Gregor Kemper*

Lineare Algebra und diskrete Strukturen 1 und 2

Vorlesungsmanuskript
Technische Universität München

2. Oktober 2018

*Verbesserungsvorschläge und Fehlermeldungen bitte an: kemper@ma.tum.de.

Inhaltsverzeichnis

Grundbegriffe

Wenn man heutzutage den Aufbau der Mathematik erklären will, kommt man um folgende zwei Elemente nicht herum: Logik und Mengenlehre. In dieser Vorlesung werden wir einen naiven, intuitiven Umgang mit der Logik pflegen und logische Strukturen und Sprechweisen im wesentlichen *en passant* kennenlernen. Die Mengenlehre werden wir ausführlicher behandeln und ihr den ersten Abschnitt der Vorlesung widmen.

Um starten zu können, erinnern wir ganz kurz an einige Sprachelemente der Logik, deren inhaltliche Bedeutung wir, wie oben angedeutet, dem „gesunden Menschenverstand“ überlassen wollen.

Sprachelemente der Logik:

- „und“ (bisweilen geschrieben als $\wedge$),
- „oder“ (bisweilen geschrieben als $\vee$),
- „nicht“ (bisweilen geschrieben als $\neg$), sowie die **Quantoren**
- „für alle“ (geschrieben als $\forall$, genannt der **Allquantor**) und
- „es gibt“ (geschrieben als $\exists$, genannt der **Existenzquantor**).

Aus diesen Sprachelementen setzt man neue zusammen:

- $A \Rightarrow B$ bedeutet: nicht A oder B.
- $A \Leftrightarrow B$ bedeutet $A \Rightarrow B$ und $B \Rightarrow A$.

Ein typisches Beispiel für die Verwendung von logischen Sprachelementen ist die bekannte Epsilon-Delta-Definition der Stetigkeit: Es seien $f\colon \mathbb{R} \to \mathbb{R}$ eine Funktion und $x_0 \in \mathbb{R}$. Dann heißt f stetig in x_0, falls

$$\forall\, \varepsilon > 0\ \exists\, \delta > 0\colon \forall x \in \mathbb{R}\colon \big[|x - x_0| < \delta \ \Rightarrow\ |f(x) - f(x_0)| < \varepsilon\big]\,.$$

1 Mengen

Alle Mathematik Lernenden haben schon mit zahlreichen Mengen zu tun gehabt: $\mathbb{R}$, $\mathbb{N}$, die Menge aller Geraden in einer Ebene, die Menge aller stetigen Funktionen $\mathbb{R} \to \mathbb{R}$, die Menge aller Paare (p, q) von Primzahlen p und q mit $q - p = 2$, und so weiter. Georg Cantor, den man als Begründer der Mengenlehre bezeichnet, formulierte 1895 folgende Definition:

> „Eine Menge ist eine Zusammenfassung von bestimmten, wohlunterschiedenen Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen."

Aus heutiger Sicht mag man diese Definition kritisieren, weil sie nicht exakt ist und weil die vorkommenden Begriffe ihrerseits einer Definition bedürfen. Schwerer wiegt jedoch die *Russelsche Antinomie*, die 1903 entdeckt wurde:

Gemäß dem Cantorschen Mengenbegriff müsste es auch die Menge aller Mengen geben, die hier mit X bezeichnet werden soll. Insbesondere gilt $X \in X$. Weiter können wir auch

$$R := \{A \in X \mid A \notin A\},$$

also die Menge aller Mengen, die sich nicht selbst als Element enthalten, bilden. (Das Symbol „:=" bedeutet hierbei: „wird definiert als".) Es gilt $R \in R$ oder $R \notin R$. Falls $R \in R$, wäre die Bedingung $A \notin A$ für $A = R$ nicht erfüllt, also definitionsgemäß $R \notin R$. Falls $R \notin R$, wäre $A \notin A$ für $A = R$ erfüllt, also definitionsgemäß $R \in R$. Wir erhalten also in beiden Fällen einen Widerspruch.

Die Entdeckung dieses Widerspruchs hat das Ende der naiven, Cantorschen Mengenlehre hervorgerufen. Aber nicht das Ende der Mathematik. Es gab mehrere Schulen, die neue Begründungen der Mengenlehre entwickelten. Hiervon hat sich die *Zermelo-Fraenkel-Mengenlehre* durchgesetzt, die wir hier in Grundzügen besprechen wollen. In der Zermelo-Fraenkel-Mengenlehre wird kein Versuch unternommen, den Mengenbegriff oder die Elementseinsbeziehung inhaltlich zu definieren. Es werden lediglich Regeln („Axiome") postuliert. Ein weiteres Merkmal ist, dass sämtliche mathematische Objekte Mengen sind. (Eine Variante lässt auch sogenannte *Urelemente* zu.) Die Zutaten der Zermelo-Fraenkel-Mengenlehre sind:

- Logik,
- das Symbol „$\in$", gelesen als „ist Element von",
- Axiome,
- vereinbarte Schreibweisen, Abkürzungen und Sprechweisen.

Die folgenden Axiome werden in der Zermelo-Fraenkel-Mengenlehre postuliert:

- Extensionalitätsaxiom (Seite 7),
- Aussonderungsaxiom (Seite 8),
- Vereinigungsmengenaxiom (Seite 9),

- Zweiermengenaxiom (Seite 10),
- Potenzmengenaxiom (Seite 10),
- Unendlichkeitsaxiom (Seite 11),
- Fundiertheitsaxiom (wird hier nicht behandelt),
- Ersetzungsaxiom (wird hier nicht behandelt),
- Auswahlaxiom (Seite 13).

In einigen Darstellungen der Zermelo-Fraenkel-Mengenlehre wird das *Leermengenaxiom* hinzugenommen oder das Auswahlaxiom als Erweiterung angesehen. Wir beginnen mit einer Schreib- und Sprechweise, die den Gleichheitsbegriffs definiert.

Definition 1.1. *Zwei Mengen A, B heißen* **gleich**, *falls sie sich bezüglich „$\in$“ identisch verhalten. Formaler: Wir schreiben $A = B$, falls gilt:*

$$\forall\, X\colon \big[X \in A \;\Leftrightarrow\; X \in B\big] \;\text{ und }\; \big[A \in X \;\Leftrightarrow\; B \in X\big]\,.$$

Aus Definition 1.1 folgt sofort:

(a) $\forall\, A\colon A = A$. („Reflexivität“),
(b) $\forall\, A, B\colon \big[A = B \;\Leftrightarrow\; B = A\big]$ („Symmetrie“),
(c) $\forall\, A, B, C\colon \big[A = B \text{ und } B = C \;\Rightarrow\; A = C\big]$ („Transitivität“).

Nun können wir das erste Axiom der Zermelo-Fraenkel-Mengenlehre formulieren.

Axiom 1.2 (Extensionalitätsaxiom)**.** *Falls zwei Mengen dieselben Elemente haben, sind sie gleich. Formaler: Für alle A, B gilt:*

$$\forall\, x\colon \big[x \in A \;\Leftrightarrow\; x \in B\big] \Rightarrow A = B.$$

Mit einem intuitiven, inhaltlichen Verständnis der Mengenlehre erscheint die Gültigkeit von Axiom 1.2 selbstverständlich. Dass es nicht inhaltsleer ist, zeigen Beispiele, in denen die Elementseinsbeziehung mit einem neuen Inhalt gefüllt ist.

Beispiel 1.3. (1) Für zwei Menschen x, y schreiben wir $x \in y$, falls x ein Kind von y ist. Es gilt also $x = y$ genau dann, wenn x und y identisch oder Geschwister sind und dieselben Kinder haben. Axiom 1.2 würde dann besagen, dass zwei Menschen, die dieselben Kinder haben, Geschwister sind—ein Unfug. Axiom 1.2 gilt in diesem Beispiel also nicht.
(2) Für zwei Menschen x, y schreiben wir $x \in y$, falls das Geburtsjahr von x nach dem von y liegt. Es gilt also $x = y$ genau dann, wenn x und y dasselbe Geburtsjahr haben. In diesem Beispiel gilt Axiom 1.2.
(3) Für zwei natürliche Zahlen x, y schreiben wir $x \in y$, falls $x < y$ gilt. Dies ergibt den gewöhnlichen Gleichheitsbegriff. Auch in diesem Beispiel gilt Axiom 1.2.
(4) Für zwei natürliche Zahlen x, y schreiben wir $x \in y$, falls $x + 1 = y$. Dies liefert den gewöhnlichen Gleichheitsbegriff. Es gilt Axiom 1.2. ◁

Wir verwenden die folgenden Schreib- und Sprechweisen:

- $x \notin A \ :\Longleftrightarrow$ nicht $x \in A$,
- $x \neq y \ :\Longleftrightarrow$ nicht $x = y$,
- $A \subseteq B$ („Teilmenge“) $:\Longleftrightarrow \ \forall\, x\colon \big[x \in A \ \Rightarrow \ x \in B\big]$,
- $A \subsetneqq B$ („echte Teilmenge“) $:\Longleftrightarrow \ A \subseteq B$ und $\exists\, x\colon \big[x \in B$ und $x \notin A\big]$.

(Hierbei deutet der Doppelpunkt vor dem Äquivalenzzeichen an, dass eine Sprechweise oder Eigenschaft definiert wird.) Aus Axiom 1.2 erhalten wir: Falls $A \subseteq B$ und $B \subseteq A$ gelten, dann $A = B$.

Um in gewohnter Weise Mengenlehre betreiben zu können, müssen wir Mengen bilden können wie

$$\{x \in \mathbb{N} \mid \exists\, y \in \mathbb{N}\colon x = y^2\}$$

oder

$$\Big\{x \in \mathbb{N} \mid x \neq 1 \text{ und } \forall\, y, z \in \mathbb{N}\colon \big[x = y \cdot z \ \Rightarrow (y = 1 \text{ oder } z = 1)\big]\Big\}\,.$$

Das folgende Axiom erlaubt es, Mengen zu konstruieren, indem wir aus einer gegebenen Menge alle Elemente, die eine gewisse Bedingung erfüllen, aussondern. Was heißt hierbei „Bedingung“? Die Antwort fällt in den Bereich der Logik. Etwas vergröbert kann man sagen, dass eine Bedingung ein Ausdruck $\mathcal{C}(x)$ ist, der aus dem Symbol „$\in$“, logischen Operatoren, mathematischen Objekten und „Variablen“ gebildet ist, und in dem x als „freie Variable“ vorkommt, während alle anderen Variablen durch Quantoren ($\forall$ und $\exists$) gebunden sind. In der Sprache der Prädikatenlogik würde man sagen: $\mathcal{C}(x)$ ist ein einstelliges Prädikat erster Stufe.

Axiom 1.4 (Aussonderungsaxiom)**.** *Für jede Bedingung $\mathcal{C}(x)$ und jede Menge A existiert eine Menge B mit:*

$$\forall\, x\colon \big[x \in B \ \Leftrightarrow \ x \in A \text{ und } \mathcal{C}(x) \text{ gilt}\big]\,.$$

Wegen Axiom 1.2 ist die Menge B aus Axiom 1.4 eindeutig bestimmt. Wir schreiben

$$B = \{x \in A \mid \mathcal{C}(x)\}\,.$$

Beispiel 1.5. Wir kommen auf die Beispiele in 1.3 zurück.

(1) Für dieses Beispiel gilt Axiom 1.4 nicht. Man betrachte die Bedingung $\mathcal{C}(x)$: $\forall\, y\colon y \notin x$, die besagt, dass x kinderlos ist. Axiom 1.4 würde nun bedeuten, dass es zu jedem Menschen A einen Menschen B gibt, dessen Kinder genau die kinderlosen Kinder zu A sind. Das ist Unfug!

(2) Auch hier gilt Axiom 1.4 nicht. Wir betrachten $\mathcal{C}(x)$: $x \notin$ Lorenz, wobei Lorenz 2010 geboren wurde. $\mathcal{C}(x)$ bedeutet, dass x im Jahr 2010 oder früher geboren wurde. Martin wurde 2008 geboren. Nach Axiom 1.4 müsste es einen Menschen B geben, so dass die Menschen,

deren Geburtsjahr nach dem von B liegt, genau diejenigen sind mit $2008 <$ Geburtsjahr ≤ 2010. Das ist Unfug.

(3) Auch hier gilt Axiom 1.4 nicht. Man betrachte $A = 5$ und die Bedingung $\mathcal{C}(x)$: $x = 4$. Axiom 1.4 würde bedeuten, dass es eine natürliche Zahl B gibt, so dass für alle natürlichen Zahlen x gilt: $x < B \Leftrightarrow x = 4$. Auch das ist Unfug!

(4) In diesem Beispiel hat jede positive natürliche Zahl A nur das einzige Element $A - 1$, und die 0 hat gar kein Element. Ist $\mathcal{C}(x)$ eine Bedingung und A eine natürliche Zahl, so können wir $B = A$ setzen, falls $\mathcal{C}(A - 1)$ gilt, und andernfalls $B = 0$. Dann wird Axiom 1.4 durch B erfüllt, es gilt also. ◁

Falls überhaupt eine Menge A existiert (dies folgt aus Axiom 1.12 auf Seite 11), dann gibt es nach Axiom 1.4 auch

$$\emptyset := \{x \in A \mid x \neq x\},$$

die **leere Menge**, die nach Axiom 1.2 eindeutig bestimmt ist, unabhängig von der Wahl von A. Weiter existiert zu Mengen A, B auch die **Schnittmenge**

$$A \cap B := \{x \in A \mid x \in B\} = \{x \in B \mid x \in A\}.$$

Zwei Mengen A, B heißen **disjunkt**, falls $A \cap B = \emptyset$. Außerdem gibt es zu Mengen A, B die **Differenzmenge**

$$A \setminus B := \{x \in A \mid x \notin B\}.$$

Ist allgemeiner M eine nicht-leere Menge, so können wir $B \in M$ wählen und die **Schnittmenge**

$$\bigcap M := \{x \in B \mid \forall\ A \in M\colon x \in A\}$$

bilden, die wegen Axiom 1.2 unabhängig von der Wahl von B ist. Eine alternative Schreibweise für $\bigcap M$ ist $\bigcap_{A \in M} A$.

Unsere bisherigen Axiome garantieren also die Existenz von Schnittmengen. Können wir auch die Existenz von Vereinigungsmengen folgern? Das Beispiel 1.3(4) zeigt, dass die Antwort nein ist. Jede Menge in diesem Beispiel hat höchstens ein Element, also kann man hier keine Vereinigungsmengen bilden, obwohl die Axiome 1.2 und 1.4 gelten. Wir benötigen also ein weiteres Axiom. Da wir nicht nur die Vereinigung zweier Mengen bilden wollen, sondern die Vereinigung beliebig vieler, fassen wir das Axiom weiter.

Axiom 1.6 (Vereinigungsmengenaxiom)**.** *Zu jeder Menge M existiert eine Menge B, so dass gilt:*

$$\forall\ x\colon \left[x \in B \Leftrightarrow \exists\ A\colon A \in M \text{ und } x \in A\right].$$

Die Menge B aus Axiom 1.6 ist wieder eindeutig bestimmt und wird mit $\bigcup M$, alternativ $\bigcup_{A \in M} A$, bezeichnet.

Können wir mit den bisherigen Axiomen die Existenz der Vereinigung *zweier* Mengen A, B garantieren? Dazu bräuchten wir eine Menge M, deren Elemente genau A und B sind. Dies liefert das folgende Axiom.

Axiom 1.7 (Zweiermengenaxiom)**.** *Für alle x, y existiert eine Menge A, so dass gilt:*

$$\forall\, z\colon \left[z \in A \;\Leftrightarrow z = x \text{ oder } z = y\right].$$

Die durch Axiom 1.7 gegebene, eindeutig bestimmte Menge wird als $A = \{x, y\}$ geschrieben, bzw. $A = \{x\}$ im Falle $x = y$. Man beachte den Unterschied zwischen x und $\{x\}$. Beispielsweise ist $\{\emptyset\} \neq \emptyset$. Ebenso beachte man den Unterschied zwischen $A \cup B$ und $\{A, B\}$. Durch Anwendung der Axiome 1.6 und 1.7 kann man auch Dreiermengen $\{x, y, z\}$ bilden und so weiter.

Axiom 1.8 (Potenzmengenaxiom)**.** *Zu jeder Menge A existiert eine Menge B, deren Elemente genau die Teilmengen von A sind, es gilt also:*

$$\forall\, x\colon \left[x \in B \;\Leftrightarrow x \subseteq A\right].$$

Die durch Axiom 1.8 gegebene Menge heißt die **Potenzmenge** von A und wird als $\mathfrak{P}(A)$ geschrieben.

Beispiel 1.9. $\mathfrak{P}(\emptyset) = \{\emptyset\}, \mathfrak{P}\left(\{\emptyset\}\right) = \left\{\emptyset, \{\emptyset\}\right\}$, und für $x \neq y$ gilt $\mathfrak{P}\left(\{x, y\}\right) = \left\{\emptyset, \{x\}, \{y\}, \{x, y\}\right\}$. ◁

Wir haben darauf verzichtet, die Gültigkeit der Axiome 1.6 bis 1.8 in unseren bisherigen Beispielen zu überprüfen. Es folgt nun ein interessantes Beispiel, in dem sie alle erfüllt sind.

Beispiel 1.10. Da dies ein Beispiel ist und nicht Teil des Aufbaus der Mathematik, ist es legitim, unser Wissen über natürliche Zahlen zu verwenden. Wir treffen wieder die Konvention, dass die natürlichen Zahlen mit 0 beginnen. Jede natürliche Zahl n hat eine Binärdarstellung $n = \sum_{i=0}^{m_n} a_i 2^i$ mit $a_i = 0$ oder $a_i = 1$ für alle i. Ist k eine weitere natürliche Zahl, so schreiben wir $k \in n$, falls $k \leq m_n$ und $a_k = 1$. (Man könnte auch sagen, dass $k \in n$ gilt, falls die größte natürliche Zahl, die $\leq \frac{n}{2^k}$ ist, ungerade ist.) Es gilt also beispielsweise $2 \in 5$, aber nicht $1 \in 5$.

Es ergibt sich der gewöhnliche Gleichheitsbegriff. Axiom 1.2 besagt, dass zwei natürliche Zahlen mit derselben Binärdarstellung gleich sind, das Axiom gilt also. Wir beobachten, dass jede natürliche Zahl endlich viele Elemente enthält. Sind umgekehrt $k_1, \ldots, k_s$ endlich viele paarweise verschiedene natürliche Zahlen, so enthält $n := \sum_{i=1}^{s} 2^{k_i}$ genau die Elemente $k_1, \ldots, k_s$.

Aus dieser Beobachtung folgt die Gültigkeit der Axiome 1.4 und 1.6 bis 1.8. (In der Tat gelten in diesem Beispiel alle Axiome der Zermelo-Fraenkel-Mengenlehre bis auf das Unendlichkeitsaxiom 1.12. Das Beispiel liefert ein Modell für die Mengenlehre endlicher Mengen.)

Wir betrachten ein paar Beispiele zu den Axiomen. Zu 2 und 5 existiert nach Axiom 1.7 die Menge $\{2,5\}$, nämlich $\{2,5\} = 2^2 + 2^5 = 36$. Die Einermenge $\{4\}$ ist $\{4\} = 16$. Was ist die Potenzmenge von 5? Es gilt $5 = \{0,2\}$, also

$$\mathfrak{P}(5) = \big\{\emptyset, \{0\}, \{2\}, \{0,2\}\big\} = \{0,1,4,5\} = 2^0 + 2^1 + 2^4 + 2^5 = 51.$$

Es sei dem Leser überlassen, die Vereinigungsmenge $\bigcup M$ von $M = 4\,294\,968\,320 = 2^{32} + 2^{10}$ zu bilden. Was ist $\Big\{ \big\{\emptyset, \{\{\emptyset\}\}\big\}, \big\{\emptyset, \{\emptyset\}\big\}, \{\emptyset\}\Big\}$? ◁

Der nächste Schritt ist die Konstruktion der natürlichen Zahlen. Damit stellen wir uns in den Gegensatz zu dem Mathematiker L. Kronecker (1823–1881), der gesagt haben soll: „Die natürlichen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk.“ Wir setzen

$$\begin{aligned}
0 &:= \emptyset,\\
1 &:= \{0\} (= \{\emptyset\}),\\
2 &:= \{0,1\} = 1 \cup \{1\} \big(= \big\{\emptyset, \{\emptyset\}\big\}\big),\\
3 &:= \{0,1,2\} = 2 \cup \{2\} \Big(= \Big\{\emptyset, \{\emptyset\}, \big\{\emptyset, \{\emptyset\}\big\}\Big\}\Big),\\
&\vdots
\end{aligned}$$

Um hieraus eine mathematische Definition zu machen und die Menge der natürlichen Zahlen konstruieren zu können, machen wir folgende Definition:

Definition 1.11. *(a) Für eine Menge A ist*

$$A^+ := A \cup \{A\}$$

der **Nachfolger** *von A.*

(b) Eine Menge M heißt **induktiv**, *falls gelten:*

(1) $\emptyset \in M$ und
(2) $\forall\, A \in M\colon A^+ \in M$.

Es folgt das nächste Axiom.

Axiom 1.12 (Unendlichkeitsaxiom)**.** *Es gibt eine induktive Menge.*

Nun können wir die Menge $\mathbb{N}$ der natürlichen Zahlen konstruieren. Zunächst beobachten wir, dass die Schnittmenge einer Menge von induktiven Mengen wieder induktiv ist. Es sei nun M eine induktive Menge, deren Existenz von Axiom 1.12 geliefert wird. Wir setzen

$$\mathcal{I}_M := \big\{M' \in \mathfrak{P}(M) \mid M' \text{ ist induktiv}\big\}.$$

Wegen $M \in \mathcal{I}_M$ ist $\mathcal{I}_M$ nicht leer, und wir können

$$\mathbb{N}_M := \bigcap \mathcal{I}_M$$

setzen. Damit ist $\mathbb{N}_M$ induktiv, genauer ist $\mathbb{N}_M$ die kleinste induktive Teilmenge von M.

Proposition 1.13. *Sind M und N induktive Mengen, so gilt $\mathbb{N}_M = \mathbb{N}_N$.*

Beweis. Die Schnittmenge $\mathbb{N}_M \cap N$ ist induktiv, also $\mathbb{N}_M \cap N \in \mathcal{I}_N$. Nach Konstruktion folgt $\mathbb{N}_N \subseteq \mathbb{N}_M \cap N \subseteq \mathbb{N}_M$. Ebenso zeigt man $\mathbb{N}_M \subseteq \mathbb{N}_N$. □

Nachdem die Unabhängigkeit von der Wahl von M geklärt ist, können und werden wir statt $\mathbb{N}_M$ auch $\mathbb{N}$ schreiben. Um die Theorie der natürlichen Zahlen weiter zu treiben, kann man nun direkt aus der Konstruktion die sogenannten *Peano-Axiome* beweisen, mit deren Hilfe sich die natürlichen Zahlen vollständig charakterisieren lassen. Danach kann man durch rekursive Definitionen die Addition und Multiplikation und die Vergleichsrelation „$\leq$“ natürlicher Zahlen erklären. Nach dem Beweis der Peano-Axiome, spätestens nach der Definition der arithmetischen Operationen, kann man die hier gegebene Definition von $\mathbb{N}$ vergessen und arbeitet nur noch mit den Eigenschaften der natürlichen Zahlen und mit den üblichen Symbolen 0, 1, 2, und so weiter. Ebenso erübrigt sich die hier gemachte Konstruktion des Nachfolgers (Definition 1.11(a)), und man schreibt statt n^+ fortan das gebräuchlichere $n + 1$.

Ein wichtiges Beweismittel ist das Prinzip der **vollständigen Induktion**, auch kurz Induktion genannt. Dies funktioniert folgendermaßen: Es sei $\mathcal{A}(n)$ eine Aussage über n (genauer: ein Prädikat erster Stufe mit n als freie Variable). Falls es gelingt, zu beweisen dass

(a) $\mathcal{A}(0)$ gilt und
(b) für alle $n \in \mathbb{N}$ gilt: $\big[\mathcal{A}(n) \Rightarrow \mathcal{A}(n+1)\big]$,

so folgt, dass $\mathcal{A}(n)$ für alle $n \in \mathbb{N}$ gilt. Intuitiv mag die Gültigkeit des Prinzips der vollständigen Induktion einleuchten, es ist aber doch beweisbedürftig. Wir geben folgenden Beweis: Die Menge

$$S := \{n \in \mathbb{N} \mid \mathcal{A}(n) \text{ gilt}\}$$

ist wegen der Voraussetzungen (a) und (b) induktiv. Nach Konstruktion ist $\mathbb{N}$ aber die kleinste induktive Menge, und es folgt $S = \mathbb{N}$. Damit ist gezeigt, dass $\mathcal{A}(n)$ für alle $n \in \mathbb{N}$ gilt.

Nachdem $\mathbb{N}$ zusammen mit den arithmetischen Operationen und der Relation „$\leq$“ konstruiert ist, kann man hieraus Schritt für Schritt die weiteren Zahlenbereiche $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ und $\mathbb{C}$ konstruieren. Hierbei sind alle Konstruktionen und Beweise im Rahmen der Zermelo-Fraenkel-Mengenlehre machbar. Wir werden es bei dieser Andeutung belassen und den Aufbau des Zahlensystems hier nicht behandeln.

Das letzte Axiom der Zermelo-Fraenkel-Mengenlehre, das wir hier besprechen wollen, ist das Auswahlaxiom. Es ist sicherlich das „prominenteste“ unter den Axiomen. Bisweilen wird es als Erweiterung der Zermelo-Fraenkel-Mengenlehre betrachtet. Man kann einen substanziellen Teil der Mathematik ohne Verwendung des Auswahlaxions betreiben. Es gibt Mathematiker, die diejenigen Teile der Mathematik, bei denen das Auswahlaxiom benötigt wird, markieren und gewissermaßen mit einem mentalen Warnschild versehen. Es gibt sogar solche, die das Auswahlaxiom ablehnen.

Axiom 1.14 (Auswahlaxiom)**.** *Es sei M eine Menge, deren Elemente nicht leere, paarweise disjunkte Mengen sind (letzteres bedeutet, dass für $A, B \in M$ mit $A \neq B$ gilt: $A \cap B = \emptyset$). Dann gibt es eine Menge X, die jedes $A \in M$ in genau einem Element schneidet, d.h.*

$$\forall\, A \in M\ \exists\, a\colon A \cap X = \{a\}.$$

Die Bezeichnung „Auswahlaxiom“ rührt daher, dass die Menge X gewissermaßen aus jeder Menge A in M ein Element „auswählt“. Man hüte sich allerdings davor, bei jedem Auftreten des Wortes „(aus-)wählen“ in einem mathematischen Beweis eine versteckte Anwendung des Auswahlaxioms zu vermuten. Ein Beispiel für die Anwendung des Auswahlaxioms werden wir im Beweis von Satz 2.6(b) sehen. Das Auswahlaxiom ist von den übrigen Axiomen der Zermelo-Fraenkel-Mengenlehre in folgendem Sinne unabhängig: Unter der Annahme, dass die übrigen Axiomen der Zermelo-Fraenkel-Mengenlehre widerspruchsfrei sind, ist sowohl die Zermelo-Fraenkel-Mengenlehre *mit* dem Auswahlaxiom also auch die Zermelo-Fraenkel-Mengenlehre mit der *Negation* des Auswahlaxioms widerspruchsfrei. Es ist also prinzipiell unmöglich, das Auswahlaxiom aus den übrigen Axiomen zu beweisen oder zu widerlegen.

Für das Auswahlaxiom selbst gibt es zahlreiche alternative Formulierungen, deren Äquivalenz (unter Voraussetzung der übrigen Axiome der Zermelo-Fraenkel-Mengenlehre) jeweils leicht einzusehen sind. (Siehe z.B. Anmerkung 2.7.) Außerdem ist das Auswahlaxiom (unter Voraussetzung der übrigen Axiome der Zermelo-Fraenkel-Mengenlehre) äquivalent zum Zornschen Lemma (siehe Satz 3.12) und zum Wohlordnungssatz (siehe Satz 3.13).

Die zwei verbleibenden Axiome der Zermelo-Fraenkel-Mengenlehre, das Fundiertheitsaxiom und das Ersetzungsaxiom, werden hier nicht behandelt, weil sich der allergrößte Teil der Mathematik ohne Benutzung dieser beiden Axiome entwickeln lässt. Mathematiker, die sich nicht mit einigen speziellen Fragen, insbesondere in der Mengenlehre selbst, beschäftigen, werden niemals mit diesen beiden Axiomen konfrontiert werden, weder explizit noch implizit.

Wir schließen diesen Abschnitt ab mit der Konstruktion von geordneten Paaren und kartesischen Produkten. Ziel ist es, zu x, y ein neues Objekt (x, y) zu konstruieren, so dass für alle x, y, x', y' die Gleichheit $(x, y) = (x', y')$ impliziert, dass $x = x'$ und $y = y'$ gelten.

Definition 1.15. *(a) Zu x, y definieren wir die Schreibweise*

$$(x, y) := \big\{\{x\}, \{x, y\}\big\}.$$

Wir nennen (x, y) *ein* **geordnetes Paar**.

