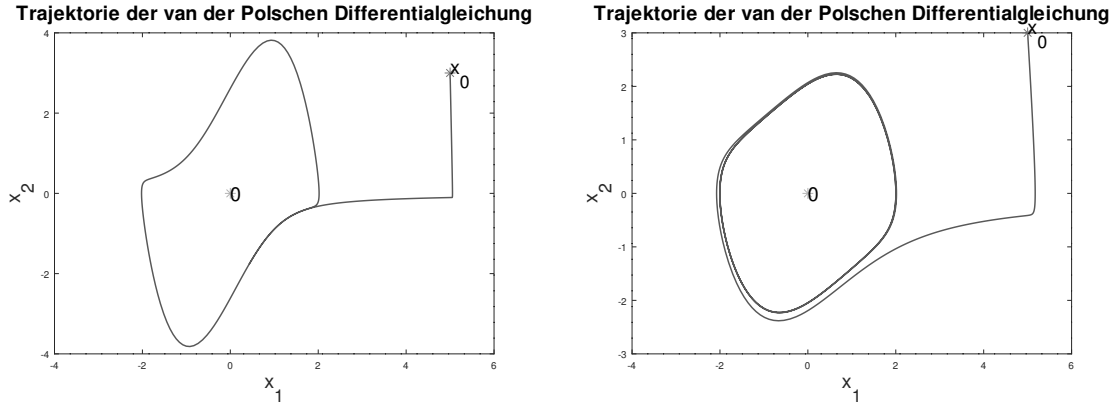


Abbildung 113: Trajektorie der van der Polschen Differentialgleichung für $\mu = 2$ bzw. $\mu = 0.5$ mit Startwert $x_0 = (5, 3)$

$$\dot{x}_2 \;=\; -x_1.$$

Dies wird beim Beweis zur Existenz geschlossener Orbits der van der Polschen Differentialgleichung ausgenutzt, siehe z. B. G. Teschl (2012, S. 219). □

Beispiel: Wir wollen das Wachstum von zwei Arten untersuchen, die sich gegenseitig beeinflussen und die wir Räuber und Beute (engl.: predator, prey) nennen. Man stelle sich etwa Raub- und Beutefische in der Adria vor. Dies war der Ausgangspunkt für die Untersuchungen von V. Volterra (1860-1940) und A. J. Lotka (1880-1949). Sei $x_1(t)$ die Population der Beute, $x_2(t)$ die Population der Räuber zur Zeit t. Falls genügend Nahrung für die Beute vorhanden ist, so dass sich diese nicht gegenseitig das Futter wegzunehmen brauchen, und keine Räuber vorhanden sind, ist in erster Näherung $\dot{x}_1 = ax_1$ mit einer positiven Konstanten a. Andererseits ist die Anzahl der "Kontakte" zwischen Räuber und Beute proportial zu x_1x_2, so dass $\dot{x}_1 = ax_1 - bx_1x_2$ mit einer positiven Konstanten b. Ist dagegen keine Beute vorhanden, so sterben die Räuber aus: $\dot{x}_2 = -cx_2$. Andererseits ist die Zuwachsrate proportional zu x_1x_2, so dass die zeitliche

Änderung der Räuberpopulation durch $\dot{x}_2 = -cx_2 + dx_1x_2$ gegeben ist. Insgesamt erhält man ein System von zwei Differentialgleichungen erster Ordnung, das sogenannte Lotka-Volterra-System:

$$\begin{aligned} \dot{x}_1 &= ax_1 - bx_1x_2, \\ \dot{x}_2 &= -cx_2 + dx_1x_2, \end{aligned}$$

wobei a, b, c, d positive Konstanten sind. Sind positive Anfangspopulationen x_{10}, x_{20} zur Zeit $t = 0$ vorgegeben, so können die zukünftigen Räuber- bzw. Beute-Populationen $x_1(t)$ bzw. $x_2(t)$ durch (numerisches) Lösen der Anfangswertaufgabe

$$\begin{aligned} \dot{x}_1 &= ax_1 - bx_1x_2, & x_1(0) &= x_{10}, \\ \dot{x}_2 &= -cx_2 + dx_1x_2, & x_2(0) &= x_{20} \end{aligned}$$

bestimmt werden. In Abbildung 114 geben wir für die Werte

$$a = 2, \quad b = 0.01; \quad c = 1, \quad d = 0.01, \qquad (x_{10}, x_{2,0}) = \begin{cases} (300, 150) \\ (400, 200) \end{cases}$$

die zugehörigen Trajektorien $(x_1(t), x_2(t))$ an. Man erkennt, dass sich geschlossene Bah-

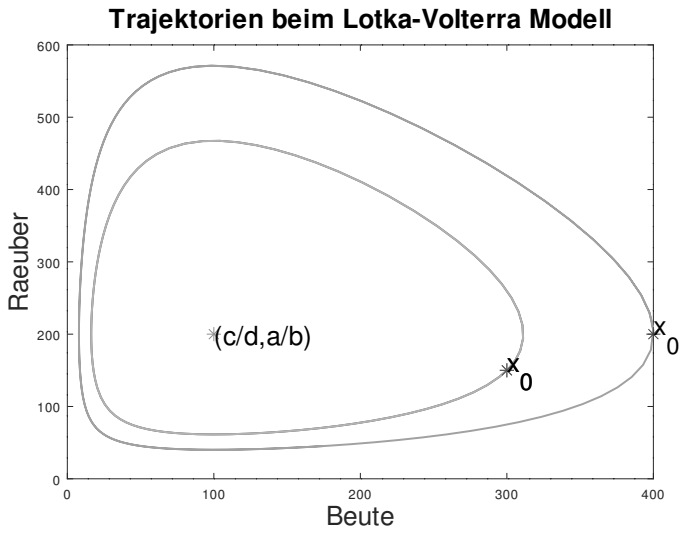


Abbildung 114: Trajektorien beim Lotka-Volterra Modell

nen ergeben. Diese enthalten den Gleichgewichtspunkt $(c/d, a/b)$ im Inneren. Bei M. W. Hirsch et al. (2004, S. 242) wird gezeigt, dass dies für jeden Startwert im offenen ersten Quadranten außer dem Gleichgewichtspunkt richtig ist. □

Beispiel: Wir betrachten das zweidimensionale autonome System

$$\begin{aligned} \dot{x}_1 &= -x_2 + x_1(1 - x_1^2 - x_2^2), \\ \dot{x}_2 &= x_1 + x_2(1 - x_1^2 - x_2^2). \end{aligned}$$

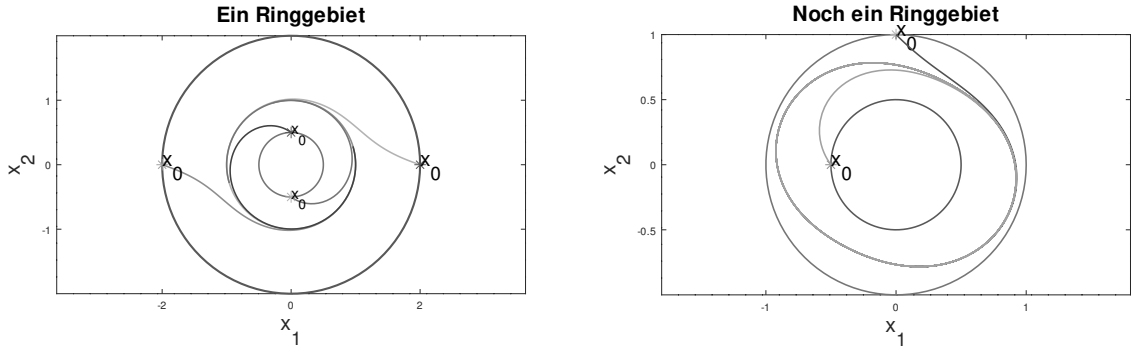


Abbildung 115: Ringgebiete für den Satz von Poincaré-Bendixson

In Abbildung 115 links geben wir zwei Kreise in der Ebene an mit dem Nullpunkt als Mittelpunkt, der eine vom Radius 2, der andere mit dem Radius $\frac{1}{2}$. Ferner stellen wir Trajektorien dar, die von diesen Kreisen in einem Punkt x_0 starten, und zwar zwei vom Kreis mit dem größeren Radius und zwei vom kleineren. Man erkennt, dass die vom größeren Kreis startenden Trajektorien ins Innere des Ringgebiete zwischen den beiden Kreisen geworfen werden, das entsprechende gilt auch für die auf dem kleineren Kreis startenden Trajektorien und dass diese sich jeweils einer geschlossenen Bahn, nämlich dem Kreis mit dem Radius 1 annähern. In der Tat ist $(x_1^*(t), x_2^*(t)) = (\cos t, \sin t)$ eine *periodische* Lösung des gegebenen Differentialgleichungssystems. Dieses besitzt offenbar $(0, 0)$ als einzigen Gleichgewichtspunkt. Daher ist insbesondere kein Gleichgewichtspunkt im Ringgebiet $R = \{x \in \mathbb{R}^2 : \frac{1}{2} < \|x\|_2 < 2\}$ enthalten.
Ein ganz ähnliches Beispiel ist durch

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2 - x_1(x_1^2 + 3x_2^2), \\ \dot{x}_2 &= -x_1 + x_2 - 2x_2^3 \end{aligned}$$

gegeben. In Abbildung 115 rechts sieht man zwei Kreise mit dem Nullpunkt als Mittelpunkt und $\frac{1}{2}$ bzw. 1 als Radius. Von jedem dieser Kreise startet in einem Punkt x_0 eine Trajektorie, die in dem durch die beiden Kreise gegebenen Ringgebiet bleiben und sich jeweils ein und demselben Grenzzyklus annähern. Dieser ist diesmal offensichtlich kein Kreis. Es stellt sich die Frage, wie man *beweisen* kann, dass eine in einem Ringgebiet startende Trajektorie dort bleibt und unter welchen Voraussetzungen dann die Existenz einer zu dem gegebenen autonomen System gehörende geschlossenen Bahn bzw. periodischen Lösung gefolgert werden kann. Im englischen heißt so ein Ringgebiet, in dem eine dort startende Trajektorie bis in alle Zukunft gefangen ist, auch eine *trapping region*. □

Beispiel: Bei M. W. Hirsch, S. Smale, R. L. Devaney (2004, S. 217), siehe auch J. Luk (2017) findet man das folgende autonome System:

$$\dot{x}_1 = \sin x_1 \left(-\frac{1}{10} \cos x_1 - \cos x_2 \right),$$

$$\dot{x}_2 \;=\; \sin x_2 \left(\cos x_1 - \frac{1}{10} \cos x_2 \right).$$

Dann sind $(\frac{\pi}{2}, \frac{\pi}{2})$, $(0,0)$, $(0,\pi)$, $(\pi,0)$ und (π,π) Gleichgewichtspunkte dieses Systems. In Abbildung 116 geben wir eine bei $(2,2)$ und eine bei $(\frac{1}{2}, \frac{1}{2})$ startende Trajektorie zu dem gegebenen System an. Beide Trajektorien nähern sich den Seiten des Quadrates

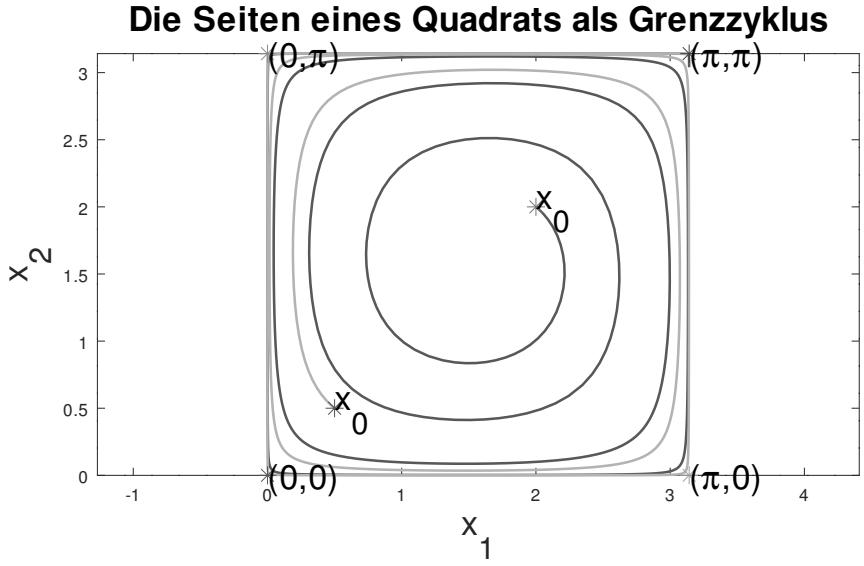


Abbildung 116: Die Seiten eines Quadrats als Grenzzyklus

$[0,\pi] \times [0,\pi]$ an, sodass diese sich als Grenzzyklus des gegebenen autonomen Systems ergeben. □

11.2 Wann existiert zu einem ebenen autonomen System keine geschlossene Bahn?

Wir beginnen mit einem einfachen Resultat, bei welchem noch nicht vorausgesetzt wird, dass das gegebene autonome System planar ist, also ein Differentialgleichungssystem von *zwei* Differentialgleichungen erster Ordnung ist.

Satz 11.1 *Sei $V\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig differenzierbar. Dann hat das sogenannte Gradientensystem*

$$\begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{pmatrix} = \dot{x} = -\nabla V(x) = - \begin{pmatrix} \dfrac{\partial V}{\partial x_1}(x_1, \ldots, x_n) \\ \vdots \\ \dfrac{\partial V}{\partial x_n}(x_1, \ldots, x_n) \end{pmatrix}$$

(das Auftreten des Minuszeichens hat nur kosmetische Gründe) keine nichttriviale periodische Lösung.

Beweis: Sei $x(\cdot)$ eine T-periodische Lösung von $\dot{x} = -\nabla V(x)$ bzw.

$$\gamma := \{(x_1(t), x_2(t)) \in \mathbb{R}^2 : t \in [0, T]\}$$

eine geschlossene Bahn. Dann ist

$$\begin{aligned} 0 &= V(x(T)) - V(x(0)) \\ &= \int_0^T \frac{d}{dt} V(x(t))\, dt \\ &= \int_0^T \nabla V(x(t))^T \dot{x}(t)\, dt \\ &= -\int_0^T \|\dot{x}(t)\|^2\, dt \\ &\leq 0. \end{aligned}$$

Also ist $x(\cdot)$ auf dem Intervall $[0, T]$ konstant und daher eine triviale periodische Lösung bzw. ein Gleichgewichtspunkt des autonomen Systems. □

In den folgenden beiden Sätzen (siehe z. B. F. Verhulst (1990, S. 39)) gehen wir auf den planaren bzw. zweidimensionalen Fall ein. Gegeben sei also das Differentialgleichungssystem

$$\text{(P)} \qquad \dot{x} = f(x) \qquad \text{bzw.} \qquad \begin{cases} \dot{x}_1 &= f_1(x_1, x_2), \\ \dot{x}_2 &= f_2(x_1, x_2). \end{cases}$$

Hierbei wird in den folgenden beiden Sätzen von Bendixson und Dulac vorausgesetzt, dass f_1 und f_2 stetig partiell differenzierbar in dem *einfach zusammenhängenden* Gebiet $D \subset \mathbb{R}^2$ sind. Grob gesagt bedeutet dies, dass in D keine Löcher enthalten sind.

Satz 11.2 (Bendixson) *Ist*

$$\operatorname{div} f(x_1, x_2) := \frac{\partial f_1}{\partial x_1}(x_1, x_2) + \frac{\partial f_2}{\partial x_2}(x_1, x_2)$$

auf D von einem Vorzeichen, ohne identisch zu verschwinden, so besitzt (P) in D keine nichttriviale geschlossene Bahn bzw. keine nichtkonstante periodische Lösung.

Beweis: Angenommen, $\gamma \subset D$ sei eine geschlossene Bahn zum planaren System (P) und G das von γ eingeschlossene Gebiet. Der Greensche Satz liefert

$$\iint_G \left(\frac{\partial f_1}{\partial x_1}(x_1, x_2) + \frac{\partial f_2}{\partial x_2}(x_1, x_2) \right) dx_1 dx_2 = \int_\gamma (-f_2(x_1, x_2) dx_1 + f_1(x_1, x_2) dx_2).$$

Die linke Seite dieser Gleichung ist nach Voraussetzung von Null verschieden, während die rechte Seite auf γ verschwindet, da dort $dx_2/dx_1 = f_2/f_1$. Damit ist der Satz bewiesen. □

Die folgende Verallgemeinerung des Bendixson-Kriteriums für die Nichtexistenz einer geschlossenen Bahn kann praktisch mit dem selben Beweis verifiziert werden.

Satz 11.3 (Dulac) *Sei $\phi\colon D \longrightarrow \mathbb{R}$ stetig partiell differenzierbar, ferner sei*

$$\operatorname{div}(\phi f)(x_1, x_2) = \frac{\partial(\phi f_1)}{\partial x_1}(x_1, x_2) + \frac{\partial(\phi f_2)}{\partial x_2}(x_1, x_2)$$

auf D von einem Vorzeichen, ohne identisch zu verschwinden. Dann besitzt (P) in D keine nichttriviale geschlossene Bahn bzw. keine nichtkonstante periodische Lösung.

Nun geben wir einige Beispiele für die Anwendung der letzten beiden Sätze an.

Beispiel: Man betrachte das System

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& f_1(x_1, x_2) \;=\; x_2, \\ \dot{x}_2 &=& f_2(x_1, x_2) \;=\; ax_1 + bx_2 - x_1^2 x_2 - x_1^3. \end{array} \right.$$

Dann ist

$$\frac{\partial f_1}{\partial x_1}(x_1, x_2) + \frac{\partial f_2}{\partial x_2}(x_1, x_2) = b - x_1^2,$$

sodass für $b < 0$ für das System keine nichtkonstante periodische Lösung existiert. □

Beispiel: Wir wollen uns überlegen, dass das System

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& f_1(x_1, x_2) \;=\; x_1(2 - x_1 - x_2), \\ \dot{x}_2 &=& f_2(x_1, x_2) \;=\; x_2(4x_1 - x_1^2 - 3) \end{array} \right.$$

in $\mathbb{R}^2_+ := \{x = (x_1, x_2) \in \mathbb{R}^2 : x_1 > 0,\, x_2 > 0\}$ keine periodische Lösung besitzt. Für $(x_1, x_2) \in \mathbb{R}^2_+$ ist

$$\operatorname{div} f(x_1, x_2) := \frac{\partial f_1}{\partial x_1}(x_1, x_2) + \frac{\partial f_2}{\partial x_2}(x_1, x_2) = \underbrace{-1 + 2x_1 - x_1^2}_{\leq 0} \underbrace{-x_2}_{<0} < 0.$$

Wegen des Bendixson-Kriteriums in Satz 11.2 gibt es in $\mathbb{R}^2_+$ keine nichttriviale geschlossene Bahn. □

Beispiel: Wir betrachten (siehe F. Verhulst (1990, S. 40)) erneut die van der Polsche Differentialgleichung

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0$$

bzw. das zugehörige System

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& x_2, \\ \dot{x}_2 &=& \mu(1 - x_1^2)x_2 - x_1. \end{array} \right.$$

Als Divergenz der Vektorfunktion

$$f(x_1, x_2) = \begin{pmatrix} x_2 \\ \mu(1 - x_1^2)x_2 - x_1 \end{pmatrix}$$

erhält man $\operatorname{div} f(x_1, x_2) = \mu(1 - x_1^2)$. Daher gibt es keine geschlossene Bahn in $D := \{(x_1, x_2) \in \mathbb{R}^2 : |x_1| < 1\}$, d. h. eine geschlossene Bahn muss $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -1\}$ oder $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = +1\}$ schneiden. □