(b) Für Mengen A, B *ist*

$$A \times B := \{(x, y) \mid x \in A \text{ und } y \in B\}$$

das **kartesische Produkt** *von* A *und* B*. Dessen Existenz und Eindeutigkeit wird durch unsere Axiome garantiert, denn*

$$A \times B = \Big\{C \in \mathfrak{P}\left(\mathfrak{P}(A \cup B)\right) \Big| \ \exists\, x \in A,\ \exists\, y \in B\colon C = \big\{\{x\}, \{x, y\}\big\}\Big\}.$$

Proposition 1.16. *Für alle* x, y, x', y' *gilt:*

$$(x, y) = (x', y') \iff x = x' \text{ und } y = y'.$$

Beweis. Es ist klar, dass die Gleichheiten $x = x'$ und $y = y'$ auch $(x, y) = (x', y')$ implizieren. Umgekehrt sei $(x, y) = (x', y')$. Mit $C := (x, y) = \big\{\{x\}, \{x, y\}\big\}$ und $C' := (x', y') = \big\{\{x'\}, \{x', y'\}\big\}$ folgt

$$\{x\} = \bigcap C = \bigcap C' = \{x'\},$$

also $x = x'$. Weiter gilt

$$\Big(\bigcup C\Big) \setminus \Big(\bigcap C\Big) = \begin{cases} \{y\} & \text{falls } x \neq y \\ \emptyset & \text{falls } x = y \end{cases}$$

und entsprechendes für C', x' und y'. Wegen $C = C'$ folgt hieraus auch $y = y'$. □

Von nun an kann man die exakte (und recht willkürliche) Definition von geordneten Paaren vergessen. Es wird nur noch die Schreibweise (x, y) benutzt und die Eigenschaft aus Proposition 1.16.

Man kann nun auch *geordnete Tripel* (x, y, z) durch $(x, y, z) := ((x, y), z)$ definieren und so weiter, entsprechend das kartesische Produkt $A \times B \times C := (A \times B) \times C$ für A, B und C Mengen. Im nächsten Abschnitt lernen wir eine alternative Konstruktion hierfür kennen (siehe Beispiel 2.3(10)).

2 Abbildungen und Mächtigkeit

Der Begriff einer Abbildung (gleichbedeutend: Funktion) ist zentral in allen Teilgebieten der Mathematik. Die Mathematik hat lange um einen tragfähigen Funktionenbegriff gerungen, beispielsweise um die Fragen, ob eine Funk-

tion durch eine Abbildungsvorschrift gegeben sein muss und inwieweit diese eindeutig sein muss. Wir benutzen die moderne Definition.

Definition 2.1. *Es seien A, B Mengen. Eine Teilmenge $f \subseteq A \times B$ heißt eine* **Abbildung** *(= **Funktion**) von A in B, falls es für jedes $x \in A$ genau ein $y \in B$ gibt mit $(x, y) \in f$. (Mit „genau ein“ ist hierbei gemeint, dass über die Existenz von y hinaus für alle $y' \in B$ gilt: $(x, y') \in f \;\Rightarrow\; y' = y$.)*

Für dieses y schreiben wir $y = f(x)$ und nennen es das **Bild** *von x (unter f). A heißt der* **Definitionsbereich**, *B der* **Bildbereich** *von f.*

Um auszudrücken, dass f eine Abbildung von A in B ist, schreiben wir $f\colon A \to B$. Falls eine Abbildungsvorschrift bekannt ist und angegeben werden soll, schreibt man $f\colon A \to B,\ x \mapsto \ldots$, wobei die Pünktchen für die Abbildungsvorschrift, die das Bild von x definiert, stehen. Diese wird in der Regel aus bereits definierten Abbildungen und anderen mathematischen Objekten („Konstanten“), bisweilen mit Fallunterscheidungen, gebildet.

Bevor wir Beispiele betrachten, machen wir ein paar Anmerkungen und eine weitere Definition.

Anmerkung. (a) In der Literatur findet man bisweilen die Schreibweise $f(x)$ für eine Funktion. Wir folgen dem Standard, dass $f(x)$ immer für das Bild eines Elements x des Definitionsbereichs steht, und schreiben f für die Funktion selbst.

(b) Es gibt keine Funktionen mit „mehreren Argumenten“. Allerdings gibt es etwa Funktionen $f\colon A \times B \to C$, deren Bilder man zweckmäßigerweise als $f(x, y)$ statt $f\big((x, y)\big)$ schreibt.

(c) Zu jeder Abbildung müssen Definitions- und Bildbereich angegeben werden. Laut unserer Definition wird allerdings B nicht eindeutig bestimmt durch $f \subseteq A \times B$. Um dies zu erreichen, wäre es besser, eine Abbildung als ein geordnetes Tripel $f = (A, B, C)$ zu definieren, wobei $C \subseteq A \times B$ die Bedingung aus Definition 2 erfüllt. Auch wenn sie formal besser wäre, würden wir mit einer solchen Definition vom gängigen Standard abweichen.

(d) Aus Definition 2.1 und Proposition 1.16 folgt folgender Gleichheitsbegriff für zwei Abbildungen $f, g\colon A \to B$:

$$f = g \quad\Longleftrightarrow\quad \forall\, x \in A\colon f(x) = g(x).$$

◁

Es folgen weitere Begriffe und Schreibweisen, die mit Abbildungen zu tun haben.

Definition 2.2. *Es seien A, B Mengen und $f\colon A \to B$ eine Abbildung.*

(a) Für eine Teilmenge $A' \subseteq A$ schreiben wir

$$f(A') := \{f(x) \mid x \in A'\} = \{y \in B \mid \exists\, x \in A'\colon y = f(x)\} \subseteq B.$$

(b) Die Teilmenge

$$\mathrm{Bild}(f) := f(A) \subseteq B$$

heißt das **Bild** *von f.*

(c) Die Abbildung f heißt **surjektiv**, *falls $f(A) = B$. Man spricht dann auch von einer Abbildung von A* **auf** *B (statt* **in** *B).*

(d) Für eine Teilmenge $B' \subseteq B$ heißt

$$f^{-1}(B') := \{x \in A \mid f(x) \in B'\} \subseteq A$$

das **Urbild** *von B' (unter f).*

(e) Die Abbildung f heißt **injektiv**, *falls für alle $x, x' \in A$ gilt:*

$$f(x) = f(x') \quad \Rightarrow \quad x = x'.$$

Gleichbedeutend ist die Bedingung, dass für $x, x' \in A$ mit $x \neq x'$ auch $f(x) \neq f(x')$ gilt, oder auch, dass für alle $y \in \mathrm{Bild}(f)$ das Urbild $f^{-1}(\{y\})$ genau ein Element hat.

(f) Die Abbildung f heißt **bijektiv**, *falls f surjektiv und injektiv ist. Gleichbedeutend ist die Bedingung, dass für alle $y \in B$ das Urbild $f^{-1}(\{y\})$ genau ein Element hat. Falls f bijektiv ist, so existiert eine* **Umkehrabbildung**

$$f^{-1}\colon B \to A,\ y \mapsto x \quad \textit{mit } f(x) = y.$$

Formaler lässt sich f^{-1} definieren als

$$f^{-1} = \{(y, x) \in B \times A \mid (x, y) \in f\}.$$

Es ist klar, dass f^{-1} dann auch bijektiv ist. Statt Umkehrabbildung sagt man bisweilen auch **inverse Abbildung** *oder* **Inverse**. *Es besteht Verwechselungsgefahr bei den Schreibweisen für das Urbild einer Menge und für die Umkehrabbildung. Eine bessere Notation wäre hier nützlich, stünde aber außerhalb jeder Tradition.*

Beispiel 2.3. (1) Die Abbildung $f\colon \mathbb{R} \to \mathbb{R},\ x \mapsto x^2$ ist weder injektiv noch surjektiv.

(2) Mit $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \mid x \geq 0\}$ definiert

$$f := \left\{(x, y) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \mid y^2 = x\right\}$$

eine Abbildung $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Erst nach Einführung der Wurzel-Symbols können wir für f die Abbildungsvorschrift $x \mapsto \sqrt{x}$ angeben, die aber nichts anderes als eine Abkürzung für $f(x)$ ist. Die Abbildung f ist bijektiv mit $f^{-1}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0},\ x \mapsto x^2$. Im Gegensatz zur Abbildung im Beispiel (1) ist f^{-1} bijektiv, weil Definitions- und Bildbereich anders festgelegt sind.

(3) Es sei A eine Menge. Die **identische Abbildung** ist definiert durch

$$\mathrm{id}_A \colon A \to A,\ x \mapsto x.$$

Sie ist bijektiv und ihre eigene Umkehrabbildung.

(4) Es sei $A = \emptyset$ und B eine beliebige Menge. Gibt es eine Abbildung $A \to B$? Das kartesische Produkt ist $A \times B = \emptyset$, also ist $\emptyset$ die einzige Teilmenge von $A \times B$. Die leere Menge erfüllt die Bedingung aus Definition 2.1 an eine Abbildung, weil nichts gefordert wird, also ist sie eine Abbildung. Es gibt also genau eine Abbildung $\emptyset \to B$. Sie ist injektiv und das Bild ist $\emptyset$. Im Kontrast hierzu gibt es nur dann eine Abbildung $A \to \emptyset$, wenn $A = \emptyset$.

(5) Die Abbildung $f\colon \mathbb{N} \to \mathbb{N},\ x \mapsto 3x$ ist injektiv, aber nicht surjektiv.

(6) Die Abbildung $f\colon \mathbb{R} \to \mathbb{R}_{\geq 0},\ x \mapsto x^2$ ist surjektiv, aber nicht injektiv.

(7) Die Exponentialfunktion $\exp\colon \mathbb{R} \to \mathbb{R}_{>0}$ ist bijektiv. Die Umkehrabbildung ist (definitionsgemäß) der natürliche Logarithmus.

(8) Die Abbildung

$$f\colon \mathbb{N} \to \{0,1\},\ x \mapsto \begin{cases} 0 & \text{falls } x \text{ gerade ist} \\ 1 & \text{sonst} \end{cases}$$

ist surjektiv, aber nicht injektiv. Das Urbild $f^{-1}(\{1\})$ ist die Menge aller ungerader Zahlen.

(9) Die Addition und Multiplikation auf $\mathbb{N}$ (und auf den weiteren Zahlenbereichen) sind durch Abbildungen $a, m\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ definiert. Statt $a(i,j)$ bzw. $m(i,j)$ benutzt man die Schreibweisen $i + j$ bzw. $i \cdot j$.

(10) Ist A eine Menge und $n \in \mathbb{N}_{>0} := \{n \in \mathbb{N} \mid n > 0\}$, so können wir ein **$n$-Tupel** von Elementen in A definieren als eine Abbildung

$$\{1, \ldots, n\} \to A,\ i \mapsto a_i,$$

wobei $\{1, \ldots, n\} := \{i \in \mathbb{N} \mid 1 \leq i \leq n\}$. Ein n-Tupel schreiben wir als $(a_1, \ldots, a_n)$. Mit

$$A^n = \big\{(a_1, \ldots, a_n) \mid \forall\, i \in \{1, \ldots, n\}\colon a_i \in A\big\}$$

bezeichnen wir die Menge aller n-Tupel. ◁

Es folgt eine weitere Definition.

Definition 2.4. *Es seien A, B Mengen und $f\colon A \to B$ eine Abbildung.*

(a) Sei $A' \subseteq A$ eine Teilmenge. Die **Einschränkung** *von f auf A' ist*

$$f|_{A'} : A' \to B,\ x \mapsto f(x).$$

Ebensogut könnte man schreiben $f|_{A'} = \{(x,y) \in f \mid x \in A'\}$.

(b) Es sei $\widehat{A}$ eine Menge mit $A \subseteq \widehat{A}$. Eine Abbildung $\widehat{f}\colon \widehat{A} \to B$ heißt eine **Fortsetzung** *von f auf $\widehat{A}$, falls $\widehat{f}|_A = f$ gilt. Man beachte, dass eine Funktion im Normalfall mehrere Fortsetzungen hat, da die Bilder der Elemente von $\widehat{A} \setminus A$ willkürlich festgelegt werden können.*

(c) *Es seien C eine Menge und $g\colon B \to C$ eine weitere Funktion. Die* **Komposition** *(= **Hintereinanderausführung**) von f und g ist definiert als*

$$g \circ f\colon A \to C,\ x \mapsto g\,(f(x))\,.$$

Ebensogut könnte man schreiben

$$g \circ f = \{(x,z) \in A \times C \mid \exists\ y \in B\colon (x,y) \in f \text{ und } (y,z) \in g\}\,.$$

Die Schreibweise $g \circ f$ sorgt manchmal für Verwirrung, weil die zweitgenannte Funktion f als erste ausgeführt wird.

Anmerkung 2.5. (a) Sind $f\colon A \to B$, $g\colon B \to C$ und $h\colon C \to D$ Abbildungen, so gilt das *Assoziativitätsgesetz*

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

(b) Es seien $f, g\colon A \to A$ Abbildungen. Obwohl $f \circ g$ und $g \circ f$ definiert sind, ist das *Kommutativitätsgesetz*

$$f \circ g = g \circ f$$

im Allgemeinen falsch. Als Beispiel betrachten wir

$$f\colon \mathbb{N} \to \mathbb{N},\ x \mapsto 2x \quad \text{und} \quad g\colon \mathbb{N} \to \mathbb{N},\ x \mapsto x+1,$$

also gilt für $x \in \mathbb{N}$:

$$(f \circ g)(x) = 2x+2 \quad \text{und} \quad (g \circ f)(x) = 2x+1.$$

Die Ungleichheit von $f \circ g$ und $g \circ f$ sieht man z.B. durch Einsetzen von $x = 0$.

(c) Ist $f\colon A \to B$ bijektiv, so gelten

$$f \circ f^{-1} = \mathrm{id}_B \quad \text{und} \quad f^{-1} \circ f = \mathrm{id}_A\,.$$

(d) Die Einschränkung einer nicht injektiven Abbildung kann injektiv sein.

(e) Fortsetzungen von Abbildungen sind vor allem dann interessant, wenn man von der Fortsetzung gewisse Eigenschaften (z.B. Stetigkeit) fordert. Dadurch kann es je nach Situation passieren, dass gar keine solche Fortsetzung existiert, oder eine Fortsetzung eindeutig bestimmt ist. ◁

Der folgende Satz stellt interessante Zusammenhänge zwischen den Begriffen injektiv und surjektiv her. Für den Beweis benötigen wir das Auswahlaxiom.

Satz 2.6. *Es seien A, B nicht leere Mengen und $f\colon A \to B$ eine Abbildung.*

(a) Genau dann ist f injektiv, wenn es eine Abbildung $g\colon B \to A$ gibt mit

$$g \circ f = \mathrm{id}_A\,.$$

(Man nennt g dann auch eine Linksinverse *von f.)*

(b) *Genau dann ist f surjektiv, wenn es eine Abbildung $g\colon B \to A$ gibt mit*

$$f \circ g = \mathrm{id}_B\,.$$

(Man nennt g dann auch eine Rechtsinverse *von f.)*

Anmerkung. Wegen (b) ist das g aus (a) surjektiv, und wegen (a) ist das g aus (b) injektiv. ◁

Beweis von Satz 2.6. (a) Wir setzen zunächst voraus, dass f injektiv ist. Wir bilden $g\colon B \to A$, indem wir jedem $y \in \mathrm{Bild}(f)$ sein eindeutig bestimmtes Urbild zuordnen und die Elemente von $B \setminus \mathrm{Bild}(f)$ auf ein willkürlich gewähltes Element von A abbilden. Formal führen wir den Beweis folgendermaßen: Wegen $A \neq \emptyset$ existiert $a \in A$, also auch

$$g := \Big\{(y,x) \in B \times A \mid (x,y) \in f \text{ oder } \big[y \notin \mathrm{Bild}(f) \text{ und } x = a\big]\Big\}\,.$$

Zu $y \in \mathrm{Bild}(f)$ existiert wegen der Injektivität von f ein eindeutiges x mit $(y,x) \in g$, und zu $y \in B \setminus \mathrm{Bild}(f)$ ist $x = a$ das eindeutige x mit $(y,x) \in g$. Also ist g eine Abbildung. Für $x \in A$ gilt $(x, f(x)) \in f$, also $(f(x), x) \in g$ und damit $(x,x) \in g \circ f$. Damit ist $g \circ f = \mathrm{id}_A$ gezeigt.
Umgekehrt nehmen wir an, dass es $g\colon B \to A$ mit $g \circ f = \mathrm{id}_A$ gibt. Für $x, x' \in A$ mit $f(x) = f(x')$ folgt dann

$$x = \mathrm{id}_A(x) = g\,(f(x)) = g\,(f(x')) = \mathrm{id}_A(x') = x',$$

also ist f injektiv.

(b) Wir nehmen zunächst an, dass f surjektiv ist. Die Idee ist, mit Hilfe des Auswahlaxioms zu jedem $y \in B$ ein Element des Urbilds $f^{-1}(\{y\})$ auszuwählen und dieses als $g(y)$ zu definieren. Formal gehen wir folgendermaßen vor: Wir bilden

$$M := \big\{f^{-1}(\{y\}) \mid y \in B\big\} = \big\{X \in \mathfrak{P}(A) \mid \exists\, y \in B\colon X = f^{-1}(\{y\})\big\}$$

wobei der zweite Ausdruck nur dazu dient zu zeigen, dass die Existenz von M durch die Axiome 1.8 und 1.4 garantiert wird. Wegen der Surjektivität von f ist jede Menge in M nicht leer. Um zu zeigen, dass die Mengen aus M paarweise disjunkt sind, betrachten wir zwei Elemente $f^{-1}(\{y\})$ und $f^{-1}(\{y'\})$ aus M. Falls deren Schnittmenge ein Element x enthält, so folgt $y = f(x) = y'$, also $f^{-1}(\{y\}) = f^{-1}(\{y'\})$. Damit ist die paarweise Disjunktheit von M bewiesen, Axiom 1.14 liefert also eine Menge X mit

$$\forall\, y \in B\ \exists\, a \in X\colon f^{-1}(\{y\}) \cap X = \{a\}. \tag{2.1}$$

Nun definieren wir

$$g := \{(y,x) \in B \times A \mid (x,y) \in f \text{ und } x \in X\}.$$

Für $y \in B$ und $x \in A$ liegt (y,x) genau dann in B, wenn $x \in f^{-1}(\{y\}) \cap X$, also ist g wegen (2.1) eine Abbildung. Für $y \in B$ sei $x := g(y)$, also $(y,x) \in g$. Es folgt $(x,y) \in f$, also $(y,y) \in f \circ g$. Damit ist $f \circ g = \mathrm{id}_B$ gezeigt.
Umgekehrt setzen wir voraus, dass $g\colon B \to A$ mit $f \circ g = \mathrm{id}_B$ existiert. Für $y \in B$ gilt dann

$$y = \mathrm{id}_B(y) = f(g(y)) \in \mathrm{Bild}(f),$$

also ist f surjektiv. □

Anmerkung 2.7. Satz 2.6(b) besagt, dass jede surjektive Abbildung eine Rechtsinverse hat. Es ist nicht schwer zu zeigen, dass diese Aussage sogar äquivalent zum Auswahlaxiom 1.14 ist. ◁

Mit Hilfe der folgenden Definition lassen sich Mengen hinsichtlich ihrer „Größe“ vergleichen.

Definition 2.8. *Es seien A, B Mengen.*

(a) A und B heißen **gleichmächtig**, *falls es eine Bijektion (= bijektive Abbildung) $f\colon A \to B$ gibt. Wir drücken dies durch die Schreibweise $A \sim B$ aus.*
(b) A heißt **höchstens so mächtig** *wie B, falls es eine Injektion (= injektive Abbildung) $f\colon A \to B$ gibt, falls A also gleichmächtig mit einer Teilmenge von B ist. Wir drücken dies durch die Schreibweise $A \lesssim B$ aus. Wegen Satz 2.6 ist $A \lesssim B$ gleichbedeutend mit der Bedingung, dass es eine Surjektion $B \to A$ gibt oder A leer ist.*
(c) B heißt **mächtiger** *als A, falls $A \lesssim B$ und A und B nicht gleichmächtig sind. Wir schreiben dann $A \prec B$.*

Bevor wir Beispiele betrachten, bringen wir einen grundlegenden Satz über die Mächtigkeit von Mengen, auf dessen Beweis wir hier verzichten müssen.

Satz 2.9. *Es seien A, B Mengen.*

(a) Es gilt $A \lesssim B$ oder $B \lesssim A$.
(b) Falls $A \lesssim B$ und $B \lesssim A$ gelten, so folgt $A \sim B$.

Anmerkung 2.10. (a) Man kann Satz 2.9 auch folgendermaßen ausdrücken: Genau einer der drei folgenden Fälle tritt ein („Trichotomie“): $A \prec B$, $A \sim B$ oder $B \prec A$.
(b) Die Aussage (a) des Satzes wird auch als „Vergleichbarkeissatz“ bezeichnet. Wir werden den Beweis mit Hilfe des Zornschen Lemmas (Satz 3.12) am Ende des nächsten Abschnitts führen. Die Aussage (b) ist bekannt als der Satz von Schröder und Bernstein. Einen Beweis kann man finden in: Paul Halmos, *Naive Mengenlehre*, Vandenhoeck & Ruprecht, Göttingen 1994. Wir lassen den Beweis aus Zeitgründen weg.

(c) Die Umkehrung von Satz 2.9(b) folgt direkt aus Definition 2.8: Falls $A \sim B$, dann $A \lesssim B$ und $B \lesssim A$.

(d) Aus Satz 2.9 folgt, dass B genau dann mächtiger als A ist, wenn es *keine* Injektion $B \to A$ gibt, oder gleichbedeutend, wenn es *keine* Surjektion $A \to B$ gibt und B nicht leer ist.

(e) Da die Komposition zweier Injektionen wieder eine Injektion ist, folgt für Mengen A, B, C aus $A \lesssim B$ und $B \lesssim C$ die Beziehung $A \lesssim C$ („Transitivität“). Außerdem gilt $A \lesssim A$ („Reflexivität“). Ebenso ist die Gleichmächtigkeitsbeziehung transitiv und reflexiv, und außerdem symmetrisch (d.h. aus $A \sim B$ folgt $B \sim A$). ◁

Beispiel 2.11. (1) Die Potenzmenge $\mathfrak{P}(\{1,2\})$ und $\{1,2,3,4\}$ sind gleichmächtig. Eine Bijektion f zwischen den beiden ist gegeben durch $f(1) = \emptyset$, $f(2) = \{1\}$, $f(3) = \{2\}$, $f(4) = \{1,2\}$.

(2) $\{1,4,5\}$ ist mächtiger als $\{3,4\}$. $\mathbb{N}$ ist mächtiger als $\mathfrak{P}(\{1,\dots,10\})$.

(3) Die „Abzählung“ $0, 1, -1, 2, -2, 3, -3, \dots$ liefert eine Bijektion $f\colon \mathbb{N} \to \mathbb{Z}$, als Formel $f(a) = (-1)^{a+1} \cdot \lfloor \frac{a+1}{2} \rfloor$, wobei für $x \in \mathbb{R}$ die größte ganze Zahl $\leq x$ mit $\lfloor x \rfloor$ bezeichnet wird. Es folgt $\mathbb{N} \sim \mathbb{Z}$.

(4) Überraschender ist, dass auch $\mathbb{N}$ und das kartesische Produkt $\mathbb{N} \times \mathbb{N}$ gleichmächtig sind. Das Schema

	0	1	2	3	4	5	⋯
0	0	1	3	6	10	15	⋯
1	2	4	7	11	16	⋯	
2	5	8	12	17	⋯		
3	9	13	18	⋯			
4	14	19	⋯				
5	20	⋯					
⋮							

liefert eine „Abzählung“ von $\mathbb{N} \times \mathbb{N}$, die man formal durch die Abbildung

$$f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N},\ (a,b) \mapsto \frac{(a+b)(a+b+1)}{2} + a$$

beschreiben kann. Es ist etwas mühsam, die Bijektivität von f, die intuitiv aus obigem Schema hervorgeht, nachzuweisen. Wie behauptet ergibt sich $\mathbb{N} \times \mathbb{N} \sim \mathbb{N}$.

(5) Die Surjektion $f\colon \mathbb{Z} \times \mathbb{N} \to \mathbb{Q},\ (a,b) \mapsto \frac{a}{b+1}$ liefert $\mathbb{Q} \lesssim \mathbb{Z} \times \mathbb{N}$. Andererseits ist $\mathbb{N}$ als Teilmenge von $\mathbb{Q}$ höchstens so mächtig wie $\mathbb{Q}$. Mit den Beispielen (3) und (4) folgt $\mathbb{N} \lesssim \mathbb{Q} \lesssim \mathbb{Z} \times \mathbb{N} \sim \mathbb{N} \times \mathbb{N} \sim \mathbb{N}$, also

$$\mathbb{Q} \sim \mathbb{N}$$

wegen Satz 2.9(b).

(6) Die Abbildung

$$f\colon \mathfrak{P}(\mathbb{N}) \to \mathbb{R},\ A \mapsto \sum_{k \in A} 10^{-k}$$

ist injektiv, denn eine Menge A wird auf eine reelle Zahl abgebildet, in deren Dezimalbruchentwicklung nur Nuller und Einser vorkommen. Dies liefert $\mathfrak{P}(\mathbb{N}) \lesssim \mathbb{R}$. Nun definieren wir eine Abbildung

$$g\colon \mathbb{R} \to \mathfrak{P}(\mathbb{Q}),\ x \mapsto \{a \in \mathbb{Q} \mid a < x\}.$$

Diese ist injektiv, denn für verschiedene reelle Zahlen x, y mit $x < y$ gibt es bekanntlich eine rationale Zahl $a \in \mathbb{Q}$ mit $x \leq a < y$, also $a \in g(y)$ aber $a \notin g(x)$, wodurch $g(x) \neq g(y)$ gezeigt ist. Wegen dem obigen Beispiel (5) ergibt sich insgesamt

$$\mathfrak{P}(\mathbb{N}) \lesssim \mathbb{R} \lesssim \mathfrak{P}(\mathbb{Q}) \sim \mathfrak{P}(\mathbb{N}),$$

gemäß Satz 2.9(b) also

$$\mathbb{R} \sim \mathfrak{P}(\mathbb{N}). \tag{2.2}$$

Aus Satz 2.12, den wir gleich beweisen, folgt, dass $\mathbb{R}$ mächtiger ist als $\mathbb{N}$.

(7) Wir haben eine Bijektion

$$f\colon \mathfrak{P}(\mathbb{N}) \times \mathfrak{P}(\mathbb{N}) \to \mathfrak{P}(\mathbb{N}),\ (A, B) \mapsto \{2x \mid x \in A\} \cup \{2x+1 \mid x \in B\}.$$

Es folgt $\mathfrak{P}(\mathbb{N}) \times \mathfrak{P}(\mathbb{N}) \sim \mathfrak{P}(\mathbb{N})$, wegen (2.2) also auch

$$\mathbb{R} \times \mathbb{R} \sim \mathbb{R}.$$

Die reelle „Ebene“ und die „Zahlengerade“ sind also gleichmächtig! ◁

Der folgende auf Georg Cantor zurückgehende Satz zeigt, dass es unendlich viele „Stufen“ der Unendlichkeit gibt.

Satz 2.12. *Sei A eine Menge. Dann ist die Potenzmenge $\mathfrak{P}(A)$ mächtiger als A.*

Beweis. Wegen Anmerkung 2.10(d) ist zu zeigen, dass es keine Surjektion $f\colon A \to \mathfrak{P}(A)$ gibt. Es sei $f\colon A \to \mathfrak{P}(A)$ irgendeine Abbildung. Um zu zeigen, dass f nicht surjektiv ist, brauchen wir eine Teilmenge $B \subseteq A$ mit $B \notin \text{Bild}(f)$. Wir setzen

$$B := \{x \in A \mid x \notin f(x)\} \subseteq A.$$

Es sei $x \in A$ beliebig. Für den Nachweis von $B \neq f(x)$ betrachten wir die Fälle $x \in B$ und $x \notin B$. Falls $x \in B$, dann folgt $x \notin f(x)$ aus der Definition von B, also $B \neq f(x)$. Falls andererseits $x \notin B$ gilt, so folgt $x \in f(x)$, also auch in diesem Fall $B \neq f(x)$.

Wie behauptet liegt B nicht im Bild von f, also ist f nicht surjektiv. □

Mit (2.2) folgt, dass $\mathbb{R}$ mächtiger als $\mathbb{N}$ ist. Die berühmte *Kontinuumshypothese* besagt, dass es keine Menge A gibt, so dass A mächtiger als $\mathbb{N}$ und $\mathbb{R}$ mächtiger als A ist, dass also nichts „zwischen $\mathbb{N}$ und $\mathbb{R}$“ liegt. In den 1960er

Jahren wurde nach langem Ringen bewiesen, dass die Kontinuumshypothese aus den Axiomen der Zermelo-Fraenkel-Mengenlehre weder beweisbar noch widerlegbar ist, in dem selben Sinne, wie dies für das Auswahlaxiom aus den übrigen Axiomen gilt.

Wir beenden den Abschnitt mit einer Definition.

Definition 2.13. *Es sei A eine Menge.*

(a) A heißt **endlich***, falls es eine natürliche Zahl $n \in \mathbb{N}$ gibt, so dass A und $\{1, \dots, n\}$ gleichmächtig sind. (Insbesondere ist $\emptyset$ endlich mit $n = 0$.) Wir schreiben dann*

$$| A |= n$$

und nennen dies die **Elementzahl** *von A. Ist A nicht endlich (also* **unendlich***), so drücken wir dies symbolisch durch $| A |= \infty$ aus.*

(b) A heißt **abzählbar unendlich***, falls A und $\mathbb{N}$ gleichmächtig sind, und* **überabzählbar***, falls A mächtiger als $\mathbb{N}$ ist.*

Anmerkung 2.14. (a) Es ist beweisbedürftig, dass die Elementanzahl einer endlichen Menge A eindeutig bestimmt ist, d.h. dass zwei „Anfangsstücke“ $\{1, \dots, n\}$ und $\{1, \dots, m\}$ mit $n, m \in \mathbb{N}$ nur dann gleichmächtig sind, wenn $n = m$ gilt. Man kann den Beweis per Induktion führen, worauf wir hier verzichten.

(b) Man kann zeigen, dass jede der folgenden zwei Bedingungen äquivalent zur Endlichkeit einer Menge A sind:

- Jede Injektion $f\colon A \to A$ ist surjektiv.
- Jede Surjektion $f\colon A \to A$ ist injektiv.

(c) In Beispiel 2.11 haben wir schon zwei Beispiele von unendlichen Mengen A gesehen, für die $A \times A \sim A$ gilt. Man kann zeigen, dass dies für *jede* unendliche Menge gilt.

Beweise zu den Aussagen (a) und (c) finden sich im oben angegebenen Buch von Halmos. ◁

3 Relationen

Ebenso wie beim Mengenbegriff unternehmen wir auch beim Begriff einer Relation keinen Versuch einer inhaltlichen Definition.

Definition 3.1. *Sei A eine Menge. Eine* **Relation** *auf A ist eine Teilmenge $R \subseteq A \times A$. Falls R eine Relation ist und $x, y \in A$, schreiben wir häufig xRy statt $(x, y) \in R$ und sagen, dass x in der Relation R zu y steht.*

Anmerkung. Bisweilen werden Relationen auch allgemeiner als Teilmengen eines kartesischen Produkts $A \times \cdots \times A$ von k Exemplaren von A definiert (*k-stellige Relation*). Eine Relation wie in Definition 3.1 nennt man auch eine *binäre Relation*.

Noch allgemeiner kann man Relationen als Teilmengen eines kartesischen Produkts $A_1 \times A_2 \times \cdots \times A_k$ mit A_i Mengen definieren. ◁

Beispiel 3.2. (1) Durch $R := \{(x, y) \in A \times A \mid x = y\}$ wird die Gleichheitsrelation auf einer Menge A definiert.
$R' := \{(x, y) \in A \times A \mid x \neq y\}$ ist die „Ungleichheitsrelation“.
(2) Beispiele für Relationen auf $\mathbb{N}$ sind:

- die Relationen „$\leq$“, „$\geq$“, „$<$“, gegeben durch

$$R = \{(x, x + a) \mid x, a \in \mathbb{N}\}$$

und so weiter;
- die *Teilbarkeitsrelation*, gegeben durch

$$x \mid y \quad :\Longleftrightarrow \quad \exists\, a \in \mathbb{N}\colon y = ax$$

(gelesen als: „x teilt y“);
- die „Parität“, gegeben durch

$$x \equiv y \quad :\Longleftrightarrow \quad 2 \mid (x - y);$$

- die „Nachfolgerrelation“, gegeben durch

$$R = \{(x, x + 1) \mid x \in \mathbb{N}\}.$$

(3) Sind A, B Mengen und $f\colon A \to B$ eine Abbildung, so ist

$$R = \{(x, y) \in A \times A \mid f(x) = f(y)\}$$

eine Relation.
(4) Für eine Menge A sind $A \times A$ bzw. $\emptyset$ immer Relationen (alles steht in Relation bzw. nichts steht in Relation). ◁

Ist R eine Relation auf einer Menge A, so lässt sich R auf eine Teilmenge $B \subseteq A$ *einschränken*, indem man $(B \times B) \cap R$ bildet.
Ebenso wie Abbildungen können auch Relationen Eigenschaften haben.

Definition 3.3. *Es sei $R \subseteq A \times A$ eine Relation.*

(a) R heißt **reflexiv**, *falls für alle $x \in A$ gilt:*

$$(x, x) \in R, \ \textit{d.h. } xRx.$$

(b) R heißt **symmetrisch**, *falls für alle $x, y \in A$ gilt:*

$$xRy \quad \Rightarrow \quad yRx.$$

(c) R heißt **antisymmetrisch**, *falls für alle $x, y \in A$ gilt:*

$$xRy \quad \textit{und} \quad yRx \quad \Rightarrow \quad x = y.$$

(d) R heißt **transitiv**, *falls für alle $x, y, z \in A$ gilt:*

$$xRy \quad \textit{und} \quad yRz \quad \Rightarrow \quad xRz.$$

(e) R heißt eine **Äquivalenzrelation**, *falls R reflexiv, symmetrisch und transitiv ist.*

(f) R heißt eine **Ordnungsrelation**, *falls R reflexiv, antisymmetrisch und transitiv ist.*

Beispiel 3.4. Wir prüfen die Eigenschaften der in Beispiel 3.2(2) betrachteter Relationen auf $\mathbb{N}$.

	reflexiv	symm.	antisymm.	transitiv	Äquiv.-/Ordnungsrel.
$=$	ja	ja	ja	ja	beides
$\neq$	nein	ja	nein	nein	weder noch
$\leq$	ja	nein	ja	ja	Ordnungsrelation
$\geq$	ja	nein	ja	ja	Ordnungsrelation
$<$	nein	nein	ja	ja	weder noch
Teilbarkeit	ja	nein	ja	ja	Ordnungsrelation
Parität	ja	ja	nein	ja	Äquivalenzrelation
Nachfolger	nein	nein	ja	nein	weder noch
$\mathbb{N} \times \mathbb{N}$	ja	ja	nein	ja	Äquivalenzrelation
$\emptyset$	nein	ja	ja	ja	weder noch

◁

Wir beschäftigen uns nun zunächst mit Äquivalenzrelationen. Ist R eine Äquivalenzrelation auf einer Menge A, so schreiben wir der besseren Lesbarkeit halber $x \sim y$ statt xRy und sprechen auch von der Äquivalenzrelation „$\sim$“.

Definition 3.5. *Wir setzen obige Situation voraus.*

(a) Für $x \in A$ heißt

$$[x]_\sim := \{y \in A \mid x \sim y\}$$

die **Äquivalenzklasse** *von x. Also ist $[x]_\sim \subseteq A$ eine Teilmenge und $x \in [x]_\sim$.*

(b) Die Menge

$$A/\!\sim \;:= \{[x]_\sim \mid x \in A\} = \{C \subseteq A \mid \exists\, x \in A\colon C = [x]_\sim\} \subseteq \mathfrak{P}(A)$$

aller Äquivalenzklassen heißt die **Faktormenge** *(= **Quotientenmenge**) von A nach $\sim$.*

(c) Für $C \in A/\!\sim$ heißt jedes $x \in C$ ein **Vertreter** *(= **Repräsentant**) der Klasse C.*

(d) Die Abbildung

$$\pi\colon A \to A/\!\sim,\ x \mapsto [x]_\sim$$

heißt die **kanonische Projektion**.

Beispiel 3.6. (1) Die Gleichheit ist eine Äquivalenzrelation. Die Äquivalenzklassen sind alle einelementig, also $[x]_= = \{x\}$ und

$$A/= = \{\{x\} \mid x \in A\},$$

was nicht dasselbe wie A ist.

(2) Die Paritätsrelation lässt sich auch auf $\mathbb{Z}$ definieren durch

$$x \equiv y \quad :\Longleftrightarrow \quad 2 \mid (x - y).$$

Es gibt zwei Klassen: $[0]_\equiv$, die Klasse aller geraden Zahlen, und $[1]_\equiv$, die Klasse aller ungeraden Zahlen. $\mathbb{Z}/\equiv$ hat zwei Elemente.

(3) Allgemeiner sei $m \in \mathbb{N}_{>0}$ fest gewählt. Für $x, y \in \mathbb{Z}$ schreiben wir

$$x \equiv y \mod m \quad :\Longleftrightarrow \quad m \mid (x - y)$$

und sagen dann, dass x *kongruent* zu y *modulo* m ist. Es ist leicht zu sehen, dass die Kongruenz modulo m eine Äquivalenzrelation ist. Die Äquivalenzklasse von $x \in \mathbb{Z}$ lässt sich schreiben als

$$[x]_\sim = \{x + km \mid k \in \mathbb{Z}\}$$

und wird auch als die *Restklasse* von x modulo m bezeichnet. Die Faktormenge wird geschrieben als $\mathbb{Z}/(m)$. Sie hat genau die m Elemente

$$\mathbb{Z}/(m) = \{[0]_\sim, [1]_\sim, \ldots, [m-1]_\sim\},$$

wobei man statt $[0]_\sim$ ebenso gut $[m]_\sim$ schreiben könnte und so weiter.

(4) Es sei $A = \{0,1\} \times \{0,1\} \times \{0,1\}$ das dreifache kartesische Produkt der Menge $\{0,1\}$. Zwei Tripel (a,b,c) und (a',b',c') aus A seien äquivalent, wenn sie bis auf die Reihenfolge übereinstimmen. Es gibt vier Äquivalenzklassen:

$$\begin{aligned}
&[(0,0,0)]_\sim = \{(0,0,0)\},\\
&[(1,1,1)]_\sim = \{(1,1,1)\},\\
&[(0,0,1)]_\sim = \{(0,0,1),(0,1,0),(1,0,0)\} \text{ und}\\
&[(1,1,0)]_\sim = \{(1,1,0),(1,0,1),(0,1,1)\}.
\end{aligned}$$

(5) Die Relation aus Beispiel 3.2(3) ist eine Äquivalenzrelation. Die Äquivalenzklassen sind die Urbilder $f^{-1}(\{y\})$ der einelementigen Teilmengen der Bildmenge $f(A)$.

(6) Für jede Menge A ist $A \times A$ eine Äquivalenzrelation. Für alle $x, y \in A$ gilt $x \sim y$. Falls A nicht leer ist, folgt

$$A/\sim = \{A\}.$$

(7) Die Gleichmächtigkeitsbeziehung ist reflexiv, symmetrisch und transitiv (siehe Anmerkung 2.10(e)). Sie ist aber keine Relation, da es die Menge aller Mengen nicht gibt. ◁

Es sei $[x]_\sim$ eine Äquivalenzklasse bezüglich einer Äquivalenzrelation auf einer Menge A. Weiter sei $y \in [x]_\sim$, also $x \sim y$. Für alle $z \in [y]_\sim$ gilt dann wegen der Transitivität von „$\sim$“ auch $x \sim z$, also $z \in [x]_\sim$. Wir erhalten $[y]_\sim \subseteq [x]_\sim$. Wegen der Symmetrie von „$\sim$“ folgt aus $y \in [x]_\sim$ auch $x \in [y]_\sim$, das gleiche Argument mit vertauschten Rollen liefert also $[x]_\sim \subseteq [y]_\sim$, und wir schließen $[x]_\sim = [y]_\sim$. Wir haben gezeigt, dass jedes Element $y \in C$ einer Äquivalenzklasse C die Klasse „vertritt“ in dem Sinne, dass $C = [y]_\sim$ gilt. Daher nennt man die Elemente von Äquivalenzklassen auch Vertreter. Alle Vertreter sind gleichberechtigt, und jede Auswahl eines bestimmten Vertreters ist ein Akt der Willkür.

Außerdem folgt, dass zwei Äquivalenzklassen, die auch nur ein Element gemeinsam haben, identisch sind. Außerdem sind Äquivalenzklassen wegen der Reflexivität nie leer, und ihre Vereinigung ergibt ganz A. Wir haben bewiesen:

Satz 3.7. *Es seien „$\sim$“ eine Äquivalenzrelation auf einer Menge A und $M := A/\sim$ die Faktormenge. Dann sind die Elemente von M nicht leer und paarweise disjunkt. Außerdem gilt $\bigcup M = A$.*

Der Satz liefert eine Steilvorlage für die Anwendung des Auswahlaxioms (Axiom 1.14). Indem man es auf $A/\sim$ anwendet und die erhaltene Menge X mit A schneidet, erhält man eine Menge $Y = A \cap X$, die zu jeder Äquivalenzklasse genau einen Vertreter enthält und die aus diesen Vertretern besteht. Eine solche Menge nennt man ein **Vertretersystem**. Es ist nicht schwer zu sehen, dass das Auswahlaxiom (unter Annahme der übrigen Axiome der Zermelo-Fraenkel-Mengenlehre) äquivalent ist zu der Aussage, dass es zu jeder Äquivalenzrelation ein Vertretersystem gibt.

Für den Rest des Abschnitts beschäftigen wir uns mit Ordnungsrelationen. Ist R eine Ordnungsrelation, so schreiben wir standardmäßig $x \leq y$ statt xRy und sprechen von der Ordnungsrelation „$\leq$“. Eine Menge mit einer Ordnungsrelation heißt auch eine **geordnete Menge**.

Beispiel 3.8. (1) Die bekannten Zahlenbereiche $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$ sind in herkömmlicher Weise geordnet. Beispielsweise gilt für $x, y \in \mathbb{Z}$ genau dann $x \leq y$, wenn $y - x \in \mathbb{N}$. Auf $\mathbb{C}$ gibt es keine natürliche Ordnungsrelation.

(2) Die Teilbarkeitsbeziehung auf $\mathbb{N}$ (siehe Beispiel 3.2(2)) ist eine Ordnungsrelation. Für $x = 3$ und $y = 5$ gilt weder $x \mid y$ noch $y \mid x$. Jede natürliche Zahl teilt 0 und ist durch 1 teilbar.

(3) Man kann die Teilbarkeitsbeziehung auch auf $\mathbb{Z}$ definieren. Dies ergibt allerdings keine Ordnungsrelation, da die Antisymmetrie fehlt. Beispielsweise gelten $-1 \mid 1$ und $1 \mid -1$. Die Teilbarkeitsbeziehung auf $\mathbb{N}$ ist die Einschränkung der Teilbarkeitsbeziehung auf $\mathbb{Z}$.

(4) Auf $A = \{1, 2, 3, 4\}$ ist eine Ordnungsrelation definiert durch

$$R = \{(3,3), (3,2), (3,1), (1,1), (1,2), (2,2), (4,4)\}.$$

Es gilt also $3 \leq 1 \leq 2$.

(5) Ist A eine Menge, so ist die Potenzmenge $\mathfrak{P}(A)$ durch die Teilmengenbeziehung geordnet, für $B, C \subseteq A$ ist also

$$B \leq C \quad :\Longleftrightarrow \quad B \subseteq C.$$

(6) Ist „$\leq$“ eine Ordnungsrelation auf einer Menge A, so erhalten wir eine neue Ordnungsrelation „$\preceq$“ auf A, indem wir für $x, y \in A$ definieren:

$$x \preceq y \quad \Longleftrightarrow \quad y \leq x.$$

(7) Auf jeder Menge A ist $\{(x, x) \mid x \in A\}$ eine Ordnungsrelation. ◁

Ist „$\leq$“ eine Ordnungsrelation auf einer Menge A, so benutzt man häufig folgende Schreib- und Sprechweisen für $x, y \in A$:

- $x \geq y \ :\Longleftrightarrow \ y \leq x$,
- $x < y \ :\Longleftrightarrow \ x \leq y$ und $x \neq y$,
- $x > y \ :\Longleftrightarrow \ y < x$,
- x und y heißen *vergleichbar*, falls $x \leq y$ oder $y \leq x$.

An den obigen Beispielen haben wir gesehen, dass in einer geordneten Menge A nicht unbedingt alle $x, y \in A$ vergleichbar sind. Dies (und Anderes) wird in folgender Definition thematisiert.

Definition 3.9. *Es sei „$\leq$“ eine Ordnungsrelation auf einer Menge A.*

(a) Die Ordnungsrelation „$\leq$“ heißt eine **totale Ordnung***, falls alle $x, y \in A$ vergleichbar sind. In diesem Fall heißt A eine* **total geordnete Menge***. Falls „$\leq$“ nicht total ist, spricht man auch von einer* partiellen Ordnung *und nennt A eine* partiell geordnete Menge*.*

(b) Eine Teilmenge $B \subseteq A$ heißt eine **Kette** *(oder auch* total geordnete Teilmenge*), falls die auf B eingeschränkte Ordnungsrelation total ist.*

(c) Ein Element $a \in A$ heißt **maximal** *bzw.* **minimal***, falls es kein $x \in A$ gibt mit $x > a$ bzw. $x < a$.*

(d) Ein Element $a \in A$ heißt **größtes** *bzw.* **kleinstes Element***, falls für alle $x \in A$ gilt: $x \leq a$ bzw. $a \leq x$.*

(e) A heißt **wohlgeordnet** *(und die Ordnungsrelation „$\leq$“ entsprechend eine* **Wohlordnung***), falls jede nicht leere Teilmenge $B \subseteq A$ ein kleinstes Element besitzt.*

(f) Eine Teilmenge $B \subseteq A$ heißt **nach oben** *bzw.* **nach unten beschränkt***, falls es ein $a \in A$ gibt, so dass $x \leq a$ bzw. $a \leq x$ für alle $x \in B$ gilt. Ein solches a heißt dann eine* **obere** *bzw.* **untere Schranke** *von B.*

Die durchaus subtilen Unterscheidungen dieser Definition illustrieren wir nun an Beispielen.

Beispiel 3.10. (1) Die herkömmlichen Ordnungsrelationen auf $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$ sind total. Damit ist auch jede Teilmenge eine Kette. Der Ausdruck „Kette“ kann irreführend sein, weil er suggeriert, dass man die Elemente einer Kette als $a_1, a_2, a_3, \ldots$ schreiben kann mit $a_1 < a_2 < a_3 < \cdots$. Das Beispiel der Kette $\mathbb{R}$ zeigt, dass dies nicht so ist, allein schon deshalb, weil $\mathbb{R}$ überabzählbar ist.
$\mathbb{N}$ hat das kleinste Element 0, sonst gibt es in den bekannten Zahlenbereichen keine kleinsten oder größten Elemente und ebensowenig maximale oder minimale Elemente.
Das offene Intervall $\{x \in \mathbb{R} \mid x < 1\}$ hat keine größten, kleinsten, maximalen oder minimalen Elemente, es ist aber nach oben beschränkt durch die obere Schranke 1. Jede Zahl ≥ 1 ist eine obere Schranke, obere Schranken sind also im Allgemeinen nicht eindeutig bestimmt.

(2) $\mathbb{N}$ ist durch die Teilbarkeitsbeziehung partiell geordnet. Die Menge aller Zweierpotenzen ist eine Kette. Das kleinste Element ist 1, das größte 0. Wenn man die Teilbarkeitsbeziehung auf $\mathbb{N} \setminus \{1\}$ einschränkt, sind die minimalen Elemente genau die Primzahlen. Minimale Elemente sind also im Allgemeinen nicht eindeutig bestimmt.

(3) Die Ordnungsrelation aus Beispiel 3.8(4) ist partiell. Die Elemente 3 und 4 sind minimal, 2 und 4 sind maximal.

(4) Das Standardbeispiel für eine wohlgeordnete Menge ist $\mathbb{N}$ mit der herkömmlichen Ordnungsrelation. Intuitiv düfte klar sein, dass $\mathbb{N}$ wohlgeordnet ist. Den Nachweis führen wir am Ende des Abschnitts (Satz 3.14). Wir merken an, dass jede wohlgeordnete Menge totalgeordnet ist, aber nicht umgekehrt, wie die Beispiele $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$ zeigen.

(5) Es seien A eine Menge und $\mathfrak{P}(A)$ die Potenzmenge mit der Teilmengenbeziehung als Ordnungsrelation (siehe Beispiel 3.8(5)). Die Ordnung ist nur dann total, wenn A höchstens ein Element enthält. Das kleinste Element von $\mathfrak{P}(A)$ ist $\emptyset$, das größte ist A. Jede Teilmenge $M \subseteq \mathfrak{P}(A)$ ist nach oben beschränkt durch $\bigcup M$ und nach unten durch $\bigcap M$ falls $M \neq \emptyset$, sonst durch jede beliebige Teilmenge.

(6) In jeder geordneten Menge A sind alle einelementigen Teilmengen und $\emptyset$ Ketten. ◁

Nur in partiell geordneten Mengen gibt es einen Unterschied zwischen größten und maximalen Elementen (bzw. zwischen kleinsten und minimalen). Die folgende Proposition handelt vom Verhältnis dieser beiden Begriffe.

Proposition 3.11. *Falls es in einer geordneten Menge A ein größtes Element a gibt, so ist dies eindeutig bestimmt, und für alle $b \in A$ gilt:*

$$b \text{ ist maximal} \quad \Longleftrightarrow \quad b = a.$$

Entsprechendes gilt für kleinste und minimale Elemente.

Beweis. Da jedes größte Element maximal ist, geht die Eindeutigkeit des größten Elements aus der zweiten Behauptung hervor, und es ist nur die Im-

plikation „⇒“ zu zeigen. Ist b maximal, so ist $a > b$ unmöglich. Andererseits gilt nach Voraussetzung $a \geq b$, also folgt $a = b$.

Der Beweis für die entsprechenden Aussagen über kleinste und minimale Elemente läuft analog. □

Wir haben nun alle Begriffe, um das Zornsche Lemma formulieren zu können.

Satz 3.12 (Zornsches Lemma)**.** *Falls in einer geordneten Menge M jede Kette nach oben beschränkt ist, so gibt es in M mindestens ein maximales Element.*

Anmerkung. Bisweilen wird zusätzlich gefordert, dass M nicht leer ist. Diese Forderung ist jedoch in den Voraussetzungen von Satz 3.12 enthalten, denn es wird insbesondere für die leere Kette die Existenz einer oberen Schranke vorausgesetzt. ◁

Wie bereits erwähnt ist das Zornsche Lemma äquivalent zum Auswahlaxiom. Der schwierigere Teil des Beweises ist die Herleitung des Zornschen Lemmas aus dem Auswahlaxiom. Wir könnten dies mit den uns zur Verfügung stehenden Mitteln durchführen, es ist jedoch sehr aufwändig und kompliziert. Der Nachweis findet sich in dem bereits erwähnten Buch von Halmos.

Als Anwendung des Zornschen Lemmas führen wir nun den Beweis des Vergleichbarkeitssatzes für Mengen.

Beweis von Satz 2.9(a). Für zwei Mengen A, B ist zu zeigen, dass es eine injektive Abbildung $A \to B$ oder eine injektive Abbildung $B \to A$ gibt. Wir nennen eine Teilmenge $C \subseteq A \times B$ des kartesischen Produkts eine *partielle Korrespondenz*, falls für alle $x, x' \in A$ und $y, y' \in B$ gelten:

$$(x, y) \in C \quad \text{und} \quad (x, y') \in C \quad \Rightarrow \quad y = y', \tag{3.1}$$

$$(x, y) \in C \quad \text{und} \quad (x', y) \in C \quad \Rightarrow \quad x = x'. \tag{3.2}$$

Nun setzen wir

$$M := \{C \subseteq A \times B \mid C \text{ ist eine partielle Korrespondenz}\}$$

und versehen M mit der durch die Teilmengenbeziehung gegebene Ordnungsrelation. Für den Nachweis der Voraussetzung des Zornschen Lemmas betrachten wir eine beliebige Kette $K \subseteq M$ und bilden die Vereinigungsmenge $Z := \bigcup K$. Falls wir nachweisen können, dass Z eine partielle Korrespondenz ist, liefert Z eine obere Schranke von K. Es seien also $x \in A$ und $y, y' \in B$ mit $(x, y) \in Z$ und $(x, y') \in Z$. Dann gibt es $C, C' \in K$ mit $(x, y) \in C$ und $(x, y') \in C'$. Da K total geordnet ist, gilt $C \subseteq C'$ oder $C' \subseteq C$. Im ersten Fall folgt $(x, y) \in C'$, also $y = y'$, da C' eine partielle Korrespondenz ist. Im zweiten Fall folgt ebenso $y = y'$. Also wird (3.1) durch Z erfüllt. Der Nachweis von (3.2) läuft entsprechend. Damit ist Z wie behauptet eine obere Schranke von K.

Das Zornsche Lemma (Satz 3.12) liefert die Existenz eines maximalen Elements $C \in M$. Wir nehmen nun an, dass es $x \in A$ gibt, so dass $(x, y') \notin C$ für alle $y' \in B$, und dass es $y \in B$ gibt, so dass $(x', y) \notin C$ für alle $x' \in A$. Dann ist $(x, y) \notin C$, aber $C \cup \{(x, y)\}$ ist eine partielle Korrespondenz. Dies steht im Widerspruch zur Maximalität von C, die Annahme ist also falsch.

Aus der Negation der Annahme erhalten wir zwei Fälle. Im ersten gibt es für alle $x \in A$ ein $y' \in B$ mit $(x, y') \in C$. Wegen (3.1) ist C dann eine Abbildung $A \to B$, die wegen (3.2) injektiv ist. Im zweiten Fall gibt es für alle $y \in B$ ein $x' \in A$ mit $(x', y) \in C$. Wegen (3.2) ist $C^* := \{(y, x) \in B \times A \mid (x, y) \in C\}$ dann eine Abbildung $B \to A$, die wegen (3.1) injektiv ist. Dies schließt den Beweis ab. □

Wir haben bereits erwähnt, dass das Auswahlaxiom äquivalent ist zum Wohlordnungssatz. Dieser wird in der Vorlesung nie verwendet, wir formulieren ihn hier aber.

Satz 3.13 (Wohlordnungssatz)**.** *Auf jeder Menge gibt es eine Wohlordnung.*

Die herkömmliche Ordnungsrelation auf $\mathbb{N}$ ist definiert durch

$$n \leq m \quad :\Longleftrightarrow \quad m = n + x \quad \text{mit} \quad x \in \mathbb{N}.$$

Satz 3.14. *Mit der herkömmlichen Ordnung ist $\mathbb{N}$ wohlgeordnet.*

Vor dem Beweis des Satzes bringen wir ein Lemma mit einem sehr seltsamen Induktionsbeweis.

Lemma 3.15. *Für jedes $n \in \mathbb{N}$ mit $n \neq 0$ gibt es ein $m \in \mathbb{N}$ mit $n = m + 1$.*

Beweis. Für die erste Behauptung benutzen wir Induktion. Für $n = 0$ ist nichts zu zeigen. Im Induktionsschritt müssen wir die Aussage für $n + 1$ anstelle von n zeigen. Sie gilt in der Tat mit $m = n$. □

Beweis von Satz 3.14. Um zu beweisen, dass jede nicht-leere Teilmenge von $A \subseteq \mathbb{N}$ ein kleinstes Element hat, zeigen per Induktion nach n, dass folgende Aussage für jedes $n \in \mathbb{N}$ gilt: Ist $A \subseteq \mathbb{N}$ eine Menge, die mindestens eine Zahl $\leq n$ enthält, so hat A ein kleinstes Element.

Der Induktionsanfang $n = 0$ funktioniert folgendermaßen: Es gibt ein $k \in A$ mit $k \leq 0$. Andererseits gilt $k = 0 + k \geq 0$, also $k = 0$. Nun gilt für jedes $m \in A$: $m = 0 + m \geq 0$, also ist 0 kleinstes Element von A.

Für den Induktionsschritt ist die Voraussetzung, dass es ein $k \in A$ mit $k \leq n+1$ gibt. Falls es auch ein $k \in A$ mit $k \leq n$ gibt, so folgt die Behauptung per Induktion. Wir dürfen also voraussetzen, dass es *kein* $k \in A$ mit $k \leq n$ gibt. Wir behaupten, dass dann $n + 1$ kleinstes Element von A ist. Es sei $m \in A$ beliebig. Die Menge $\{n, m\}$ hat nach Induktionsvoraussetzung ein kleinstes Element, und da $m \leq n$ *nicht* gilt, muss dieses n sein, also $n < m$. Dies bedeutet $m = n + x$ mit $0 \neq x \in \mathbb{N}$, also nach Lemma 3.15 $x = y + 1$ mit $y \in \mathbb{N}$. Wir erhalten

$$m = n + y + 1 = (n + 1) + y \geq n + 1.$$

Dies zeigt, dass $n + 1$ eine untere Schranke von A ist. Da A aber auch eine Zahl $\leq n + 1$ enthält, muss diese gleich $n + 1$ sein, also $n + 1 \in A$, und damit ist $n + 1$ kleinstes Element. □

Auf Satz 3.14 beruht das Prinzip der **starken Induktion**, das wir nun vorstellen: Es sei $\mathcal{A}(n)$ eine Aussage über eine natürliche Zahl n. Man darf nun voraussetzen, dass $\mathcal{A}(k)$ für alle natürlichen Zahlen $k < n$ gilt (Induktionsannahme), und muss daraus folgern, dass $\mathcal{A}(n)$ gilt. Dann ist $\mathcal{A}(n)$ für alle $n \in \mathbb{N}$ bewiesen.

Für den Beweis, dass dies tatsächlich zutrifft, nehmen wir an, dass es natürliche Zahlen n gibt, für die $\mathcal{A}(n)$ nicht gilt. Dann ist die Menge

$$M := \{n \in \mathbb{N} \mid \mathcal{A}(n) \text{ gilt nicht}\} \subseteq \mathbb{N}$$

nicht leer. Nach Satz 3.14 hat M ein kleinstes Element $n_0 \in M$. Für $k \in \mathbb{N}$ mit $k < n_0$ folgt $k \notin M$, also gilt $\mathcal{A}(k)$ für diese k. Da man hieraus schließen kann, dass auch $\mathcal{A}(n_0)$ gilt, folgt $n_0 \notin M$, ein Widerspruch.

Ein typisches Beispiel für starke Induktion ist der Beweis des folgenden wichtigen Satzes.

Satz 3.16. *Jede natürliche Zahl $n \geq 2$ lässt sich als Produkt von Primzahlen schreiben.*

Beweis. Es sei $n \in \mathbb{N}$. Falls $n < 2$, so ist nichts zu zeigen, wir nehmen also $n \geq 2$ an. Ist n eine Primzahl so sind wir fertig. Andernfalls gibt es eine Zerlegung $n = a \cdot b$ mit $2 \leq a, b < n$. Gemäß der Induktionsannahme sind a und b Produkte von Primzahlen, also auch n. □

Der Satz sagt nicht, dass die Zerlegung als Produkt von Primzahlen bis auf die Reihenfolge eindeutig ist. Dies beweisen wir (wesentlich) später, siehe Satz 18.14.

Es fällt auf, dass das Prinzip der starken Induktion keinen Induktionsanfang benötigt.

Diskrete Strukturen: Graphen

4 Wege und Bäume

Graphen sind diskrete Objekte, die vielseitig zur Beschreibung realer Situationen einsetzbar sind. Wir beginnen mit der Definition.

Definition 4.1. *Ein* **Graph** *ist ein geordnetes Paar $G = (V, E)$, bestehend aus einer nicht-leeren, endlichen Menge V und einer Menge*

$$E \subseteq \big\{\{x, y\} \mid x, y \in V,\ x \neq y\big\}$$

von zweielementigen Teilmengen von V. Die Elemente von V werden **Knoten** *oder auch* **Ecken** *genannt, die von E werden* **Kanten** *genannt.*

Oft werden Graphen durch Diagramme gekennzeichnet oder gegeben, wie durch folgendes Beispiel gezeigt wird.