Beispiel: Wir wollen uns überlegen, dass das System

$$\left\{\begin{array}{rcccl} \dot{x}_1 & = & f_1(x_1,x_2) & = & x_2, \\ \dot{x}_2 & = & f_2(x_1,x_2) & = & -x_1 - x_2 + x_1^2 + x_2^2 \end{array}\right.$$

in $\mathbb{R}^2$ keine geschlossene Bahn besitzt. Es ist $\operatorname{div} f(x_1,x_2) = -1 + 2x_2$, sodass wir mit Satz 11.2 erhalten, dass in den Halbräumen $\{(x_1,x_2) \in \mathbb{R}^2 : x_2 < \frac{1}{2}\}$ und $\{(x_1,x_2) \in \mathbb{R}^2 : x_2 > \frac{1}{2}\}$ keine nichttriviale geschlossene Bahn liegt. Damit ist aber die gestellte Behauptung nicht bewiesen. Wir wenden nun Satz 11.3 mit $D := \mathbb{R}^2$ und $\phi(x_1,x_2) := e^{\alpha x_1}$ mit geeignetem $\alpha \in \mathbb{R}$ an. Nun ist

$$\operatorname{div}(\phi f)(x_1,x_2) = (\alpha+2)e^{\alpha x_1}x_2 - e^{\alpha x_1}.$$

Wählt man daher $\alpha := -2$, so ist $\operatorname{div}(\phi f)(x_1,x_2) = -e^{\alpha x_1} < 0$ auf $\mathbb{R}^2$ und der Satz 11.3 liefert, dass es keine nichttriviale geschlossene Bahn im $\mathbb{R}^2$ gibt. □

11.3 Hilfsmittel zum Beweis des Satzes von Poincaré-Bendixson

Der Satz von Poincaré-Bendixson macht Aussagen über Lösungen *planarer* Differentialgleichungssysteme. Bei den folgenden Definitionen und Bezeichnungen (wir halten uns im wesentlichen an F. Verhulst (1990, S. 41 ff.), J. K. Hale (1969, S. 46 ff.) und G. Teschl (2012, S. 192 ff.)) ist es allerdings *nicht* nötig, sich auf den planaren Fall zu spezialisieren. Hilfsmittel für den uns eigentlich interessierenden planaren Fall werden im nächsten Unterabschnitt bereit gestellt. Wir gehen daher jetzt von einem autonomen Differentialgleichungssystem

$$\dot{x} = f(x)$$

aus, bei dem $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ einmal stetig partiell differenzierbar ist und eine Lösung $x(t) = x(t;x_0)$ der Anfangswertaufgabe $\dot{x} = f(x)$, $x(0) = x_0$ *für alle* $t \in \mathbb{R}$ existiert. Mit

$$\gamma(x_0) := \{x(t;x_0) : t \in \mathbb{R}\}$$

bezeichnen wir die zugehörige *Bahn* oder *Trajektorie* im zugehörigen *Phasenraum* $\mathbb{R}^n$. Ist also $x(t_1;x_0) = x_1$, so ist $\gamma(x_0) = \gamma(x_1)$. Gelegentlich unterscheiden wir zwischen der *positiven* und der *negativen Bahn*. Naheliegenderweise (x ist nach wie vor die Lösung der Anfangswertaufgabe $\dot{x} = f(x)$, $x(0) = x_0$) definieren wir

$$\begin{array}{rcll} \gamma^+(x_0) & := & \{x(t) : t \geq 0\} & \text{(positive Bahn)}, \\ \gamma^-(x_0) & := & \{x(t) : t \leq 0\} & \text{(negative Bahn)}. \end{array}$$

Weiter heißt eine Menge $M \subset \mathbb{R}^n$ *invariant*, wenn mit einem $x_0 \in M$ die gesamte Bahn durch x_0 in M liegt, also $\gamma = \{x(t;x_0) : t \in \mathbb{R}\} \subset M$ gilt. Entsprechend nennen wir eine Menge M *positiv* bzw. *negativ invariant*, wenn mit $x_0 \in M$ die gesamte positive Bahn $\gamma^+(x_0)$ bzw. negative Bahn $\gamma^-(x_0)$ in M enthalten ist.

Definition 11.4 Sei $\gamma = \{x(t) : t \in \mathbb{R}\}$ eine Bahn, wobei $x(\cdot)$ eine Lösung von $\dot{x} = f(x)$ ist. Dann heißt

$$\omega(\gamma) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} \text{Es existiert eine Folge } \{t_k\} \subset \mathbb{R} \text{ mit} \\ \lim_{k\to\infty} t_k = \infty,\ p = \lim_{k\to\infty} x(t_k) \end{array} \right\}$$

die *ω-Grenzmenge* (ω-limitset) zur Bahn γ. Elemente von $\omega(\gamma)$ heißen *positive Grenzpunkte* (positive limitpoints). Entsprechend heißt

$$\alpha(\gamma) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} \text{Es existiert eine Folge } \{t_k\} \subset \mathbb{R} \text{ mit} \\ \lim_{k\to\infty} t_k = -\infty,\ p = \lim_{k\to\infty} x(t_k) \end{array} \right\}$$

die *α-Grenzmenge* (α-limitset) zur Bahn γ. Elemente von $\alpha(\gamma)$ heißen *negative Grenzpunkte* (negative limitpoints). Die ω-Grenzmenge $\omega(\gamma^+)$ einer positiven Bahn γ^+ und dir α-Grenzmenge $\alpha(\gamma^-)$ einer negativen Bahn γ^- sind natürlich ganz entsprechend definiert.

Bemerkung: Die Bezeichnungen für die Grenzmengen $\alpha(\gamma)$ und $\omega(\gamma)$ bei z. B. J. K. Hale (1969) und F. Verhulst (1990) finde ich schon fast ein wenig witzig: Man denkt an α und ω, Anfang und Ende. Dagegen sind die entsprechenden Bezeichnungen $\omega_-(\gamma)$ und $\omega_+(\gamma)$ bzw. $\omega_\sigma(\gamma)$ mit $\sigma \in \{+,-\}$ bei G. Teschl (2012) m. E. nur aus formalen Gründen sinnvoll. □

Im folgenden Satz (siehe z. B. J. K. Hale (1969, S. 47), F. Verhulst (1990, S. 43), G. Teschl (2012, S. 193 ff.)) werden Eigenschaften von Grenzmengen formuliert.

Satz 11.5 *Die Grenzmengen $\alpha(\gamma)$ und $\omega(\gamma)$ zu einer Bahn γ sind abgeschlossen und invariant. Mit $x(\cdot; x_0)$ bezeichnen wir die Lösung der Anfangswertaufgabe $\dot{x} = f(x)$, $x(0) = x_0$. Ist die positive Bahn $\gamma^+ = \gamma^+(x_0) := \{x(t; x_0) : t \geq 0\}$ beschränkt, so ist die ω-Grenzmenge $\omega(\gamma)$ kompakt, nichtleer und zusammenhängend. Ferner gilt*

$$\lim_{t\to\infty} d(x(t; x_0), \omega(\gamma)) = 0,$$

wobei

$$d(x(t; x_0), \omega(\gamma)) := \inf_{p\in\omega(\gamma)} \|x(t; x_0) - p\|_2$$

der Abstand von $x(t; x_0)$ zu $\omega(\gamma)$ ist. Eine entsprechende Aussage gilt für die α-Grenzmenge $\alpha(\gamma)$, wenn die negative Bahn $\gamma^- = \gamma^-(x_0)$ beschränkt ist.

Beweis: Sei $p \in \mathrm{cl}\,(\omega(\gamma))$. Dann existiert eine Folge $\{p_j\}_{j\in\mathbb{N}} \subset \omega(\gamma)$ mit

$$\|p_j - p\| \leq \frac{1}{j}, \quad j \in \mathbb{N}.$$

Wegen $\{p_j\}_{j\in\mathbb{N}} \subset \omega(\gamma)$ existiert für jedes $j \in \mathbb{N}$ eine Folge $\{t_{kj}\}_{k\in\mathbb{N}}$ mit

$$\lim_{k\to\infty} t_{kj} = \infty, \qquad \lim_{k\to\infty} x(t_{kj}; x_0) = p_j.$$

Man gebe sich ein $j \in \mathbb{N}$ vor. Dann existiert $k(j) \in \mathbb{N}$ mit

$$t_{k(j)j} \geq j, \qquad \|x(t_{k(j)j}; x_0) - p_j\|_2 \leq \frac{1}{j}.$$

Für jedes $j \in \mathbb{N}$ ist daher

$$\|x(t_{k(j)j}; x_0) - p\|_2 \leq \|x(t_{k(j)j}; x_0) - p_j\| + \|p_j - p\|_2 \leq \frac{2}{j}.$$

Definiert man daher die Folge $\{t_j\}_{j\in\mathbb{N}}$ durch $t_j := t_{k(j)j}$, $j \in \mathbb{N}$, so ist $t_j \geq j$ und $\|x(t_j; x_0) - p\|_2 \leq 2/j$ und folglich $p \in \omega(\gamma)$. Damit ist nachgewiesen, dass $\omega(\gamma)$ *abgeschlossen* ist. Um die *Invarianz* von $\omega(\gamma)$ nachzuweisen, haben wir zu zeigen, dass mit $p \in \omega(\gamma)$ die gesamte Bahn $\{x(t; p) : t \in \mathbb{R}\}$ in $\omega(\gamma)$ enthalten ist. Zu $p \in \omega(\gamma)$ existiert nach Definition der ω-Grenzmenge $\omega(\gamma)$ eine Folge $\{t_k\}$ mit $t_k \to \infty$ und $x(t_k; x_0) \to p$. Für ein beliebiges $t \in \mathbb{R}$ ist

$$x(t + t_k; x_0) = x(t; \underbrace{x(t_k; x_0)}_{\to p}) \to x(t; p),$$

woraus $x(t; p) \in \omega(\gamma)$ folgt und damit die Invarianz von $\omega(\gamma)$ bewiesen ist. Entsprechend kann für die α-Grenzmenge $\alpha(\gamma)$ argumentiert werden.

Nun wird zusätzlich vorausgesetzt, dass die positive Bahn γ^+ beschränkt ist, also in einer kompakten Menge $K \subset \mathbb{R}^n$ enthalten ist. Dann ist auch $\omega(\gamma)$ beschränkt und damit insgesamt *kompakt*. Wählt man eine beliebige Folge $\{t_k\} \subset \mathbb{R}$ mit $t_k \to \infty$, so ist $x(t_k; x_0) \in \gamma^+$ für alle hinreichend großen k und folglich $\{x(t_k; x_0\}$ beschränkt. Ein Häufungspunkt dieser Folge liegt in $\omega(\gamma)$, also ist $\omega(\gamma)$ *nichtleer*. Schließlich zeigen wir, dass $\omega(\gamma)$ zusammenhängend ist. Angenommen, dies wäre nicht der Fall. Dann existieren disjunkte offene Mengen $A, B \subset \mathbb{R}^n$ mit der Eigenschaft, dass $A \cap \omega(\gamma)$ und $B \cap \omega(\gamma)$ nichtleer sind und $\omega(\gamma) \subset A \cup B$. Dann gibt es Folgen $\{t_k\}_{k\in\mathbb{N}}$ und $\{s_k\}_{k\in\mathbb{N}}$ mit $t_k < s_k < t_{k+1}$, $k \in \mathbb{N}$, sowie $t_k \to \infty$ (und $s_k \to \infty$) sowie $\{x(t_k; x_0)\}_{k\in\mathbb{N}} \subset A$ und $\{x(s_k; x_0)\}_{k\in\mathbb{N}} \subset B$. Sei nun $k \in \mathbb{N}$ fest. Da $\{x(t; x_0) : t \in [t_k, s_k]\}$ als stetiges Bild der zusammenhängenden Menge $[t_k, s_k]$ selbst zusammenhängend ist, existiert ein $r_k \in (t_k, s_k)$ mit $x(r_k; x_0) \in K \setminus (A \cup B)$. Offenbar ist auch $r_k \to \infty$. Da $K \setminus (A \cup B)$ kompakt ist, kann aus $\{x(r_k; x_0)\}_{k\in\mathbb{N}}$ eine konvergente Teilfolge ausgewählt werden. Deren Limes p liegt einerseits in $\omega(\gamma)$ und andererseits in $K \setminus (A \cup B)$. Dies ist aber ein Widerspruch zu $\omega(\gamma) \subset A \cup B$. Damit ist gezeigt, dass $\omega(\gamma)$ *zusammenhängend* ist. Nun zeigen wir, dass $\lim_{t\to\infty} d(x(t; x_0), \omega(\gamma)) = 0$. Dies ist aber klar. Denn ist $\{t_k\} \subset \mathbb{R}$ eine beliebige Folge mit $t_k \to \infty$, so besitzt $\{x(t_k; x_0\} \subset \gamma^+$ eine gegen ein $p \in \omega(\gamma)$ konvergente Teilfolge. Entsprechend kann für die α-Grenzmenge $\alpha(\gamma)$ argumentiert werden. Damit ist der gesamte Satz bewiesen. □

Beispiel: Die ω-Grenzmenge $\omega(\gamma)$ einer Bahn γ ist wegen des ersten Teiles von Satz 11.5 in jedem Falle abgeschlossen und bezüglich des Systems $\dot{x} = f(x)$ invariant. Ist die positive Bahn $\gamma^+ = \gamma^+(x_0)$ *nicht* beschränkt, so kann die ω-Grenzmenge $\omega(\gamma)$ leer, nicht kompakt oder nicht zusammenhängend sein. Für $\dot{x} = 1$ ist z. B. $x(t; x_0) = x_0 + t$, folglich ist $\gamma^+(x_0) = \{x_0 + t : t \geq 0\}$ unbeschränkt und $\omega(\gamma) = \emptyset$. Nun betrachten wir

die Anfangswertaufgabe

$$\begin{aligned} \dot{x}_1 &= x_1(1 - x_1^2 + 5x_1x_2) - 5x_2, & x_1(0) &= \tfrac{1}{2}, \\ \dot{x}_2 &= x_2(1 - x_1^2 + 5x_1x_2) + 5x_1, & x_2(0) &= 1. \end{aligned}$$

In Abbildung 117 haben wir die entsprechende Trajektorie für das Zeitintervall $[0, 3.5]$ angegeben. Man kann zeigen, dass jede in dem Streifen $(-1, 1) \times \mathbb{R}$ startende Tra-

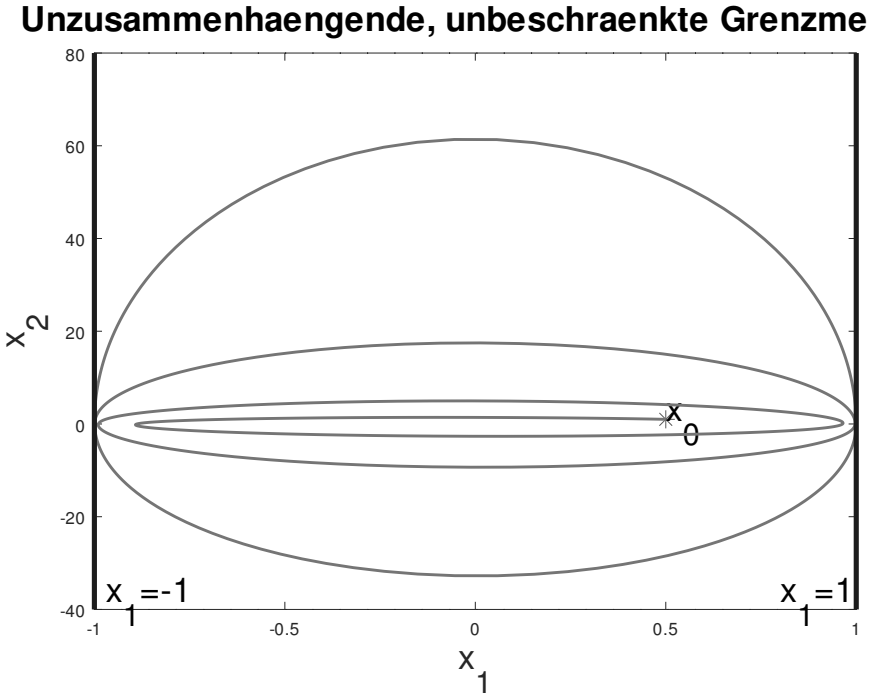


Abbildung 117: Unbeschränkte positive Bahn

jektorie bzw. positive Bahn γ in diesem Streifen verbleibt und die zugehörige ω-Grenzmenge $\omega(\gamma)$ durch $(\{-1\} \times \mathbb{R}) \cup (\{1\} \times \mathbb{R})$ gegeben ist, eine unbeschränkte und unzusammenhängende Menge. Wir verweisen auf das Youtube-Video unter `https://www.youtube.com/watch?v=Hkz3TR2ylCg`. □

Beispiel: Gegeben sei das System

$$\begin{aligned} \dot{x}_1 &= -x_1, \\ \dot{x}_2 &= -2x_2. \end{aligned}$$

Mit $x_0 = (x_{10}, x_{20})$ ist eine zugehörige Bahn durch

$$\gamma = \{x(t; x_0) : t \in \mathbb{R}\} = \{(x_{10}e^{-t}, x_{20}e^{-2t}) : t \in \mathbb{R}\}$$

gegeben. Offenbar ist $\omega(\gamma) = \{0\}$ der Ursprung des Phasenraums. In Abbildung 118 geben wir einige positive Bahnen an, die mit $t \to \infty$ jeweils zum Nullpunkt streben. □

Beispiel: Gegeben sei das System (harmonischer Oszillator)

$$\dot{x}_1 = x_2,$$

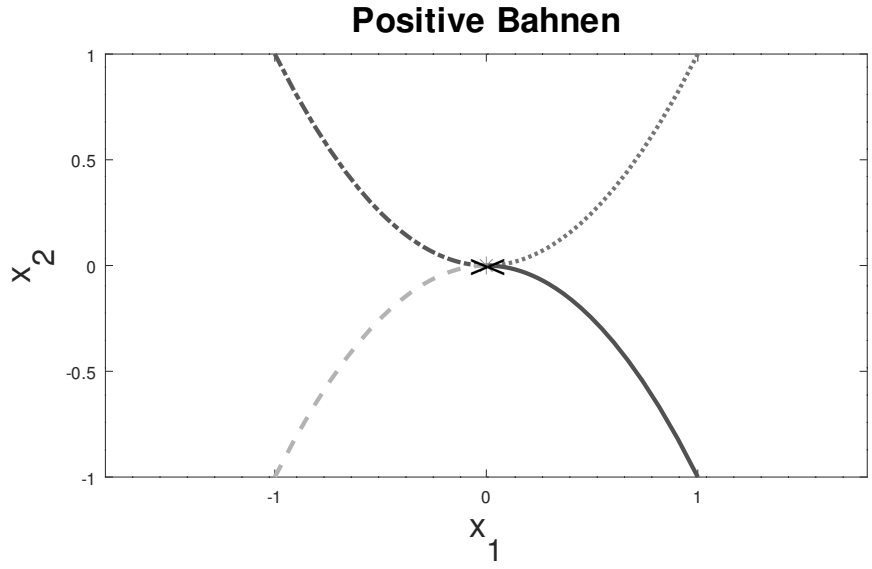


Abbildung 118: Positive Bahnen zum System $\dot{x}_1 = -x_1$, $\dot{x}_2 = -2x_2$

$$\dot{x}_2 \;=\; -x_1.$$

Mit $x_0 = (x_{10}, x_{20})$ ist eine zugehörige Bahn durch

$$\gamma = \{x(t; x_0) : t \in \mathbb{R}\} = \{(x_{10}\cos t + x_{20}\sin t, -x_{10}\sin t + x_{20}\cos t) : t \in \mathbb{R}\}$$

gegeben. Alle Bahnen sind geschlossen, es sind nämlich Kreise durch x_0 mit dem Nullpunkt als Mittelpunkt, und für jede Bahn γ ist $\omega(\gamma) = \alpha(\gamma) = \gamma$. □

Definition 11.6 Eine Menge $M \subset \mathbb{R}^n$ heißt *minimal* für das System $\dot{x} = f(x)$, falls M abgeschlossen, invariant (d. h. mit $x_0 \in M$ ist $x(t; x_0) \in M$ für alle $t \in \mathbb{R}$) und nichtleer ist und es keine echte Teilmenge von M mit diesen drei Eigenschaften gibt.

Für den Beweis der Existenz minimaler Mengen benötigen wir das *Zornsche Lemma*:

- Eine halbgeordnete Menge, in der jede Kette eine untere Schranke besitzt, enthält mindestens ein minimales Element.

Hier müssen die vier Vokabeln *halbgeordnete Menge*, *Kette*, *untere Schranke* und schließlich *minimales Element* erläutert werden.