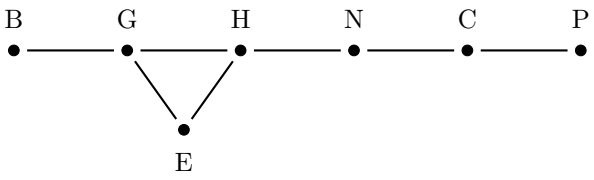
Beispiel 4.2. (1) Die Knotenmenge sei gegeben durch die Länder Mittelamerikas (gekennzeichnet durch ihre Anfangsbuchstaben), also

$$V = \{B, C, E, G, H, N, P\}.$$

Falls zwei dieser Länder aneinander grenzen, seien sie durch eine Kante verbunden. Wir erhalten

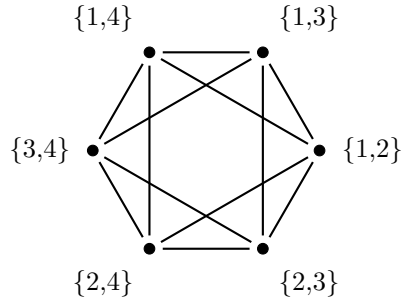
$$E = \big\{\{B, G\}, \{C, N\}, \{C, P\}, \{E, G\}, \{E, H\}, \{G, H\}, \{H, N\}\big\},$$

was sich als das Diagramm



darstellt.

(2) Das folgende Diagramm stellt den Graphen mit den zweielementigen Teilmengen der Menge $\{1, 2, 3, 4\}$ als Knoten dar, wobei zwei Knoten eine Kante haben, falls ihre Schnittmenge nicht leer ist.



(3) Auch interessant ist der Graph, dessen Knoten alle Teilnehmer bei Facebook sind, mit Kanten zwischen Facebook-Freunden. Diesen Graphen hier zu zeichnen würde den Umfang des Skrips sprengen.

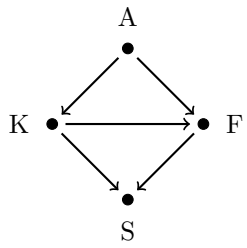
◁

Anmerkung 4.3. Es gibt einige Varianten des Begriffs eines Graphen. Die wichtigsten hiervon wollen wir hier vorstellen.

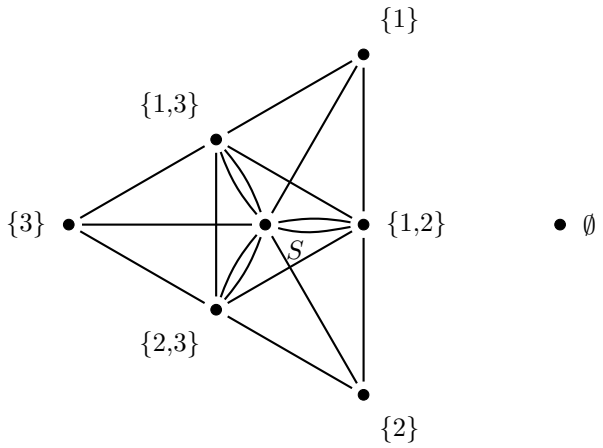
(a) Zunächst werden häufig auch unendliche Graphen betrachtet, d.h. die Bedingung der Endlichkeit an V wird weggelassen.

(b) Manchmal werden in Graphen auch Kanten von einem Knoten zu sich selbst („Schleifen“) zugelassen, definiert als einelementige Teilmengen von V.

(c) **Gerichtete Graphen**: Die Kanten haben eine Richtung und werden durch Pfeile gekennzeichet. Mathematisch definiert man dies, indem man sagt, dass die Kantenmenge eine Teilmenge des kartesischen Produkts $V \times V$ ist, wobei Schleifen (also Kanten der Form (x, x)) meist nicht zugelassen werden. Ein Beispiel ist die Nahrungskette verschiedener Tierarten, die eben im Allgemeinen keine Kette, sondern ein gerichteter Graph ist. Hier betrachten wir: Kormoran (K) und Forelle (F) fressen Steinkrebs (S), Adler (A) und Kormoran fressen Forelle, und Adler frisst Komoran. Der Graph ist

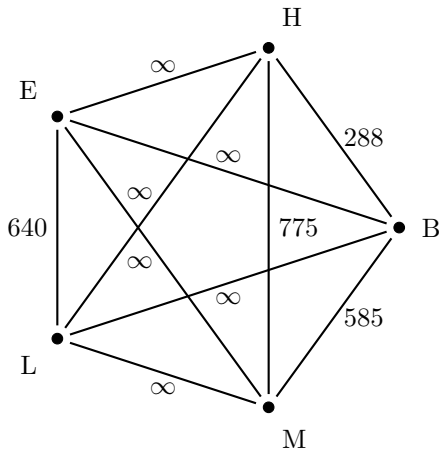


(d) **Multigraphen** : Zwischen zwei Knoten sind mehrere Kanten erlaubt. Die exakte mathematische Definition geben wir später (Definition 5.1). Als Beispiel zeichnen wir den Graphen, dessen Knoten die Teilmengen von $S := \{1, 2, 3\}$ sind, wobei jedes gemeinsame Element von zwei Teilmengen für eine Kante sorgt.



Man betrachtet auch gerichtete Mutigraphen.

(e) **Gewichtete Graphen**: Dies sind Graphen, deren Kanten mit Elementen aus einer Menge (oft $\mathbb{R}$ oder $\mathbb{R} \cup \{\infty\}$) „gewichtet“ sind. Man kann sie definieren, indem man die Kantenmenge E durch eine Funktion ersetzt, die jeder zweielementigen Menge von Knoten das Gewicht der Kante zwischen ihnen zuordnet. Hierbei kann ein bestimmtes Gewicht (typischerweise 0 oder ∞) als nicht-existente Kante gedeutet werden. Ein typisches Beispiel ist der Entfernungsgraph zwischen Städten, dessen Kanten die Entfernung (Straßenverbindung auf dem Landweg) angibt. Für Berlin (B), Edinburgh (E), Hamburg (H), London (L) und München (M) ergibt sich



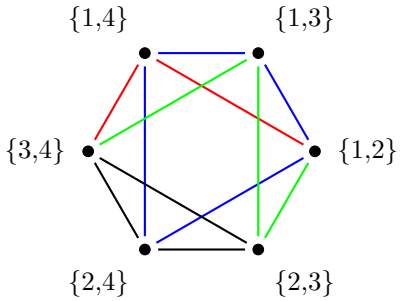
Die mit ∞ gewichteten Kanten bedeuten, dass es keinen Landweg gibt, sie können auch weggelassen werden.
Man betrachtet auch gewichtete gerichtete Graphen sowie Graphen, deren Knoten gewichtet sind.

Im Lichte dieser Varianten spricht man bisweilen von einem **einfachen Graph**, um zu spezifizieren, dass ein Graph gemäß Definition 4.1 gemeint ist. ◁

Im diesem Abschnitt sein $G = (V, E)$ immer ein Graph (gemäß Definition 4.1).

Definition 4.4. *(a) Ein* **Weg** *ist ein $(n+1)$-Tupel $(x_0, \ldots, x_n)$ von Knoten, so dass $\{x_{i-1}, x_i\} \in E$ für $1 \leq i \leq n$ und außerdem $x_i \neq x_j$ für alle $i, j \in \{0, \ldots, n\}$ mit $0 < i - j < n$. (Alle Knoten müssen verschieden sein bis auf die mögliche Ausnahme von x_0 und x_n.) Genauer spricht man von einem Weg der Länge n von x_0 nach x_n.*
(b) Ein Weg heißt ein **Kreis**, *falls $x_0 = x_n$ und $n \geq 3$.*
(c) G heißt **zusammenhängend**, *falls es für alle Knoten $x, y \in V$ mit $x \neq y$ einen Weg von x nach y gibt.*
(d) G heißt **kreisfrei**, *falls G keine Kreise hat.*

Beispiel 4.5. In dem Graphen aus Beispiel 4.2(2) sind zwei Wege von $\{3, 4\}$ nach $\{1, 2\}$ rot und grün gefärbt. Ein Kreis ist blau gefärbt.



Der Graph ist zusammenhängend. Dies triff nicht auf das Beispiel in Anmerkung 4.3(d) zu. ◁

Anmerkung 4.6. Für zwei Knoten x, y von G können wir $x \sim y$ schreiben, falls es einen Weg von x nach y gibt oder $x = y$. Dies ergibt eine Relation auf V, die reflexiv und symmetrisch ist. Um die Transitivität einzusehen, müssen wir Wege von Knoten x nach y und von y nach z zusammenhängen. Das Resultat ist ein Tupel wie in Definition 4.4(a), aber ohne die Verschiedenheit der x_i. Treten in dem Tupel aber zwei gleiche x_i auf, so kann man es verkürzen, indem man das Zwischenstück und eines der x_i herausnimmt. So bekommt man schließlich einen Weg von x nach z. Damit ist gezeigt, dass „$\sim$“ eine Äquivalenzrelation ist. Die Äquivalenzklassen, zusammen mit den Kanten zwischen ihren Knoten, heißen die **Zusammenhangskomponenten** von G. Diese sind zusammenhängend, und G selbst ist genau dann zusammenhängend, falls es nur eine Zusammenhangskomponente gibt. ◁

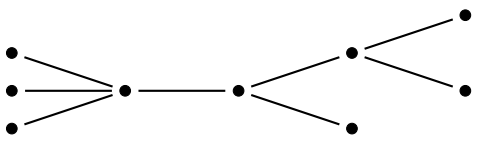
Zum Thema Zusammenhang und Kreisfreiheit werden wir etwas später beweisen:

Satz 4.7. *(a) Ist G zusammenhängend, so folgt $|E| \geq |V| - 1$.*
(b) Ist G kreisfrei, so folgt $|E| \leq |V| - 1$.
(c) G ist genau dann kreisfrei, wenn es für zwei verschiedene Knoten x, y von G höchstens einen Weg von x nach y gibt.

Ein wichtiger Typ von Graphen wird durch die folgende Definition gegeben.

Definition 4.8. *Der Graph G heißt ein* **Baum**, *falls er zusammenhängend und kreisfrei ist. In diesem Zusammenhang nennt man einen kreisfreien Graph auch einen* **Wald**, *da seine Zusammenhangskomponenten Bäume sind.*

Beispielsweise ist der Graph



ein Baum. Über Bäume werden wir beweisen:

Satz 4.9. *Die folgen Aussagen sind äquivalent:*

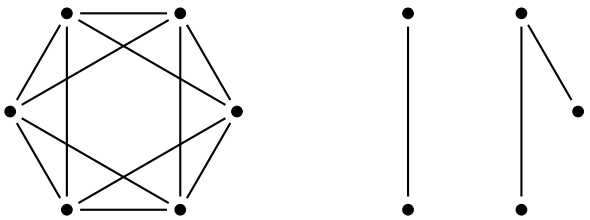
(a) G *ist ein Baum.*
(b) Für zwei verschiedene Knoten x, y *von* G *gibt es genau einen Weg von* x *nach* y*.*
(c) Es gilt $|E| = |V| - 1$*, und* G *ist zusammenhängend oder kreisfrei.*

Hat G also die „richtige“ Kantenzahl (nämlich $|V|-1$), so reicht der Nachweis des Zusammenhangs *oder* der Kreisfreiheit, um die andere dieser Eigenschaften zu garantieren.

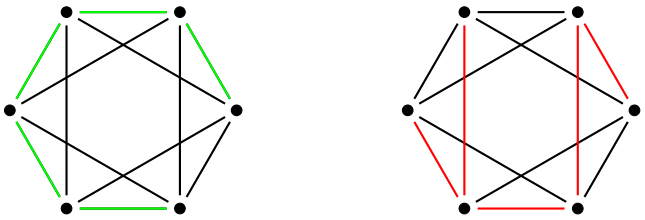
Wir beweisen die Sätze 4.7 und 4.9 nach dem folgenden Satz 4.12, für dessen Formulierung wir eine (auch sonst wichtige) Definition brauchen.

Definition 4.10. *(a) Ein Graph* $H = (W, F)$ *heißt* **Teilgraph** *von* G*, falls* $W \subseteq V$ *und* $F \subseteq E$*. Wir drücken dies durch* $H \leq G$ *aus. Ist* $W = V$*, so heißt* H *ein* **aufspannender Teilgraph***. Gleichbedeutend mit „Teilgraph“ sprechen wir auch von Untergraphen und Subgraphen.*
(b) Ein aufspannender Teilgraph B *von* G *heißt ein* **Spannbaum** *von* G*, falls* B *ein Baum ist.*

Beispiel 4.11. (1) Die Zusammenhangskomponenten eines Graphen sind Teilgraphen.
(2) Das folgende Diagramm stellt einen Graphen mit einem (nicht aufspannenden) Teilgraphen dar.



Der Teilgraph ist kreisfrei, aber nicht zusammenhängend. Im folgenden Diagramm sind zwei Spannbäume des linken Graphen farbig markiert.



Wir sehen, dass Spannbäume nicht eindeutig bestimmt sind. Aber aus Satz 4.9(c) wissen wir, dass alle Spannbäume dieselbe Kantenzahl (hier 5) haben müssen. ◁

Satz 4.12. *Der Graph G sei zusammenhängend, und $H \leq G$ sei ein kreisfreier Teilgraph. Dann gibt es einen Spannbaum B von G mit $H \leq B$. Insbesondere hat jeder zusammenhängende Graph einen Spannbaum.*

Beweis. Wir benutzen Induktion nach der Kantenzahl $|E|$. Falls G bereits kreisfrei ist, gibt es nichts zu zeigen. Wir nehmen also an, dass G einen Kreis K hat. Da H kreisfrei ist, gibt es in diesem Kreis zwei aufeinander folgende Knoten x, y, so dass die Kante $\{x, y\} \in E$ *nicht* Kante von H ist. Durch Entfernen dieser Kante bilden wir den aufspannenden Teilgraph

$$G' := \big(V, E \setminus \big\{\{x, y\}\big\}\big) .$$

Es folgt $H \leq G'$. Wir behaupten, dass G' zusammenhängend ist. Auch in G' gibt es einen Weg von x nach y, also $x \sim y$ (siehe Anmerkung 4.6 für die verwendete Notation). Ist nun $(x_0, \ldots, x_n)$ irgendein Weg in G, so gilt $x_i \sim x_{i+1}$ für alle i auch in G', also wegen der Transitivität $x_0 \sim x_n$. Nachdem wir wissen, dass G' zusammenhängend ist, liefert die Induktionsannahme einen Spannbaum B von G' mit $H \leq B$. Da G' in G und B in G' aufspannend sind, folgt, dass B auch ein Spannbaum von G ist. □

Vor dem Beweis der Sätze 4.7 und 4.9 schieben wir eine Definition und zwei Lemmata ein.

Definition 4.13. *Der* **Grad** *eines Knotens $x \in V$ ist die Anzahl der Kanten, die x mit anderen Knoten verbinden. Er wird mit* $\deg(x)$ *bezeichnet, also*

$$\deg(x) := \big|\big\{\{x, y\} \mid \{x, y\} \in E\big\}\big| .$$

Knoten vom Grad 0 nennt man auch **isolierte Knoten**.

Lemma 4.14. *Falls G kreisfrei ist und $E \neq \emptyset$, so hat G mindestens zwei Knoten vom Grad* 1.

Beweis. Da es Kanten gibt, gibt es auch Wege. Wir wählen einen Weg $(x_0, \ldots, x_n)$ maximaler Länge n. Also gibt es eine Kante zwischen x_0 und x_1 und damit $\deg(x_0) \geq 1$. Um zu zeigen, dass der Grad nicht größer als 1 ist, nehmen wir an, dass es eine Kante $\{x_0, y\} \in E$ mit $y \neq x_1$ gibt. Falls $y = x_i$ für ein i, dann wäre $i \geq 2$ und damit $(y, x_0, \ldots, x_i)$ ein Kreis, im Widerspruch zur Kreisfreiheit von G. Falls aber $y \neq x_i$ für alle i, so wäre $(y, x_0, \ldots, x_n)$ ein Weg der Länge $n+1$ im Widerspruch zur Maximalität von n. Da dasselbe auch mit x_n anstelle von x_0 gilt, ist das Lemma bewiesen. □

In einem Baum nennt man Knoten von Grad 1 auch **Blätter**.

Lemma 4.15. *Ist G ein Baum, so folgt $|E| = |V| - 1$.*

Beweis. Wir führen den Beweis per Induktion nach der Knotenzahl $|V|$. Für $|V| = 1$ ist nichts zu zeigen, wir setzen also $|V| > 1$ voraus. Weil G zusammenhängend ist, gibt es Kanten, also nach Lemma 4.14 auch einen Knoten x_0

vom Grad 1. Es sei $x_1 \in V$ der (einzige) mit x_0 verbundene Knoten. Wir entfernen nun x_0 aus dem Graphen, d.h. wir bilden den (nicht aufspannenden) Teilgraph

$$G' := \left(V \setminus \{x_0\}, E \setminus \big\{\{x_0, x_1\}\big\}\right).$$

G' ist zusammenhängend, denn für verschiedene Knoten x, y von G' gibt es einen Weg in G von x nach y. Weil x_0 nur mit einem einzigen Knoten eine Kante hat, kann in diesem Weg x_0 wegen der Verschiedenheit der Knoten im Weg nicht vorkommen, also liegt der Weg in G'. Da außerdem jeder Teilgraph eines kreisfreien Graphen selbst kreisfrei ist, gilt dies auch für G', also ist G' ein Baum. Die Induktionsannahme liefert nun

$$\big|E \setminus \big\{\{x_0, x_1\}\big\}\big| = |V \setminus \{x_0\}| - 1,$$

woraus die Behauptung folgt. □

Beweis von Satz 4.7. (a) Nach Satz 4.12 hat G einen Spannbaum $B = (V, F)$, für den nach Lemma 4.15 $|F| = |V| - 1$ gilt. Wegen $F \subseteq E$ folgt $|E| \geq |V| - 1$.

(b) Durch Hinzufügen von Kanten können wir aus G einen zusammenhängenden Graph G' machen. Nach Satz 4.12 hat G' einen Spannbaum $B = (V, F)$ mit $G \leq B$, also $E \subseteq F$. Mit Lemma 4.15 folgt $|E| \leq |V| - 1$.

(c) Wir setzen zunächst voraus, dass G kreisfrei ist und nehmen an, dass es zwei verschiedene Wege $(x_0, \ldots, x_n)$ und $(y_0, \ldots, y_m)$ gibt mit $x_0 = y_0 \neq x_n = y_m$. Sei $k \geq 0$ maximal mit $x_i = y_i$ für $i \leq k$, d.h. die Wege trennen sich nach dem Knoten $x_k = y_k$. Wegen der Verschiedenheit der x_i bzw. der y_i folgt $k < \min\{n, m\}$. Wegen $x_n = y_m$ gibt es auch ein minimales $l > k$, so dass x_l mit einem der y_i übereinstimmt, etwa $x_l = y_j$, d.h. die Wege laufen bei x_l wieder zusammen. Wir erhalten den Kreis

$$(x_k = y_k, y_{k+1} \ldots y_{j-1}, y_j = x_l, x_{l-1}, \ldots, x_k)$$

und damit einen Widerspruch zur Kreisfreiheit von G.
Da umgekehrt jeder Kreis $(x_0, \ldots, x_n = x_0)$ zu zwei verschiedenen Wegen (x_0, x_{n-1}) und $(x_0, x_1, \ldots, x_{n-1})$ führt, folgt auch, dass ein Graph mit höchstens einem Weg zwischen zwei Knoten kreisfrei ist. □

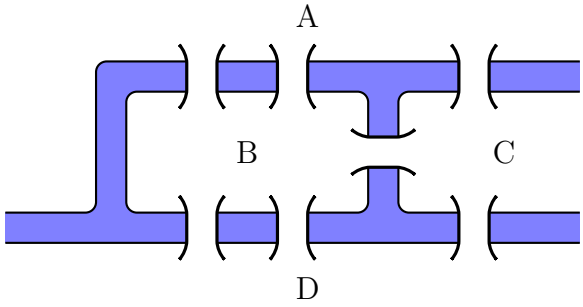
Beweis von Satz 4.9. Die Äquivalenz von (a) und (b) ergibt sich aus Satz 4.7(c) und der Definition von „zusammenhängend“ für Graphen. Die Implikation „(a) ⇒ (c)“ folgt aus der Definition eines Baumes und Lemma 4.15.

Es bleibt zu zeigen, dass (a) aus (c) folgt, wir haben also die Fälle zu betrachten, dass G zusammenhängend oder kreisfrei ist. Im ersten Fall hat G nach Satz 4.12 einen Spannbaum $B = (V, F)$. Nach Lemma 4.15 folgt $|F| = |V| - 1 = |E|$, also $F = E$ und G ist somit selbst ein Baum.

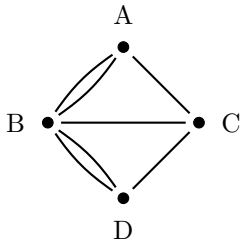
Sei nun G kreisfrei. Wie im Beweis von Satz 4.7(b) finden wir einen Baum $B = (V, F)$ mit $E \subseteq F$, und Lemma 4.15 mit der Voraussetzung $|E| = |V| - 1$ liefert wieder $F = E$. □

5 Multigraphen und eulersche Graphen

Im 18en Jahrhundert gab es in Königsberg (heute: Kaliningrad) sieben Brücken über den Fluss Pregel. Sie verbanden die Königsberger Stadtgebiete (gekennzeichnet durch die Buchstaben A-D) wie folgt:



Als das **Königsberger Brückenproblem** bezeichnet man die Frage, ob ein Spaziergang möglich ist, auf dem man jede Brücke genau einmal benutzt. Im besten Fall sollte dieser Weg sogar geschlossen sein. Dabei dürfen die Stadtgebiete mehrmals besucht werden. Möglicherweise hat Leonhard Euler (1707-1783) als erster erkannt, dass sich das Problem auf eine Graphentheoretische Frage reduziert. Da Stadtgebiete durch Brücken verbunden werden, stellt man sie als Knoten und die Brücken als Kanten dar. So erhält man den folgenden Multigraph



Die Frage ist nun, ob man sich so durch den Graph bewegen kann, dass man jede Kante genau einmal benutzt. Um diese anzugehen, müssen wir zunächst eine exakte Definition von Multigraphen geben. Es gibt verschiedene Möglichkeiten, dies zu tun. Wir folgen der Idee, die Kanten nicht nur als Zweiermengen von Knoten zu definieren, sondern ihnen zusätzlich eine Nummer zu geben, so dass man verschiedene Kanten zwischen denselben beiden Knoten unterscheiden kann.

Definition 5.1. *Ein* **Multigraph** *ist ein geordnetes Paar $G = (V, E)$, bestehend aus einer nicht-leeren, endlichen Menge V und einer endlichen Menge E, deren Elemente die Form*

$$K = \big(\{x, y\}, n\big)$$

mit $x, y \in V$, $x \neq y$, und $n \in \mathbb{N}$ haben. Ein solches K steht für eine Kante zwischen den Knoten x und y.

Der obige obige Graph wäre also gegeben durch die Kantenmenge

$$E = \big\{\big(\{A, B\}, 1\big), \big(\{A, B\}, 2\big), \big(\{B, C\}, 1\big), \big(\{B, D\}, 1\big), \big(\{B, D\}, 2\big), \big(\{A, C\}, 1\big), \big(\{C, D\}, 1\big)\big\}.$$

Die Multigraphen stellen eine Verallgemeinerung der einfachen Graphen (gemäß Definition 4.1) dar. Die Begriffe und Resultate aus Abschnitt 4 übertragen sich direkt auf den Fall von Multigraphen, wobei nur ein Begriff geschärft werden muss: Gibt es zwischen zwei Knoten x und y mehr als eine Kante, so sieht man (definitionsgemäß) den Weg von x nach y über eine der Kanten und zurück nach y über eine andere Kante als **Kreis** an, so dass ein Multigraph, der überhaupt mehrfache Kanten hat, niemals kreisfrei ist.

Für den Rest des Abschnits sei $G = (V, E)$ ein Multigraph, auch wenn wir bisweilen einfach von dem „Graph“ G sprechen werden.

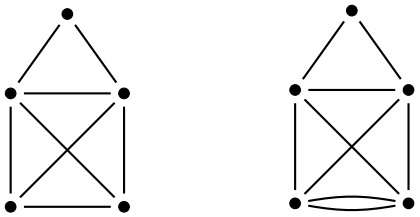
Da Leonhard Euler das Königsberger Brückenproblem gelöst hat, sind die Begriffe, die das Problem präzisieren, nach ihm benannt.

Definition 5.2. *(a) Ein* **Kantenzug** *ist ein m-Tupel $Z = (K_1, \ldots, K_m)$ mit $K_i = \big(\{x_{i-1}, x_i\}, n_i\big) \in E$, wobei $x_0, \ldots, x_m \in V$ (nicht notwendigerweise verschiedene) Knoten sind, die K_i aber paarweise verschieden sein müssen. Wir sagen, dass der Kantenzug die Kanten $K_1, \ldots, K_m$* benutzt *und die Knoten $x_0, \ldots, x_m$* besucht*. Ein Kantenzug ist also eine „Tour“, bei der jede Kante höchstens einmal benutzt werden darf.*

(b) Ein Kantenzug Z wie oben heißt **geschlossen***, falls $x_0 = x_m$. Er heißt* **eulersch***, falls $m = |E|$, d.h. falls* jede Kante *des Graphen benutzt wird.*

(c) Der Graph G heißt **eulersch***, falls es einen geschlossenen eulerschen Kantenzug gibt. Er heißt* **semi-eulersch***, falls es einen (nicht notwendig geschlossenen) eulerschen Kantenzug gibt.*

Anschaulich gesprochen ist ein Graph semi-eulersch, wenn man seine Kanten in einem Zug, also ohne abzusetzen, durchzeichnen kann. Ziel dieses Abschnittes ist es, einfache Kriterien herzuleiten für die Entscheidung, ob G (semi-)eulersch ist (Sätze 5.4 und 5.6). Beispiele für einen semi-eulerschen und einen eulerschen Graph sind das „Haus des Nikolaus“ und das „Haus des Nikolaus mit Fundament“.



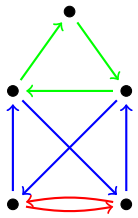
Hierbei ist momentan offen, ob das Haus des Nikolaus sogar eulersch und nicht nur semi-eulersch ist. Aus Satz 5.4, dessen Beweis wir nun angehen werden, ergibt sich jedoch, dass dies nicht der Fall ist.

Wie bei einfachen Graphen ist auch bei Multigraphen der **Grad** eines Knotens $x \in V$ als die Anzahl der von dem Knoten ausgehenden Kanten definiert. Beispielsweise haben im „Haus des Nikolaus mit Fundament“ alle Knoten den Grad 4 bis auf den obersten, der Grad 2 hat.

Proposition 5.3. *Die folgenden Aussagen sind äquivalent:*

(a) Es gibt geschlossene Kantenzüge $Z_1, \ldots, Z_r$, so dass jede Kante von E in genau einem der Z_i benutzt wird.
(b) Sämtliche Knoten von G haben eine gerade Zahl als Grad.

Bevor wir die Proposition beweisen, illustrieren wir die Situation der Aussage (a) bei dem „Haus des Nikolaus mit Fundament“, wobei die Kantenzüge Z_i durch verschiedene Farben dargestellt sind.



Es gibt viele andere mögliche Wahlen für die Kantenzüge Z_i.

Beweis von Proposition 5.3. Wir setzen zunächst die Aussage (a) voraus. Bei einem geschlossenen Kantenzug wird jeder Knoten, der besucht wird, auch wieder verlassen, und dabei werden verschiedene Kanten benutzt. Hieraus ergibt sich (b).

Nun setzen wir (b) voraus und beweisen (a) mit (starker) Induktion nach der Kantenzahl $|E|$. Im Falle $E = \emptyset$ ist nichts zu zeigen ($r = 0$). Im Falle $E \neq \emptyset$ gibt es Kantenzüge, und wir können einen Kantenzug $Z = (K_1, \ldots, K_m)$ mit maximaler Länge m wählen. Die von Z besuchten Knoten seien $x_0, \ldots, x_m$. Wir nehmen an, dass Z nicht geschlossen sei, also $x_0 \neq x_m$. Dann leisten die Kanten von Z zu dem Grad von x_m (ebenso von x_0) einen ungeraden Beitrag. Wegen (b) folgt, dass von x_m eine von Z

nicht benutzte Kante ausgeht. Diese können wir an Z anhängen, im Widerspruch zur Maximalität der Länge von Z. Also ist Z doch geschlossen.

Wir wissen bereits, dass die Kanten von Z zu jedem Grad eines Knotens einen geraden Beitag leisten. Also gilt (b) auch für den Teilgraph

$$G' := \big(V, E \setminus \{K_1, \ldots, K_m\}\big).$$

Per Induktion folgt nun (a) für G' und damit, durch Hinzufügen von Z zu den Kantenzügen von G', auch für G. □

Für die Formulierung der nächsten beiden Sätze benutzen wir folgende ad hoc Notation: Mit G^0 bezeichnen wir den Teilgraphen, der aus G durch das Entfernen aller isolierter Knoten aber Beibehalten aller Kanten entsteht.

Satz 5.4. *Falls $E \neq \emptyset$, so sind die folgenden Aussagen äquivalent:*

(a) G ist eulersch.
(b) G^0 ist zusammenhängend, und sämtliche Knoten von G haben eine gerade Zahl als Grad.

In diesem Fall ist jeder eulersche Kantenzug geschlossen.

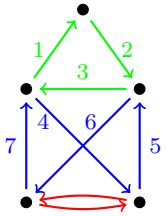
Beweis. Falls G eulersch ist, gilt die Aussage (a) aus Proposition 5.3 mit $r = 1$, also haben gemäß der Proposition alle Knoten geraden Grad. Außerdem besucht ein eulerscher Kantenzug jeden nicht isolierten Knoten, woraus der Zusammenhang von G^0 folgt.

Gilt umgekehrt (b), so liefert Proposition 5.3 geschlossene Kantenzüge $Z_1, \ldots, Z_r$ mit den dort genannten Eigenschaften. Im Falle $r = 1$ ist (a) gezeigt, wir setzen also $r \geq 2$ voraus. Nun nehmen wir an, dass es für kein $i \in \{2, \ldots, r\}$ einen Knoten gibt, der sowohl von Z_i als auch von Z_1 besucht wird. Dann gehen von den von Z_1 besuchten Knoten nur die Kanten aus Z_1 aus. Diese Knoten bilden also eine Zusammenhangskomponente, im Widerspruch zum Zusammenhang von G^0. Es folgt, dass es ein $i \geq 2$ gibt, so dass mindestens ein Knoten sowohl von Z_1 als auch von Z_i besucht wird.

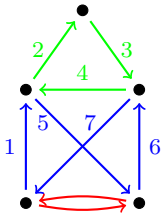
Es sei x ein solcher gemeinsam von Z_1 und Z_i besuchter Knoten. Wir können die Kanten in Z_1 und Z_i so umnummerieren, dass beide bei x beginnen und enden. Nun hängen wir Z_1 und Z_i zusammen, indem wir die entsprechenden Kanten hintereinander schreiben. Dies ergibt einen geschlossenen Kantenzug, der alle Kanten von Z_1 und von Z_i genau einmal benutzt. Nun können wir Z_1 durch den neuen Kantenzug ersetzen und Z_i streichen. Indem wir so fortfahren, erreichen wir schließlich $r = 1$.

Die letzte Behauptung folgt aus der Beobachtung, dass die Endknoten eines nicht geschlossenen eulerschen Kantenzugs ungeraden Grad haben. □

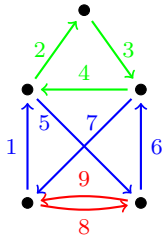
Beispiel 5.5. Das Aneinanderhängen der Kantenzüge aus dem obigen Beweis ist hier anhand des „Hauses des Nikolaus mit Fundament“ illustriert:



Aneinanderhängen der grünen und blauen Kantenzüge



Umnummerieren für Start links unten



Anhängen des roten Kantenzugs

◁

Nach Satz 5.4 ist das Haus des Nikolaus also nicht eulersch, aber gemäß dem folgenden Satz semi-eulersch.

Satz 5.6. *Falls $E \neq \emptyset$, so sind die folgenden Aussagen äquivalent:*

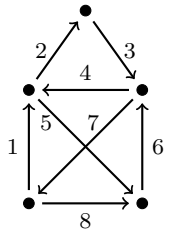
(a) G ist semi-eulersch aber nicht eulersch.
(b) G^0 ist zusammenhängend, und G hat genau zwei Knoten mit einer ungeraden Zahl als Grad.

In diesem Fall hat jeder eulersche Kantenzug die beiden Knoten mit ungeradem Grad als Endknoten.

Beweis. Falls die Aussage (a) gilt, so gibt es einen nicht geschlossenen eulerschen Kantenzug. Wir haben schon im Beweis von Satz 5.4 gesehen, dass hieraus der Zusammenhang von G^0 folgt. Außerdem haben die Endknoten des eulerschen Kantenzugs ungeraden Grad, alle anderen Knoten aber geraden Grad, es folgt also (b).

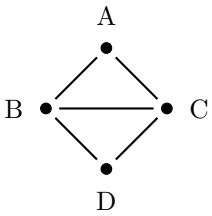
Nun setzen wir umgekehrt die Aussage (b) voraus. Hieraus folgt, dass *jeder* eulersche Kantenzug die beiden Knoten mit ungeradem Grad als Endknoten hat, die letzte Behauptung des Satzes. Insbesondere gibt es keinen geschlossenen eulerschen Kantenzug, G ist also nicht eulersch. Um einzusehen, dass G semi-eulersch ist, verbinden wir die beiden Knoten mit ungeradem Grad durch eine zusätzliche Kante K. Dadurch entsteht ein Graph G', bei dem alle Knoten geraden Grad haben. Nach Satz 5.4 ist G' eulersch, wir haben also einen geschlossenen eulerschen Kantenzug. Dessen Kanten können wir so anordnen, dass die zusätzliche Kante K als *letzte* Kante benutzt wird. Nun streichen wir diese Kante und erhalten so einen eulerschen Kantenzug in G. Die Aussage (a) gilt also. □

Das „Haus des Nikolaus“ ist demnach semi-eulersch, mit den unteren Knoten vom Grad 3. Einen eulerschen Kantenzug erhält man, indem zwischen diesen beiden Knoten eine weitere Kante hinzufügt und nun einen geschlossenenen eulerschen Kantenzug konstruiert mit der neuen Kante als letzte. Dies wurde in Beispiel 5.5 durchgeführt. Durch Entfernen dieser Kante erhält man folgenden eulerschen Kantenzug für das „Haus des Nikolaus“:

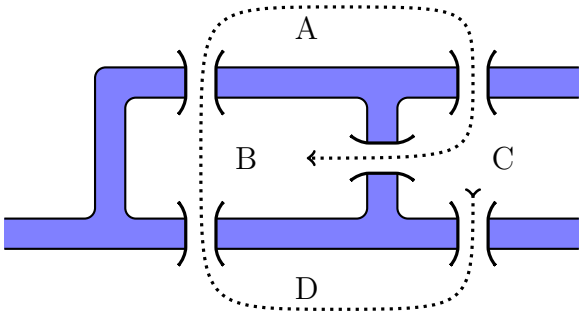


Der obige eulersche Kantenzug ist bei weitem nicht der einzig mögliche.

Nun können wir zurückkommen auf unsere Ausgangsfrage, das Königsberger Brückenproblem. Bei dem entsprechenden Multigraph (siehe zu Beginn des Abschnitts) haben sämtliche Knoten ungeraden Grad. Der Graph ist daher nicht semi-eulersch, also hat das Problem eine negative Antwort. Wir können noch mehr sagen: Sobald man eine Brücke abreißt oder hinzubaut, ändert man den Grad von genau zwei Knoten um Eins, also wird der Graph semi-eulersch. Per Wikipedia oder Google Maps erfährt man, dass in der heutigen Innenstadt von Kaliningrad zwei der Brücken fehlen: Es gibt nur noch je eine Brücke zwischen der Insel und den nördlichen und südlichen Stadtgebieten. Der heutige Graph ist also



und damit semi-eulersch. Ein Spaziergang, der jede Brücke genau einmal benutzt, ist in der folgenden Skizze eingezeichnet.



Es gibt aber keinen Rundgang, der jede Brücke genau einmal benutzt.

Algebraische Strukturen

Wir beschäftigen uns nun mit den grundlegenden algebraischen Strukturen: Gruppen, Ringe und Körper. Für diese werden wir jeweils die Grundbegriffe und einige Beispiele besprechen.

6 Gruppen

Definition 6.1. *Eine* **Gruppe** *ist eine Menge G zusammen mit einer Abbildung $p\colon G\times G\to G$ (die wir* **Produkt** *nennen und für die wir die Schreibweise $p(a,b)=a\cdot b=ab$ verwenden), so dass die folgenden Axiome gelten:*

$$\forall\ a,b,c\in G:\quad (a\cdot b)\cdot c=a\cdot(b\cdot c), \tag{AG}$$

$$\exists\ e\in G:\quad \forall\ a\in G:\quad e\cdot a=a, \tag{NE}$$

$$\forall\ a\in G:\quad \exists\ a'\in G:\quad a'\cdot a=e. \tag{IE}$$

(Hierbei ist (IE) eigentlich eine weitere Eigenschaft von e.)

Eine Gruppe G heißt **abelsch** *(oder auch kommutativ), falls außerdem gilt:*

$$\forall\ a,b\in G:\quad a\cdot b=b\cdot a. \tag{KG}$$

Anmerkung. Unsere Ausdrucksweise „eine Menge … zusammen mit einer Abbildung“ ist eigentlich ungenau. Formal befriedigender wäre es, eine Gruppe als ein geordnetes Paar (G,p) zu definieren, wobei G eine Menge und $p\colon G\times G\to G$ eine Abbildung ist, so dass die obigen Axiome gelten. ◁

Bevor wir Beispiele von Gruppen anschauen, beweisen wir das folgende Resultat:

Satz 6.2. *Für jede Gruppe G gelten:*

(a) *Es gibt genau ein $e \in G$, das* (NE) *erfüllt. Dieses e heißt das* **neutrale Element** *von G.*
(b) *Für jedes $a \in G$ gibt es genau ein $a' \in G$, das* (IE) *erfüllt. Dieses a' heißt das* **inverse Element** *zu a und wird mit $a' = a^{-1}$ bezeichnet.*
(c) *Für jedes $a \in G$ gelten*

$$ae = a \quad \text{und} \quad aa^{-1} = e.$$

Beweis. Wir beginnen mit (c). Für $a \in G$ gibt es wegen (IE) $a' \in G$ mit $a'a = e$ und $a'' \in G$ mit $a''a' = e$. Es folgt

$$\begin{aligned} aa' &\underset{(NE)}{=} e(aa') \underset{(IE)}{=} (a''a')(aa') \underset{(AG)}{=} a''(a'(aa')) \\ &\underset{(AG)}{=} a''((a'a)a') \underset{(IE)}{=} a''(ea') \underset{(NE)}{=} a''a' \underset{(IE)}{=} e, \end{aligned} \tag{6.1}$$

und weiter

$$ae \underset{(IE)}{=} a(a'a) \underset{(AG)}{=} (aa')a \underset{(6.1)}{=} ea \underset{(NE)}{=} a. \tag{6.2}$$

Damit ist (c) nachgewiesen. Zum Beweis von (a) sei $\widetilde{e} \in G$ ein weiteres Element, das (NE) erfüllt. Dann folgt

$$\widetilde{e} \underset{(6.2)}{=} \widetilde{e}e \underset{(NE)}{=} e,$$

was die behauptete Eindeutigkeit liefert. Zum Beweis von (b) sei $\widetilde{a} \in G$ ein weiteres Element mit $\widetilde{a}a = e$. Dann folgt

$$\widetilde{a} \underset{(6.2)}{=} \widetilde{a}e \underset{(6.1)}{=} \widetilde{a}(aa') \underset{(AG)}{=} (\widetilde{a}a)a' = ea' \underset{(NE)}{=} a'.$$

Dies schließt den Beweis ab. □

Beispiel 6.3. (1) Die Mengen $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$ zusammen mit der gewöhnlichen Addition als Produkt sind abelsche Gruppen mit 0 als neutralem Element.
(2) Die Mengen $\mathbb{Q}\setminus\{0\}$ und $\mathbb{R}\setminus\{0\}$ zusammen mit dem gewöhnlichen Produkt sind abelsche Gruppen mit 1 als neutralem Element.
(3) Die Menge $\mathbb{Z} \setminus \{0\}$ mit dem gewöhnlichen Produkt ist keine Gruppe, da (IE) verletzt ist. Aber $\{1, -1\} \subseteq \mathbb{Z}$ ist mit dem gewöhnlichen Produkt eine Gruppe.
(4) Auf der Menge

$$G = \{(a_1, a_2) \in \mathbb{R}^2 \mid a_1 \neq 0\}$$

definieren wir ein Produkt durch

$$(a_1, a_2) \cdot (b_1, b_2) = (a_1b_1, a_1b_2 + a_2),$$

wobei wir in den Formeln die gewöhnliche Addition und Multiplikation von $\mathbb{R}$ verwenden. Für den Nachweis von (AG) nehmen wir $(a_1, a_2), (b_1, b_2), (c_1, c_2) \in G$ und bilden

$$\begin{aligned}\Big((a_1, a_2) \cdot (b_1, b_2)\Big) \cdot (c_1, c_2) &= (a_1 b_1, a_1 b_2 + a_2) \cdot (c_1, c_2) \\ &= \Big(a_1 b_1 c_1, a_1 b_1 c_2 + a_1 b_2 + a_2\Big)\end{aligned}$$

und

$$\begin{aligned}(a_1, a_2) \cdot \Big((b_1, b_2) \cdot (c_1, c_2)\Big) &= (a_1, a_2) \cdot (b_1 c_1, b_1 c_2 + b_2) \\ &= \Big(a_1 b_1 c_1, a_1 (b_1 c_2 + b_2) + a_2\Big).\end{aligned}$$

Durch Vergleich erkennt man die Gültigkeit von (AG). Mit $e := (1, 0)$ gilt für alle $(a_1, a_2) \in G$:

$$e \cdot (a_1, a_2) = (a_1, a_2).$$

Außerdem gilt für $(a_1, a_2) \in G$:

$$(a_1^{-1}, -a_1^{-1} a_2) \cdot (a_1, a_2) = (1, 0) = e$$

(wobei a_1^{-1} das reelle Inverse ist). Also ist G eine Gruppe. Ist G abelsch? Das Beispiel $(1, 1) \cdot (2, 1) = (2, 2)$ und $(2, 1) \cdot (1, 1) = (2, 3)$ zeigt, dass dies nicht der Fall ist.

(5) Die Menge $G = \{e\}$ mit $e \cdot e = e$ bildet eine Gruppe, die *triviale Gruppe*.

(6) Die Menge aller Drehungen, die ein Quadrat in sich selbst überführen, ist mit der Komposition eine Gruppe. Sie hat 4 Elemente. Man nennt G die *Symmetriegruppe* des Quadrates. Auch andere geometrische Objekte haben Symmetriegruppen, ebenso Kristalle oder Moleküle. ◁

Für eine Gruppe G gelten die folgenden Rechenregeln:

- $\forall\ a \in G: \quad (a^{-1})^{-1} = a,$
- $\forall\ a, b \in G: \quad (ab)^{-1} = b^{-1} a^{-1}.$

Wir verwenden die folgenden Schreibweisen:

- Statt $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ schreiben wir $a \cdot b \cdot c$, und entsprechend $a \cdot b \cdot c \cdot d$ und so weiter.
- Für $n \in \mathbb{N}_{>0}$: $a^n = \underbrace{a \cdots a}_{n \text{ mal}}$, $a^0 = e$ und $a^{-n} = (a^n)^{-1}$.
- Abelsche Gruppen schreiben wir oft *additiv*: Statt $a \cdot b$ schreiben wir $a + b$. In diesem Fall schreiben wir 0 für das neutrale Element und $-a$ für das inverse Element von $a \in G$.

Das für uns wichtigste Beispiel einer Gruppe ist die symmetrische Gruppe, die wir nun einführen.

Definition 6.4. *Für eine Menge A wird*

$$S_A := \{f\colon A \to A \mid f \text{ ist bijektiv}\}$$

durch $f \cdot g := f \circ g$ (Komposition) eine Gruppe. (Die Gültigkeit von (AG) ist klar, die Identität ist das neutrale Element, und zu $f \in S_A$ ist die Umkehrabbildung das inverse Element.) S_A heißt die **symmetrische Gruppe** *auf A. Die Elemente von S_A heißen* **Permutationen**. *Besonders wichtig ist der Fall $A = \{1, \dots, n\}$ mit $n \in \mathbb{N}$. Hier schreiben wir S_n statt S_A und sprechen von der symmetrischen Gruppe auf n Ziffern.*

Beispiel 6.5. (1) Für $n = 2$ ist

$$S_n = \{\mathrm{id}, \sigma\}$$

mit $\sigma(1) = 2$ und $\sigma(2) = 1$. Es gilt $\sigma^2 = \mathrm{id}$. S_2 ist abelsch.

(2) Die S_3 hat 6 Elemente, denn es gibt $6 = 3!$ bijektive Abbildungen $\{1,2,3\} \to \{1,2,3\}$. Wir benutzen folgende Schreibweise: $(1,2,3)$ steht für die Permutation aus S_3 mit $1 \mapsto 2 \mapsto 3 \mapsto 1$, und $(1,2)$ steht für die Permutation mit $1 \mapsto 2 \mapsto 1$ und $3 \mapsto 3$ (und entsprechend für andere Ziffern). Dann gilt

$$S_3 = \big\{\mathrm{id}, \underbrace{(1,2,3)}_{=:\sigma}, (3,2,1), \underbrace{(1,2)}_{=:\tau}, (1,3), (2,3)\big\}.$$

Es gilt

$$\sigma \cdot \tau = (1,3),$$

aber

$$\tau \cdot \sigma = (2,3).$$

(Man beachte, dass man für die Bildung von $\sigma \cdot \tau$ zuerst τ und dann σ ausführen muss.) S_3 ist also nicht abelsch. ◁

Das obige Beispiel zeigt, dass S_n für $n \geq 3$ nicht abelsch ist. Es gilt allgemein

$$|S_n| = n!,$$

wobei $n! = n(n-1)\cdots 2 \cdot 1$ wie immer für die **Fakultät** von n steht.

Anmerkung 6.6. Wie schon im obigen Beispiel gezeigt, benutzt man für Elemente der symmetrischen Gruppe S_n oft eine Darstellung durch *elementfremde Zykel*, die hier kurz erklärt werden soll. Zunächst ist ein **Zykel** eine Permutation, die gewisse Zahlen $a_1, \dots, a_r \in \{1, \dots, n\}$ zyklisch vertauscht, d.h. a_i wird auf a_{i+1} abgebildet $(1 \leq i \leq r-1)$, a_r wird auf a_1 abgebildet, und alle anderen Zahlen bleiben fest. Man schreibt diese Permutation als $(a_1, \dots, a_r)$. Durch einen Induktionsbeweis kann man einsehen, dass sich jede Permutation $\sigma \in S_n$ schreiben lässt als ein Produkt

$$\sigma = (a_{1,1}, a_{1,2}, \dots, a_{1,r_1})(a_{2,1}, \dots, a_{2,r_2}) \cdots (a_{s,1}, \dots, a_{s,r_s}), \tag{6.3}$$

wobei die $a_{i,j}$ paarweise verschieden sind. Aufgrund dieser Verschiedenheit nennt man die vorkommenden Zykel *elementfremd*. Wegen der Elementfremdheit spielt die Reihenfolge der Zykel in (6.3) keine Rolle.

Beipielsweise hat die Permutation $\sigma \in S_5$ mit $\sigma(1) = 4, \sigma(2) = 5, \sigma(3) = 1, \sigma(4) = 3$ und $\sigma(5) = 2$ die Darstellung $\sigma = (1, 4, 3)(2, 5)$. ◁

Wir behandeln in diesem Abschnitt noch drei wichtige Begriffe aus der Gruppentheorie: Untergruppen, Erzeugung und Homomorphismen.

Definition 6.7. *Eine nicht leere Teilmenge $H \subseteq G$ einer Gruppe heißt* **Untergruppe***, falls für alle $a, b \in H$ auch das Produkt $a \cdot b$ und das Inverse a^{-1} Elemente von H sind. Insbesondere liegt das neutrale Element von G in H, und H ist dann selbst eine Gruppe.*

Beispiel 6.8. (1) Für jede Gruppe G sind $\{e\} \subseteq G$ und $G \subseteq G$ Untergruppen.
(2) In $\mathbb{Z}$ (als Gruppe zusammen mit der Addition) ist $n \cdot \mathbb{Z} := \{nx \mid x \in \mathbb{Z}\}$ für jedes $n \in \mathbb{Z}$ eine Untergruppe.
(3) In $\mathbb{R} \setminus \{0\}$ (zusammen mit dem herkömmlichen Produkt) ist $\{1, -1\}$ eine Untergruppe. Aber $\{1, 2, -1, -2\}$ ist keine Untergruppe.
(4) Die Gruppe G aus Beispiel 6.3(4) hat die Untergruppen

$$H = \{(a, 0) \mid a \in \mathbb{R} \setminus \{0\}\}.$$

und

$$N = \{(1, a) \mid a \in \mathbb{R}\}.$$

(5) In S_3 sind

$$A_3 = \{\mathrm{id}, (1, 2, 3), (3, 2, 1)\}$$

und

$$H = \{\mathrm{id}, (1, 2)\}$$

Untergruppen, und ebenso $H' = \{\mathrm{id}, (1, 3)\}$ und $H'' = \{\mathrm{id}, (2, 3)\}$. ◁

Anmerkung. Es ist leicht zu zeigen, dass der Schnitt zweier Untergruppen einer Gruppe G wieder eine Untergruppe ist. Dies gilt auch für den Schnitt beliebig vieler Untergruppen.

Allerdings ist die Vereinigung von Untergruppen in der Regel keine Untergruppe, wie man etwa anhand der Untergruppe A_3 und H aus Beispiel 6.8(5) sieht. ◁

Definition 6.9. *Es seien G eine Gruppe und $M \subseteq G$ eine Teilmenge. Die von M* **erzeugte Untergruppe** *von G ist die Menge aller Elemente von G, die sich als Produkt $a_1 a_2 \cdots a_k$ beliebiger Länge k schreiben lassen, wobei für jedes i gilt: $a_i \in M$ oder $a_i^{-1} \in M$. Die Faktoren a_i in einem solchen Produkt müssen nicht verschieden sein. Die von M erzeugte Untergruppe ist tatsächlich eine Untergruppe, genauer gesagt die kleinste Untergruppe, die alle Elemente von M enthält.*

Falls die von M erzeugt Untergruppe ganz G ist, so sagen wir, dass G von M erzeugt wird.

Beispiel 6.10. (1) $\mathbb{Z}$ mit der gewöhnlichen Addition wird durch $M = \{1\}$ (man sagt auch: durch das Element 1) erzeugt.
(2) Die Symmetriegruppe des Quadrats (siehe Beispiel 6.3(6)) wird durch eine Drehung um 90^0 erzeugt.
(3) Die von der Permutation $(1, 2, 3)$ erzeugte Untergruppe der S_3 ist die A_3 (siehe Beispiel 6.8(5)).
(4) Die S_3 wird von $\sigma = (1, 2, 3)$ und $\tau = (1, 2)$ erzeugt. Dies kann man leicht nachrechnen. ◁

Anmerkung. Die von einer Teilmenge $M \subseteq G$ erzeugte Untergruppe lässt sich auch als der Schnitt aller Untergruppen $H \subseteq G$ mit $M \subseteq H$ definieren. Es kommt dabei dasselbe heraus wir in Definition 6.9. ◁

Die folgende Proposition gibt ein Erzeugendensystem der symmetrischen Gruppe S_n an. Als eine **Transposition** bezeichnen wir eine Permutation mit Zykeldarstellung von der Form (i, j): Zwei Zahlen werden vertauscht, alle anderen festgelassen. Transpositionen sind ihre eigenen Inversen.

Proposition 6.11. *Die Gruppe S_n wird von Transpositionen erzeugt.*

Beweis. Wir benutzen Induktion nach n. Für $n \leq 1$ ist $|S_n| = 1$, also erzeugt durch die leere Menge. Wir setzen ab jetzt $n \geq 2$ voraus und müssen zeigen, dass jede Permutation $\sigma \in S_n$ ein Produkt von Transpositionen ist. Zunächst betrachten wir den Fall $\sigma(n) = n$. Dann liefert die Einschränkung von σ auf $\{1, \ldots, n-1\}$ ein Element von S_{n-1}, welches nach Induktion ein Produkt von Transpositionen ist. Also ist auch σ ein Produkt von Transpositionen.

Schließlich betrachten wir den Fall $\sigma(n) \neq n$. Wir setzen $k := \sigma(n)$ und bilden

$$\tau := (k, n) \circ \sigma.$$

Es folgt $\tau(n) = n$, also ist τ nach dem obigen Fall ein Produkt von Transpositionen, und $\sigma = (k, n) \circ \tau$ auch. □

Anmerkung. Man kann zeigen, dass die S_n auch von den beiden Permutationen $\sigma = (1, 2, \ldots, n)$ und $\tau = (1, 2)$ erzeugt wird. ◁

Definition 6.12. *Es seien G und H Gruppen. Eine Abbildung $\varphi\colon G \to H$ heißt ein* **Homomorphismus** *(von Gruppen), falls für alle $a, b \in G$ gilt:*

$$\varphi(ab) = \varphi(a)\varphi(b).$$

Für einen Homomorphismus $\varphi\colon G \to H$ heißt

$$\mathrm{Kern}(\varphi) := \{a \in G \mid \varphi(a) = e_H\}$$

der **Kern** *von φ. (Hierbei ist e_H das neutrale Element von H.)*

Beispiel 6.13. (1) Die Exponentialfunktion liefert einen Homomorphismus von $\mathbb{R}$ mit der Addition in $\mathbb{R} \setminus \{0\}$ mit der Multiplikation. Der Kern ist $\{0\}$ und das Bild ist $\mathbb{R}_{>0}$. Auch die Exponentialfunktion von $\mathbb{C}$ liefert

einen Homomorphismus von der additiven Gruppe von $\mathbb{C}$ in $\mathbb{C} \setminus \{0\}$. Der Kern ist $\mathbb{Z} \cdot 2\pi i$.

(2) Die Abbildung $\varphi\colon \mathbb{Z} \to \{1, -1\},\ i \mapsto (-1)^i$ ist ein Homomorphismus von der additiven Gruppe von $\mathbb{Z}$ in die multiplikative Gruppe $\{\pm 1\}$. Der Kern besteht aus allen geraden Zahlen.

(3) Für eine positive natürliche Zahl n ist $\varphi_n\colon \mathbb{Z} \to \mathbb{Z},\ x \mapsto nx$ ein injektiver Homomorphismus.

(4) Es sei G die Gruppe aus Beispiel 6.3(4). Dann ist

$$\varphi\colon G \to \mathbb{R} \setminus \{0\},\ (a_1, a_2) \mapsto a_1$$

ein Homomorphismus in die multiplikative Gruppe von $\mathbb{R}$. Der Kern ist die Untergruppe N aus Beispiel 6.8(4). Allerdings ist

$$\psi\colon G \to \mathbb{R},\ (a_1, a_2) \mapsto a_2$$

kein Homomorphismus in die additive Gruppe.

(5) Sind G und H Gruppen, so ist $\varphi\colon G \to H,\ a \mapsto e_H$ (das neutrale Element von H) ein Homomorphismus.

(6) Sei G eine Gruppe. Die Abbildung $\varphi\colon G \to G,\ a \mapsto a^{-1}$ ist nur dann ein Homomorphismus, wenn G abelsch ist.

(7) Sei G eine Gruppe und $a \in G$. Dann ist

$$\varphi_a\colon G \to G,\ x \mapsto axa^{-1}$$

ein Homomorphismus. ◁

Proposition 6.14. *Es seien G, H Gruppen und $\varphi\colon G \to H$ ein Homomorphismus. Dann gelten:*

(a) $\varphi(e_G) = e_H$ *(mit der offensichtlichen Bezeichnung für die neutralen Elemente der beiden Gruppen).*

(b) Für alle $a \in G$ gilt $\varphi(a^{-1}) = \varphi(a)^{-1}$.

(c) $\mathrm{Bild}(\varphi) \subseteq H$ *ist eine Untergruppe.*

(d) $\mathrm{Kern}(\varphi) \subseteq G$ *ist eine Untergruppe.*

(e) Genau dann ist φ injektiv, wenn $\mathrm{Kern}(\varphi) = \{e_G\}$.

Beweis. (a) Es gilt

$$\varphi(e_G) = \varphi(e_G \cdot e_G) = \varphi(e_G) \cdot \varphi(e_G).$$

Durch Multiplikation mit $\varphi(e_G)^{-1}$ ergibt sich die Behauptung.

(b) Für $a \in G$ gilt:

$$\varphi(a^{-1}) \cdot \varphi(a) = \varphi(a^{-1}a) = \varphi(e_G) \underset{\text{(a)}}{=} e_H.$$

Hieraus folgt die Behauptung.

(c) Es seien $x, y \in \text{Bild}(\varphi)$. Dazu gibt es $a, b \in G$ mit $x = \varphi(a)$ und $y = \varphi(b)$. Also

$$xy = \varphi(a)\varphi(b) = \varphi(ab) \in \text{Bild}(\varphi)$$

und

$$x^{-1} = \varphi(a)^{-1} \underset{\text{(b)}}{=} \varphi(a^{-1}) \in \text{Bild}(\varphi).$$

(d) Wegen (a) gilt $e_A \in \text{Kern}(\varphi)$, also $\text{Kern}(\varphi) \neq \emptyset$. Weiter gilt für $a, b \in \text{Kern}(\varphi)$:

$$\varphi(ab) = \varphi(a)\varphi(b) = e_H e_H = e_H \quad \text{und} \quad \varphi(a^{-1}) \underset{\text{(b)}}{=} e_H^{-1} = e_H,$$

also $ab \in \text{Kern}(\varphi)$ und $a^{-1} \in \text{Kern}(\varphi)$.

(e) Wir nehmen zunächst an, dass φ injektiv sei. Für $a \in \text{Kern}(\varphi)$ gilt dann

$$\varphi(a) = e_h \underset{\text{(a)}}{=} \varphi(e_G) \quad \Longrightarrow \quad a = e_G.$$

Da e_G wegen (a) immer ein Element von $\text{Kern}(\varphi)$ ist, folgt $\text{Kern}(\varphi) = \{e_G\}$.
Wir nehmen nun umgekehrt $\text{Kern}(\varphi) = \{e_G\}$ an. Es seien $a, b \in G$ mit $\varphi(a) = \varphi(b)$. Dann folgt

$$e_H = \varphi(a)\varphi(b)^{-1} \underset{\text{(b)}}{=} \varphi(a)\varphi(b^{-1}) = \varphi(ab^{-1}),$$

also $ab^{-1} \in \text{Kern}(\varphi)$. Nach Voraussetzung folgt $ab^{-1} = e_G$, also $a = b$. Die Injektivität von φ ist damit nachgewiesen. □

Anmerkung. Ist $a \in \text{Kern}(\varphi)$ im Kern eines Homomorphismus $\varphi\colon G \to H$, so gilt für alle $b \in G$:

$$\varphi(bab^{-1}) = \varphi(b)\varphi(a)\varphi(b)^{-1} = \varphi(b)\varphi(b)^{-1} = e_H,$$

also $bab^{-1} \in \text{Kern}(\varphi)$. Man sagt, dass $\text{Kern}(\varphi)$ ein **Normalteiler** von G ist, also eine Untergruppe H, bei der für jedes Element $a \in H$ auch die *konjugierten* Elemente bab^{-1} ($b \in G$) in H liegen. ◁

Ein bijektiver Homomorphismus $G \to H$ zwischen zwei Gruppen heißt auch ein **Isomorphismus**. Zwei Gruppen G und H heißen **isomorph**, falls es einen Isomorphismus $G \to H$ gibt.