1. Eine *halbgeordnete Menge* ist ein Paar $(P, \leq)$, wobei P eine Menge und $\leq$ eine Relation auf P ist, die *transitiv* (aus $x, y, z \in P$ mit $x \leq y$ und $y \leq z$ folgt $x \leq z$), *reflexiv* (es ist $x \leq x$ für alle $x \in P$) und *antisymmetrisch* (aus $x, y \in P$ mit $x \leq y$ und $y \leq x$ folgt $x = y$) ist.

2. Eine *Kette* in der halbgeordneten Menge $(P, \leq)$ ist eine Teilmenge $T \subset P$ mit der Eigenschaft, dass $x \leq y$ oder $y \leq x$ für alle $x, y \in T$. Eine Kette T ist also *total geordnete* Teilmenge der halbgeordneten Menge P.

3. Von einer Kette wird nun noch gefordert, dass sie eine *untere Schranke* besitzt. D. h. zu jeder Kette $T \subset P$ existiert ein $s \in P$ (untere Schranke) mit $s \leq t$ für alle $t \in T$. Man beachte, dass die untere Schranke nicht zu T gehören muss!

4. Ein Element $m \in P$ heißt *minimal*, wenn aus $x \leq m$ für ein $x \in P$ folgt, dass $x = m$.

Einen Beweis des folgenden Satzes findet man auch bei J. K. Hale (1969, S. 48).

Satz 11.7 *Sei $A \subset \mathbb{R}^n$ eine nichtleere, kompakte und bezüglich des Systems $\dot{x} = f(x)$ invariante Menge. Dann existiert eine für dieses System minimale Menge $M \subset A$.*

Beweis: Sei

$$P := \{B \subset \mathbb{R}^n : B \subset A,\ B \text{ nichtleer, kompakt und invariant bezüglich } \dot{x} = f(x)\}.$$

Auf P wird eine Halbordnung $\leq$ dadurch definiert, dass $B_2 \leq B_1$ für $B_1, B_2 \in P$ falls $B_2 \subset B_1$. Um nachzuweisen, dass in P ein minimales Element existiert, wenden wir das Zornsche Lemma an. Hierzu ist nachzuweisen, dass eine Kette $T \subset P$, also eine total geordnete Teilmenge T von P, eine untere Schranke besitzt. Sei also $T \subset P$ eine Kette. Man definiere $S := \bigcap_{B \in T} B$. Wir zeigen, dass $S \in P$ eine untere Schranke der Kette T ist. Natürlich ist S in der kompakten Menge A enthalten. Als Durchschnitt kompakter Mengen ist S selbst kompakt. Offensichtlich ist S auch invariant bezüglich $\dot{x} = f(x)$. Denn ist $x_0 \in S$, so ist $x_0 \in B$ für alle $B \in T$, wegen der Invarianz von B ist also $x(t; x_0) \in B$ für alle $t \in \mathbb{R}$ und alle $B \in T$ und folglich $x(t; x_0) \in S$ für alle $t \in \mathbb{R}$ und damit S invariant. Hierbei ist $x(\cdot; x_0)$ natürlich die Lösung von $\dot{x} = f(x)$, $x(0) = x_0$. Um $S \in P$ zu zeigen, muss also nur noch $S \neq \emptyset$ nachgewiesen werden. Hierzu zeigen wir, dass der Durchschnitt endlich vieler Elemente der Kette T nichtleer ist. Sind $B_1, B_2 \in T$, so ist $B_1 \leq B_2$ oder $B_2 \leq B_1$ bzw. $B_1 \subset B_2$ oder $B_2 \subset B_1$. Mit $B_1, B_2 \in T$ ist also $B_1 \cap B_2$ eine nichtleere, kompakte und invariante Menge und selbst ein Element von T. Die entsprechende Aussage gilt für den Durchschnitt endlich vieler Elemente von T, d. h. T besitzt die sogenannte *finite intersection property*. Daher ist S nichtleer. Denn wäre $S = \emptyset$ bzw. $\bigcap_{B \in T} B = \emptyset$, so wäre $A \subset \bigcup_{B \in T} B^c$, wobei B^c das Komplent von B ist. Wegen der Kompaktheit von A kann aus der offenen Überdeckung $\bigcup_{B \in T} B^c$ eine endliche Teilüberdeckung ausgewählt werden. Es existieren also $\{B_1, \ldots, B_k\} \subset T$ mit $A \subset \bigcup_{i=1}^k B_i^c$. Wegen $B_i \subset A$, $i = 1, \ldots, k$, folgt hieraus $\bigcap_{i=1}^k B_i = \emptyset$, ein Widerspruch zur finite intersection property von T. Also ist $S \neq \emptyset$. Ferner ist $S \leq B$ für alle $B \in T$, d. h. $S \in P$ ist eine untere Schranke der Kette T. Wegen des Zornschen Lemmas gibt es in P ein minimales Element $M \in P$. Insbesondere ist $M \subset A$. Wir wollen uns überlegen, dass M eine minimale Menge im Sinne von Definition 11.6 ist. Hierzu nehmen wir an, es sei $N \subset M$ abgeschlossen, invariant und nichtleer. Offenbar ist dann $N \in P$. Da M in P ein minimales Element ist, ist $N = M$. Also ist M die gesuchte minimale Menge. Der Satz ist bewiesen. □

Beispiele: Für das System $\dot{x}_1 = -x_1$, $\dot{x}_2 = -2x_2$ ist $\omega(\gamma) = \{0\}$. Natürlich ist $\omega(\gamma)$ abgeschlossen, invariant und nichtleer. Da keine echte Teilmenge diese Eigenschaften besitzt, ist $\omega(\gamma)$ minimal.

Für den harmonischen Oszillator $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1$ sind alle Bahnen γ geschlossen und damit $\omega(\gamma) = \alpha(\gamma) = \gamma$. Offenbar sind diese Mengen auch minimal. □

11.4 Der Satz von Poincaré-Bendixson und sein Beweis

In diesem Unterabschnitt beschränken wir uns auf *planare* autonome Differentialgleichungssysteme, betrachten also die Aufgabe

$$\dot{x} = \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} = f(x),$$

wobei $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ stetige erste partielle Ableitungen besitzt. Ferner nehmen wir an, dass von uns betrachtete Lösungen dieses Systems auf $\mathbb{R}$ existieren.

Ein wesentlicher Unterschied zwischen dem $\mathbb{R}^2$ und einem $\mathbb{R}^n$ mit $n \geq 3$ besteht darin, dass im $\mathbb{R}^2$ der *Jordansche Kurvensatz* gilt, den wir jetzt ohne Beweis zitieren.

Satz 11.8 (Jordanscher Kurvensatz) *Sei $J \subset \mathbb{R}^2$ eine Jordankurve, d. h. es ist*

$$J = \{\phi(t) : t \in [0, 1)\},$$

wobei

- $\phi\colon [0, 1] \longrightarrow \mathbb{R}^2$ *ist stetig,*
- $\phi(0) = \phi(1)$,
- $\phi|_{[0,1)}$ *ist injektiv.*

Dann ist $\mathbb{R}^2 \setminus J$ die Vereinigung von zwei offenen Mengen $\mathrm{int}(J)$ *und* $\mathrm{ext}(J)$*, die jeweils J als Rand besitzen. Hierbei heißt* $\mathrm{int}(J)$ *das Innere von J und ist beschränkt, während* $\mathrm{ext}(J)$ *unbeschränkt ist und das Äußere von J heißt.*

Wir formulieren jetzt den Satz von Poincaré-Bendicson und werden im Anschluss daran auf die zum Beweis nötigen Hilfsmittel eingehen, die sich auf den planaren Fall beziehen. Der Beweisaufbau folgt dem bei F. Verhulst (1990, S. 45 ff.), der sich wiederum eng an die Darstellung bei J. K. Hale (1969, S,. 51 ff.) hält.

Satz 11.9 (Poincaré-Bendixson) *Gegeben sei das planare autonome System $\dot{x} = f(x)$. Die positive Bahn γ^+ sei beschränkt und die ω-Grenzmenge $\omega(\gamma^+)$ dieser Bahn enthalte keine kritischen (bzw. stationären) Punkte des Systems. Dann ist $\omega(\gamma^+)$ eine geschlossene Bahn bzw. das Bild einer periodischen Lösung von $\dot{x} = f(x)$. Ist $\omega(\gamma^+) \neq \gamma^+$, so nennt man $\omega(\gamma^+)$ einen Grenzzyklus.*

Ein wichtiger Begriff beim Beweis des Satzes von Poincaré-Bendixson ist der Begriff der *Transversalen*, der jetzt definiert wird. Diese Definition findet man z. B. bei E. A. Coddington, N. Levinson (1955, S. 392), L. Cesari (1970, S. 164).

Definition 11.10 Ein kompaktes Streckensegment

$$l := \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\} \subset \mathbb{R}^2$$

mit $x_0 \neq x_1$ heißt eine *Transversale* zu f bzw. der durch f bestimmten Bahnen γ, wenn $f(x)$ für alle $x \in l$ nicht parallel zu l ist, also $f(x) \neq \lambda(x_1 - x_0)$ für alle $\lambda \in \mathbb{R}$ gilt bzw. $f(x)$ und $x_1 - x_0$ linear unabhängig sind.

Insbesondere ist $f(x) \neq 0$ für alle Punkte x einer Transversalen l zu f, d. h. alle Punkte einer Transversalen zu f sind nichtkritisch für das System $\dot{x} = f(x)$. Im folgenden Satz sammeln wir einige einfache Eigenschaften einer Transversalen zu f, siehe L. Cesari (1970, S. 164) und auch J. M. Melenk (2020).

Satz 11.11 *Es gelten die folgenden Aussagen über Transversalen zu f.*

1. *Sei p ein nichtkritischer Punkt zu f. Dann existiert eine Transversale l zu f, die p im (relativen) Inneren enthält, etwa als Mittelpunkt. Diese Transversale kann außer $\pm f(p)$ jede Richtung annehmen.*

2. *Seien $x_0, x_1 \in \mathbb{R}^2$ mit $x_0 \neq x_1$ gegeben und $l := \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ das durch x_0 und x_1 gegebene kompakte Streckensegment. Ist $z \neq 0$ ein beliebiger Punkt des $\mathbb{R}^2$ mit $z^T(x_1 - x_0) = 0$, so ist l genau dann eine Transversale zu f, wenn $z^T f(x) \neq 0$ für alle $x \in l$.*

3. *Sei $l = \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ eine Transversale zu f und $p_0 \in l$. Mit $x(\cdot; p_0)$ sei die Lösung von $\dot{x} = f(x)$ mit $x(0) = p_0$ bezeichnet. Dann ist $x(0; p_0) \in l$ und es existiert ein $\epsilon > 0$ mit $x(t; p_0) \notin l$ für alle $t \in \mathbb{R}$ mit $|t| \leq \epsilon$. Bahnen kreuzen eine Transversale zu f.*

4. *Sei $l = \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ eine Transversale zu f und $z \in \mathbb{R}^2 \setminus \{0\}$ mit $z^T(x_1 - x_0) = 0$ gegeben. Dann ist das Vorzeichen von $z^T \dot{x}(0; p_0) = z^T f(p_0)$ für jedes $p_0 \in l$ dasselbe, d. h. alle Bahnen zu f kreuzen die Transversale l zu f in der gleichen Richtung. Insbesondere kreuzt eine geschlossene Bahn eine Transversale in höchstens einem Punkt.*

5. *Sei $l = \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ eine Transversale zu f und $p_0 \in l \setminus \{x_0, x_1\}$ ein (relativ) innerer Punkt von l. Dann existiert zu jedem $\epsilon > 0$ ein positives $\delta = \delta(\epsilon, p_0)$ derart, dass jede in $p \in B(p_0; \delta)$ startende Bahn $\{x(t; p) : t \in \mathbb{R}\}$ die Transversale in einer Zeit $t = t(p) \in [-\epsilon, \epsilon]$ kreuzt bzw. $x(t; p) \in l$ gilt.*

6. *Sei $l = \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ eine Transversale zu f. Sei $\eta := \{x(t; p_0) : t \in [a,b]\}$ eine "Teilbahn" der Bahn $\gamma := \{x(t; p_0) : t \in \mathbb{R}\}$. Dann gilt:*

   (a) *Auf der Transversalen l liegen nur endlich viele Punkte der Teilbahn η, d. h. es ist $l \cap \eta = \{p_1, \ldots, p_m\}$ endlich.*

   (b) *Die Anordnung der Punkte $p_1, \ldots, p_m$ auf der Transversalen l ist dieselbe wie die dieser Punkte auf der Teilbahn η. Genauer gilt für die Punkte*

$$p_i = (1-\tau_i)x_0 + \tau_i x_1 = x(t_i; p_0), \quad i = 1, \ldots, m,$$

mit "Durchstoßzeiten" t_i, $i = 1, \ldots, m$, dass

$$t_1 < t_2 < \cdots < t_m \Longrightarrow \tau_1 < \tau_2 < \cdots < \tau_m \quad \textit{oder} \quad \tau_1 > \tau_2 > \cdots > \tau_m.$$

Beweis: Sei p ein nichtkritischer Punkt zu f, also $f(p) \neq 0$. Dann existiert eine offene Kreisscheibe $B(p;\epsilon)$ um p mit dem Radius $\epsilon > 0$ mit $f(x) \neq 0$ für alle $x \in B(p;\epsilon)$. Nun gebe man sich ein $x_1 \in B(p;\epsilon) \setminus \{p\}$ vor, für welches $x_1 - p$ und $f(p)$ linear unabhängig sind. Indem wir ϵ notfalls verkleinern, können wir annehmen, dass $x_1 - p$ und $f(x)$ für alle $x \in B(p;\epsilon)$ linear unabhängig sind. Mit $x_0 := 2p - x_1$ ist

$$l := \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$$

eine Transversale zu f, die p als Mittelpunkt enthält. Denn ist $x = (1-\tau)x_0 + \tau x_1$ mit $\tau \in [0,1]$ ein Element von l, so ist

$$\|x - p\|_2 = \underbrace{|2\tau - 1|}_{\leq 1} \, \underbrace{\|x_1 - p\|_2}_{<\epsilon} < \epsilon,$$

also $l \subset B(p;\epsilon)$. Für alle $x \in l$ sind daher $f(x)$ und $x_1 - x_0 = 2(x_1 - p)$ linear unabhängig, also l eine Transversale zu f. Damit ist der erste Teil des Satzes bewiesen. Nun sei $l := \{(1-\tau)x_0 + \tau x_1 : \tau \in [0,1]\}$ mit $x_0 \neq x_1$ und $z \neq 0$ ein Element des $\mathbb{R}^2$ mit $z^T(x_1 - x_0) = 0$. Dann ist l genau dann eine Transversale zu f, wenn

$$f(x) \notin \text{span}\,\{x_1 - x_0\} = \text{span}\,\{z\}^\perp \quad \text{für alle } x \in l,$$

womit auch der zweite Teil des Satzes bewiesen ist.

Wir wollen Teil 2 dieses Satzes anwenden und geben uns ein beliebiges $z \neq 0$ mit $z^T(x_1 - x_0) = 0$ vor. Wegen $p_0 \in l$ ist einerseits $z^T f(p_0) \neq 0$. Wir machen einen Widerspruchsbeweis und nehmen an, die Behauptung sei nicht richtig. Dann existiert eine Folge $\{t_k\} \subset \mathbb{R}$ mit $t_k \to 0$ und $x(t_k; p_0) \in l$ für alle $k \in \mathbb{N}$. Dann ist

$$x(t_k; p_0) - p_0 = x(t_k; p_0) - x(0; p_0) \in \text{span}\,\{x_1 - x_0\}$$

und daher andererseits

$$z^T f(p_0) = z^T \dot{x}(0; p_0) = z^T \left(\lim_{k\to\infty} \frac{x(t_k; p_0) - x(0; p_0)}{t_k} \right) = \lim_{k\to\infty} \underbrace{z^T \left(\frac{x(t_k; p_0) - p_0}{t_k} \right)}_{=0} = 0.$$

Damit haben wir den gewünschten Widerspruch erhalten. Aus

$$0 \neq z^T f(p_0) = \lim_{t\to 0} z^T \left(\frac{x(t; p_0) - p_0}{t} \right)$$

erkennen wir ferner, dass die Funktion

$$t \mapsto z^T \left(\frac{x(t; p_0) - p_0}{t} \right)$$

für kleine $|t|$ von einem Vorzeichen ist bzw. die Funktion

$$t \mapsto z^T(x(t;p_0) - p_0)$$

bei $t = 0$ das Vorzeichen wechselt. Dies bedeutet, dass die Bahn $\gamma = \{x(t;p_0) : t \in \mathbb{R}\}$ die Transversale l in p_0 kreuzt. Damit ist auch die dritte Aussage des Satzes bewiesen. Wir definieren die stetige Funktion $\phi\colon [0,1] \longrightarrow \mathbb{R}$ durch

$$\phi(\tau) := z^T f(\underbrace{(1-\tau)x_0 + \tau x_1}_{\in l}).$$

Wegen Teil 2 dieses Satzes hat ϕ auf $[0,1]$ keine Nullstelle und ist daher von einem Vorzeichen. Damit ist auch die vierte Aussage bewiesen.

Wir machen einen Widerspruchsbeweis, nehmen also an, die Aussage sei nicht richtig. Dann existiert ein $\epsilon_0 > 0$ und eine Folge $\{p_k\}_{k\in\mathbb{N}}$ mit $p_k \to p_0$ derart, dass $x(t;p_k) \notin L$ für alle $t \in [-\epsilon_0, \epsilon_0]$. Da p_0 ein (relativ) innerer Punkt der Transversalen l ist, gibt es ein $\delta_0 > 0$ mit der Eigenschaft, dass

$$x \in p_0 + \text{span}\,\{x_1 - x_0\}, \quad \|x - p_0\|_2 \le \delta_0 \Longrightarrow x \in l.$$

Das veranschaulichen wir uns in Abbildung 119. Wir können $\epsilon_0 > 0$ so klein wählen,

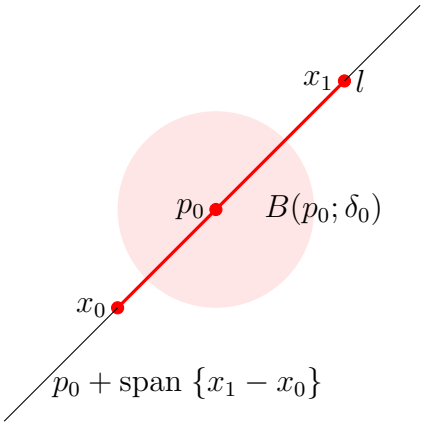


Abbildung 119: $B(p_0;\delta_0) \cap (p_0 + \text{span}\,\{x_1 - x_0\}) \subset l$

dass

$$\|x(t;p_k) - p_0\| \le \delta_0 \quad \text{für alle } t \in [-\epsilon_0, \epsilon_0] \text{ und alle } k \ge K_0$$

mit einem $K_0 \in \mathbb{N}$. Man wähle ein $z \in \mathbb{R}^2 \setminus \{0\}$ mit $z^T(x_1 - x_0) = 0$ und definiere

$$\phi_k(t) := z^T(x(t;p_k) - p_0), \qquad \phi_\infty(t) := z^T(x(t;p_0) - p_0).$$

Dann konvergiert die Folge $\{\phi_k\}_{k\in\mathbb{N}}$ auf $[-\epsilon_0, \epsilon_0]$ gleichmäßig gegen ϕ_∞. Die Funktion ϕ_∞ hat wegen

$$\phi_\infty(0) = 0, \qquad \dot\phi_\infty(0) = z^T f(p_0) \neq 0$$

(Teil 2 des Satzes) einen Vorzeichenwechsel bei $t = 0$. Für alle hinreichend großen k kann daher ϕ_k nicht von einem Vorzeichen auf $[-\epsilon_0, \epsilon_0]$ sein, hat also für diese k eine Nullstelle in $[-\epsilon_0, \epsilon_0]$. Andererseits hat aber ϕ_k für alle $k \geq K_0$ keine Nullstelle und damit auch keinen Zeichenwechsel auf $[-\epsilon_0, \epsilon_0]$. Denn wäre $\phi_k(t) = 0$ für ein $t \in [-\epsilon_0, \epsilon_0]$ und ein $k \geq K_0$, so wäre $z^T(x(t; p_k) - p_0) = 0$, also

$$x(t; p_k) - p_0 \in \text{span } \{z\}^\perp = \text{span } \{x_1 - x_0\}$$

und nach Wahl von ϵ_0 wäre $x(t; p_k) \in l$, ein Widerspruch. Damit ist auch die fünfte Aussage des Satzes bewiesen.