Beipielsweise sind die Gruppen S_2 und $\{1, -1\}$ isomorph. Nicht isomorph sind aber die 4-elementigen Untergruppen H_1 und H_2 von S_4 erzeugt durch $(1, 2, 3, 4)$ bzw. durch $(1, 2)$ und $(3, 4)$, denn H_2 kann nicht durch ein einziges Element erzeugt werden. Isomorphe Gruppen haben exakt die selben gruppentheoretischen Eigenschaften.

7 Ringe und Körper

Definition 7.1. *Ein* **Ring** *ist eine Menge R zusammen mit zwei Abbildungen $R \times R \to R$, $(a, b) \mapsto a + b$ („Summe“) und $R \times R \to R$, $(a, b) \mapsto a \cdot b$ („Produkt“), so dass gelten:*

(a) Zusammen mit der Addition ist R eine abelsche Gruppe. (Wir benutzen additive Notation und schreiben 0 für das neutrale Element.)

(b) Für $a, b, c \in R$ gilt

$$(a \cdot b) \cdot c = a \cdot (b \cdot c).$$

(c) Es gibt $1 \in R$, so dass für alle $a \in R$ gilt:

$$1 \cdot a = a \cdot 1 = a.$$

(d) Für alle $a, b, c \in R$ gelten:

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \textit{und} \quad (a + b) \cdot c = a \cdot c + b \cdot c.$$

Ein Ring R heißt **kommutativ**, *falls für alle $a, b \in R$ gilt:*

$$a \cdot b = b \cdot a.$$

Ein kommutativer Ring R heißt ein **Körper**, *falls $0 \neq 1$ und zu jedem $a \in R$ mit $a \neq 0$ ein $a^{-1} \in R \setminus \{0\}$ existiert mit $a^{-1}a = 1$. Dies ist gleichbedeutend damit, dass $R \setminus \{0\}$ mit dem Produkt eine Gruppe bildet.*

Anmerkung. Manchmal wird die Forderung (c) weggelassen und zwischen „Ringen mit Eins“ und „Ringen ohne Eins“ unterschieden. ◁

Bevor wir Beispiele von Ringen anschauen, beweisen wir ein paar wichtige Rechenregeln in Ringen.

Satz 7.2. *Es sei R ein Ring.*

(a) Für alle $a \in R$ gilt:

$$0 \cdot a = a \cdot 0 = 0.$$

(b) Für alle $a, b \in R$ gilt:

$$(-a) \cdot b = a \cdot (-b) = -(a \cdot b).$$

Beweis. (a) Wir haben

$$0 \cdot a = 0 \cdot a + a - a = 0 \cdot a + 1 \cdot a - a = (0+1) \cdot a - a = 1 \cdot a - a = a - a = 0,$$

und ebenso folgt $a \cdot 0 = 0$.

(b) Es gilt

$$(-a) \cdot b = (-a) \cdot b + a \cdot b - (a \cdot b) = (-a+a) \cdot b - (a \cdot b) = 0 \cdot b - (a \cdot b) \underset{\text{(a)}}{=} -(a \cdot b),$$

und ebenso folgt $a \cdot (-b) = -(a \cdot b)$. □

Beispiel 7.3. (1) $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ und $\mathbb{C}$ sind kommutative Ringe. $\mathbb{Q}$, $\mathbb{R}$ und $\mathbb{C}$ sind Körper.

(2) Der kleinste Ring ist $R = \{0\}$ mit $0+0=0$ und $0\cdot 0 = 0$. In diesem Ring gilt $1 = 0$.

(3) Es seien S eine Menge und A ein (kommutativer) Ring. Dann wird

$$R = A^S := \{f\colon S \to A \mid f \text{ ist eine Abbildung}\}$$

mit

$$f \underset{+}{\cdot} g\colon S \to A,\ x \mapsto f(x) \underset{+}{\cdot} g(x)$$

(also *punktweiser* Addition und Multiplikation) ein (kommutativer) Ring. Das Nullelement ist die Nullabbildung $S \to A$, $s \mapsto 0$, und das Einselement ist die Einsabbildung.

(4) Wir versehen $R := \mathbb{R}^3$ mit einer Summe und einem Produkt durch

$$(a_1, a_2, a_3) + (b_1, b_2, b_3) := (a_1 + b_1, a_2 + b_2, a_3 + b_3)$$

und

$$(a_1, a_2, a_3) \cdot (b_1, b_2, b_3) := (a_1 \cdot b_1, a_2 \cdot b_2, a_1 b_3 + a_3 b_2).$$

(Hierbei werden auf den rechten Seiten der Gleichungen die herkömmlichen Operationen von $\mathbb{R}$ benutzt.) Die Bedingungen (a) und (d) aus Definition 7.1 sind unmittelbar klar. Das Assoziativitätsgesetz in (b) bestätigt man durch Nachrechnen. Weiter gilt für $(a_1, a_2, a_3) \in \mathbb{R}^3$:

$$(1, 1, 0) \cdot (a_1, a_2, a_3) = (a_1, a_2, a_3)$$

und

$$(a_1, a_2, a_3) \cdot (1, 1, 0) = (a_1, a_2, a_3),$$

also gilt auch (c). Ist R kommutativ? Die Antwort lautet *nein*, denn

$$(1, 0, 0) \cdot (0, 0, 1) = (0, 0, 1), \quad \text{aber} \quad (0, 0, 1) \cdot (1, 0, 0) = (0, 0, 0).$$

An der letzten Gleichung sieht man, dass das Produkt zweier Ringelemente, die beide ungleich 0 sind, trotzdem 0 sein kann. Dies Phänomen kann auch bei kommutativen Ringen auftreten (siehe Beispiel 7.5(2)). ◁

Wir haben in Beispielen schon verschiedentlich über Teilbarkeit von ganzen Zahlen gesprochen. Dies verallgemeinern wir auf allgemeine kommutative Ringe R, indem wir für $a, b \in R$ sagen, dass a ein **Teiler** von b ist (gleichbedeutend: a teilt b, oder auch: b ist Vielfaches von a), falls es $c \in R$ gibt mit

$$b = ac.$$

Wir benutzen hierfür die Schreibweise $a \mid b$. Man beachte, dass die Teilbarkeit von dem gewählten Ring abhängt. In $R = \mathbb{Q}$ gilt beispielsweise $2 \mid 3$. Der folgende Satz ist zugleich auch eine Definition.

Satz 7.4. *Es seien R ein kommutativer Ring und $a \in R$.*

(a) Durch

$$x \equiv y \mod a \quad :\Longleftrightarrow \quad a \mid (x-y) \quad \textit{für} \quad x, y \in R$$

wird eine Äquivalenzrelation auf R definiert. Falls $x \equiv y \mod a$, so sagen wir, dass x und y **kongruent modulo** *a sind.*

(b) Die Äquivalenzklasse eines $x \in R$ ist

$$[x]_{\equiv} = \{x + ya \mid y \in R\} =: x + Ra$$

und wird auch eine **Restklasse** *modulo a genannt. Die Faktormenge schreiben wir als*

$$R/(a) := R/\equiv = \{x + Ra \mid x \in R\}.$$

(c) Die Faktormenge $R/(a)$ wird ein kommutativer Ring durch folgende Definition der Summe und des Produkts: Für $C_1, C_2 \in R/(a)$ wählen wir $x, y \in R$ mit $x \in C_1$ und $y \in C_2$ und setzen

$$C_1 + C_2 := (x+y) + Ra \quad \textit{und} \quad C_1 \cdot C_2 = xy + Ra.$$

$R/(a)$ heißt der **Restklassenring** *modulo a.*

Beweis. (a) Für alle $x \in R$ ist $x - x = 0 = a \cdot 0$ (wegen Satz 7.2(a)), also gilt die Reflexivität. Zum Nachweis der Symmetrie seien $x, y \in R$ mit $x \sim y \mod a$, also $x - y = ac$ mit $c \in R$. Dann folgt

$$y - x = -(ac) = a(-c)$$

(wegen Satz 7.2(b)), also gilt die Symmetrie. Zum Nachweis der Transitivität seien $x, y, z \in R$ mit $x \sim y \mod a$ und $y \sim z \mod a$, also $x - y = ac$ und $y - z = ad$ mit $c, d \in R$. Dann folgt

$$x - z = (x-y) + (y-z) = ac + ad = a(c+d),$$

also $x \sim z \mod a$. Damit gilt auch die Transitivität.

(b) Für $y \in R$ sind äquivalent:

$$y \in [x]_{\sim} \quad \Longleftrightarrow \quad \exists\, z \in R\colon y - x = za \quad \Longleftrightarrow \quad y \in x + Ra.$$

Dies zeigt die behauptete Gleichheit.

(c) Das Entscheidende ist hier der Nachweis der *Wohldefiniertheit*, also dass $C_1 + C_2$ und $C_1 \cdot C_2$ nicht von der Wahl der Vertreter x, y abhängen. Es

seien also $x' \in C_1$ und $y' \in C_2$ weitere Vertreter. Wir haben also $c, d \in R$ mit $x' - x = ca$ und $y' - y = da$. Es folgt

$$(x' + y') - (x + y) = (c + d) \cdot a, \quad \text{also} \quad (x' + y') + Ra = (x + y) + Ra,$$

und weiter

$$x'y' - xy = x'y' - x'y + x'y - xy = x'da + cay = (x'd + cy) \cdot a,$$

also $x'y' + Ra = xy + Ra$. Damit ist die Wohldefiniertheit gezeigt. Die Ringaxiome vererben sich von R auf $R/(a)$. Exemplarisch rechnen dies anhand des Assoziativgesetzes der Multiplikation nach: Es seien $C_1, C_2, C_3 \in R/(a)$ und $x \in C_1$, $y \in C_2$ und $z \in C_3$. Dann gelten $xy \in C_1 \cdot C_2$ und $yz \in C_2 \cdot C_3$, also

$$(C_1 \cdot C_2) \cdot C_3 = (xy)z + Ra \quad \text{und} \quad C_1 \cdot (C_2 \cdot C_3) = x(yz) + Ra,$$

also $(C_1 \cdot C_2) \cdot C_3 = C_1 \cdot (C_2 \cdot C_3)$. Das Nullelement von $R/(a)$ ist $0 + Ra = Ra$, und das Einselement ist $1 + Ra$. □

Wir beschäftigen uns nun mit dem Ring $R = \mathbb{Z}/(m)$, wobei $m \in \mathbb{N}_{>0}$ eine fest gewählte positive natürliche Zahl ist. Für $x \in \mathbb{Z}$ schreiben wir $\overline{x} = x + \mathbb{Z}m \in \mathbb{Z}/(m)$. Es gilt also

$$\mathbb{Z}/(m) = \left\{\overline{0}, \overline{1}, \ldots, \overline{m-1}\right\}.$$

Beispiel 7.5. (1) Für $m = 3$ werden Summe und Produkt in folgenden Tabellen gegeben:

$+$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{0}$	$\overline{1}$

und

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{1}$

Wir sehen hieran, dass $\mathbb{Z}/(3)$ ein Körper ist. Es gilt $\overline{1} + \overline{1} + \overline{1} = \overline{0}$.

(2) Für $m = 4$ ergibt sich folgende Multiplikationstabelle:

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{0}$	$\overline{2}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

$\mathbb{Z}/(4)$ ist kein Körper, denn $\overline{2}$ ist nicht invertierbar. Es gilt $\overline{2} \cdot \overline{2} = \overline{0}$.

(3) Für $m = 1$ ist $\mathbb{Z}/(m) = \{\overline{0}\}$ der Nullring. ◁

Im Beispiel haben wir beobachtet, dass $\mathbb{Z}/(3)$ ein Körper ist, $\mathbb{Z}/(4)$ aber nicht. Dies sind Instanzen des folgenden Satzes. Wir erinnern daran, dass eine natürliche Zahl $n \in \mathbb{N}$ eine **Primzahl** heißt, falls $n > 1$ und n nur die Teiler 1 und n hat.

Satz 7.6. *Für $m \in \mathbb{N}_{>0}$ ist $\mathbb{Z}/(m)$ genau dann ein Körper, wenn m eine Primzahl ist.*

Beweis. Wir setzen zunächst voraus, dass $\mathbb{Z}/(m)$ ein Körper ist. Aus $\overline{1} \neq \overline{0}$ folgt dann $m > 1$. Es sei $m = xy$ mit $x, y \in \mathbb{N}$ und $y > 1$. Wir müssen $y = m$ zeigen. Wegen $1 \leq x < m$ ist $\overline{x} \neq \overline{0}$, also ist $\overline{x}$ nach Voraussetzung invertierbar. Wir erhalten

$$\overline{y} = \overline{x}^{-1} \cdot \overline{x} \cdot \overline{y} = \overline{x}^{-1} \cdot \overline{m} = \overline{x}^{-1} \cdot \overline{0} = \overline{0}.$$

Es folgt $m \mid y$, also $y = m$.

Nun sei umgekehrt m eine Primzahl. Aus $m > 1$ folgt dann $\overline{1} \neq \overline{0}$. Es sei $\overline{y} \in \mathbb{Z}/(m) \setminus \{\overline{0}\}$. Die Abbildung

$$\varphi\colon \mathbb{Z}/(m) \to \mathbb{Z}/(m),\ \overline{x} \mapsto \overline{x} \cdot \overline{y}$$

ist (wegen des Distributivgesetzes) ein Homomorphismus der additiven Gruppe von $\mathbb{Z}/(m)$. Wir wollen das Kriterium aus Proposition 6.14(e) benutzen, um die Injektivität von φ zu zeigen. Es sei also $\varphi(\overline{x}) = \overline{0}$. Dies bedeutet $m \mid (x \cdot y)$. Weil m eine Primzahl ist und $m \nmid y$, folgt $m \mid x$, also $\overline{x} = \overline{0}$. Nach Proposition 6.14(e) folgt die Injektivität von φ. Als injektive Selbstabbildung einer endlichen Menge ist φ also auch surjektiv (siehe Anmerkung 2.14(b)). Insbesondere existiert $\overline{x} \in \mathbb{Z}/(m)$ mit $\varphi(\overline{x}) = \overline{1}$, also $\overline{x} \cdot \overline{y} = \overline{1}$. Damit ist jedes $\overline{y} \in \mathbb{Z}/(m) \setminus \{\overline{0}\}$ invertierbar, und damit ist $\mathbb{Z}/(m)$ ein Körper. □

Anmerkung 7.7. (a) Im obigen Beweis kam folgender Schluss vor: Falls eine Primzahl ein Produkt ganzer Zahlen teilt, so teilt sie mindestens einen der Faktoren. Für diesen Schluss haben wir stillschweigend den Satz über eindeutige Primzerlegung in $\mathbb{N}$ benutzt. Dieser wir im Abschnitt 18 bewiesen (siehe Satz 18.14).

(b) Ist p eine Primzahl, so schreiben wir standardmäßig $\mathbb{F}_p$ statt $\mathbb{Z}/(p)$.

(c) Die effiziente Berechnung von Inversen in $\mathbb{F}_p$ lässt sich mit Hilfe des *euklidischen Algorithmus* durchführen, den wir hier nicht besprechen.

(d) Zu jeder Primzahlpotenz $q = p^n$ (mit $n \in \mathbb{N}_{>0}$) gibt es einen Körper $\mathbb{F}_q$ mit q Elementen. Es handelt sich dabei *nicht* um $\mathbb{Z}/(q)$, die Konstruktion ist komplizierter. ◁

Definition 7.8. *Es sei R ein Ring. Falls es ein $m \in \mathbb{N}_{>0}$ gibt mit*

$$\underbrace{1 + \cdots + 1}_{m \text{ mal}} = 0,$$

so heißt das kleinste m mit dieser Eigenschaft die **Charakteristik** *von R, geschrieben als* $\mathrm{char}(R)$. *Falls es kein solches m gibt, setzen wir* $\mathrm{char}(R) := 0$.

Beispiel 7.9. (1) $\mathrm{char}(\mathbb{Z}) = \mathrm{char}(\mathbb{Q}) = \mathrm{char}(\mathbb{R}) = \mathrm{char}(\mathbb{C}) = 0$.
(2) $\mathrm{char}\,(\mathbb{Z}/(m)) = m$, $\mathrm{char}(\mathbb{F}_p) = p$. ◁

Anmerkung. Die Charakteristik eines Körpers ist eine Primzahl oder 0. ◁

Im Rest dieses Abschnitts beschäftigen wir uns mit Polynomen. Nach dem naiven Polynombegriff sind Polynome Funktionen von einer bestimmten Form, nämlich

$$f = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0.$$

Wenn wir das Polynom $f = x^2 - x$ als Polynom mit Koeffizienten in $\mathbb{F}_2$ anschauen, sehen wir, dass $f(0) = f(1) = 0$, also müsste f nach diesem Polynombegriff das Nullpolynom sein. Wir möchten aber auch Elemente aus größeren Ringen in Polynome einsetzen können, und dabei können z.B. bei dem obigen Polynom Werte ungleich Null herauskommen. Wir benötigen also einen anderen Polynombegriff. Die Idee ist, dass Polynome durch die Folgen ihrer Koeffizienten $a_0, a_1, \ldots$ gegeben sein sollen. Es ist naheliegend, sie entsprechend als nichts anderes als Koeffizientenfolgen zu definieren.

Definition 7.10. *Es sei R ein kommutativer Ring.*

(a) Ein **Polynom** *über R ist eine Abbildung $f\colon \mathbb{N} \to R,\ i \mapsto a_i$ (d.h. ein R-wertige Folge), bei der höchstens endlich viele der a_i ungleich 0 sind. Die a_i heißen die* **Koeffizienten** *von f.*

(b) Falls bei einem Polynom f mindestens eines der a_i ungleich 0 ist, so heißt das maximale i mit $a_i \neq 0$ der **Grad** *von f, geschrieben als* $\deg(f)$. *Falls alle a_i gleich 0 sind, so setzen wir* $\deg(f) = -\infty$.

(c) Für zwei Polynome $f\colon \mathbb{N} \to R,\ i \mapsto a_i$ und $g\colon \mathbb{N} \to R,\ i \mapsto b_i$ definieren wir

$$f + g\colon \mathbb{N} \to R,\ i \mapsto a_i + b_i$$

und

$$f \cdot g\colon \mathbb{N} \to R,\ i \mapsto \sum_{j=0}^{i} a_j b_{i-j} = \sum_{\substack{j,k \in \mathbb{N} \\ mit\ j+k=i}} a_j \cdot b_k.$$

(d) Mit x bezeichnen wir das spezielle Polynom, bei dem $1 \in \mathbb{N}$ auf $1 \in R$ und alle anderen $i \in \mathbb{N}$ auf $0 \in R$ abgebildet werden. Für $a \in R$ bezeichnen wir das Polynom mit $0 \mapsto a$ und $i \mapsto 0$ für $i > 0$ mit a. (Anders gesagt: Wir fassen die Elemente von R als spezielle Polynome auf.)

(e) Die Menge aller Polynome über R heißt der **Polynomring** *über R und wird mit $R[x]$ bezeichnet.*

Satz 7.11. *Es sei R ein kommutativer Ring.*

(a) Der Polynomring $R[x]$ ist ein kommutativer Ring.

(b) Für ein Polynom $f\colon \mathbb{N} \to R,\ i \mapsto a_i$ *mit* $a_i = 0$ *für* $i > n$ *gilt*

$$f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i \tag{7.1}$$

(mit $x^0 := 1$*).*

Beweis. (a) Es ist klar, dass $R[x]$ mit der Summe aus Definition 7.10(c) eine abelsche Gruppe bildet mit der Nullfolge als Nullelement. Für den Nachweis der weiteren Ringaxiome seien $f\colon \mathbb{N} \to R,\ i \mapsto a_i$, $g\colon \mathbb{N} \to R,\ i \mapsto b_i$ und $h\colon \mathbb{N} \to R,\ i \mapsto c_i$ drei Polynome. Der i-te Koeffizient von $(f \cdot g) \cdot h$ ist

$$\sum_{j=0}^{i} (j\text{-ter Koeffizient von } f \cdot g) \cdot c_{i-j} =$$

$$\sum_{j=0}^{i} \Big(\sum_{k=0}^{j} a_k b_{j-k} \Big) c_{i-j} = \sum_{j=0}^{i} \sum_{k=0}^{j} a_k b_{j-k} c_{i-j} = \sum_{\substack{j,k,l \in \mathbb{N} \\ \text{mit } j+k+l=i}} a_j b_k c_l.$$

Da die entsprechende Rechnung für $f \cdot (g \cdot h)$ zu demselben Ergebnis führt, folgt die Bedingung (b) von Definition 7.1. Man sieht sofort, dass das Kommutativgesetz $f \cdot g = g \cdot f$ gilt. Weiter ergibt sich der i-te Koeffizient von $f \cdot (g + h)$ zu

$$\sum_{j=0}^{i} a_j (b_{i-j} + c_{i-j}) = \sum_{j=0}^{i} a_j b_{i-j} + \sum_{j=0}^{i} a_j c_{i-j},$$

welches auch der i-te Koeffizient von $f \cdot g + f \cdot h$ ist. Zusammen mit dem Kommutativgesetz ergibt dies Definition 7.1(d). Das Polynom mit $0 \mapsto 1$ und $i \mapsto 0$ für $i > 0$ liefert ein Einselement. Insgesamt ist $R[x]$ ein kommutativer Ring.

(b) Wir schreiben

$$\delta_{i,j} := \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{sonst} \end{cases}.$$

Also ist x definiert als die Folge $j \mapsto \delta_{1,j}$. Für $i \in \mathbb{N}$ behaupten wir, dass x^i die Folge $j \mapsto \delta_{i,j}$ ist. Für den Beweis benutzen wir Induktion nach i. Für $i = 0$ ist die Behauptung korrekt. Falls sie für ein i gilt, so ist $x^{i+1} = x \cdot x^i$ die Folge

$$j \mapsto \sum_{k=0}^{j} \delta_{1,k} \delta_{i,j-k} = \delta_{i,j-1} = \delta_{i+1,j},$$

also gilt die Behauptung auch für $i+1$. Für $a \in R$ bezeichnen wir die Folge $j \mapsto a \cdot \delta_{0,j}$ mit a. Also ist $a \cdot x^i$ die Folge

$$j \mapsto \sum_{k=0}^{j} a \cdot \delta_{0,k}\delta_{i,j-k} = a \cdot \delta_{i,j},$$

und für $a_0, \ldots, a_n \in R$ ist $\sum_{i=0}^{n} a_i x^i$ die Folge

$$j \mapsto \sum_{i=0}^{n} a_i \cdot \delta_{i,j} = a_j.$$

Es folgt (7.1). □

Von nun an schreiben wir Polynome nur noch in der Form (7.1).

Die folgende Definition erlaubt es, Elemente eines Rings in Polynome einzusetzen.

Definition 7.12. *Es seien R ein kommutativer Ring, $f = \sum_{i=0}^{n} a_i x^i \in R[x]$ ein Polynom und $c \in R$.*

(a) Das Element

$$f(c) := \sum_{i=0}^{n} a_i c^i \in R$$

heißt die **Auswertung** *von f bei c.*

(b) Falls $f(c) = 0$, so heißt c eine **Nullstelle** *von f.*

(c) Die Abbildung

$$R \to R,\ c \mapsto f(c)$$

heißt die zu f gehörige **Polynomfunktion**.

Anmerkung 7.13. (a) Wir können ein Polynom aus $R[x]$ auch bei Elementen aus einem Ring S, der R umfasst, auswerten. S muss dafür nicht kommutativ sein.

(b) Für $f, g \in R[x]$ und $c \in R$ gelten

$$(f+g)(c) = f(c) + g(c) \quad \text{und} \quad (f \cdot g)(c) = f(c) \cdot g(c).$$

Dies kann man auch ausdrücken, indem man sagt, dass die Abbildung $R[x] \to R$, $f \mapsto f(c)$ ein *Ring-Homomorphismus* ist.

(c) Ist $f \in R[x]$ ein Polynom vom Grad 0 oder $-\infty$, so ist die zugehörige Polynomfunktion konstant. Man nennt f ein *konstantes Polynom* , falls $\deg(f) \leq 0$. ◁

Von nun an beschäftigen wir uns mit Polynomen über Körpern. In diesem Fall kann man Polynome nicht nur addieren und multiplizieren, sondern man hat auch eine *Division mit Rest*, die im folgenden Satz behandelt wird.

Satz 7.14. *Es seien K ein Körper und $f, g \in K[x]$ Polynome mit $g \neq 0$. Dann gibt es $q, r \in K[x]$ mit*

$$f = g \cdot q + r \quad \textit{und} \quad \deg(r) < \deg(g).$$

Beweis. Wir schreiben

$$f = \sum_{i=0}^{n} a_i x^i \quad \text{und} \quad g = \sum_{i=0}^{m} b_i x^i$$

mit $a_i, b_i \in K$, $b_m \neq 0$, und benutzen Induktion nach n. Im Fall $n < m$ stimmt der Satz mit $q = 0$ und $r = f$. Falls $n \geq m$, bilden wir

$$\widetilde{f} := f - b_m^{-1} a_n x^{n-m} \cdot g.$$

Dann gilt $\widetilde{f} = \sum_{i=0}^{n-1} c_i x^i$ mit $c_i \in K$. Nach Induktion gibt es $\widetilde{q}, r \in K[x]$ mit

$$\widetilde{f} = \widetilde{q} \cdot g + r \quad \text{und} \quad \deg(r) < \deg(g).$$

Es folgt

$$f = \widetilde{f} + b_m^{-1} a_n x^{n-m} \cdot g = \underbrace{\left(\widetilde{q} + b_m^{-1} a_n x^{n-m}\right)}_{=:q} \cdot g + r.$$

Dies schließt den Beweis ab. □

Beispiel 7.15. Für $f = x^4$ und $g = x^2 + 1$ ergibt sich

$$x^4 = (x^2 + 1)(x^2 - 1) + 1,$$

also $q = x^2 - 1$ und $r = 1$. ◁

Wir bemerken, dass für zwei Polynome $f, g \in K[x]$ über einem Körper die Formel

$$\deg(f \cdot g) = \deg(f) + \deg(g) \tag{7.2}$$

gilt. (Die Konvention $\deg(0) := -\infty$ war dadurch motiviert, dass diese Gleichung auch für das Nullpolynom gelten sollte.) Die obige Formel kann schiefgehen über Ringen, in denen zwei Elemente ungleich Null trotzdem das Produkt 0 haben können.

Korollar 7.16. *Es sei $f \in K[x] \setminus \{0\}$ ein Polynom über einem Körper K und $c \in K$ eine Nullstelle. Dann gilt*

$$f = (x - c) \cdot g \tag{7.3}$$

mit $g \in K[x]$ und $\deg(g) = \deg(f) - 1$.

Beweis. Division mit Rest liefert

$$f = (x - c) \cdot g + r$$

mit $g, r \in K[x]$, $\deg(r) < \deg(x-c) = 1$. Also ist r konstant. Einsetzen von c liefert

$$0 = f(c) = (c-c) \cdot g(c) + r(c) = r.$$

Hieraus folgt (7.3). Die Aussage über den Grad von g folgt aus (7.2). □

Korollar 7.17. *Es sei $f \in K[x] \setminus \{0\}$ ein Polynom über einem Körper. Dann hat f höchstens $\deg(f)$ Nullstellen (in K).*

Beweis. Wir führen den Beweis durch Induktion nach $n := \deg(f)$. Im Falle $n = 0$ ist f konstant und ungleich Null, also gibt es keine Nullstellen.

Im Weiteren sei $n > 0$ und $c \in K$ eine Nullstelle von f. Nach Korollar 7.16 gilt $f = (x-c) \cdot g$ mit $g \in K[x]$ und $\deg(g) = n-1$. Für jede weitere Nullstelle $b \in K$ von f gilt

$$0 = f(b) = (b-c)g(b).$$

Falls $b \neq c$, liefert Multiplikation mit $(b-c)^{-1}$, dass $g(b) = 0$ sein muss. Nach Induktion hat aber g höchstens $n-1$ Nullstellen, und es folgt die Behauptung. □

Beispiel 7.18. (1) Wir betrachten $f = x^4 - 1 \in \mathbb{R}[x]$. Wegen $f(1) = 0$ ist f durch $x - 1$ teilbar:

$$x^4 - 1 = (x-1) \underbrace{(x^3 + x^2 + x + 1)}_{=:g}.$$

Für g finden wir die Nullstelle -1, und es gilt

$$g = (x+1)(x^2+1),$$

also

$$f = (x-1)(x+1)(x^2+1).$$

Das Polynom $x^2 + 1$ hat keine Nullstelle (in $\mathbb{R}$).

(2) Um zu sehen, dass die Voraussetzung in Korollar 7.17, dass K ein Körper ist, nicht weggelassen werden kann, betrachten wir den Ring $R = \mathbb{Z}/(8)$ und das Polynom $f = x^2 - 1 \in R[x]$. Wir finden die Nullstellen $\overline{1}, \overline{3}, \overline{5}$ und $\overline{7}$ von f, also mehr, als der Grad angibt. ◁

Ist $f \in K[x] \setminus \{0\}$ ein Polynom über einem Körper und c eine Nullstelle, so gilt $f = (x-c) \cdot g$ mit $g \in K[x]$ (Korollar 7.16). Man nennt den Faktor $x - c$ auch einen *Linearfaktor*. Nun kann es passieren, dass c auch eine Nullstelle von g ist. In diesem Fall folgt $f = (x-c)^2 h$ mit $h \in K[x]$, und man kann fortfahren, bis das verbleibende Polynom c nicht mehr als Nullstelle hat. Der höchste Exponent e, so dass $(x-c)^e$ ein Teiler von f ist, heißt die **Vielfachheit** der Nullstelle c von f. Insbesondere spricht man von *einfachen* ($e = 1$) und *mehrfachen* ($e > 1$) Nullstellen.

Nachdem man alle Linearfaktoren $(x-c)$ zur Nullstelle c von f abgespalten hat, kann man weitere Nullstellen des verbleibenden Polynoms suchen und

die entsprechenden Linearfaktoren abspalten. Falls dieser Prozess mit einem konstanten Polynom endet, also

$$f = a \cdot \prod_{i=1}^{n} (x - c_i)$$

mit $a, c_i \in K$, $a \neq 0$ (wobei die c_i nicht unbedingt verschieden sein müssen), so sagen wir, dass f (über K) *in Linearfaktoren zerfällt.*

Beispiel 7.19. Wir setzen $K = \mathbb{R}$.

(1) Das Polynom

$$f = x^5 - 2x^3 + x = x(x^2 - 1)^2 = x(x - 1)^2(x + 1)^2$$

zerfällt in Linearfaktoren. Es hat die Nullstellen ± 1 mit der Vielfachheit 2 und 0 als einfache Nullstelle.

(2) Das Polynom $x^4 - 1$ aus Beispiel 7.18(1) zerfällt nicht in Linearfaktoren. ◁

Definition 7.20. *Ein Körper K heißt* **algebraisch abgeschlossen***, falls jedes nicht-konstante Polynom $f \in K[x]$ eine Nullstelle in K hat.*

Ist K algebraisch abgeschlossen, so zerfällt jedes nicht-konstante Polynom $f \in K[x]$ in Linearfaktoren.

$\mathbb{R}$ ist nicht algebraisch abgeschlossen, z.B. fehlt dem Polynom $x^2 + 1$ eine Nullstelle in $\mathbb{R}$. Das wichtigste Beispiel für einen algebraisch abgeschlossenen Körper ist $\mathbb{C}$:

Satz 7.21 (Fundamentalsatz der Algebra)**.** *Der Körper $\mathbb{C}$ der komplexen Zahlen ist algebraisch abgeschlossen.*

Wir können den Beweis hier nicht führen, da er Methoden aus der Funktionentheorie (oder der Algebra) benötigt.

Aus der obigen Betrachtung folgt, dass jedes Polynom über einem algebraisch abgeschlossenen Körper in Linearfaktoren zerfällt.

Lineare Algebra: Vektorräume

In diesem Kapitel kommen wir zu den Kernthemen der linearen Algebra: den Vektorräumen, ihren Abbildungen und den Matrizen.

8 Vektorräume und Unterräume

In diesem Abschnitt steht K immer für einen Körper. Man verliert nichts Wesentliches, wenn man sich $K = \mathbb{R}$ oder $K = \mathbb{C}$ vorstellt.

Definition 8.1. *Ein K-***Vektorraum** *(auch: Vektorraum über K) ist eine Menge V zusammen mit zwei Abbildungen $\boxplus$: $V \times V \to V$, $(v, w) \mapsto v \boxplus w$ und $\boxdot$: $K \times V \to V$, $(a, v) \mapsto a \boxdot v$, so dass folgende Axiome gelten:*

(1) V ist mit $\boxplus$ als Verknüpfung eine abelsche Gruppe. Man verwendet additive Schreibweise.

(2) Für alle $a \in K$ und $v, w \in V$ gilt

$$a \boxdot (v \boxplus w) = a \boxdot v \boxplus a \boxdot w$$

(mit der Konvention Punkt vor Strich, also $a \boxdot v \boxplus a \boxdot w = (a \boxdot v) \boxplus (a \boxdot w)$).

(3) Für alle $a, b \in K$ und $v \in V$ gilt

$$(a + b) \boxdot v = a \boxdot v \boxplus b \boxdot v.$$

(4) Für alle $a, b \in K$ und $v \in V$ gilt

$$(a \cdot b) \boxdot v = a \boxdot (b \boxdot v).$$

(5) Für alle $v \in V$ gilt

$$1 \boxdot v = v.$$

Die Elemente eines Vektorraums heißen **Vektoren**. *Die Elemente von K werden (in diesem Zusammenhang) oft* **Skalare** *genannt. Wir haben die Symbole „$\boxplus$“ und „$\boxdot$“ für die Unterscheidung von der Addition und Multiplikation im Körper K verwendet. Ab jetzt werden wir immer $v+w$ für $v \boxplus w$ und $a \cdot v$ oder av für $a \boxdot v$ schreiben.*

Wir hätten einen Vektorraum auch formaler als ein Tripel $(V, \boxplus, \boxdot)$ definieren können. Wir verwenden jedoch den etwas laxeren Sprachgebrauch „eine Menge ... zusammen mit Abbildungen ...“.

Beispiel 8.2. (1) Es sei $n \in \mathbb{N}_{>0}$ fest und

$$K^n = \underbrace{K \times \cdots \times K}_{n \text{ mal}}$$

das n-fache kartesische Produkt. K^n wird zu einem K-Vektorraum durch

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n) \quad \text{für} \quad x_i, y_i \in K$$

und

$$a \cdot (x_1, \ldots, x_n) := (ax_1, \ldots, ax_n) \quad \text{für} \quad a, x_i \in K.$$

Dies sieht man sofort durch Nachprüfen von Definition 8.1. Der Nullvektor ist $(0, \ldots, 0)$. Man nennt K^n auch den den **n-dimensionalen Standardraum**.

(2) $V = \{0\}$ (abelsche Gruppe mit nur einem Element 0) wird mit $a \cdot 0 := 0$ für $a \in K$ ein K-Vektorraum. Dieser Vektorraum heißt der **Nullraum**.

(3) K selbst ist ein K-Vektorraum (mit der Addition und Multiplikation von K).

(4) $\mathbb{C}$ ist ein $\mathbb{R}$-Vektorraum; $\mathbb{R}$ ist ein $\mathbb{Q}$-Vektorraum.

(5) Der Polynomring $K[x]$ ist ein K-Vektorraum (mit der üblichen Polynomaddition und dem üblichen Produkt einer Konstanten aus K und eines Polynoms).

(6) Für (festes) $d \in \mathbb{N}$ ist $\{f \in K[x] \mid \deg(f) < d\}$ ein K-Vektorraum.

(7) S sei irgendeine Menge und

$$V := K^S = \{f \colon S \to K \mid f \text{ Abbildung}\}.$$

Für $f, g \in V$ und $a \in K$ definieren wir $f + g$ und $a \cdot f \in V$ durch

$$f + g \colon S \to K,\ x \mapsto f(x) + g(x) \quad \text{und} \quad a \cdot f \colon S \to K,\ x \mapsto a \cdot f(x).$$

(Man sagt auch, dass die Summe von Funktionen und das skalare Vielfache einer Funktion *punktweise* definiert werden.) Durch stures Nachrechen sieht man, dass V ein K-Vektorraum ist. Der Nullvektor ist die sogenannte *Nullfunktion* f_0, definiert durch $f_0(x) = 0$ für alle $x \in S$.

(8) Gegenbeispiel: Es sei V eine abelsche Gruppe mit neutralem Element 0, aber $V \neq \{0\}$. Wir setzen $a \cdot v := 0$ für alle $a \in K$ und $v \in V$. Dann

sind die Axiome (1) bis (4) in Definition 8.1 erfüllt, aber (5) nicht. Der mögliche Verdacht, dass (5) überflüssig sein könnte, erweist sich also als unbegründet. ◁

Anmerkung 8.3. Man kann in Definition 8.1 auch K durch einen Ring R ersetzen. Dadurch wird der Begriff eines R-**Moduls** definiert. Man könnte sagen, dass ein Modul dasselbe ist wie ein Vektorraum, nur über einem Ring statt über einem Körper.

Beispielsweise wird jede (additiv geschriebene) abelsche Gruppe G ein $\mathbb{Z}$-Modul, indem wir für $n \in \mathbb{N}$ und $x \in G$

$$n \cdot x := \underbrace{x + \cdots + x}_{n \text{ mal}} \quad \text{und} \quad (-n) \cdot x := -(n \cdot x)$$

setzen. ◁

Aus den Vektorraumaxiomen ergeben sich ein paar Rechenregeln:

Proposition 8.4. *Es seien V ein K-Vektorraum und $a \in K$, $v \in V$. Dann gelten:*

(a) $a \cdot 0 = 0$ *und* $0 \cdot v = 0$ *(in der ersten Gleichung bezeichnet die linke 0 den Nullvektor, in der zweiten das Nullelement von K);*
(b) $(-a) \cdot v = a \cdot (-v) = -(a \cdot v)$*;*
(c) aus $a \cdot v = 0$ *folgt* $a = 0$ *oder* $v = 0$*.*

Beweis. Wir verwenden nur die Vektorraum- (und Körper-)Axiome.

(a) Es gelten

$$a \cdot 0 \underset{(1)}{=} a \cdot 0 + a \cdot 0 - (a \cdot 0) \underset{(2)}{=} a \cdot (0 + 0) - (a \cdot 0) \underset{(1)}{=} a \cdot 0 - (a \cdot 0) \underset{(1)}{=} 0$$

und

$$0 \cdot v \underset{(1)}{=} 0 \cdot v + 0 \cdot v - (0 \cdot v) \underset{(3)}{=} (0 + 0) \cdot v - (0 \cdot v) = 0 \cdot v - (0 \cdot v) \underset{(1)}{=} 0.$$

(b) Es gelten

$$(-a)v \underset{(1)}{=} (-a)v + av - (av) \underset{(3)}{=} (-a + a)v - (av) = 0v - (av) \underset{(a)}{=} -(av)$$

und

$$a(-v) \underset{(1)}{=} a(-v) + av - (av) \underset{(2)}{=} a(-v + v) - (av) \underset{(1)}{=} a0 - (av) \underset{(a)}{=} -(av).$$

(c) Es sei $a \cdot v = 0$ aber $a \neq 0$. Dann folgt

$$v \underset{(5)}{=} 1 \cdot v = (a^{-1}a) \cdot v \underset{(4)}{=} a^{-1} \cdot (av) = a^{-1} \cdot 0 \underset{(a)}{=} 0.$$

□

Definition 8.5. *Sei V ein K-Vektorraum. Eine Teilmenge $U \subseteq V$ heißt ein* **Unterraum** *(auch: Untervektorraum, Teilraum), falls gelten:*

(1) $U \neq \emptyset$;
(2) Für $v, w \in U$ ist auch $v + w \in U$ (also ist U mit $+$ eine Untergruppe);
(3) Für $a \in K$ und $v \in U$ gilt $a \cdot v \in U$.

Aus der Definition folgt sofort:

- Jeder Unterraum enthält den Nullvektor.
- Mit den Operationen „+“ und „·“ von V wird ein Unterraum U selbst ein K-Vektorraum.
- Für den Nachweis, dass eine nicht-leere Teilmenge $U \subseteq V$ ein Unterraum ist, genügt es zu zeigen, dass für $v, w \in U$ und $a \in K$ auch $av + w$ in U liegt.

Beispiel 8.6. (1) $V = \mathbb{R}^2$. Jede Gerade durch den Nullpunkt ist ein Unterraum. Formaler: Wähle $v \in V$. Dann ist $K \cdot v := \{a \cdot v \mid a \in K\} \subseteq V$ ein Unterraum. Dies gilt sogar für jeden Vektorraum V und $v \in V$. Geraden im $\mathbb{R}^2$, die nicht durch den Nullpunkt gehen, sind keine Unterräume.
(2) $U = \{0\}$ und V selbst sind Unterräume eines Vektorraums V.
(3) Sei $V = K[x]$ der Polynomring und $d \in \mathbb{N}$ fest. Dann ist

$$U = \{f \in V \mid \deg(f) < d\} \subseteq V$$

ein Unterraum (siehe Beispiel 8.2(5) und (6)).
(4) Sei S eine Menge und $V = K^S$ (siehe Beispiel 8.2(7)). Wähle $x \in S$ fest. Dann ist

$$U := \{f \in V \mid f(x) = 0\} \subseteq V$$

ein Unterraum. (Die Bedingung $f(x) = 1$ würde nicht zu einem Unterraum führen!)
(5) Die Menge aller stetigen (diffenrenzierbaren) Funktionen $\mathbb{R} \to \mathbb{R}$ bildet einen Unterraum von $\mathbb{R}^{\mathbb{R}}$.
(6) Die Vereinigungsmenge zweier Geraden $U_1, U_2 \subseteq \mathbb{R}^2$ durch den Nullpunkt ist kein Unterraum (es sei denn $U_1 = U_2$). ◁

Das letzte Beispiel zeigt, dass Vereinigungen von Unterräumen im Allgemeinen keine Unterräume sind. Die folgende Proposition beschäftigt sich mit Schnitten von Unterräumen.

Proposition 8.7. *Es seien V ein K-Vektorraum und $U_1, U_2 \subseteq V$ Unterräume. Dann gelten:*

(a) $U_1 \cap U_2 \subseteq V$ ist ein Unterraum.
(b) $U_1 + U_2 := \{v + w \mid v \in U_1,\ w \in U_2\} \subseteq V$ ist ein Unterraum.
(c) Ist $\mathcal{M} \neq \emptyset$ eine nicht-leere Menge, deren Elemente Unterräume von V sind, so ist auch der Schnitt

$$\bigcap \mathcal{M} = \bigcap_{U \in \mathcal{M}} U \subseteq V$$

ein Unterraum.

Beweis. Wir müssen nur (b) und (c) zeigen, da (a) ein Spezialfall von (c) ist.

(b) Es gilt $U_1 + U_2 \neq \emptyset$. Seien $v + w$ und $v' + w'$ Elemente von $U_1 + U_2$ mit $v, v' \in U_1$, $w, w' \in U_2$. Dann folgt

$$(v + w) + (v' + w') = (v + v') + (w + w') \in U_1 + U_2,$$

und für $a \in K$ folgt $a \cdot (v + w) = av + aw \in U_1 + U_2$. Also ist $U_1 + U_2$ ein Unterraum.

(c) Wir schreiben $W := \bigcap_{U \in \mathcal{M}} U$. Für alle $U \in \mathcal{M}$ gilt $0 \in U$, also $0 \in W$. Weiter gilt für $v, w \in W$, dass v und w in allen $U \in \mathcal{M}$ liegen. Damit auch $v + w \in U$ für alle $U \in \mathcal{M}$, also $v + w \in W$. Ebenso folgt $a \cdot v \in W$ für $a \in K$ und $v \in W$. damit ist gezeigt, dass W ein Unterraum ist. □

Der Unterraum $U_1 + U_2$ aus Proposition 8.7(b) heißt der **Summenraum** von U_1 und U_2. Man kann auch aus mehr als zwei Unterräumen den Summenraum bilden. Proposition 8.7(c) drückt man manchmal aus, indem man sagt, dass die Menge der Unterräume eines Vektorraums ein *durchschnittsabgeschlossenes System* bilden. Proposition 8.7(c) macht die folgende Definition möglich.

Definition 8.8. *Es seien V ein K-Vektorraum und $S \subseteq V$ eine Teilmenge. (Wir setzen nicht voraus, dass S ein Unterraum ist.) Wir betrachten die Menge $\mathcal{M} := \{U \subseteq V \mid U \text{ ist ein Unterraum und } S \subseteq U\}$ und bilden*

$$\langle S \rangle := \bigcap_{U \in \mathcal{M}} U. \tag{8.1}$$

$\langle S \rangle$ heißt der von S **erzeugte Unterraum** *(auch: aufgespannter Unterraum, Erzeugnis) von V. Falls $S = \{v_1, \ldots, v_n\}$ endlich ist, schreiben wir $\langle S \rangle$ auch als*

$$\langle v_1, \ldots, v_n \rangle.$$

Man sieht sofort, dass $\langle S \rangle$ der kleinste Unterraum von V ist, der S (als Teilmenge) enthält. Genauer: Jeder Unterraum von V, der S enthält, enthält auch $\langle S \rangle$.

Die obige Definition ist konzeptionell elegant. Sie wirft jedoch die Frage auf, wie sich der von S erzeugte Unterraum explizit beschreiben lässt. Dieser Frage wenden wir uns jetzt und zu Beginn des folgenden Abschnitts zu.

Beispiel 8.9. (1) Sei $v \in V$ ein Vektor. Wie sieht $\langle v \rangle$ aus? Die Antwort lautet: $\langle v \rangle = K \cdot v = \{a \cdot v \mid a \in K\}$. Denn $K \cdot v$ ist ein Unterraum, der v enthält, und andererseits ist $K \cdot v$ in jedem Unterraum U mit $v \in U$ enthalten.

(2) Noch einfacher ist der Fall $S = \emptyset$: $\langle \emptyset \rangle = \{0\}$, der Nullraum. ◁

Wir betrachten nun den Fall, dass S die Vereinigung zweier Unterräume ist.

Satz 8.10. *Es seien V ein K-Vektorraum, U_1 und U_2 Unterräume und $S := U_1 \cup U_2$. Dann gilt*

$$\langle S \rangle = U_1 + U_2.$$

Beweis. Nach Proposition 8.7(b) ist $U_1 + U_2$ ein Unterraum. Außerdem liegt jedes $v \in U_1$ (als $v+0$) und jedes $w \in U_2$ (als $0+w$) in U_1+U_2. U_1+U_2 ist also einer der Räume U, die in (8.1) zum Schnitt kommen, also $\langle S \rangle \subseteq U_1 + U_2$.

Umgekehrt sei $U \subseteq V$ ein Unterraum mit $S \subseteq U$. Für $v \in U_1$ und $w \in U_2$ folgt dann $v + w \in U$, also $U_1 + U_2 \subseteq U$. Wegen (8.1) impliziert dies $U_1 + U_2 \subseteq \langle S \rangle$. □

Beispiel 8.11. Es seien $U_1, U_2 \subseteq \mathbb{R}^3$ zwei verschiedene Geraden durch den Nullpunkt. Dann ist $U_1 + U_2$ eine Ebene. ◁

Um eine allgemeingültige Antwort auf die Frage nach einer expliziten Beschreibung des erzeugten Unterraums $\langle S \rangle$ einer Teilmenge $S \subseteq V$ zu geben, benötigen wir eine Definition.

Definition 8.12. *(a) Es seien $v_1, \ldots, v_n \in V$ Vektoren. Ein Vektor $v \in V$ heißt* **Linearkombination** *von $v_1, \ldots, v_n$, falls es Skalare $a_1, \ldots, a_n \in K$ gibt mit*

$$v = a_1 v_1 + \cdots + a_n v_n.$$

(b) Es sei $S \subseteq V$ eine Teilmenge. Ein Vektor $v \in V$ heißt **Linearkombination** *von S, falls es $n \in \mathbb{N}$ und $v_1, \ldots, v_n \in S$ gibt, so dass v eine Linearkombination von $v_1, \ldots, v_n$ ist. Falls $S = \emptyset$, so sagen wir, dass der Nullvektor 0 (die einzige) Linearkombination von S ist. (0 wird als leere Summe aufgefasst.)*

Es ist klar, dass die Teile (a) und (b) der Definition für endliche Mengen $S = \{v_1, \ldots, v_n\}$ übereinstimmen. In (b) geht man über endliche Auswahlen von Vektoren, da es in der linearen Algebra nur endliche Summen gibt (ebenso wie in der Analysis, in der man Grenzwerte von endlichen Teilsummen betrachtet).

Nun beantworten wir die Frage nach dem erzeugten Unterraum.

Satz 8.13. *Für eine Teilmenge $S \subseteq V$ ist der erzeugte Unterraum $\langle S \rangle$ die Menge aller Linearkombinationen von S:*

$$\langle S \rangle = \{ v \in V \mid v \text{ ist Linearkombination von } S \}.$$

Insbesondere gilt für $v_1, \ldots, v_n \in V$:

$$\langle v_1, \ldots, v_n \rangle = \left\{ \sum_{i=1}^{n} a_i v_i \mid a_1, \ldots, a_n \in K \right\}.$$

Beweis. Es sei $W \subseteq V$ die Menge aller Linearkombinationen von S. Es gilt $0 \in W$. Da die Summe zweier Linearkombinationen und ein skalares Vielfaches einer Linearkombination wieder Linearkombinationen sind, folgt, dass W ein Unterraum ist. Außerdem liegt jedes $v \in S$ in W. Damit ist W einer der Unterräume U, die in (8.1) zum Schnitt kommen. Es folgt $\langle S\rangle \subseteq W$.

Andererseits sei $U \subseteq V$ ein Unterraum mit $S \subseteq U$. Für $v_1, \ldots, v_n \in S$ und $a_1, \ldots, a_n \in K$ liegen dann alle v_i in U und damit auch $\sum_{i=1}^n a_i v_i$. Also enthält U alle Linearkombinationen von S, d.h. $W \subseteq U$. Dies impliziert $W \subseteq \langle S\rangle$, und der Beweis ist abgeschlossen. □

Beispiel 8.14. (1) Die Vektoren $v = (1,-1),\ w = (0,1) \in \mathbb{R}^2$ haben die Linearkombination

$$1 \cdot (1,-1) + 3 \cdot (0,1) = (1,2).$$

Die Menge aller Linearkombinationen ist

$$\langle v, w\rangle = \{a \cdot (1,-1) + b \cdot (0,1) = (a, -a+b) \mid a, b \in \mathbb{R}\} = \mathbb{R}^2.$$

(2) Die Vektoren $v = (1,-1),\ w = (-1,1) \in \mathbb{R}^2$ haben die Linearkombination

$$1 \cdot v + 3 \cdot w = (-2,2) = -2 \cdot v.$$

Die Menge aller Linearkombinationen ist

$$\langle v, w\rangle = \{a \cdot v + b \cdot w = (a-b, -a+b) \mid a, b \in \mathbb{R}\} = \langle v\rangle = \langle w\rangle \subsetneqq \mathbb{R}^2.$$

(3) Mit

$$e_1 := (1,0,0),\ e_2 := (0,1,0),\ e_3 := (0,0,1) \in \mathbb{R}^3$$

gilt

$$\mathbb{R}^3 = \langle e_1, e_2, e_3\rangle.$$

Es ist klar, dass sich dies von $\mathbb{R}^3$ auf K^n verallgemeinern lässt.

(4) Es seien $V = \mathbb{R}^{\mathbb{R}}$ und $f, g \in V$ mit $f(x) = \sin(x)$ und $g(x) = \cos(x)$. Es sei $h \in \langle f, g\rangle$, also $h(x) = a\sin(x) + b\cos(x)$ mit $a, b \in \mathbb{R}$. Es gibt ein $x_0 \in \mathbb{R}$ mit

$$a = \sqrt{a^2+b^2} \cdot \cos(x_0) \quad \text{und} \quad b = \sqrt{a^2+b^2} \cdot \sin(x_0).$$

Es folgt

$$h(x) = \sqrt{a^2+b^2}\,(\cos(x_0)\sin(x) + \sin(x_0)\cos(x)) = \sqrt{a^2+b^2} \cdot \sin(x_0+x),$$

also sind alle Linearkombinationen von f und g „phasenverschobene“ Sinus-Funktionen verschiedener „Amplitude“.

(5) Es seien $V = K[x]$ der Polynomring über einem Körper und

$$S = \{x^i \mid i \in \mathbb{N}\} = \{1, x, x^2, \ldots\}.$$

Dann gilt

$$V = \langle S \rangle,$$

denn jedes Polynom ist eine Linearkombination von Potenzen x^i. Die Exponentialfunktion $\sum_{i=0}^{\infty} \frac{1}{i!} x^i$ liegt jedoch nicht in $\langle S \rangle$, da nur endliche Summen enthalten sind. ◁

9 Lineare Gleichungssysteme und Matrizen

Auch in diesem Abschnitt steht K immer für einen Körper. Wir entwickeln Rechentechniken, die bei fast allen rechnerischen Problemen der linearen Algebra zum Einsatz kommen.

Wir untersuchen Gleichungssysteme von der Art

$$\begin{array}{rcrcrcr} x_1 & & + & 2x_3 & + & x_4 & = & -3 \\ 2x_1 & & + & 4x_3 & - & 2x_4 & = & 2 \\ & x_2 & & & - & x_4 & = & 2 \\ x_1 & & + & 2x_3 & + & 2x_4 & = & -5 \end{array} . \tag{9.1}$$

Solche Gleichungssysteme nennt man **lineare Gleichungssysteme** (kurz: LGS). Wir verfolgen dabei folgende Idee: Das Addieren eines Vielfachen einer Gleichung zu einer andern ändert die Lösungsmenge nicht, es kann aber das Gleichungssystem vereinfachen. Wenn wir beispielsweise in (9.1) die erste Gleichung von der vierten subtrahieren, ergibt sich $x_4 = -2$. Um die Handhabung zu vereinfachen, werden wir lineare Gleichungssysteme in sogenannte Matrizen zusammenfassen. Zunächst definieren wir, was wir unter einer Matrix verstehen wollen.

Definition 9.1. *Es seien $m, n \in \mathbb{N}_{>0}$ positive natürliche Zahlen. Eine $m \times n$-***Matrix** *ist eine „rechteckige Anordnung“*

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

mit $a_{i,j} \in K$. Formaler definieren wir eine $m \times n$-Matrix als eine Abbildung $\{1, \ldots, m\} \times \{1, \ldots, n\} \to K$, wobei das Bild von (i, j) mit $a_{i,j}$ bezeichnet wird.

Das Element $a_{i,j}$ einer Matrix A heißt der (i, j)-te **Eintrag** *von A. Wir benutzen verschiedene Schreibweisen für Matrizen:*

$$A = (a_{i,j})_{\substack{i=1,\dots,m\\ j=1,\dots,n}} = (a_{i,j})_{\substack{1\le i\le m\\ 1\le j\le n}} = (a_{i,j})_{i,j} = (a_{i,j}),$$

wobei die beiden letzten benutzt werden, wenn m und n aus dem Kontext klar sind. Durch die Definition einer Matrix ergibt sich automatisch der Gleichheitsbegriff von Matrizen: Zwei $m \times n$-Matrizen $A = (a_{i,j})$ und $B = (b_{i,j})$ sind gleich, falls $a_{i,j} = b_{i,j}$ für alle i und j gilt.

Die Menge aller $m \times n$-Matrizen wird mit $K^{m\times n}$ bezeichnet.

Eine $1 \times n$-Matrix $(a_1, \dots, a_n) \in K^{1\times n}$ wird als **Zeilenvektor***, eine $n \times 1$-Matrix $\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in K^{n\times 1}$ als* **Spaltenvektor** *bezeichnet. Elemente des n-dimensionalen Standardraums werden wir meist als Spaltenvektoren schreiben. Es wird sich bald zeigen, warum dies praktisch ist.*

Für $A = (a_{i,j}) \in K^{m\times n}$ und $i \in \{1, \dots, m\}$ ist $(a_{i,1}, \dots, a_{i,n}) \in K^{1\times n}$ die **i-te Zeile** *von A. Für $j \in \{1, \dots, n\}$ ist $\begin{pmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{pmatrix} \in K^{m\times 1}$ die* **j-te Spalte** *von A.*

Eine Matrix $A \in K^{m\times n}$ mit $m = n$ heißt **quadratisch***. Für $A = (a_{i,j}) \in K^{m\times n}$ ist $A^T := (a_{j,i}) \in K^{n\times m}$ die* **transponierte Matrix***; also z.B.*

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

Eine quadratische Matrix heißt **symmetrisch***, falls $A^T = A$ gilt.*

Zu einem linearen Gleichungssystem mit m Gleichungen und n Unbekannten $x_1, \dots, x_n$ bilden wir nun die **Koeffizientenmatrix**, indem wir den Koeffizienten von x_j in der i-ten Gleichung als (i, j)-ten Eintrag nehmen. Dies ergibt eine $m \times n$-Matrix. Das Gleichungssystem heißt **homogen**, falls auf der rechten Seite der Gleichungen lauter Nullen stehen, und andernfalls **inhomogen**. Falls das lineare Gleichungssystem inhomogen ist, erweitert man die Koeffizientenmatrix, indem man eine Spalte mit den rechten Seiten der Gleichungen anhängt. Die so gebildete $m \times (n + 1)$-Matrix nennt man die **erweiterte Koeffizientenmatrix**. Sie kodiert die gesamte Information des LGS. Beispielsweise gehört zu dem System (9.1) die erweiterte Koeffizientenmatrix

$$\left(\begin{array}{cccc|c} 1 & 0 & 2 & 1 & -3 \\ 2 & 0 & 4 & -2 & 2 \\ 0 & 1 & 0 & -1 & 2 \\ 1 & 0 & 2 & 2 & -5 \end{array}\right).$$

Die Trennlinie vor der letzten Spalte hat keine mathematische Bedeutung, sie dient nur als Gedächtnisstütze.

Unser Ziel ist es, einen Algorithmus zur Bestimmung der **Lösungsmenge** (also die Menge aller $x \in K^n$, für die alle Gleichungen eines LGS gelten) zu entwickeln. Hierfür definieren wir zunächst einige Manipulationen, die auf Matrizen allgemein und im besonderen auf die erweiterte Koeffizientenmatrix eines LGS angewandt werden können. Diese Manipulationen heißen **elementare Zeilenoperationen** und gliedern sich in drei Typen:

Typ I: Vertauschen zweier Zeilen;
Typ II: Multiplizieren einer Zeile mit einem Skalar $a \in K \setminus \{0\}$;
Typ III: Addieren des a-fachen einer Zeile zu einer anderen, wobei $a \in K$.

Es ist unmittelbar klar, dass das Anwenden von elementaren Zeilenoperationen auf die erweiterte Koeffizientenmatrix eines LGS die Lösungsmenge unverändert lässt. Wir können ein LGS also mit diesen Operationen manipulieren mit dem Ziel, es auf eine so einfache Gestalt zu bringen, dass man die Lösungsmenge direkt ablesen kann. Die angestrebte Gestalt ist die *Zeilenstufenform* gemäß der folgenden Definition.