Für den Beweis des ersten Teils der sechsten Aussage des Satzes machen wir einen Widerspruchsbeweis und nehmen an, dass es eine Folge $\{t_k\}_{k\in\mathbb{N}} \subset [a, b]$ mit $x(t_k; p_0) \in l \cap \eta$, $k \in \mathbb{N}$, gibt. Indem man notfalls zu einer Teilfolge übergeht, können wir annehmen, dass $t_k \to t_\infty \in [a, b]$. Sei $x_\infty := x(t_\infty; p_0)$. Da $l \cap \eta$ abgeschlossen ist, ist $x_\infty \in l \cap \eta$. Mit einen beliebigen $z \in \mathbb{R}^2 \setminus \{0\}$ mit $z^T(x_1 - x_0) = 0$ ist dann

$$x(t_k; p_0) - x_\infty \in \text{span } \{x_1 - x_0\} = \text{span } \{z\}^\perp.$$

Daher ist einerseits

$$z^T \left(\lim_{k\to\infty} \frac{x(t_k; p_0) - x_\infty}{t_k - t_\infty} \right) = z^T \dot{x}(t_\infty; p_0) = z^T f(p_0) \neq 0$$

wegen Teil 2 dieses Satzes. Anderseits ist

$$z^T \left(\lim_{k\to\infty} \frac{x(t_k; p_0) - x_\infty}{t_k - t_\infty} \right) = \lim_{k\to\infty} z^T \Bigg(\underbrace{\frac{x(t_k; p_0) - x_\infty}{t_k - t_\infty}}_{\in \text{span } \{x_1 - x_0\} = \text{span } \{z\}^\perp} \Bigg) = 0,$$

ein Widerspruch. Seien nun $p_1 = x(t_1; p_0)$ und $p_2 = x(t_2; p_0)$ mit $t_1 < t_2$ zwei aufeinander folgende Schnittpunkte der Teilbahn η (mit dem Anfangspunkt $A = x(a; p_0)$ und dem Endpunkt $B = x(b; p_0)$) der Bahn γ. Wir nehmen an, es sei $p_1 \neq p_2$. Andernfalls wäre η und damit auch γ eine geschlossene Bahn, die mit einer zugehörigen Transversalen höchstens einen Punkt gemeinsam haben kann, in diesem Fall also genau einen Punkt gemein hat. Die Kurve J, die aus aus dem Bogen

$$\widehat{p_1 p_2} := \{x(t; p_0) : t \in [t_1, t_2]\}$$

und dem in der Transversale l enthaltenen Geradensegment $\overline{p_2 p_1}$ besteht, ist eine Jordankurve. Wir veranschaulichen dies in Abbildung 120. Nun betrachten wir Punkte $q = x(s_q; p_0) \in \gamma$ mit $s_q < t_1$ und $r = x(s_r; p_0) \in \gamma$ mit $s_r > t_2$. Sind s_q und s_r hinreichend nahe bei t_1 bzw. t_2, so liegen diese beiden Punkte in unterschiedlichen durch die Jordankurve gegebenen Gebieten, d.h. ein Punkt liegt im Innengebiet $\text{int}(J)$ und der andere im Außengebiet $\text{ext}(J)$. In Abbildung 120 links geben wir den Fall an, dass $q \in \text{ext}(J)$ und $r \in \text{int}(J)$, in Abbildung 120 rechts ist $q \in \text{int}(J)$ und $r \in \text{ext}(J)$. Wir betrachten den ersten Fall, der zweite kann entsprechend behandelt werden. Von $r \in \text{int}(J)$ startend kann die Bahn γ nur dann in den Außenbereich $\text{ext}(J)$ gelangen,

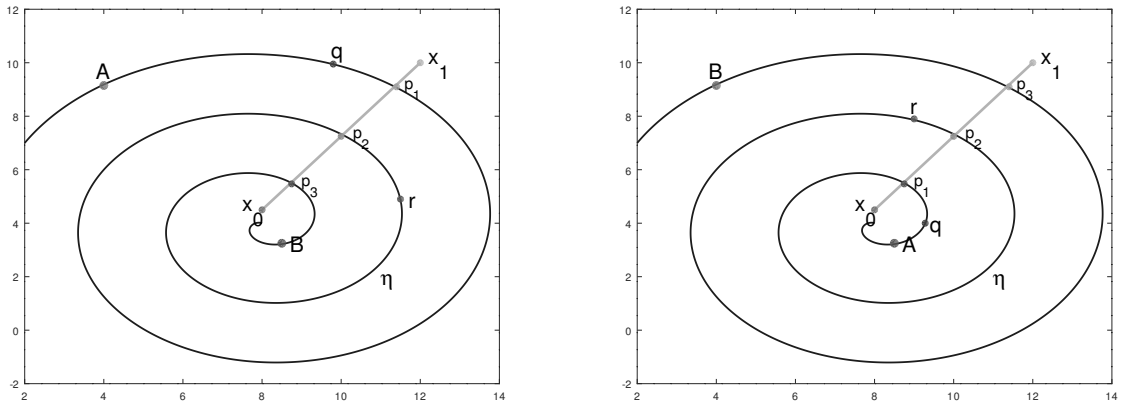


Abbildung 120: Veranschaulichung des sechsten Teiles von Satz 11.11

wenn sie den Bogen $\widehat{p_1p_2}$ oder das Segment $\overline{p_2p_1}$ kreuzt. Ersteres ist nicht möglich, da wegen der eindeutigen Lösbarkeit von Anfangswertaufgaben sich eine Bahn nicht kreuzen kann. Das Segment $\overline{p_2p_1}$ als Teil der Transversale kann nicht gekreuzt werden, da die Kreuzungen der Bahn mit der Transversalen sämtlich in derselben Richtung erfolgen. Daher ist $\gamma \subset \text{int}(J)$ für alle $t > t_2$. Der nächste (nach p_2) Schnittpunkt $p_3 = x(t_3; p_0)$ (mit $t_3 > t_2$) von γ mit l liegt daher in $\text{int}(J)$ und ist von p_2 verschieden. Daher liegt p_2 *zwischen* p_1 und p_3 auf l. Daher ist $\tau_1 < \tau_2 < \tau_3$ oder $\tau_1 > \tau_2 > \tau_3$. Damit ist auch der sechste Teil des Satzes bewiesen. □

Bemerkung: In Abbildung 121 geben wir eine Transversale zu Bahnen der van der Polschen Differentialgleichung an. Entsprechend der Aussage von Teil 4 von Satz 11.11 kreuzen diese die Transversale sämtlich in der gleichen Richtung. □

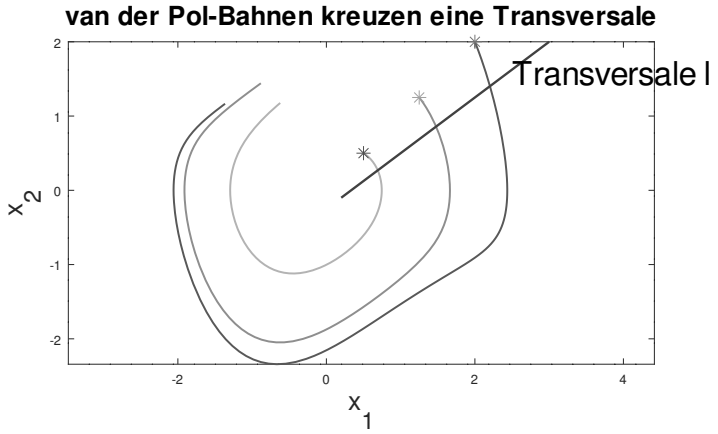


Abbildung 121: van der Pol Bahnen und eine Transversale

Satz 11.12 *Ist $M \subset \mathbb{R}^2$ eine für das planare System $\dot{x} = f(x)$ minimale und beschränkte Menge, so ist M ein kritischer Punkt oder eine geschlossene Bahn.*

Beweis: Als für das System $\dot{x} = f(x)$ minimale Menge ist M abgeschlossen (und wegen der vorausgesetzten Beschränktheit sogar kompakt), invariant und nichtleer. Es existiert also ein $x_0 \in M$ und wegen der Invarianz von M ist die gesamte durch x_0 verlaufende Bahn $\gamma = \gamma(x_0)$ in M enthalten. Die α- bzw. ω-Grenzmenge $\alpha(\gamma)$ bzw. $\omega(\gamma)$ ist dann ebenfalls in M enthalten. Denn ist etwa $p \in \omega(\gamma)$, so existiert nach Definition eine Folge $\{t_k\} \subset \mathbb{R}$ mit $t_k \to \infty$ und $x(t_k; x_0) \to p$. Da aber $\{x(t_k; x_0)\} \subset \gamma \subset M$, ist $p \in \mathrm{cl}\,(M) = M$ und damit $\omega(\gamma) \subset M$. Wenn M einen kritischen Punkt enthält, so folgt aus der Minimalität von M, dass M genau dieser kritische Punkt ist. Daher nehmen wir jetzt an, dass M keine kritischen Punkte enthält und haben zu zeigen, dass M eine geschlossene Bahn ist. Da sowohl γ als auch $\omega(\gamma)$ in M enthalten sind, ist $M = \omega(\gamma)$ wegen der Minimalität von M und daher $\gamma \subset \omega(\gamma) = M$. Nun wähle man einen Punkt $p_0 \in \gamma$ und eine Transversale l zu γ, die p_0 im (relativen) Inneren enthält. Dies ist wegen des ersten Teiles von Satz 11.11 möglich. Also ist $p_0 \in l \cap \omega(\gamma)$. Wir überlegen uns nun:

(1) Zu $p_0 \in l \cap \omega(\gamma)$ existiert eine monoton wachsende Folge $\{t_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ mit $t_k \to \infty$ und $\lim_{k\to\infty} x(t_k; x_0) = p_0$ sowie $x(t_k; x_0) \in l$ für alle k.

Denn: Man wähle eine monoton fallende Nullfolge $\{\epsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$. Teil 5 von Satz 11.11 liefert:

($*$) Da $p_0 \in l$ (relativ) innerer Punkt[55] der Transversalen l ist, existiert zu $\epsilon_k > 0$ ein $\delta_k > 0$ mit der Eigenschaft, dass es zu jedem $p_k \in B(p_0; \delta_k)$ ein $t(p_k) \in [-\epsilon_k, \epsilon_k]$ mit $x(t(p_k); p_k) \in l$ gibt.

O. B. d. A. ist auch $\{\delta_k\}_{k\in\mathbb{N}}$ eine Nullfolge. Wegen $p_0 \in \omega(\gamma)$ existiert eine gegen ∞ konvergierende Folge $\{t_k^{(1)}\}_{k\in\mathbb{N}} \subset \mathbb{R}$ mit $p_k := x(t_k^{(1)}; x_0) \in B(p_0; \delta_k)$. Die Folge $\{t_k^{(1)}\}_{k\in\mathbb{N}}$ kann so gewählt werden, dass $t_k^{(1)} + \epsilon_k < t_{k+1}^{(1)} - \epsilon_k$, $k \in \mathbb{N}$. Die gesuchte Folge $\{t_k\}_{k\in\mathbb{N}}$ definieren wir durch $t_k := t_k^{(1)} + t(p_k)$, $k \in \mathbb{N}$. Dann ist die Folge $\{t_k\}_{k\in\mathbb{N}}$ wegen

$$\begin{aligned} t_{k+1} - t_k &= t_{k+1}^{(1)} + t(p_{k+1}) - (t_k^{(1)} + t(p_k)) \\ &\geq t_{k+1}^{(1)} - \epsilon_{k+1} - t_k^{(1)} - \epsilon_k \\ &\geq t_{k+1}^{(1)} - t_k^{(1)} - 2\epsilon_k \\ &\qquad\qquad (\text{wegen } \epsilon_{k+1} \leq \epsilon_k) \\ &> 0 \end{aligned}$$

monoton wachsend und natürlich gilt $\lim_{k\to\infty} t_k = \infty$. Weiter ist

$$x(t_k; x_0) = x(t_k^{(1)} + t(p_k); x_0) = x(t(p_k); x(t_k^{(1)}; x_0)) = x(t(p_k); p_k) \in l, \quad k \in \mathbb{N}.$$

[55] Ist p_0 einer der beiden Endpunkte der Transversalen l, also kein (relativ) innerer Punkt von l, so kann diese ein klein wenig verlängert werden und bleibt immer noch eine Transversale. Es ist also nicht entscheidend, dass p_0 im (relativen) Inneren von l liegt. Dies nutzen wir im Beweisschritt (2) aus.

Zu zeigen bleibt, dass $\lim_{k\to\infty} x(t_k;x_0) = p_0$. Um dies nachzuweisen, beachten wir, dass

$$\begin{aligned} x(t_k;x_0) &= x(t_k^{(1)} + t(p_k);x_0) \\ &= x(t_k^{(1)};x_0) + \left(\int_0^1 \dot{x}(t_k^{(1)} + st(p_k);x_0)\,ds\right)t(p_k) \\ &= x(t_k^{(1)};x_0) + \left(\int_0^1 f(x(t_k^{(1)} + st(p_k);x_0)\,ds\right)t(p_k) \end{aligned}$$

Daher ist

$$\begin{aligned} \|x(t_k;x_0) - p_0\|_2 &\leq \underbrace{\|x(t_k^{(1)};x_0) - p_0\|_2}_{\leq\delta_k} + \left(\int_0^1 \|f(\underbrace{x(t_k^{(1)} + st(p_k);x_0)}_{\in\gamma})\|_2\,ds\right)\underbrace{|t(p_k)|}_{\leq\epsilon_k} \\ &\leq \delta_k + C\epsilon_k, \end{aligned}$$

wobei C eine Konstante mit $\|f(x)\|_2 \leq C$ für alle $x \in \gamma \subset M$. Damit ist auch $\lim_{k\to\infty} x(t_k;x_0) = p_0$ und insgesamt die Zwischenbehauptung (1) nachgewiesen. Die Idee des Beweises besteht also im wesentlichen darin, durch einen *großen* Schritt $t_k^{(1)}$ in die Nähe des Grenzpunktes $p_0 \in l \cap \omega(\gamma)$ und anschließend durch einen *kleinen* Korrekturschritt auf der Transversalen l zu gelangen.

Nun kommt die zweite Zwischenbehauptung.

(2) Es ist $l \cap \omega(\gamma) = \{p_0\}$.

Denn: Da $p_0 \in l \cap \omega(\gamma)$, haben wir zu zeigen, dass es neben p_0 keinen weiteren Punkt in $l \cap \omega(\gamma)$ gibt. Angenommen, es ist auch $p_0' \in l \cap \omega(\gamma)$. Wegen der Zwischenbehauptung (1) existieren monoton wachsende, gegen ∞ konvergierende Folgen $\{t_k\}_{k\in\mathbb{N}}$ und $\{t_k'\}_{k\in\mathbb{N}}$ mit $\lim_{k\to\infty} x(t_k;x_0) = p_0$, $\lim_{k\to\infty} x(t_k';x_0) = p_0'$ sowie $x(t_k;x_0) \in l$, $x(t_k';x_0) \in l$ für alle $k \in \mathbb{N}$. O. B. d. A. ist $t_k \leq t_k' \leq t_{k+1}$, $k \in \mathbb{N}$. Hat die Transversale l die Darstellung

$$l = \{(1-\tau)y_0 + \tau y_1 : \tau \in [0,1]\}$$

mit $y_0 \neq y_1$, so haben die Schnittpunkte $x(t_k;x_0)$ und $x(t_k';x_0)$ der Bahn γ mit der Transversale l eine Darstellung

$$x(t_k;x_0) = (1-\tau_k)y_0 + \tau_k y_1, \qquad x(t_k';x_0) = (1-\tau_k')y_0 + \tau_k' y_1,$$

wobei wegen des sechsten Teiles von Satz 11.11 sich die Anordnung der Schnittzeitpunkte auf die Anordnung der Schnittpunkte auf der Transversalen überträgt. D. h. es ist o. B. d. A. $\tau_k \leq \tau_k' \leq \tau_{k+1}$, $k \in \mathbb{N}$. Hieraus folgt die Konvergenz der Folgen $\{\tau_k\}_{k\in\mathbb{N}}$ und $\{\tau_k'\}_{k\in\mathbb{N}}$ gegen ein und denselben Limes und damit $p_0 = p_0'$ und die Zwischenbehauptung (2) ist bewiesen.

Bisher wissen wir, dass $\gamma \subset \omega(\gamma) = M$ und M keinen kritischen Punkt enthält. Nun zeigen wir, dass $\gamma = \omega(\gamma)$ eine geschlossene Bahn ist. Das ist aber jetzt einfach. Wir wählen uns ein beliebiges $p_0 \in \gamma$ und eine Transversale l zum System $\dot{x} = f(x)$, die p_0 im relativen Inneren enthält. Wegen $\gamma \subset \omega(\gamma)$ ist $p_0 \in l \cap \omega(\gamma)$. Wegen der Zwischenbehauptung (1) existiert eine monoton wachsende Folge $\{t_k\}_{k\in\mathbb{R}}$ mit $t_k \to \infty$,

$\lim_{k\to\infty} x(t_k;x_0) = p_0$ sowie $x(t_k;x_0) \in l$ für alle $k \in \mathbb{N}$. Wegen $\gamma \subset \omega(\gamma)$ ist $x(t_k;x_0) \in l \cap \gamma \subset l \cap \omega(\gamma)$. Wegen Zwischenbehauptung (2) besteht $l \cap \omega(\gamma)$ aus genau einem Punkt. Daher ist $x(t_k;x_0) = p_0$ für alle $k \in \mathbb{N}$. Folglich ist $\gamma = \omega(\gamma) = M$ eine geschlossene Bahn und der Satz ist bewiesen. □

Beweis des Satzes von Poincaré-Bendixson: Da die positive Bahn γ^+ als beschränkt vorausgesetzt wird, liefert Satz 11.5, dass $\omega(\gamma^+)$ kompakt, zusammenhängend, invariant und nichtleer ist. Wegen Satz 11.7 (setze dort $A := \omega(\gamma^+)$) existiert eine für das System $\dot{x} = f(x)$ im Sinne von Definition 11.6 minimale Menge $M \subset \omega(\gamma^+)$. Da $\omega(\gamma^+)$ nach Voraussetzung keine kritischen Punkte enthält, sind auch in M keine kritischen Punkte enthalten. Satz 11.12 sagt aus, dass eine für das planare System $\dot{x} = f(x)$ minimale und beschränkte Menge ein kritischer Punkt oder eine geschlossene Bahn ist. Da $M \subset \omega(\gamma^+)$ beschränkt und minimal ist, ferner keinen kritischen Punkt enthält, ist M also wegen Satz 11.12 eine geschlossene Bahn. Wir überlegen uns, dass $M = \omega(\gamma^+)$. Angenommen, $\omega(\gamma^+) \setminus M$ sei nichtleer. Da

$$\omega(\gamma^+) = \mathrm{cl}\,(\omega(\gamma^+) \setminus M) \cup M$$

zusammenhängend ist (Satz 11.5), ist

$$\mathrm{cl}\,(\omega(\gamma^+) \setminus M) \cap M \neq \emptyset.$$

Daher existiert eine Folge $\{p_k\} \subset \omega(\gamma^+) \setminus M$ und ein $p_0 \in M$ mit $\lim_{k\to\infty} p_k = p_0$. Da p_0 kein kritischer Punkt ist, ist p_0 (relativ) innerer Punkt einer Transversalen durch p_0, siehe erster Teil von Satz 11.11. Die Zwischenbehauptung (2) aus dem letzten Satz 11.12 liefert $l \cap \omega(\gamma^+) = \{p_0\}$. Wegen des fünften Teiles von Satz 11.11 existiert eine Umgebung U von p_0 mit der Eigenschaft, dass jede in einem $p \in U$ startende Bahn die Transversale l in einem Punkt q trifft bzw. kreuzt. Da $p_k \in U$ für alle hinreichend großen k, trifft die von p_k ausgehende Bahn in einem Punkt q_k die Transversale l. Wegen der Invarianz der Grenzmenge $\omega(\gamma^+)$ ist mit p_k auch die gesamte von p_k ausgehende Bahn und insbesondere auch q_k in $\omega(\gamma^+)$ enthalten. Für alle hinreichend großen k ist also $q_k \in l \cap \omega(\gamma^+)$ und daher $q_k = p_0$. Daher liegen für alle hinreichend großen k die Punkte q_k auf der geschlossenen Bahn M und daher auch die Punkte p_k. Dies ist aber ein Widerspruch zu $\{p_k\} \subset \omega(\gamma^+) \setminus M$. Also ist $M = \omega(\gamma^+)$ eine geschlossene Bahn. Der Satz von Poincaré-Bendixson ist bewiesen. □

Als ein einfaches Korollar, in welchem der Begriff der ω-Grenzmenge nicht vorkommt, formulieren wir:

Korollar 11.13 *Gegeben sei das planare autonome System $\dot{x} = f(x)$. Die positive Bahn γ^+ sei in der kompakten Menge K enthalten. Die Menge K enthalte keine kritischen (bzw. stationären) Punkte des Systems. Dann ist in K eine (nichttriviale) geschlossene Bahn zu $\dot{x} = f(x)$ enthalten.*

Beweis: Aus $\gamma^+ \subset K$ und der Kompaktheit von K folgt $\omega(\gamma^+) \subset K$. Da in K und damit auch in $\omega(\gamma^+)$ kein kritischer Punkt von f enthalten ist, liefert der Satz von Poincaré-Bendixson, dass $\omega(\gamma^+) \subset K$ eine nichttriviale geschlossene Bahn ist. □

Im folgenden Satz (siehe z.B. J. K. Hale (1969, S. 55), L. Cesari (1971, S. 166), M. Farkas (1994, S. 92)) lassen wir, verglichen mit dem obigen Korollar zum Satz von

Poincaré-Bendixson, in der kompakten Menge K, die den positiven orbit γ^+ enthält, eine *endliche Zahl* kritischer Punkte zu.

Satz 11.14 *Gegeben sei das planare autonome System $\dot{x} = f(x)$. Die positive Bahn γ^+ sei in der kompakten Menge K enthalten. Die Menge K enthalte nur eine endliche Zahl kritischer Punkte des Systems. Dann gilt für $\omega(\gamma^+)$ genau eine der folgenden drei Aussagen:*

1. *$\omega(\gamma^+)$ ist ein kritischer Punkt.*
2. *$\omega(\gamma^+)$ ist eine (nichttriviale) geschlossene Bahn.*
3. *$\omega(\gamma^+)$ enthält eine endliche Zahl kritischer Punkte und eine Menge von Bahnen mit der Eigenschaft, dass deren α- und ω-Grenzmengen jeweils aus genau einem kritischen Punkt bestehen.*

Beweis: Als Teilmenge von K enthält $\omega(\gamma^+)$ höchstens eine endliche Anzahl von kritischen Punkten. Wenn $\omega(\gamma^+)$ keine nichtkritischen Punkte enthält, so besteht $\omega(\gamma^+)$ als zusammenhängende Menge aus genau einem kritischen Punkt. Dies ist der erste Fall. Wenn $\omega(\gamma^+)$ eine nichttriviale geschlossene Bahn enthält, so stimmt $\omega(\gamma^+)$ mit dieser Bahn überein. Dies haben wir beim Beweis[56] des Satzes von Poincaré-Bendixson nachgewiesen. Das ist der zweite Fall. Nun nehmen wir an, dass $\omega(\gamma^+)$ nichtkritische Punkte und keine geschlossene Bahn enthalte. Sei $\gamma_0 \subset \omega(\gamma^+)$ eine Bahn in $\omega(\gamma^+)$. Da $\omega(\gamma^+)$ kompakt ist, ist $\omega(\gamma_0) \subset \omega(\gamma^+)$. Wir nehmen an, $p_0 \in \omega(\gamma_0)$ sei ein nichtkritischer Punkt und werden diese Annahme zum Widerspruch führen. Sei l eine Transversale, die p_0 als (relativ) inneren Punkt besitzt. Wie wir beim Beweis von Satz 11.12 gezeigt haben, ist dann

$$l \cap \omega(\gamma^+) = l \cap \omega(\gamma_0) = \{p_0\}.$$

Die Bahn γ_0 muss l in einem Punkt q_0 treffen. Da $\gamma_0 \subset \omega(\gamma^+)$, ist $q_0 = p_0$ und γ_0 eine geschlossene Bahn im Widerspruch zur Voraussetzung. Also enthält $\omega(\gamma_0)$ keine nichtkritischen Punkte und besteht daher, da die Menge zusammenhängend ist, aus genau einem kritischen Punkt. Entsprechendes gilt auch für die α-Grenzmenge. Dies ist der dritte Fall. □

Der Satz von Poincaré-Bendixson ist das wichtigste Hilfsmittel zum Nachweis dafür, dass ein gegebenes planares autonomes Differentialgleichungssystem eine periodische Lösung bzw. eine nichttriviale geschlossene Bahn besitzt. Die Idee besteht dabei i. Allg. darin, ein sogenanntes Bendixsonsches Ringgebiet zu konstruieren. Bahnen, die in diesem Ringebiet starten, sollten in diesem bleiben. Der äußere Rand sollte so beschaffen sein, dass auf diese treffende Bahnen ins Innere zurückgeworfen werden. Entsprechendes sollte für den inneren Rand gelten. Wenn außerdem in diesem Ringgebiet kein kritischer Punkt des gegebenen Systems liegt, ist die Existenz einer nichttrivialen geschlossenen Bahn in diesem Gebiet durch den Satz von Poincaré-Bendixson gesichert. Man kann sich natürlich fragen, weshalb es unbedingt ein *Ringgebiet* sein sollte, in dem sich alles abspielt. Der Grund ist einfach der, dass im Inneren einer geschlossenen Bahn *notwendigerweise* ein kritischer Punkt von f enthalten ist. Wir folgen beim Beweis für diese Aussage (mit kleinen Lücken) der Darstellung bei G. Teschl (2012, S. 226).

[56] Die geschlossene Bahn war dort mit M bezeichnet.

Satz 11.15 *Das planare autonome Differentialgleichungssystem besitze eine nichttriviale periodische Lösung bzw. eine nichttriviale geschlossene Bahn γ. Im Inneren* $\mathrm{int}(\gamma)$ *ist dann ein kritischer Punkt von f enthalten.*

Beweis: Wegen des Jordanschen Kurvensatzes ist $\mathrm{int}(\gamma)$ einfach zusammenhängend und damit wegen des *Riemannschen Abbildungssatzes* konform äquivalent zur offenen Einheitskreisscheibe. Wegen des Satzes von Osgood-Carathéodory (wir haben im Abschnitt über konforme Abbildungen nur einen Spezialfall bewiesen, hier ist also die Lücke!) lässt sich die konforme Abbildung von $\mathrm{int}(\gamma)$ mit der Jordankurve γ als Rand zu einem Homöomorphismus von $C := \mathrm{cl}\,(\mathrm{int}(\gamma))$ auf die abgeschlossene Einheitskreisscheibe fortsetzen. Daher sind C und die abgeschlossene Einheitskreisscheibe homöomorph. Dann ist C invariant, denn eine in $\mathrm{int}(\gamma)$ startende Bahn kann nicht ins Äußere $\mathrm{ext}(\gamma)$ gelangen, denn dazu müsste sie die geschlossene Bahn γ kreuzen, was nicht möglich ist. Eine auf dem Rand γ von C startende Bahn bleibt natürlich sowieso auf dieser Bahn.

Nun gebe man sich eine Folge $\{t_j\}_{j\in\mathbb{N}} \subset \mathbb{R}_+$ mit $\lim_{j\to\infty} t_j = 0$ vor und definiere die Abbildung $F_j\colon C \longrightarrow C$ durch $F_j(x_0) := x(t_j; x_0)$, $j \in \mathbb{N}$. Wie stets ist hierbei $x(t; x_0)$ die Lösung der Anfangswertaufgabe $\dot{x} = f(x)$, $x(0) = x_0$, zur Zeit t. Die stetige Abbildung F_j bildet die zur abgeschlossenen Einheitskreisscheibe homöomorphe Menge C in sich ab. Der *Brouwersche Fixpunktsatz* liefert die Existenz eines Fixpunktes $p_j \in C$ von F_j, $j \in \mathbb{N}$. Indem man notfalls zu einer Teilfolge übergeht, können wir annehmen, dass die Folge $\{p_j\}_{j\in\mathbb{N}} \subset C$ gegen ein $p \in C$ konvergiert. Unser Ziel besteht darin nachzuweisen, dass $x(t; p) = p$ für alle $t > 0$ bzw. $f(p) = 0$ ist, womit die Behauptung bewiesen wäre.

Man gebe sich ein $t > 0$ vor und bestimme $n_j \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ mit $0 \le t - n_j t_j < t_j$, $j \in \mathbb{N}$. Dann ist $\lim_{j\to\infty} n_j t_j = t$ und daher einerseits

$$\lim_{j\to\infty} x(n_j t_j; x_j) = x(t; x).$$

Weiter ist $x(n_j t_j; x_j) = x_j$, $j \in \mathbb{N}$. Diese Aussage ist trivialerweise richtig, falls $n_j = 0$. Da x_j ein Fixpunkt von F_j ist, ist $x(t_j; x_j) = x_j$ und daher die Aussage für $n_j = 1$ richtig. Da aber

$$x(n_j t_j; x_j) = x(t_j + (n_j - 1)t_j; x_j) = x((n_j - 1)t_j; x(t_j; x_j)) = x((n_j - 1)t_j; x_j)$$

ist die Aussage allgemein richtig. Daher ist andererseits

$$\lim_{j\to\infty} x(n_j t_j; x_j) = \lim_{j\to\infty} x_j = x,$$

insgesamt ist der Satz bewiesen. □

Bemerkung: Eine Variante zu Satz 11.15, der aussagt, dass im Inneren einer geschlossenen Bahn eines planaren, autonomen Differentialgleichungssystems ein kritischer Punkt liegt, ist die folgende Aussage (siehe z. B. F. Verhulst (1990, S. 58)).

- *Gegeben sei das autonome Differentialgleichungssystem $\dot{x} = f(x)$, wobei f eine stetig partiell differenzierbare Abbildung des $\mathbb{R}^n$ in sich ist. Sei $V \subset \mathbb{R}^n$ eine bezüglich dieses Systems positiv invariante, kompakte und konvexe Menge. Dann ist in V ein kritischer Punkt von f enthalten.*

Denn: Der Beweis ist fast derselbe wie der von Satz 11.15. Wir wiederholen nur den Anfang. Sei $\{t_j\}_{j\in\mathbb{N}} \subset \mathbb{R}_+$ eine Nullfolge und $F_j\colon V \longrightarrow \mathbb{R}^n$ definiert durch $F_j(x_0) := x(t_j; x_0)$, $j \in \mathbb{N}$. Da V eine bezüglich des Systems $\dot{x} = f(x)$ positiv invariante Menge ist, ist $F_j(V) \subset V$. Der Brouwersche Fixpunktsatz liefert die Existenz eines Fixpunktes $p_j \in V$ von F_j, $j \in \mathbb{N}$. Indem man notfalls zu einer Teilfolge übergeht, können wir annehmen, dass die Folge $\{p_j\}_{j\in\mathbb{N}} \subset V$ gegen ein $p \in V$ konvergiert. Unser Ziel besteht darin nachzuweisen, dass $x(t; p) = p$ für alle $t > 0$ bzw. $f(p) = 0$ ist, womit die Behauptung bewiesen wäre. Dies kann genau wie im Beweis von Satz 11.15 erfolgen. □

11.5 Anwendungen des Satzes von Poincaré-Bendixson

Wir geben zunächst einige spezielle Beispiele an.

Beispiel: Wir betrachten das planare Differentialgleichungssystem (siehe F. Verhulst (1990, S. 49))

$$\begin{aligned} \dot{x}_1 &= x_1 - x_2 - x_1(x_1^2 + x_2^2), \\ \dot{x}_2 &= x_1 + x_2 - x_2(x_1^2 + x_2^2). \end{aligned}$$

Der einzige kritische Punkt dieses Systems ist $(0, 0)$, wie man leicht nachrechnet. Sei $x(t) = (x_1(t), x_2(t))$ eine Lösung dieses Systems und

$$v(t) := \frac{1}{2}[x_1(t)^2 + x_2(t)^2].$$

Dann ist

$$\begin{aligned} \dot{v} &= x_1\dot{x}_1 + x_2\dot{x}_2 \\ &= x_1[x_1 - x_2 - x_1(x_1^2 + x_2^2)] + x_2[x_1 + x_2 - x_2(x_1^2 + x_2^2)] \\ &= (x_1^2 + x_2^2)[1 - (x_1^2 + x_2^2)]. \end{aligned}$$

Man betrachte nun ein Ringgebiet A_{r_1,r_2} mit dem Mittelpunkt $(0, 0)$, einem inneren Radius $r_1 \in (0, 1)$ und einem äußeren Radius $r_2 > 1$. Dieses Ringgebiet ist offenbar wegen obiger Rechnung positiv invariant. Denn ist $(x_1(0), x_2(0)) \in \partial B(0; r_1)$, so ist $\dot{v}(0) = r_1^2(1-r_1^2) > 0$. D. h. eine Bahn, die auf dem Rand des inneren Kreises startet, gelangt zunächst ins Innere des Ringgebietes. Entsprechendes gilt für den äußeren Rand. Da das Ringgebiet auch keinen kritischen Punkt enthält, folgt aus dem Satz von Poincaré-Bendixson die Existenz mindestens einer geschlossenen Bahn in dem Ringgebiet. Für $r_1 = 0.5$, $r_2 = 2$ sowie $x_0 = (0.5, 0)$ und $x_0 = (0, 2)$ geben wir die entsprechenden Bahnen in Abbildung 122 an. □

Beispiel: Jetzt betrachten wir das planare System (siehe F. Verhulst (1990, S. 50), wir ersetzen die rechte Seite durch das Negative!)

$$\begin{aligned} \dot{x}_1 &= -x_1(x_1^2 + x_2^2 - 2x_1 - 3) + x_2, \\ \dot{x}_2 &= -x_2(x_1^2 + x_2^2 - 2x_1 - 3) - x_1 \end{aligned}$$

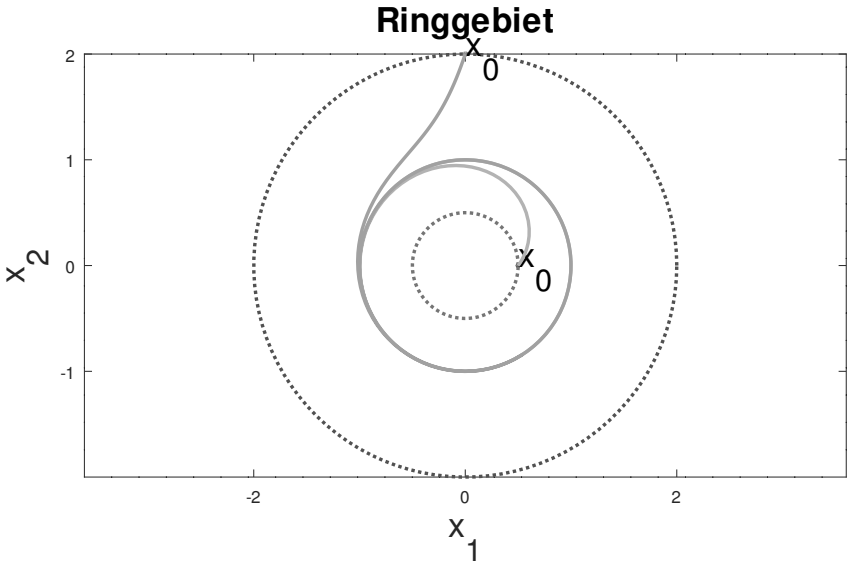


Abbildung 122: Ein Grenzzyklus in einem Ringgebiet I

bzw. $\dot{x} = f(x)$. Wieder ist $(0,0)$ der einzige kritische Punkt. Durch eine Anwendung von Satz 11.2 überlegen wir uns zunächst, in welchem Gebiet garantiert *keine* geschlossene Bahn existieren kann. Es ist

$$\mathrm{div} f(x_1,x_2) = -4x_1^2 - 4x_2^2 + 6x_1 + 6 = -4\left[\left(x_1 - \frac{3}{4}\right)^2 + x_2^2 - \frac{33}{16}\right].$$

Im Kreis mit dem Mittelpunkt $(\frac{3}{4},0)$ und dem Radius $\sqrt{33}/4$ ist $\mathrm{div} f$ von einem Vorzeichen ohne zu verschwinden, sodass im Inneren dieses Kreises keine geschlossene Bahn zu dem gegebenen Differentialgleichungssystem existieren kann. Wie im vorigen Beispiel sei $x(t) = (x_1(t), x_2(t))$ eine Lösung des betrachteten Differentialgleichungssystems und

$$v(t) := \frac{1}{2}[x_1(t)^2 + x_2(t)^2].$$

Dann ist

$$\begin{aligned} \dot{v} &= x_1[-x_1(x_1^2 + x_2^2 - 2x_1 - 3) + x_2] + x_2[-x_2(x_1^2 + x_2^2 - 2x_1 - 3) - x_1] \\ &= -(x_1^2 + x_2^2)[x_1^2 + x_2^2 - 2x_1 - 3]. \end{aligned}$$

Sei nun $r > 0$ und $(x_1(0), x_2(0)) \in \partial B(0;r)$. Offenbar ist dann

$$r^2(3 - 2r - r^2) \le \dot{v}(0) \le r^2(3 + 2r - r^2).$$

Hieraus erkennt man: Ist $r \in (0,1)$, so ist $\dot{v}(0) > 0$, für $r > 3$ ist dagegen $\dot{v}(0) < 0$. Daher ist in dem Ringgebiet

$$A_{1,3} := \{x \in \mathbb{R}^2 : 1 \le \|x\|_2 \le 3\}$$

wegen des Satzes von Poincaré-Bendixson wenigstens eine geschlossene Bahn enthalten.

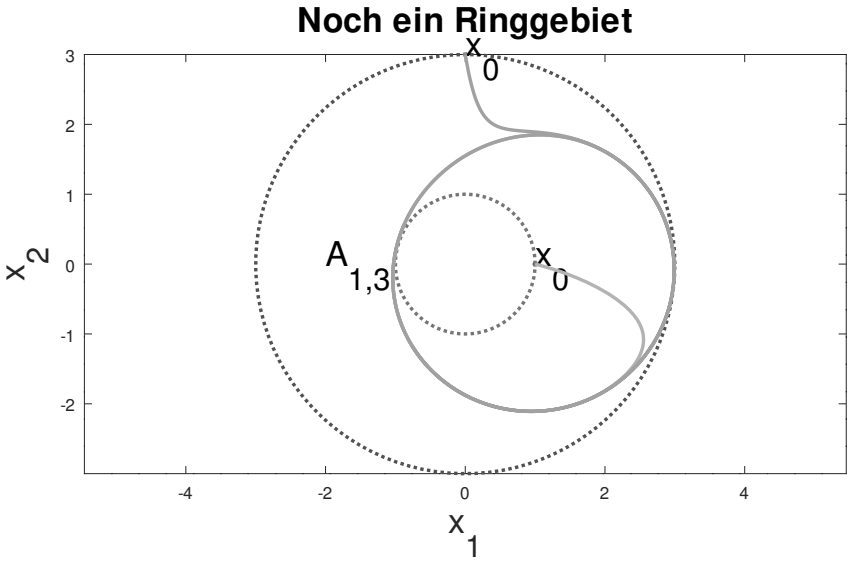


Abbildung 123: Ein Grenzzyklus in einem Ringgebiet II

In Abbildung 123 geben wir das Ringgebiet $A_{1,3}$ sowie in $x_0 = (1, 0)$ bzw. $x_0 = (0, 3)$ startende Bahnen an. □

Beispiel: Bei K. Ciesielski (2012) findet man als Beispiel für die Anwendung des Satzes von Poincaré-Bendixson die autonome Differentialfleichung zweiter Ordnung

$$\ddot{x} - \dot{x}(1 - 3x^2 - 2\dot{x}^2) + x = 0.$$

Eine Überführung auf ein System von zwei Differentialgleichungen erster Ordnung liefert

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_2(1 - 3x_1^2 - 2x_2^2) - x_1 \end{aligned}$$

bzw. $\dot{x} = f(x)$. Dann ist $\mathrm{div} f(x_1, x_2) = 1 - 3x_1^2 - 6x_2^2$, sodass wegen Satz 11.2 im Inneren $\mathrm{int}(E_{3,6})$ der Ellipse

$$E_{3,6} := \{(x_1, x_2) : 3x_1^2 + 6x_2^2 = 1\}$$

keine geschlossene Bahn liegen kann. Wieder sei $x(t) = (x_1(t), x_2(t))$ eine Lösung des betrachteten Differentialgleichungssystems und

$$v(t) := \frac{1}{2}[x_1(t)^2 + x_2(t)^2].$$

Dann ist

$$\dot{v} = x_1\dot{x}_1 + x_2\dot{x}_2 = x_1x_2 + x_2[x_2(1 - 3x_1^2 - 2x_2^2) - x_1] = x_2^2(1 - 3x_1^2 - 2x_2^2).$$

Hieraus erkennt man: Liegt $(x_1(t), x_2(t))$ im Inneren $\mathrm{int}(E_{3,2})$ der Ellipse

$$E_{3,2} := \{(x_1, x_2) : 3x_1^2 + 2x_2^2 = 1\},$$

so ist $\dot{v}(t) \geq 0$, während $\dot{v}(t) \leq 0$ für $(x_1(t), x_2(t)) \in \mathrm{ext}(E_{3,2})$. Jetzt bestimmen wir die größte in $\mathrm{int}(E_{3,2}) \cup E_{3,2}$ enthalte Kreisscheibe mit 0 als Mittelpunkt, ferner entsprechend die kleinste $\mathrm{int}(E_{3,2}) \cup E_{3,2}$ umfassende Kreisscheibe mit 0 als Mittelpunkt. Als Lösungen erhält man offenbar $B[0; 1/\sqrt{3}]$ bzw. $B[0; 1/\sqrt{2}]$. Wir veranschaulichen uns die Situation in Abbildung 124 links. Wegen des Satzes von Poincaré-Bendixson

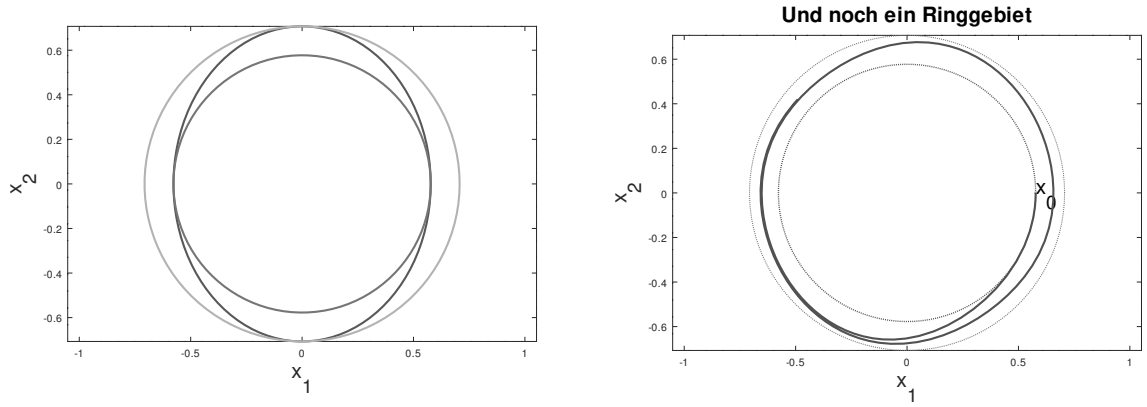


Abbildung 124: Ein Ringgebiet für $\ddot{x} - \dot{x}(1 - 3x^2 - 2\dot{x}^2) + x = 0$

ist in dem Ringgebiet

$$A_{1/\sqrt{3},1/\sqrt{2}} := \left\{ x \in \mathbb{R}^2 : \frac{1}{\sqrt{3}} \leq \|x\|_2 \leq \frac{1}{\sqrt{2}} \right\}$$

mindestens eine geschlossene Bahn enthalten. In Abbildung 124 rechts geben wir eine in $x_0 = (1/\sqrt{3}, 0)$ startende Bahn an. □

Die Konstruktion eines Bendixsonschen Ringgebietes zu den bisherigen Beispielen war relativ einfach. Das wird nun schwieriger. Unser Ziel ist es, den folgenden Satz zu beweisen (siehe z. B. J. K. Hale (1969, S. 57 ff.), F. Verhulst (1990, S. 51) und G. Teschl (2012, S. 216 ff.)). Zugrunde gelegt wird hierbei eine spezielle *Liénardsche Differentialgleichung*.

Satz 11.16 *Gegeben sei die autonome Differentialgleichung zweiter Ordnung*

$$\ddot{x} + f(x)\dot{x} + x = 0, \tag{P}$$

wobei f stetig auf $\mathbb{R}$ ist. Die folgenden Bedingungen seien erfüllt:

(a) *$F(x) := \int_0^x f(s)\,ds$ ist eine in x ungerade Funktion.*

(b) *Es ist $\lim_{x\to\infty} F(x) = +\infty$. Ferner existiert $\beta \in \mathbb{R}$ derart, dass F auf (β, ∞) positiv und monoton wachsend ist. O. B. d. A. ist $F(\beta) = 0$.*

(c) *Es existiert $\alpha > 0$ derart, dass $F(x) < 0$ für $x \in (0, \alpha)$. O. B. d. A. ist $F(\alpha) = 0$.*

Dann besitzt (P) mindestens eine nichttriviale periodische Lösung. Ist $\alpha = \beta$, so besitzt (P) genau eine nichttriviale periodische Lösung.

Bemerkung: Mit $f(x) := -\mu(1 - x^2)$ und $\mu > 0$ ist die van der Polsche Differentialgleichung ein Spezialfall der Gleichung (P) in Satz 11.16. Dann ist

$$F(x) := \int_0^x f(s)\,ds = \mu\left(\frac{1}{3}x^3 - x\right).$$

Offensichtlich ist F eine in x ungerade Funktion. Offensichtlich ist $\lim_{x\to+\infty} F(x) = +\infty$. Auf $(\sqrt{3}, \infty)$ ist F positiv und monoton wachsend. Schließlich ist $F(x) < 0$ für $x \in (0, \sqrt{3})$. Die Voraussetzungen von Satz 11.16 sind also mit $\alpha = \beta = \sqrt{3}$ erfüllt. Der Satz 11.16 liefert also, dass die van der Polsche Differentialgleichung genau eine nichttriviale periodische Lösung besitzt. □

Beweis von Satz 11.16: Der Differentialgleichung (P) zweiter Ordnung wird das System (wir ändern hier die Bezeichnungen ein wenig)

$$(*) \qquad \begin{aligned} \dot{x} &= y - F(x), \\ \dot{y} &= -x \end{aligned}$$

von zwei Differentialgleichungen erster Ordnung zugeordnet. Dieses hat $(0, 0)$ als einzigen kritischen Punkt. Ist $\gamma = \{(x(t), y(t)) : t \in \mathbb{R}\}$ eine (nichttriviale) geschlossene Bahn zu diesem System, so ist $x(\cdot)$ eine (nichttriviale) periodische Lösung von (P). Wir versuchen daher, zu dem System $(*)$ ein Bendixsonsches Ringgebiet zu konstruieren. Als "inneren" Rand können wir hierbei einen Kreis um den Nullpunkt mit einem hinreichend kleinen Radius wählen. Denn sei (x, y) eine Lösung von $(*)$ und

$$V(x, y) := \frac{1}{2}(x^2 + y^2).$$

Dann ist

$$\dot{V} = x\dot{x} + y\dot{y} = x(y - F(x)) + y(-x) = -xF(x).$$

Wegen der Voraussetzungen (a) und (c) ist $\dot{V}(t) = -x(t)F(x(t)) \geq 0$, falls $|x(t)| < \alpha$. Bahnen, die auf dem Rand einer Kreisscheibe mit einem Radius kleiner als α können daher nicht in diese Kreisscheibe eintreten. Die Konstruktion des "äußeren" Randes ist etwas komplizierter. Wir betrachten eine Bahn zum System $(*)$, welche in $A := (0, y_0)$ startet, wobei wir $y_0 > 0$ hinreichend groß wählen werden. Diese Bahn verfolgen wir, bis sie in einem Punkt $D = (0, y_1)$ die y-Achse wieder erreicht. Mit (x, y) ist auch $(-x, -y)$ eine Lösung von $(*)$, da F eine ungerade Funktion ist. Die bei $(0, -y_0)$ startende und in $(0, -y_1)$ die y-Achse wieder erreichende Bahn erhält man also durch Spiegeln der ersten Bahn am Ursprung. Dies verdeutlichen wir uns in Abbildung 125. Ein Kandidat

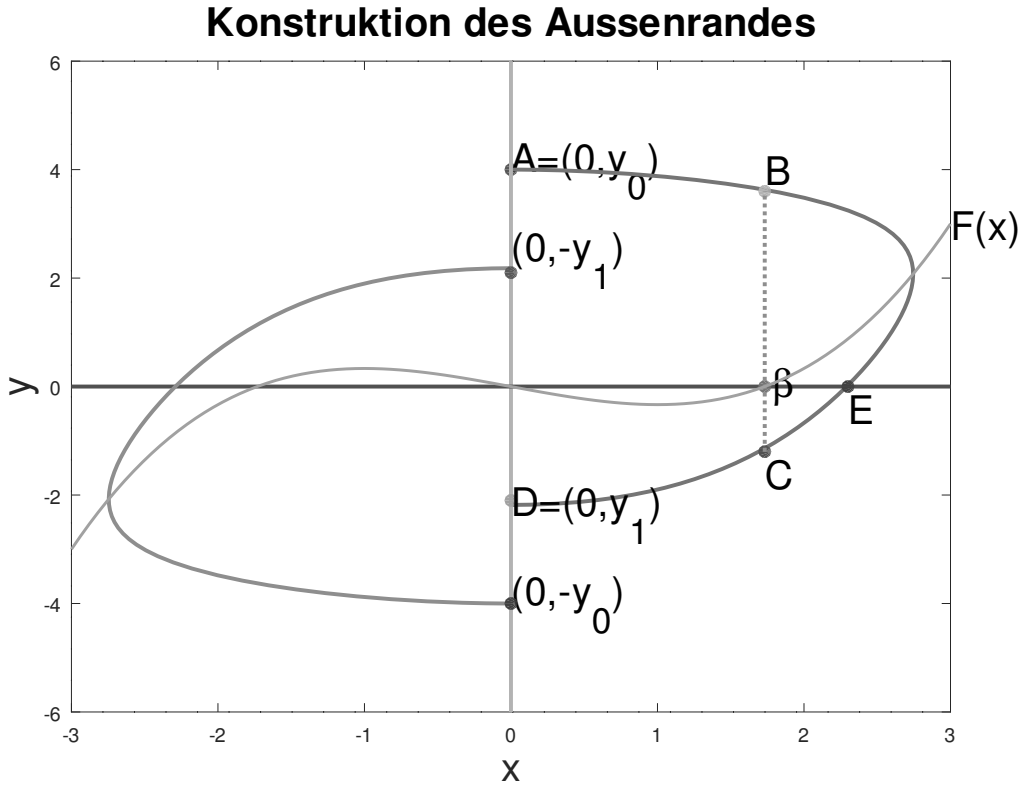


Abbildung 125: Konstruktion des äußeren Randes eines Bendixsonschen Ringgebiets

für den Außenrand eines Bendixsonschen Ringgebietes besteht also aus der Bahn von $A = (0, y_0)$ bis $D = (0, y_1)$, dem Segment von $(0, y_1)$ bis $(0, -y_0)$, der Bahn von $(0, -y_0)$ bis $(0, -y_1)$ (die sich durch Spiegeln der Bahn von A bis D am Ursprung ergibt) und schließlich dem Segment von $(0, -y_1)$ bis $(0, y_0)$. Dies ist aber nur dann ein geeigneter Außenrand, wenn (wie in Abbildung 125) $|y_1| < y_0$. Wir haben uns daher zu überlegen, dass dies für alle hinreichend großen y_0 der Fall ist. Bevor wir dies tun, haben wir aber zu zeigen, dass die in $A = (0, y_0)$ startende Halbbahn $\gamma^+ = \{(x(t), y(t)) : t \geq 0\}$ zumindest für hinreichend große $y_0 > 0$ wirklich so aussieht, wie in Abbildung 125 angegeben. Hierzu verweisen wir auf G. Teschl (2012, S. 216 ff.), wo gezeigt wird:

- Es existiert $r > 0$ derart, dass die zur Zeit $t = 0$ in $(0, r)$ startende Bahn den Graphen von F in $(\beta, 0)$ schneidet.

- Sei $y_0 > r$ und $\gamma^+ = \{(x(t), y(t)) : t \geq 0\}$ die zur Zeit $t = 0$ in $(0, y_0)$ startende Halbbahn. Dann existieren $0 < t_1 < t_2 < T$ mit $x(t_1) = x(t_2) = \beta$ und $(x(T), y(T)) = (0, y_1)$ mit $y_1 < 0$. Ferner ist $y(t) > F(x(t))$ für $t \in [0, t_1]$ und $y(t) < F(x(t))$ für $t \in [t_2, T]$. Insbesondere ist $x(\cdot)$ monoton wachsend und $y(\cdot)$ monoton fallend auf $[0, t_1]$. Entsprechend ist $x(\cdot)$ monoton fallend und $y(\cdot)$ monoton wachsend auf $[t_2, T]$. Auf diesen Intervallen sind also jeweils die Umkehrfunktionen erklärt. Z. B. existiert zu der Abbildung $x\colon [0, t_1] \longrightarrow [0, \beta]$ die

Umkehrabbildung $x^{-1}\colon [0,\beta] \longrightarrow [0,t_1]$, welche also einem $x \in [0,\beta]$ den Wert $t = x^{-1}(x) \in [0,t_1]$ zuordnet, für welchen $x(t) = x$ ist.

In Abbildung 126 veranschaulichen wir uns diese beiden Aussagen. Die erste Aussage

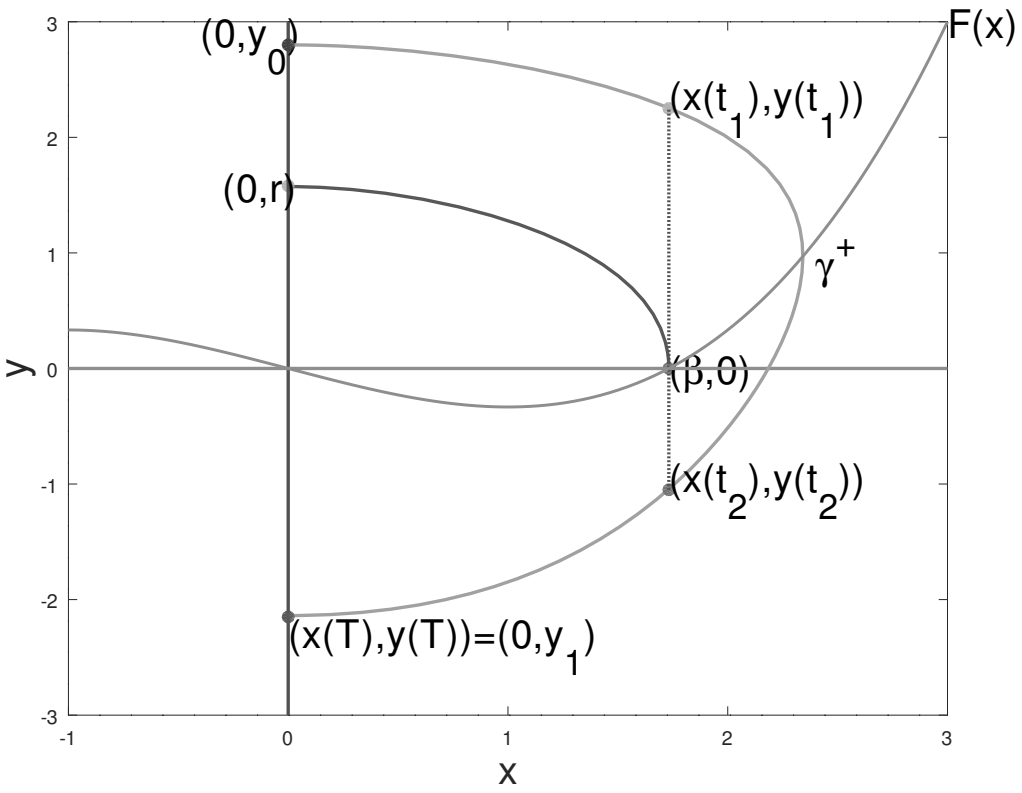


Abbildung 126: Veranschaulichung des Beweises von Satz 11.16

erhält man hierbei, indem man ausgehend von $(\beta, 0)$ die *negative* Halbbahn zum System $(*)$ verfolgt und sich überlegt, dass diese $L_+ := \{(0,y) : y > 0\}$ in einem Punkt $(0,r)$ schneidet.

Zu einem gegebenen $y_0 > r$ betrachten wir nun das von $A = (0,y_0)$ nach $D = (0,y_1)$ verlaufende Bahnsegment $\{(x(t),y(t)) : t \in [0,T]\}$. Mit

$$V(x,y) = \frac{1}{2}(x^2 + y^2)$$

ist dann

$$\begin{aligned}
\Delta(y_0) &:= \frac{1}{2}(y_1^2 - y_0^2) \\
&= V(0,y_1) - V(0,y_0) \\
&= V(x(T),y(T)) - V(x(0),y(0)) \\
&= \int_0^T \frac{d}{dt} V(x(t),y(t))\, ds
\end{aligned}$$

$$
\begin{aligned}
&= -\int_0^T x(t)F(x(t))\,dt \\
&= \underbrace{-\int_0^{t_1} x(t)F(x(t))\,dt}_{=:\Delta_1(y_0)} \underbrace{-\int_{t_1}^{t_2} x(t)F(x(t))\,dt}_{=:\Delta_2(y_0)} \underbrace{-\int_{t_2}^{T} x(t)F(x(t))\,dt}_{=:\Delta_3(y_0)}.
\end{aligned}
$$

Wir wollen nun zeigen, dass $\Delta(y_0) < 0$ bzw. $|y_1| < y_0$ für alle hinreichend großen $y_0 > 0$ ist. Es ist

$$
\begin{aligned}
\Delta_1(y_0) &= -\int_0^{t_1} x(t)F(x(t))\,dt \\
&= -\int_0^{\beta} \frac{xF(x)}{\dot{x}(x^{-1}(x))}\,dx \\
&\qquad \text{(Substitution } x(t) = x\text{)} \\
&= \int_0^{\beta} \frac{-xF(x)}{y(x^{-1}(x)) - F(x)}\,dx.
\end{aligned}
$$

Im Integranden, dessen Nenner positiv ist, hängt nur $y(x^{-1}(x))$ von y_0 ab. Mit wachsendem y_0 wächst auch $y(x^{-1}(x))$, da Bahnen sich nicht schneiden können. Daher ist in der Ungleichung

$$
|\Delta_1(y_0)| \le \int_0^{\beta} \frac{x|F(x)|}{y(x^{-1}(x)) - F(x)}\,dx
$$

die rechte Seite in y_0 monoton fallend. Insbesondere ist $\{\Delta_1(y_0) : y_0 > r\}$ beschränkt. Weiter ist

$$
\begin{aligned}
\Delta_2(y_0) &= -\int_{t_1}^{t_2} x(t)F(x(t))\,dt \\
&= -\int_{y(t_1)}^{y(t_2)} \frac{x(y^{-1}(y))F(x(y^{-1}(y)))}{\dot{y}(y^{-1}(y))}\,dy \\
&\qquad \text{(Substitution } y(t) = y\text{)} \\
&= \int_{y(t_1)}^{y(t_2)} F(x(y^{-1}(y)))\,dy \\
&= -\int_{y(t_2)}^{y(t_1)} F(x(y^{-1}(y)))\,dy.
\end{aligned}
$$

Wegen Bedingung (b) sowie $y(t_1) \to \infty$ mit $y_0 \to \infty$ gilt $\Delta_2(y_0) \to -\infty$ mit $y_0 \to \infty$. Schließlich ist

$$
\begin{aligned}
\Delta_3(y_0) &= -\int_{t_2}^{T} x(t)F(x(t))\,dt \\
&= -\int_{\beta}^{0} \frac{xF(x)}{\dot{x}(x^{-1}(x))}\,dx \\
&\qquad \text{(Substitution } x(t) = x\text{)}
\end{aligned}
$$

$$= \int_0^\beta \frac{-xF(x)}{F(x) - y(x^{-1}(x))}\,dx.$$

Hier können wir zunächst wörtlich so argumentieren wie bei der Abschätzung von $\Delta_1(y_0)$: Im Integranden, dessen Nenner positiv ist, hängt nur $y(x^{-1}(x))$ von y_0 ab. Mit wachsendem y_0 *fällt* $y(x^{-1}(x))$, da Bahnen sich nicht schneiden können. Daher ist in der Ungleichung

$$|\Delta_3(y_0)| \le \int_0^\beta \frac{x|F(x)|}{F(x) - y(x^{-1}(x))}\,dx$$

die rechte Seite in y_0 monoton fallend. Insbesondere ist $\{\Delta_3(y_0) : y_0 > r\}$ beschränkt. Insgesamt gilt $\Delta(y_0) \to -\infty$ mit $y_0 \to \infty$, womit schließlich gezeigt ist, dass $\Delta(y_0) < 0$ bzw. $|y_1| < y_0$ für alle hinreichend großen y_0. Damit ist die Existenz eines Bendixsonschen Ringgebietes nachgewiesen. Wegen des Satzes von Poincaré-Bendixson besitzt das System $(*)$ mindestens eine nichttriviale geschlossene Bahn bzw. die Liénardsche Differentialgleichung (P) wenigstens eine nichttriviale periodische Lösung.

Nun sei $\alpha = \beta$ in den Voraussetzungen (b) und (c) von Satz 11.16. Insbesondere ist dann $F(x) < 0$ für alle $x \in (0, \beta)$. In

$$\Delta_1(y_0) = \int_0^\beta \frac{-xF(x)}{y(x^{-1}(x)) - F(x)}\,dx, \qquad \Delta_3(y_0) = \int_0^\beta \frac{-xF(x)}{F(x) - y(x^{-1}(x))}\,dx$$

sind nicht nur die Nenner der Integranden, sondern wegen $\alpha = \beta$ auch die Zähler auf $(0, \beta)$ positiv. Daher sind $\Delta_1(y_0)$ und $\Delta_3(y_0)$ in y_0 monoton fallende Funktionen. Aber auch

$$\Delta_2(y_0) = -\int_{y(t_2)}^{y(t_1)} F(x(y^{-1}(y)))\,dy$$

ist eine monoton fallende Funktion in y_0, da Bahnen zum System $(*)$ sich nicht kreuzen können. Insgesamt ist $\Delta(\cdot)$ als Summe der monoton fallenden Funktionen Δ_1, Δ_2 und Δ_3 eine auf (r, ∞) monoton fallende Funktion, die wegen $\Delta(y_0) \to -\infty$ mit $y_0 \to \infty$ auf (r, ∞) genau eine Nullstelle besitzt. Dies impliziert, dass $(*)$ genau eine nichttriviale geschlossene Bahn bzw. (P) genau eine nichttriviale periodische Lösung besitzt. □

Beispiel: In den Abbildungen 125 und 126 haben wir die Konstruktion des Außenrandes eines Bendixsonschen Ringgebietes dargestellt, wobei in dem System

$$(*) \qquad \begin{aligned} \dot{x} &= y - F(x), \\ \dot{y} &= -x \end{aligned}$$

die Funktion F durch

$$F(x) := \mu\left(\frac{1}{3}x^3 - x\right)$$

mit $\mu > 0$ gegeben ist (wir wählten $\mu = 0.5$), es sich bei $(*)$ also im wesentlichen um die van der Polsche Differentialgleichung handelt. In Satz 11.16 ist $\alpha = \beta$ und dies impliziert die Existenz genau einer nichttrivialen geschlossenen Bahn. Bei W. S. Koon (2009) ist ein Beispiel angegeben, bei welchem zwei Grenzzyklen existieren. Und zwar sei

$$F(x) := \frac{8}{25}x^5 - \frac{4}{3}x^3 + \frac{4}{5}x$$

im System $(*)$. In Abbildung 127 geben wir die Funktion F mit den beiden positiven Nullstellen $\alpha \approx 0.85251$ und $\beta \approx 1.8547$ wieder. Man beachte, dass die Voraussetzungen

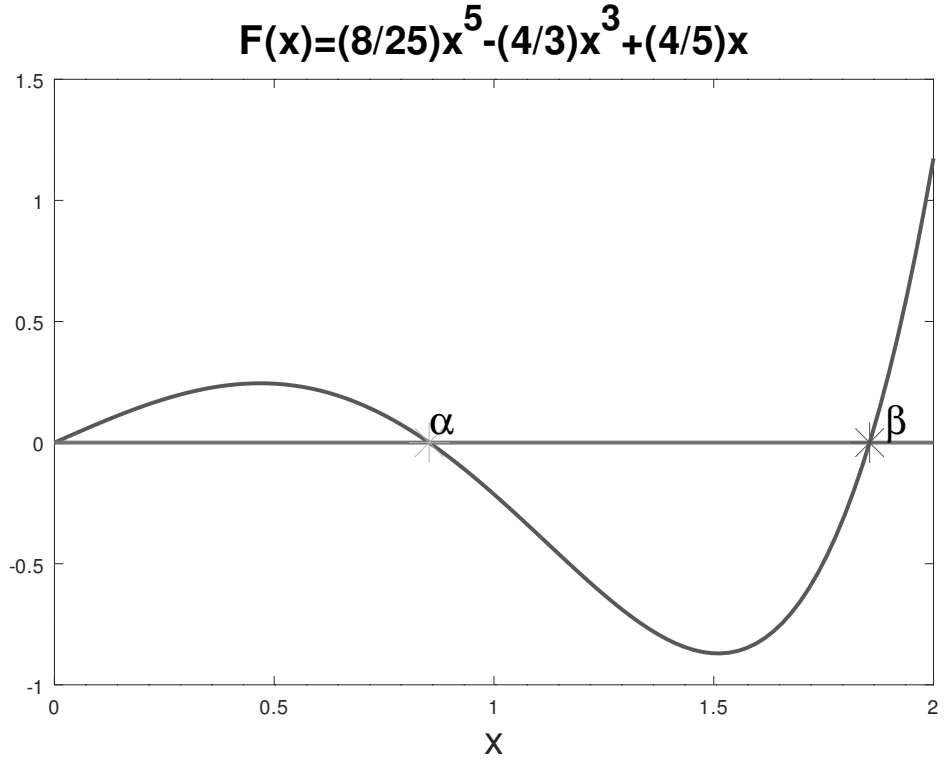


Abbildung 127: Die Funktion $F(x) = \frac{8}{25}x^5 - \frac{4}{3}x^3 + \frac{4}{5}x$

von Satz 11.16 *nicht* erfüllt sind, da F auf $(0, \alpha)$ positiv ist. In Abbildung 128 sieht man Teile von zwei Bahnen. Die äußere ist eine in $x_0 = (3, 0)$ startende positive Bahn, die innere ist eine in $x_0 = (0.2, 0)$ startende *negative* Bahn. Offenbar gibt es zwei Grenzzyklen, wobei ein genauer Beweis dafür schwierig sein dürfte. □

Literatur

[1] Ahlfors, L. V. (1966) *Complex Analysis. Second Edition*. McGraw Hill, New York.

[2] Ahrens, W. (1901) *Mathematische Unterhaltungen und Spiele*. B. G. Teubner, Leipzig. Im Internet unter `https://ia902606.us.archive.org/27/items/mathematischeun02ahregoog/mathematischeun02ahregoog.pdf` einzusehen.

[3] Aigner, M. und G. M. Ziegler (2018) *Das BUCH der Beweise. 5. Auflage*. Springer, Berlin-Heidelberg-New York.

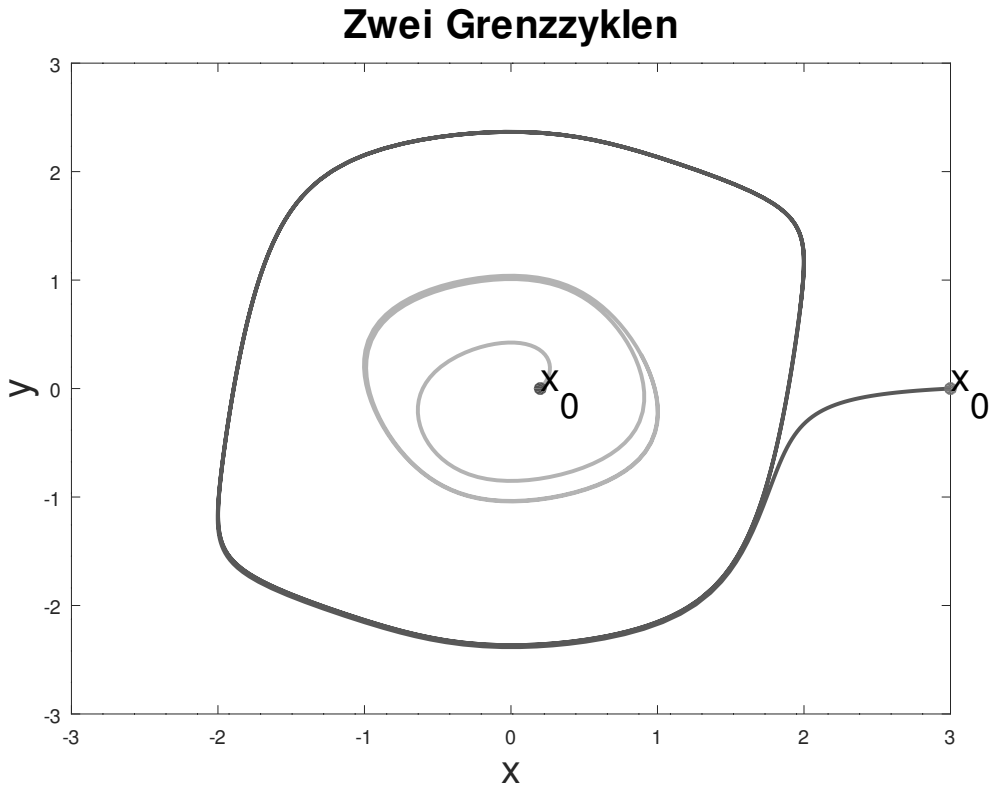


Abbildung 128: Zwei Grenzzyklen

[4] Anderson, I. (1990) *Combinatorial Designs: Construction Methods*. John Wiley & Sons, New York-Chichester-Brisbane-Toronto.

[5] Angeltveit, V. and B. D. McKay (2017) $R(5,5) \leq 48$. Im Internet unter `https://arxiv.org/pdf/1703.08768.pdf` einzusehen.

[6] Bak, J. and D. J. Newman (2010) Complex Analysis. Third Edition. Springer, New York-Heidelberg-London.

[7] Baker, M. and D. Clark (2002) Prime Number Theorem Lecture Notes. Im Internet unter `http://www.math.harvard.edu/archive/113_fall_01/113pnt.pdf` einzusehen.

[8] Ball, K. (1992) Ellipsoids of maximal volume in convex bodies. Im Internet unter `https://arxiv.org/pdf/math/9201217.pdf` einzusehen.

[9] Ball, K. (1997) *An Elementary Introduction to Modern Convex Geometry*. Flavors of Geometry. MSRI Publications Volume 31. Im Internet unter `https://people.eecs.berkeley.edu/~wainwrig/Reading/ball97.pdf` einzusehen.

[10] Ball, S. and Z. Weiner (2011) An Introduction to Finite Geometry. Im Internet unter https://mat-web.upc.edu/people/simeon.michael.ball/IFG.pdf einzusehen.

[11] Bateman, H. and H. Diamond (1996) A hundred years of prime numbers. Amer. Math. Monthly 103, 729—741. Im Internet unter https://www.math.fsu.edu/~quine/ANT/The%20American%20mathematical%20monthly%201996%20Bateman.pdf einzusehen.

[12] Behrend, F. (1938) Über die kleinste umbeschriebene und die größte einbeschriebene Ellipse eines konvexen Bereichs. Math. Ann. 115, 379–411.

[13] Beutelspacher, A. und U. Rosenbaum (2004) *Projektive Geometrie. von den Grundlagen bis zu den Anwendungen. 2. Auflage*. Friedr. Vieweg & Sohn Verlag, Wiesbaden.

[14] Bishop, C. J. (2010) The Riemann Mapping Theorem. Im Internet unter https://www.math.stonybrook.edu/~bishop/classes/math626.F08/rmt.pdf einzusehen.

[15] Bogosel, B. (2017) A geometric proof of the Siebeck-Marden theorem. American Mathematical Monthly 124, 459–463. Im Internet unter http://www.cmap.polytechnique.fr/~beniamin.bogosel/pdfs/mardenthm.pdf einzusehen.

[16] Bollobás, B. (1998) *Modern Graph Theory*. Springer, New York-Berlin-Heidelberg.

[17] Bose, A. C. (1939) On the construction of balanced incomplete block designs. Ann. Eugenics 9, 353–399. Im Internet unter https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-1809.1939.tb02219.x einzusehen.

[18] Bruck, R. H. and H. J. Ryser (1949) The nonexistence of certain finite projective planes. Canadian Journal of Mathematics 1, 88–93. Im Internet unter https://cms.math.ca/openaccess/cjm/v1/cjm1949v01.0088-0093.pdf einzusehen.

[19] Bundschuh, P. (2002) *Einführung in die Zahlentheorie. 5., überarbeitete und aktualisierte Auflage*. Springer, Berlin-Heidelberg-New York.

[20] Cameron, P. J. (1994) *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, Cambridge.

[21] Casarotto, C. (2006) Graph Theory and Cayley's Formula. In Internet unter https://www.math.uchicago.edu/~may/VIGRE/VIGRE2006/PAPERS/Casarotto.pdf einzusehen.

[22] Cayley, A. (1889) A theorem on trees. Quart. J. Pure Appl. Math. 23, 376–378. Collected Mathematical Papers Vol. 13, Cambridge University Press 1897, 26–28. Im Internet unter https://books.google.de/books?id=M7c4AAAAIAAJ&pg=PA26&redir_esc=y&hl=de#v=onepage&q&f=false einzusehen.

[23] Cesari, L. (1970) *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations. Third Edition*. Springer-Verlag. Berlin-Heidelberg-New York.

[24] Ciesielski, K. (2012) The Poincaré-Bendixson Theorem: from Poincaré to the XXIst century. Central European Journal of Mathematics 10, 2110—2128. Im Internet unter `https://www.degruyter.com/view/journals/math/10/6/article-p2110.xml` einzusehen.

[25] Coddington, E. A. and N. Levinson (1955) *Theory of Ordinary Differential Equations*. Tata McGraw-Hill. New Delhi. Im Internet unter `https://ptvtpqa.files.wordpress.com/2013/12/coddington-e-levinson-n-theory-of-ordinary-differential-equations.pdf` einzusehen.

[26] Colbourn, C. J. and A. Rosa (1999) *Triple Systems*. Clarendon Press, Oxford.

[27] Conway, J. B. (1973) *Functions of One Complex Variable*. Springer-Verlag, New York-Heidelberg-Berlin.

[28] Danzer, L., D. Laugwitz und H. Lenz, H. (1957) Über das Löwnersche Ellipsoid und sein Analogon unter den einem Eikörper einbeschriebenen Ellipsoiden. Arch. Math. 8, 214–219.

[29] Deuber, W. (1973) Partitionen und lineare Gleichungssysteme. Math. Z. 133, 109–123.

[30] Diener, I., E. Schmitt and H. L. de Vries (1985) All 80 Steiner triple systems on 15 elements are derived. Discrete Mathematics 55, 13–19. Im Internet unter `https://ac.els-cdn.com/S0012365X85800170/1-s2.0-S0012365X85800170-main.pdf?_tid=4b58a669-da75-4bba-8975-c3dcac423eeb&acdnat=1531808395_08bfc93d9b30d9e84a0208e1138eda39` einzusehen.

[31] Eberlein, P. J. and G. S. Mudholkar (1968) Some remarks on the van der Waerden conjecture. Journal of Combinatorial Theory 5, 386–396. Im Internet unter `https://core.ac.uk/download/pdf/82268738.pdf` einzusehen.

[32] Eberlein, P. J. (1969) Remarks on the van der Waerden conjecture, II. Linear Algebra and its Applications 2, 311—320. Im Internet unter `https://core.ac.uk/download/pdf/82035950.pdf` einzusehen.

[33] Egorychev, G. P. (1981) The solution of van der Waerden's problem for permanents. Advances in Mathematics 42, 299–305. Im Internet unter `https://ac.els-cdn.com/000187088190044X/1-s2.0-000187088190044X-main.pdf?_tid=554e346f-eee7-4cee-b614-a4e0b6150d65&acdnat=1537975865_f529c8b1b3cdc69aac0519ef0fb2d8ef` einzusehen.

[34] Erdös, P. and G. Szekeres (1935) A combinatorial problem in geometry. Compositio Math. 2, 463–470. Im Internet unter https://renyi.hu/~p_erdos/1935-01.pdf einzusehen.

[35] Erdös, P. and G. Szekeres (1960) On some extremum problems in elementary geometry. Ann. Univ. Sci. Budapest. Eötvös Sect. Math. 3–4, 53–62. Im Internet unter http://renyi.hu/~p_erdos/1960-09.pdf einzusehen.

[36] Erdös, P. (1947) Some remarks on the theory of graphs. Bulletin Amer. Math. Soc. 53, 292–294. Im Internet unter http://www.ams.org/journals/bull/1947-53-04/S0002-9904-1947-08785-1/S0002-9904-1947-08785-1.pdf einzusehen.

[37] Falikman, D. I. (1981) Proof of the van der Wawerden conjecture regarding the permanent of a doubly stochastic matrix (Russian). Mat. Zametki 29, 931–938. Im Internet unter http://www.mathnet.ru/links/113dec35fc2516671299ab0a13dccba4/mzm6253.pdf einzusehen.

[38] Farkas, M. (1994) *Periodic Motions*. Springer-Verlag. New York.

[39] Gaier, D. (1964) *Konstruktive Methoden der konformen Abbildung*. Springer-Verlag, Berlin-Göttingen-Heidelberg.

[40] Gessel, I. and G.-C. Rota (Ed.) (1987) *Classic Papers in Combinatorics*. Birkhäuser, Boston-Basel-Stuttgart.

[41] Graham, R. L. and B. L. Rothschild (1974) A short proof of van der Waerden's theorem on arithmetic progressions. Proceedings of the American Mathematical Society 42, 285–386. Im Internet unter http://www.math.ucsd.edu/~ronspubs/74_01_van_der_waerden.pdf einzusehen.

[42] Graham, R. L. (1981) *Rudiments of Ramsey Theory.* Regional Conference Series in Mathematics Number 45. Published by the American Mathematical Society. Im Internet unter http://www.math.ntu.edu.tw/~gjchang/courses/2015-02-Combinatorics-II/[09]Graham-Rediments-RT-chp-3-4.pdf einzusehen.

[43] Graham, R. L., B. L. Rothschild and J. H. Spencer (1990) *Ramsey Theory. Second Edition*. John Wiley & Sons. New York-Chichester-Brisbane-Toronto-Singapore.

[44] Granell, M. and T. Griggs (1994) An introduction to Steiner systems. Mathematical Spectrum 26 no. 3, 74–80. Im Internet unter http://mcs.open.ac.uk/mjg47/papers/IntroSteiner.pdf einzusehen.

[45] Güler, O. and F. Gürtuna (2007) The extremal volume ellipsoids of convex bodies, their symmetry properties, and their determination in some special cases. Im Internet unter http://www.optimization-online.org/DB_FILE/2007/09/1771.pdf einzusehen.

[46] Gurvits, L. (2008) Van der Waerden/Schrijver like conjectures and stable (aka homogeneous) polynomials: one theorem for all. Im Internet unter https://arxiv.org/pdf/0711.3496.pdf einzusehen.

[47] Gutknecht, M. H. (1977) Existence of a solution to the discrete Theodorsen equation for conformal mapping. Math. Comp. 31, 478—480.

[48] Gutknecht, M. H. (1979) Fast algorithms for the conjugate periodic function. Computing 22, 79—91. Im Internet unter http://www.sam.math.ethz.ch/~mhg/pub/mhg-published/07-Gut79-Computing22.pdf einzusehen.

[49] Gutknecht, H. M. (1981) Solving Theodorsen's integral equation for conformal maps with fast Fourier transform and various iterative methods. Numer. Math. 36, 405—429. Im Internet unter http://www.sam.math.ethz.ch/~mhg/pub/mhg-published/Gut80-NM36-sTiefcm.pdf einzusehen.

[50] Hale, J. K. (1969) *Ordinary Differential Equations*. Wiley-Interscience. New York-London-Sydney-Toronto.

[51] Hall, M. Jr. (1986) *Combinatorial Theory*. John Wiley & Sons. New York-Chichester-Brisbane-Toronto-Singapore.

[52] Hanani, H. (1960) On quadruple systems. Canadian J. Math. 12, 145–157. Im Internet unter https://cms.math.ca/openaccess/cjm/v12/cjm1960v12.0145-0157.pdf einzusehen.

[53] Hartman, P. (1964) *Ordinary Differential Equations*. John Wiley & Sons, Inc. New York-London-Sydney.

[54] Hartsfield, N. and G. Ringel (1990) *Pearls in Graph Theory. A comprehensive introduction*. Academic-Press, Boston-San Diego-New York-London-Sydney-Tokyo-Toronto.

[55] Heffter, L. (1897) Ueber Tripelsysteme. Math. Annalen 49, 101–112.

[56] Henk, M. (2012) Löwner-John ellipsoids. Documenta Mathematica. Extra Volume ISMP, 95–106. Im Internet unter http://www.math.uiuc.edu/documenta/vol-ismp/24_henk-martin.pdf einzusehen.

[57] Hirsch, M. W., S. Smale and R. L. Devaney (2004) *Differential Equations, Dynamical Systems and an Introduction to Chaos*. Elsevier Academic Press. Amsterdam-Boston-Heidelberg. Im Internet unter https://thalis.math.upatras.gr/~bountis/files/def-eq.pdf einzusehen.

[58] Howard, R (1997) The John ellipsoid theorem. Im Internet unter http://people.math.sc.edu/howard/Notes/john.pdf einzusehen.

[59] Hughes, D. R. and F. C. Piper (1973) *Projective Planes*. Springer-Verlag, New York-Heidelberg-Berlin.

[60] Hughes, D. R. and F. C. Piper (1985) *Design Theory*. Cambridge University Press, Cambridge.

[61] Jacobs, K. (1983) *Einführung in die Kombinatorik*. Walter de Gruyter, Berlin-New York.

[62] Joyal, A. Une théorie combinatoire des séries formelles. Advances in Math. 42, 1–82. Im Internet unter `https://ac.els-cdn.com/0001870881900529/1-s2.0-0001870881900529-main.pdf?_tid=cadab38a-0997-11e8-9b5e-00000aacb35e&acdnat=1517740998_ae61fa279a48af6e37cc23385d2d641a` einzusehen.

[63] John, F. (1948) Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*. Interscience Publishers, Inc., New York, N. Y., pp. 187–204.

[64] John, F. (1985) *Collected Papers. Volume 2.* J. Moser (Ed.). Birkhäuser, Boston-Basel-Stuttgart.

[65] Juhnke, F. (1994) Embedded maximal ellipsoids and semi-infinite optimization. Beiträge Algebra Geom. 35, 163–171. Im Internet unter `https://www.emis.de/journals/BAG/vol.35/no.2/b35h2juh.ps.gz` einzusehen.

[66] Kalman, D. (2008) An elementary proof of Marden's Theorem. American Mathematical Monthly 115, 330-338. Im Internet unter `http://dankalman.net/AUhome/pdffiles/mardenAMM.pdf` einzusehen.

[67] Kalman, D. (2008) The most marvelous theorem in mathematics. Im Internet unter `http://www.maa.org/external_archive/joma/Volume8/Kalman/index.html` einzusehen.

[68] Knuth, D. E. (1981) A permanent inequality. Amer. Math. Monthly 88, 731–740.

[69] Korevaar, J. (1982) On Newman's quick way to the prime number theorem. Math. Intelligencer 4, 108—115. Im Internet unter `https://staff.fnwi.uva.nl/j.korevaar/KorNewmanPNT.pdf` einzusehen.

[70] Koon, W. S. (2009) Lectures on periodic orbits. Im Internet unter `http://www.cds.caltech.edu/archive/help/uploads/wiki/files/224/cds140b-perorb.pdf` einzusehen.

[71] Kress, R. (1997) *Numerical Analysis*. Springer, New York-Berlin-Heidelberg.

[72] Lam, C. W. H. (1996) The search for a finite projective plane of order 10. Im Internet einzusehen unter `http://www.cecm.sfu.ca/organics/papers/lam/paper/html/paper.html`.

[73] Landman, B. M. and A. Robertson (2014) *Ramsey Theory on the Integers. Second Edition*. AMS, Providence, Rhode Island.

[74] Laurent, M. and A. Schrijver (2010) On Leonid Gurvit's proof for permanents. Amer. Math. Monthly 117, 903–911. Im Internet unter https://homepages.cwi.nl/~monique/files/monthly903-911-schrijver.pdf einzusehen.

[75] Leader, I. (2013) Part III Ramsey Theory. Im Internet unter https://maths.ucd.ie/~stiofainf/lecture-notes/ramsey.pdf einzusehen.

[76] Lee, G. E. and D. Zeilberger (2012) Joyal's proof of Cayley's formula. Im Internet unter http://sites.math.rutgers.edu/~zeilberg/mamarim/mamarimPDF/JoyalCayley.pdf einzusehen.

[77] Lindner, C. C. and C. A. Rodger (1997) *Design Theory*. CRC Press, Boca Raton-New York.

[78] Lindner, C. C. and A. Rosa (1978) Steiner quadruple systems-a survey. Discrete Mathematics 22, 147–182. Im Internet unter https://ac.els-cdn.com/0012365X7890122X/1-s2.0-0012365X7890122X-main.pdf?_tid=2e922d36-5c08-436e-bc17-54ad44ab0997&acdnat=1531562409_56e921d6ab3b0701ab75ea57732f2764 einzusehen.

[79] Lint, J. H. van and R. M. Wilson (1992) *A Course in Combinatorics*. Cambridge University Press, Cambridge.

[80] Liu, C. (2016) Ramsey Theory. Im Internet unter http://math.uchicago.edu/~may/REU2016/REUPapers/Liu,C.pdf einzusehen.

[81] London, D (1971) Some notes on the van der Waerden conjecture. Linear Alg. Appl. 4, 155–160. Im Internet unter https://ac.els-cdn.com/002437957190036X/1-s2.0-002437957190036X-main.pdf?_tid=4a4e2f64-e938-4100-abb5-41cdcf39d445&acdnat=1541028534_197a95b7a4549d769bfa29c0e01a5eb9 einzusehen.

[82] Luk, J. (2017) Notes on the Poincaré-Bendixson theorem. Im Internet unter https://pdfs.semanticscholar.org/946d/d86b1c3b9be34f75638a9abc09027e039b9e.pdf?_ga=2.266605198.1051640702.1592325751-610147527.1592325751 einzusehen.

[83] Marcus, M. and H. Minc (1962) Some results on doubly stochastic matrices. Proc. Amer. Math. Soc. 13, 571–579.

[84] Marcus, M. and M. Newman (1959) On the minimum of the permanent of a doubly stochastic matrix. Duke Math. J. 26, 61–72.

[85] Marden, M. (1945) A note on the zeros of sections of a partial fraction. Bull. Amer. Math. Soc./ 51, 935–940. Im Internet unter http://projecteuclid.org/download/pdf_1/euclid.bams/1183507539 einzusehen.

[86] Matoušek, J. und J. Nešetřil (2002). *Diskrete Mathematik. Eine Entdeckungsreise*. Springer, Berlin-Heidelberg-New York-Barcelona-Hongkong-London-Mailand-Paris-Tokio.

[87] McShane, E. J. (1937) On the Osgood-Carathéodory theorem. Amer. Math. Monthly 44, 288—291.

[88] Melenk, J. M. (2016) Satz von Poincaré-Bendixson. Im Internet unter `https://www.asc.tuwien.ac.at/~melenk/teach/ode_SS16/poincare-bendixson.pdf` einzusehen.

[89] Minc, H. (1978) *Permanents*. Encyclopedia of Mathematics and its Applications Volume 6. Addison-Wesley Publishing Company. London-Amsterdam-Don Mills, Ontario-Sydney-Tokyo.

[90] Minc, H. (1983) The van der Waerden permanent conjecture. In: General Inequalities 3 (Edited by E. F. Beckenbach, W. Walter), 23–40. Birkhäuser Verlag, Basel-Boston-Stuttgart.

[91] Minda, D. and S. Phelps (2008) Triangles, ellipses and cubic polynomials. American Mathematical Monthly 115, 679–689. Im Internet unter `https://pdfs.semanticscholar.org/fd3b/7d1fadd19f0677f6f6faa4426115d4b57ef7.pdf` einzusehen.

[92] Monard, F. (2017) Lecture 11-Montel's theorem, Riemann's mapping theorem. Im Internet unter `https://people.ucsc.edu/~fmonard/Sp17_Math207/lecture11.pdf` einzusehen.

[93] Moreno, C. J. and S. S. Wagstaff Jr. (2005) *Sums of Squares of Integers*. CRC Press, Baton Rouge, Florida.

[94] Morris, W. and V. Soltan (2000) The Erdös-Szekeres problem on points in convex position-a survey. Bull. Amer. Math. Soc. (N. S.) 37, 437–458. Im Internet unter `http://www.ams.org/journals/bull/2000-37-04/S0273-0979-00-00877-6/S0273-0979-00-00877-6.pdf` einzusehen.

[95] Nešetřil, J. (1995) Ramsey Theory. In: *Handbook of Combinatorics II*. Edited by R. L. Graham, M. Grötschel, L. Lovász. Elsevier, Amsterdam-Lausanne-New York.

[96] Newman, D. J. (1980) Simple analytic proof of the prime number theorem. Amer. Math. Monthly 87, 693—696. Im Internet unter `http://www.math.stonybrook.edu/~moira/mat331-spr10/papers/1980%20NewmanSimple%20Analytic%20Proof%20of%20the.pdf` einzusehen.

[97] Newman, D. J. (1998) *Analytic Number Theory*. Springer, New York-Berlin-Heidelberg.

[98] Novinger, W. P. (1975) An elementary approach to the problem of extending conformal maps to the boundary. Amer. Math. Monthly 82, 279—282.

[99] O'Rourke, C. (2013) The prime number theorem: Analytic and elementary proofs. Im Internet unter http://eprints.maynoothuniversity.ie/4470/1/finaldraftmsc.pdf einzusehen.

[100] Ostrowski, A. (1952) On a discontinuous analogue of Theodorsen's and Garrick's method. Nat. Bur. Standards Appl. Math. Series 18, 165-174.

[101] Peltesohn, R. (1939) Eine Lösung der beiden Heffterschen Differenzenprobleme. Compositio Mathematica 6, 251–257. Im Internet unter http://archive.numdam.org/article/CM_1939__6__251_0.pdf einzusehen.

[102] Phelps, K. T. (1976) Some sufficient conditions for a Steiner triple system to be a derived triple system. Journal of Combinatorial Theory (A) 20, 393–397. im Internet unter https://ac.els-cdn.com/0097316576900388/1-s2.0-0097316576900388-main.pdf?_tid=5fb0aee7-651b-4083-88e4-dfc7ebb34c1c&acdnat=1533906401_f309ae0ac869860b395b2ef1de124c01 einzusehen.

[103] Pitman, J. (1999) Coalescent random forests. J. Combinatorial Theory, Ser. A 85, 165–193. Im Internet unter https://www.stat.berkeley.edu/~pitman/457.pdf einzusehen.

[104] Prüfer, H. (1918) Neuer Beweis über Permutationen. Archiv der Mathematik und Physik 27, 142–144.

[105] Rado, R. (1933) Studien zur Kombinatorik. Mathematische Zeitschrift 36, 424–480. Im Internet unter https://edoc.hu-berlin.de/bitstream/handle/18452/795/27037.pdf?sequence=1&isAllowed=y einzusehen.

[106] Ramsey, F. P. (1930) On a problem of formal logic. In: London Math. Soc. 30, 264–286. Im Internet unter http://www-lb.cs.umd.edu/~gasarch/TOPICS/ramsey/ramseyorig.pdf einzusehen.

[107] Ray-Chaudhuri, D. K. and R. M. Wilson (1971) Solution of Kirkman's schoolgirl problem. Amer. Math. Soc. Symp. Pure Math. 19, 187–204.

[108] Rohe, K. (2015) Visualizing Marden's theorem with Scilab. Im Internet unter https://arxiv.org/ftp/arxiv/papers/1502/1502.01367.pdf einzusehen.

[109] Rouse Ball, W. W. (1905) *Mathematical Recreations and Essays. Fourth Edition.* MacMillan, London. Im Internet unter http://www.gutenberg.org/files/26839/26839-pdf.pdf?session_id=8bd08a02a183ea88b768cb0502075bf57197204f einzusehen.

[110] Sandifer, E. (2006) How Euler did it. Infinite many primes. Im Internet unter http://eulerarchive.maa.org/hedi/HEDI-2006-03.pdf einzusehen.

[111] Schrijver, A. (1998) Counting 1-factors in regular bipartite graphs. Journal of Combinatoriasl Theory, Series B 72, 122—135. Im Internet unter https://ac.els-cdn.com/S0095895697917986/1-s2.0-S0095895697917986-main.pdf?_tid=084c6b7b-a090-4790-8db6-cc84a744fea9&acdnat=1542823128_cd8cc56013a80d974e324832360cffdf einzusehen.

[112] Schrijver, A. and W. G. Valiant (1980) On lower bounds for permanents. Indag. Math. 42, 425—427. Im Internet unter https://ac.els-cdn.com/1385725880900438/1-s2.0-1385725880900438-main.pdf?_tid=67593182-0889-44ea-a9e6-fc26e4594177&acdnat=1542805753_4ac533161f705dbcd129486379fdd74b einzusehen.

[113] Schur, I. (1917) Über die Kongruenz $x^m + y^m \equiv z^m \pmod p$. Jahresbericht der DMV 25, 114–117.

[114] Siebeck, J. (1864) Über eine neue analytische Behandlung der Brennpunkte. J. Reine Angew. Math. 64, 175–182.

[115] Skolem, Th. (1958) Some remarks on the triple systems of Steiner. Math. Scand. 6, 187–280. Im Internet unter http://www.mscand.dk/article/view/10551/8572 einzusehen.

[116] Sodhi, A. (2007) Cyclical diversions from Kirkman's scoolgirl problem. Im internet unter https://cms.math.ca/crux/v33/n4/public_page211-213.pdf einzusehen.

[117] Soifer, A. (2009) *The Mathematical Coloring Book. Mathematics of Coloring and the Colorful Life of its Creators*. Springer, New York.

[118] Sutherland, A. (2015) Number Theory I. Lecture # 15. The Riemann zeta function and prime number theorem. Im Internet unter http://math.mit.edu/classes/18.785/2015fa/LectureNotes15.pdf einzusehen.

[119] Stein, E. M. and R. Shakarchi (2003) *Complex Analysis.* Princeton University Press, Princeton. Im Internet unter https://www.fing.edu.uy/~cerminar/Complex_Analysis.pdf einzusehen. bibitemjsSteiner, J. (1853) Combinatorische Aufgabe. J. Reine Angew. Math. 45, 181–182. Im Internet unter https://gdz.sub.uni-goettingen.de/id/PPN243919689_0045?tify={%22pages%22:[189],%22panX%22:0.531,%22panY%22:0.762,%22view%22:%22info%22,%22zoom%22:0.554} einzusehe.

[120] Stinson, D. R. (2004) *Combinatorial designs. Constructions and Analysis*. Springer, New York-Berlin-Heidelberg-Hong Kong-London-Milan-Paris-Tokyo.

[121] Szekeres, G. and L. Peters (2006) Computer solution to the 17-point Erdös-Szekeres problem. Anziam J. 48, 151–164. Im Internet unter http://www.austms.org.au/Publ/Jamsb/V48P2/pdf/2409.pdf einzusehen.

[122] Takács, L. (1990) On Cayley's formula for counting forests. Journal of Combinatorial Theory Series A53, 321–323. Im Internet unter https://core.ac.uk/download/pdf/82105567.pdf einzusehen.

[123] Teschl, G. (2012) *Ordinary Differential Equations and Dynamical Systems*. American Mathematical Society. Im Internet unter https://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf einzusehen.

[124] Tverberg, H. (1963) On the permanent of a bistochsatic matrix. Math. Scand. 12, 25–35.

[125] van der Waerden, B. L. (1926) Aufgabe 45. Jahresberichte der DMV 35, 117.

[126] van der Waerden, B. L. (1965) Wie der Beweis der Vermutung von Baudet gefunden wurde. Abh. Math. Sem. Univ. Hamb. 28, 6–15. Ein reprint ist bei Elem. Math. 53 (1998) 139 – 148 wiedergegeben. Dieser kann über http://www.ems-ph.org/journals/show_abstract.php?issn=0013-6018&vol=53&iss=4&rank=2 eingesehen werden.

[127] van Lint, J. H. (1981) Note on Egoritsjev's proof of the van der Waerden conjecture. Memorandum 1981-01. Dept. of Math., Eindhoven University of Technology, Eindhoven. Im Internet unter https://pure.tue.nl/ws/files/4254844/696880.pdf einzusehen.

[128] van Lint, J. H. (1982) The van der Waerden conjecture: Two proofs in one year. Math. Intelligencer 4, 72—77.

[129] van Lint, J. H. and R. M. Wilson (1992) *A Course in Combinatorics*. Cambridge University Press, Cambridge.

[130] Verhulst, F. (1990) *Nonlinear Differential Equations and Dynamical Systems*. Springer-Verlag. Berlin-Heidelberg-New York.

[131] Voorhoeve (1979) A lower bound for the permanents of certain $(0,1)$-matrices. Indag. Math. 41, 83—86. Im Internet unter https://ac.els-cdn.com/138572587990012X/1-s2.0-138572587990012X-main.pdf?_tid=e6ae87f7-562b-40fd-a8ba-2b789fca5df8&acdnat=1542819091_5a748619c8da9327ee0a9eb0e2da48f0 einzusehe.

[132] Wallis, W. D. (1988) *Combinatorial Designs*. Marcel Dekker, Inc., New York-Basel.

[133] Weisstein, E. W. (2018) Projective Plane. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/ProjectivePlane.html.

[134] Werner, J. (1984) *Optimization. Theory and Applications*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.

[135] Werner, J. (1992a) *Numerische Mathematik 1. Lineare und nichtlineare Gleichungssysteme, Interpolation, numerische Integration*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.

[136] Werner, J. (1992b) *Numerische Mathematik 2. Eigenwertaufgaben, lineare Optimierungsaufgaben, unrestringierte Optimierungsaufgaben*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.

[137] Werner, J. (2000) *Operations Research*. Im Internet unter `https://num.math.uni-goettingen.de/werner/opres.pdf` einzusehen.

[138] Werner, J. (2002) *Unrestringierte Optimierungsaufgaben*. Im Internet unter `http://num.math.uni-goettingen.de/werner/uncopt.pdf` einzusehen.

[139] Werner, J. (2013) *Merkwürdige Mathematik*. Im Internet unter `http://num.math.uni-goettingen.de/werner/schmankerl.pdf` einzusehen.

[140] Werner, J. (2017) *Mehr merkwürdige Mathematik*. Im Internet unter `http://num.math.uni-goettingen.de/werner/schmankerl2.pdf` einzusehen.

[141] Zagier, D. (1997) Newman's short proof of the prime number theorem. Amer. Math. Monthly 104, 705—708. Im Internet unter `https://people.mpim-bonn.mpg.de/zagier/files/doi/10.2307/2975232/fulltext.pdf` einzusehen.

[142] Zeilmann, A. (2013) Beweis des Primzahlsatzes nach Newman. Im Internet unter `https://www-m3.ma.tum.de/foswiki/pub/M3/Allgemeines/FunktTheo13/Zeilmann-Primzahlsatz.pdf` einzusehen.

[143] Zygmund, A. (1935) Trigonometrical Series. Warschau. Im Internet unter `http://matwbn.icm.edu.pl/ksiazki/mon/mon05/mon0504.pdf` einzusehen.