Definition 9.2. *Es sei $A \in K^{m\times n}$. Wir sagen, dass A in* **Zeilenstufenform** *ist, falls gelten:*

(a) Beginnt eine Zeile mit k Nullen, so stehen unter diesen Nullen lauter weitere Nullen.
(b) Unter dem ersten Eintrag $\neq 0$ einer jeden Zeile (falls diese nicht nur aus Nullen besteht) stehen lauter Nullen.

Wir sagen, dass A in **strenger Zeilenstufenform** *ist, falls zusätzlich gilt:*

(c) Über dem ersten Eintrag $\neq 0$ einer jeden Zeile (falls diese nicht nur aus Nullen besteht) stehen lauter Nullen.

Beispiel 9.3. Zur Illustration mögen folgende Beispiele dienen:

(1) Die Matrix $\begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ ist *nicht* in Zeilenstufenform.

(2) Die Matrix $\begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ ist *nicht* in Zeilenstufenform.

(3) Die Matrix $\begin{pmatrix} 1 & 2 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$ ist in Zeilenstufenform, aber nicht in strenger Zeilenstufenform.

(4) Die Matrix $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$ ist in strenger Zeilenstufenform. ◁

Beispiel 9.4. Wir wenden elementare Zeilenoperationen auf die erweiterte Koeffizientenmatrix des LGS (9.1) an mit dem Ziel, die Matrix auf strenge Zeilenstufenform zu bringen.

$$\left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 2&0&4&-2&2\\ 0&1&0&-1&2\\ 1&0&2&2&-5 \end{array}\right) \begin{array}{l}\\ -2\\ \\ -1\end{array} \xrightarrow[\text{Typ III}]{} \left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&0&0&-4&8\\ 0&1&0&-1&2\\ 0&0&0&1&-2 \end{array}\right) \xrightarrow[\text{Typ I}]{} \left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&1&0&-1&2\\ 0&0&0&-4&8\\ 0&0&0&1&-2 \end{array}\right)$$

$$\left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&1&0&-1&2\\ 0&0&0&-4&8\\ 0&0&0&1&-2 \end{array}\right) \cdot\left(-\tfrac{1}{4}\right) \xrightarrow[\text{Typ II}]{} \left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&1&0&-1&2\\ 0&0&0&1&-2\\ 0&0&0&1&-2 \end{array}\right) \xrightarrow[\text{Typ III}]{} \left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&1&0&-1&2\\ 0&0&0&1&-2\\ 0&0&0&0&0 \end{array}\right) \begin{array}{l}\\ \\ \\ -1\end{array}$$

$$\left(\begin{array}{cccc|c} 1&0&2&1&-3\\ 0&1&0&-1&2\\ 0&0&0&1&-2\\ 0&0&0&0&0 \end{array}\right) \begin{array}{l}-1\\ \\ 1\\ \\ \end{array} \xrightarrow[\text{Typ III}]{} \left(\begin{array}{cccc|c} 1&0&2&0&-1\\ 0&1&0&0&0\\ 0&0&0&1&-2\\ 0&0&0&0&0 \end{array}\right)$$

Hierbei haben wir jeweils gekennzeichnet, wie wir von einer Matrix zur nächsten gekommen sind. Dies ist sehr zu empfehlen, damit die Rechnung nachvollziehbar und Fehler korrigierbar sind. ◁

Nun können wir das Verfahren formalisieren. Wir erhalten den berühmten Gauß-Algorithmus.

Algorithmus 9.5 (Gauß)**.**

Eingabe: Eine Matrix $A \in K^{m\times n}$.
Ausgabe: Eine Matrix $B \in K^{m\times n}$ in (strenger) Zeilenstufenform, die aus A durch elementare Zeilenoperationen hervorgeht.

(1) Setze $B := A$.
(2) B sei bis zur r-ten Zeile in Zeilenstufenform, d.h. (a) und (b) aus Definition 9.2 seien bis zur r-ten Zeile erfüllt. (Hierbei ist $r = 0$ möglich!)
(3) Falls $r = m$, so ist B in Zeilenstufenform. Falls strenge Zeilenstufenform gewünscht ist, gehe zu (8).
(4) Suche den am weitesten links stehenden Eintrag $\neq 0$ von B unterhalb der r-ten Zeile. (Falls es mehrere solche Einträge gibt, wähle einen aus.)
(5) Bringe diesen Eintrag in die $(r+1)$-te Zeile (Operation Typ I).
(6) Erzeuge unterhalb dieses Eintrags lauter Nullen (Operationen Typ III, optional auch II).
(7) Gehe zu (2).
(8) Bringe B auf strenge Zeilenstufenform (Operationen Typ III).

Der Gaußalgorithmus ist das „rechnerische Herz“ der linearen Algebra. Wir werden noch sehen, dass er für viele rechnerische Aufgaben eingesetzt wird. Wir haben ihn im Zusammenhang mit linearen Gleichungssystemen eingeführt. Da wir bereits gesehen haben, dass sich bei elementaren Zeilenoperationen die Lösungsmenge nicht ändert, müssen wir uns nur noch überzeugen, dass wir anhand einer (strengen) Zeilenstufenform des Systems die Lösungsmenge besonders leicht ablesen können.

Beispiel 9.6. Wir setzen das Beispiel des in (9.1) gegebenen LGS fort. In Beispiel 9.4 wurde die erweiterte Koeffizientenmatrix auf strenge Zeilenstufenform gebracht, wodurch wir das äquivalente LGS mit Matrix

$$\left(\begin{array}{cccc|c} 1 & 0 & 2 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

erhalten. In ausführlicher Schreibweise liest sich dies als

$$\begin{aligned} x_1 + 2x_3 &= -1, \\ x_2 &= 0, \\ x_4 &= -2. \end{aligned}$$

Die Lösungsmenge lässt sich ablesen:

$$L = \left\{ \begin{pmatrix} -2x_3 - 1 \\ 0 \\ x_3 \\ -2 \end{pmatrix} \;\middle|\; x_3 \in K \text{ beliebig} \right\}.$$

Man kann den Parameter x_3 hierbei natürlich durch einen anderen Buchstaben ersetzen. ◁

Jetzt geben wir unser Lösungsverfahren für LGS in formalerer Weise an.

Algorithmus 9.7 (Lösen von LGS)**.**

Eingabe: Ein LGS mit der erweiterten Koeffizientenmatrix $(A|b)$ mit $A \in K^{m\times n}$ und $b \in K^m$ (also m Gleichungen mit n Unbekannten).
Ausgabe: Die Lösungsmenge L.

(1) Bringe die erweiterte Koeffizientenmatrix $(A|b) \in K^{m\times(n+1)}$ auf strenge Zeilenstufenform. Ab jetzt setzen wir voraus, dass $(A|b)$ bereits in strenger Zeilenstufenform ist.
(2) Es sei r die Anzahl der Zeilen, die mindestens einen Eintag $\neq 0$ haben. Für $i = 1, \ldots, r$ sei $j_i \in \{1, \ldots, n+1\}$ die Position (= Spalte), in der der erste Eintrag $\neq 0$ der i-ten Zeile steht.
(3) Falls $j_r = n+1$, so ist das LGS unlösbar, also $L = \emptyset$. (Die r-te Zeile lautet dann nämlich $(0 \cdots 0|b_r)$ mit $b_r \neq 0$, was der Gleichung $0\cdot x_1 + \cdots + 0\cdot x_n = b_r$ entspricht.)
(4) Andernfalls seien $k_1, \ldots, k_{n-r}$ diejenigen Zahlen in $\{1, \ldots, n\}$, die nicht eines der j_i sind. Also $\{1, \ldots, n\} \setminus \{j_1, \ldots, j_r\} = \{k_1, \ldots, k_{n-r}\}$.
(5) Die Lösungsmenge ist

$$L = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \;\middle|\; x_{k_1}, \ldots, x_{k_{n-r}} \in K \text{ beliebig}, \right.$$
$$\left. x_{j_i} = a_{i,j_i}^{-1} \cdot \left(b_i - \sum_{j=1}^{n-r} a_{i,k_j} \cdot x_{k_j} \right) \text{ für } i = 1, \ldots, r \right\}. \quad (9.2)$$

(Diese Formel ergibt sich durch Auflösen der i-ten Gleichung nach x_{j_i}.) Die Lösungsmenge wird also parametrisiert durch die „freien“ Variablen x_{k_i}, während die x_{j_i} von diesen abhängig sind.

Es ist fast unmöglich, sich die Formel (9.2) zu merken, und noch unmöglicher, sie tatsächlich anzuwenden, es sei denn, man ist ein Computer und kein Mensch. Man ist also weiterhin darauf angewiesen, die Lösungsmenge eines LGS anhand der strengen Zeilenstufenform mit Hilfe von mathematisch-handwerklichen Grundfertigkeiten abzulesen.

Bei LGS können drei „Hauptfälle“ für die Lösungsmenge L eintreten:

(1) Unlösbarkeit: $L = \emptyset \;\Leftrightarrow\; j_r = n + 1$.

(2) Eindeutige Lösbarkeit: $|L| = 1 \;\Leftrightarrow\; r = n$ und $j_r = n$. In diesem Fall gilt automatisch $j_i = i$ für alle i, und die strenge Zeilenstufenform hat die übersichtliche Gestalt

$$\left(\begin{array}{ccccc|c} a_{1,1} & 0 & & \cdots & 0 & b_1 \\ 0 & a_{2,2} & & & \vdots & \vdots \\ & & \ddots & & \vdots & \vdots \\ \vdots & & & a_{n-1,n-1} & 0 & b_{n-1} \\ 0 & \cdots & & 0 & a_{n,n} & b_n \\ 0 & \cdots & & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ 0 & \cdots & & \cdots & 0 & 0 \end{array}\right).$$

Die (einzige) Lösung ergibt sich dann als $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1/a_{1,1} \\ \vdots \\ b_n/a_{n,n} \end{pmatrix}$.

(3) Uneindeutige Lösbarkeit: $|L| > 1 \;\Leftrightarrow\; r < n$ und $j_r \neq n+1$. Dann hat die Lösungsmenge $n - r$ freie Parameter. Insbesondere folgt $|L| = \infty$, falls K unendlich viele Elemente hat (der Standardfall).

Allein aus der Anzahl der Gleichungen und der Unbekannten kann man nicht auf den Eintritt einer der Hauptfälle schließen. Als Einziges lässt sich sagen, dass eindeutige Lösbarkeit nur dann eintreten kann, wenn mindestens so viele Gleichungen wie Unbekannte vorhanden sind.

Die Zahl r aus Algorithmus 9.7 spielt eine wichtige Rolle. Daher geben wir ihr einen Namen.

Definition 9.8. *Es sei $A \in K^{m\times n}$, und $A' \in K^{m\times n}$ sei eine Matrix in Zeilenstufenform, die durch elementare Zeilenoperationen aus A hervorgegangen ist. Dann ist der* **Rang** *von A die Anzahl r der Zeilen in A', die mindestens einen Eintrag $\neq 0$ haben. Wir schreiben $r =: \mathrm{rg}(A)$.*

Eine quadratische Matrix $A \in K^{n\times n}$ heißt **regulär**, *falls $rg(A) = n$.*

Das Problem bei dieser Definition ist, dass es verschiedene Matrizen A' gibt, die in Zeilenstufenform sind und die durch elementare Zeilenoperationen aus A hervorgegangen sind. Es ist (bisher) nicht klar, dass all diese A' dieselbe Anzahl von Zeilen $\neq 0$ haben. Nur wenn dies klar ist, ist $\mathrm{rg}(A)$ eindeutig definiert. Wir werden dies in Abschnitt 10 nachtragen.

Wir sehen sofort, dass für $A \in K^{m\times n}$ die Ungleichung $\mathrm{rg}(A) \leq \min\{m, n\}$ gilt. Unser Lösbarkeitskriterium für LGS können wir nun so formulieren:

Satz 9.9. *Ein LGS mit erweiterter Koeffizientenmatrix $(A|b)$ ist genau dann lösbar, wenn A denselben Rang hat wie $(A|b)$.*

In diesem Zusammenhang ist das folgende Resultat interessant:

Proposition 9.10. *Es seien $A, A' \in K^{m\times n}$, wobei A' durch elementare Zeilenoperationen aus A hervorgegangen ist. Dann erzeugen die Zeilen von A denselben Unterraum von $K^{1\times n}$ wie die Zeilen von A'.*

Beweis. Wir müssen zeigen, dass elementare Zeilenoperationen den von den Zeilen $v_1, \ldots, v_m$ erzeugten Raum U nicht ändern.

Typ I: Offenbar ändert sich U nicht.
Typ II: ebenso.
Typ III: Nach Umnummerieren der Zeilen ersetzt die Operation v_1 durch $v_1 + cv_2$, $c \in K$. Die neuen Zeilen erzeugen

$$\langle v_1 + sv_2, v_2, \ldots, v_m\rangle = \left\{ a_1(v_1 + cv_2) + \sum_{i=2}^{m} a_i v_i \mid a_i \in K \right\} = U,$$

also auch hier keine Änderung. □

Zum Schluss des Abschnitts sei erwähnt, dass die Lösungsmengen von homogenen LGS mit n Unbekannten immer Unterräume des K^n sind.

10 Lineare Unabhängigkeit und Basen

In diesem Abschnitt führen wir einige zentrale Begriffe der linearen Algebra ein. Wie zuvor bezeichnet K immer einen Körper und V einen Vektorraum.

Bei Beispiel 8.14(1),(3),(4) und (5) fällt auf, dass jeder Vektor aus dem erzeugten Unterraum *eindeutig* als Linearkombination darstellbar ist, d.h. es gibt nur eine Wahl für die Koeffizienten a_i. Beim Beispiel 8.14(2) ist dies nicht der Fall. Diese Beobachtung gibt Anlass zu folgender Definition.

Definition 10.1. *(a) Vektoren $v_1, \dots, v_n \in V$ heißen* **linear unabhängig**, *falls für alle $a_1, \dots, a_n$ folgende Implikation gilt:*

$$a_1 v_1 + \cdots + a_n v_n = 0 \quad \Rightarrow \quad a_1 = 0,\ a_2 = 0, \dots, a_n = 0.$$

*Gleichbedeutend damit ist: Für jede Linearkombination $v \in \langle v_1, \dots, v_n \rangle$ gibt es eindeutig bestimmte $a_1, \dots, a_n \in K$ mit $v = \sum_{i=1}^n a_i v_i$ („*eindeutige Darstellungseigenschaft*"). Der Beweis, dass lineare Unabhängigkeit und die eindeutige Darstellungseigenschaft gleichbedeutend sind, sei dem Leser überlassen. Die Vektoren $v_1, \dots, v_n$ heißen* **linear abhängig**, *falls sie nicht linear unabhängig sind. Wir betonen, dass es sich hierbei nicht um Eigenschaften von einzelnen Vektoren handelt (außer im Fall $n = 1$), sondern um Eigenschaften eines „Ensembles" von Vektoren.*

(b) Eine Teilmenge $S \subseteq V$ heißt **linear unabhängig**, *falls für alle $n \in \mathbb{N}$ und alle paarweise verschiedenen $v_1, \dots, v_n \in S$ gilt, dass $v_1, \dots, v_n$ linear unabhängig ist. Andernfalls heißt S* **linear abhängig**. *$S = \emptyset$ ist (per definitionem) linear unabhängig.*

Beispiel 10.2. (1) Seien $V = \mathbb{R}^2$, $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ und $v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Wir testen auf lineare Unabhängigkeit. Es gelte also $a_1 v_1 + a_2 v_2 = 0$ mit $a_1, a_2 \in \mathbb{R}$. Hieraus ergibt sich das homogene LGS $a_1 + a_2 = 0$, $a_1 - a_2 = 0$. Die einzige Lösung ist $a_1 = a_2 = 0$, also sind v_1, v_2 linear unabhängig.

(2) Nun betrachten wir $v_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ und $v_2 = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix} \in \mathbb{R}^3$. Wenn wir wie oben auf lineare Unabhängigkeit testen, erhalten wir das homogene LGS $a_1 + 2a_2 = 0$, $-a_1 - 2a_2 = 0$, $0 = 0$, das (unter anderen) die nicht-triviale Lösung $a_1 = 2$, $a_2 = -1$ hat. Es folgt $2v_1 - v_2 = 0$, also sind v_1, v_2 linear abhängig.

(3) Es seien $V = K[x]$ und $S = \{x^i \mid i \in \mathbb{N}\}$. Wir behaupten, dass S linear unabhängig ist. Zum Nachweis nehmen wir beliebige, paarweise verschiedene $x^{i_1}, \dots, x^{i_n} \in S$ und setzen $\sum_{j=1}^n a_j x^{i_j} = 0$ mit $a_j \in K$ voraus. Hieraus folgt (mit dem üblichen Identitätsbegriff für Polynome) direkt, dass $a_j = 0$ für alle j. Also ist S linear unabhängig.

(4) Der Fall $n = 1$: Ein einzelner Vektor $v \in V$ ist genau dann linear unabhängig, wenn $v \neq 0$. Dies folgt aus Proposition 8.4(c). ◁

Für Vektoren $v_1, \dots, v_n \in K^m$ haben wir folgenden Test auf lineare Unabhängigkeit: Man bilde die Matrix $A := (v_1 | v_2 | \cdots | v_n) \in K^{m \times n}$ mit den v_i als Spalten. (Die senkrechten Linien sollen nur der Verdeutlichung dienen.) Dann gilt:

$$\boxed{v_1, \dots, v_n \text{ sind linear unabhängig} \quad \Longleftrightarrow \quad \operatorname{rg}(A) = n.}$$

Begründung: Die v_i sind genau dann linear unabhängig, wenn das homogene LGS mit Koeffizientenmatrix A als einzige Lösung den Nullvektor hat (siehe auch Beispiel 10.2(1) und (2)). Nach (2) auf Seite 79 und Definition 9.8 trifft dies genau dann ein, wenn $\operatorname{rg}(A) = n$.

Wegen $\operatorname{rg}(A) \leq \min\{m, n\}$ (siehe nach Definition 9.8) folgt aus unserem Test sofort, dass im K^m höchstens m Vektoren linear unabhängig sein können. Hat man mehr als m Vektoren, so sind diese automatisch linear abhängig.

Definition 10.3. *Es sei $S \subseteq V$ eine Teilmenge.*

(a) S heißt ein **Erzeugendensystem** *von V, falls $\langle S \rangle = V$.*
(b) S heißt eine **Basis** *von V, falls S ein linear unabhängiges Erzeugendensystem von V ist. Anders gesagt: S ist Basis, falls jedes $v \in V$ in eindeutiger Weise als Linearkombination von S darstellbar ist.*

Beispiel 10.4. (1) Die Vektoren

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \; e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

bilden eine Basis von K^3.

(2) Auch die Vektoren

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \; v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad v_3 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}$$

bilden eine Basis von K^3. Wir sehen also, dass ein Vektorraum mehrere Basen haben kann. (In der Tat haben „fast alle“ Vektorräume „sehr viele“ verschiedene Basen.)

(3) In Verallgemeinerung von (1) sei

$$(i\text{-te Position}) \; \to \; \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} =: e_i \in K^n.$$

Dann ist $S = \{e_1, \ldots, e_n\}$ eine Basis von K^n. S heißt die **Standardbasis** des K^n.

(4) Für $V = K[x]$ ist $S = \{x^i \mid i \in \mathbb{N}\}$ eine Basis. Dies geht aus Beispiel 8.14(5) und aus Beispiel 10.2(3) hervor. Wir haben es hier mit einer unendlichen Basis zu tun.

(5) Der Nullraum $V = \{0\}$ hat die leere Menge $S = \emptyset$ als Basis. Dies ist einer der exotischen Fälle, in denen es nur eine Basis gibt.
(6) Wir betrachten das homogene LGS mit der Koeffizientenmatrix

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 2 & 0 & 4 & -2 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 2 & 2 \end{pmatrix}.$$

Wir können A in Zeilenstufenform B bringen, indem wir uns an Beispiel 9.4 orientieren, und erhalten

$$B = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Hieraus lesen wir die Lösungsmenge

$$L = \left\{ \begin{pmatrix} -2a \\ 0 \\ a \\ 0 \end{pmatrix} \mid a \in \mathbb{R} \right\} = \left\langle \begin{pmatrix} -2 \\ 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle$$

ab. (Wir könnten auch das formale Lösungsverfahren 9.7 benutzen.) Der angegebene erzeugende Vektor bildet eine einelementige Basis von L. ◁

Allgemein sei ein homogenes LGS mit der Koeffizientenmatrix $A \in K^{m\times n}$ gegeben. Es seien $k_1, \ldots, k_{n-r} \in \{1, \ldots, n\}$ die im Lösungsverfahren 9.7(4) bestimmten Indizes. Für $j = 1, \ldots, n-r$ sei v_j der durch (9.2) gewonnene Lösungsvektor mit $x_{k_j} = 1$ und $x_{k_i} = 0$ für $i \neq j$. In v_j ist die j_i-te Komponente also $-a_{i,j_i}^{-1} \cdot a_{i,k_j}$ $(i = 1, \ldots, r)$. Dann ist $\{v_1, \ldots, v_{n-r}\}$ eine Basis des Lösungsraums L. Die Erzeugereigenschaft ergibt sich direkt aus (9.2), und diese Gleichung zeigt außerdem, dass die k_j-te Koordinate von $\sum_{i=1}^{n-r} a_i v_i$ (mit $a_i \in K$) genau a_j ist, woraus die lineare Unabhängigkeit folgt. Wir haben also ein Verfahren, um für den Lösungsraum eines homogenen LGS eine Basis zu finden.

Wir geben nun zwei (zur Definition alternative) Charakterisierungen von Basen an.

Satz 10.5. *Für eine Teilmenge $S \subseteq V$ sind äquivalent:*

(a) S ist eine Basis von V.
(b) S ist eine maximal linear unabhängige Teilmenge von V (d.h. S ist linear unabhängig, aber für jedes $v \in V \setminus S$ wird $S \cup \{v\}$ linear abhängig).
(c) S ist ein minimales Erzeugendensystem von V (d.h. $V = \langle S \rangle$, aber für alle $v \in S$ ist $S \setminus \{v\}$ kein Erzeugendensystem).

Beweis. Wir beginnen mit der Implikation „(a) ⇒ (b)“. Sei also S eine Basis von V. Dann ist S linear unabhängig, es ist also nur die Maximalität zu zeigen. Hierzu sei $v \in V \setminus S$. Da S ein Erzeugendensystem ist, gibt es $v_1, \dots, v_n \in S$ und $a_1, \dots, a_n \in K$ mit

$$v = \sum_{i=1}^{n} a_i v_i,$$

also

$$(-1) \cdot v + \sum_{i=1}^{n} a_i v_i = 0.$$

Hierbei können wir die v_i als paarweise verschieden annehmen. Dies zeigt, dass $\{v, v_1, \dots, v_n\}$ linear abhängig ist, also auch $S \cup \{v\}$.

Nun zeigen wir „(b) ⇒ (c)“. Es sei also S maximal linear unabhängig. Wir zeigen zunächst, dass S ein Erzeugendensystem ist. Hierzu sei $v \in V$. Falls $v \in S$, so gilt auch $v \in \langle S \rangle$, und wir sind fertig. Wir dürfen also $v \notin S$ annehmen. Nach Voraussetzung ist $S \cup \{v\}$ linear abhängig, also gibt es paarweise verschiedene $v_1, \dots, v_n \in S$ und $a, a_1, \dots, a_n \in K$, die nicht alle 0 sind, so dass

$$av + \sum_{i=1}^{n} a_i v_i = 0.$$

(Selbst falls v in einer solchen Darstellung des Nullvektors nicht vorkäme, könnten wir es „künstlich“ durch $a := 0$ hinzufügen.) Falls $a = 0$, so wären $v_1, \dots, v_n$ linear abhängig, im Widerspruch zur linearen Unabhängigkeit von S. Es folgt $a \neq 0$, also

$$v = -\sum_{i=1}^{n} a^{-1} a_i v_i \in \langle S \rangle.$$

Nun ist noch die Minimalität von S als Erzeugendensystem zu zeigen. Hierzu sei $v \in S$. Falls $S \setminus \{v\}$ ein Erzeugendensystem wäre, dann gäbe es insbesondere $v_1, \dots, v_n \in S \setminus \{v\}$ und $a_1, \dots, a_n \in K$ mit

$$v = \sum_{i=1}^{n} a_i v_i.$$

Hierbei können wir die v_i als paarweise verschieden annehmen. Es folgt $(-1) \cdot v + \sum_{i=1}^{n} a_i v_i = 0$, im Widerspruch zur linearen Unabhängigkeit von S. Also ist S tatsächlich ein minimales Erzeugendensystem.

Schließlich zeigen wir „(c) ⇒ (a)“. Es sei also S ein minimales Erzeugendensystem. Wir müssen die lineare Unabhängigkeit von S zeigen. Es seien also $v_1, \dots, v_n \in S$ paarweise verschieden und $a_1, \dots, a_n \in K$ mit $\sum_{i=1}^{n} a_i v_i = 0$. Wir nehmen an, dass nicht alle a_i Null sind. Durch Umnummerieren können

wir $a_1 \neq 0$ erreichen. Es folgt

$$v_1 = \sum_{i=2}^{n} -a_1^{-1} a_i v_i \in \langle S' \rangle$$

mit $S' := S \setminus \{v_1\}$. Alle Elemente von S liegen also in $\langle S' \rangle$, also $V = \langle S' \rangle$, im Widerspruch zur Minimalität von S. Somit ist S linear unabhängig. □

Die Frage, ob jeder Vektorraum eine Basis hat, wird durch den folgenden Satz mit „ja“ beantwortet, den wir mit Hilfe des Zornschen Lemmas beweisen werden.

Satz 10.6 (Basisergänzungssatz)**.** *Es seien $S \subseteq V$ ein Erzeugendensystem (z.B. $S = V$) und $A \subseteq S$ eine linear unabhängige Teilmenge (z.B. $A = \emptyset$). Dann gibt es eine Basis B von V mit $A \subseteq B \subseteq S$.*

Beweis. Wir betrachten die Menge

$$M := \{X \subseteq V \mid X \text{ ist linear unabhängig und } A \subseteq X \subseteq S\}.$$

Die Menge M ist geordnet durch $X \leq Y \iff X \subseteq Y$. Wir prüfen die Voraussetzung des Zornschen Lemmas (Satz 3.12). Es sei $C \subseteq M$ also eine Kette. Falls $C = \emptyset$, so liefert $A \in M$ eine obere Schranke von C. Andernfalls setzen wir

$$Y := \bigcup C = \bigcup_{X \in C} X$$

und behaupten $Y \in M$. (Hieraus folgt, dass Y eine obere Schranke von C ist.) Es ist klar, dass $A \subseteq Y \subseteq S$ gilt. Zum Nachweis der linearen Unabhängigkeit von Y nehmen wir paarweise verschiedene $v_1, \ldots, v_n \in Y$. Für jedes i gibt es ein $X_i \in C$ mit $v_i \in X_i$. Da C totalgeordnet ist, gibt es ein X_i, das alle anderen umfasst. Damit sind $v_1, \ldots, v_n$ Elemente von diesem X_i. Wegen der linearen Unabhängigkeit von X_i folgt, dass $v_1, \ldots, v_n$ linear unabhängig ist. Also ist Y linear unabhängig und damit ein Element von M.

Das Zornsche Lemma liefert nun die Existenz eines maximalen Elements $B \in M$. Es folgt sofort, dass B linear unabhängig ist und $A \subseteq B \subseteq S$. Zum Nachweis der Erzeugereigenschaft von B nehmen wir zunächst einen Vektor $v \in S$. Falls $v \in B$, so folgt $v \in \langle B \rangle$. Andernfalls gilt

$$A \subseteq B \subsetneq B \cup \{v\} \subseteq S.$$

Wegen der Maximalität von B muss $B \cup \{v\}$ also linear abhängig sein, d.h. es gibt paarweise verschiedene $v_1, \ldots, v_n \in B$ und $a, a_1, \ldots, a_n \in K$, die nicht alle 0 sind, so dass

$$av + \sum_{i=1}^{n} a_i v_i = 0.$$

Wegen der linearen Unabhängigkeit von B folgt $a \neq 0$, also

$$v = -a^{-1} \sum_{i=1}^{n} a_i v_i \in \langle S \rangle.$$

Es ergibt sich $S \subseteq \langle B \rangle$, also

$$V = \langle S \rangle \subseteq \langle B \rangle \subseteq V.$$

Damit ist B ein linear unabhängiges Erzeugendensystem von V, und der Satz ist bewiesen. □

Durch Anwendung von Satz 10.6 auf $S = V$ und $B = \emptyset$ ergibt sich:

Korollar 10.7 (Basissatz)**.** *Jeder Vektorraum hat eine Basis.*

Anmerkung. Man kann die Begriffe Linearkombination, Erzeugendensystem und lineare Unabhängigkeit auch auf Moduln anwenden und somit den Basissatz für Moduln formulieren. Er ist jedoch für Moduln im Allgemeinen *falsch*. Beispielsweise hat keine nicht-triviale, endliche abelsche Gruppe als $\mathbb{Z}$-Modul (siehe Anmerkung 8.3) eine Basis. ◁

Beispiel 10.8. Es sei M eine unendliche Menge und $V = K^M$. Für V ist keine Basis bekannt, auch wenn Satz 10.6 die Existenz garantiert! Auch in Spezialfällen oder für viele interessante Unterräume ist keine Basis bekannt. Beispielsweise ist keine Basis für den Vektorraum der konvergenten reellen Folgen bekannt.

Für jedes $x \in M$ kann man die Abbildung $\delta_x \in V$ mit $\delta_x(y) = 1$ für $y = x$, 0 sonst, betrachten. Dann ist $S := \{\delta_x \mid x \in M\}$ linear unabhängig. S ist jedoch keine Erzeugendensystem, da es in der linearen Algebra keine unendlichen Summen gibt. ◁

Wir haben gesehen, dass ein Vektorraum (sehr viele) verschiedene Basen haben kann. Unser nächstes Ziel ist der Nachweis, dass alle Basen gleich viele Elemente haben (sofern sie endlich sind). Der Schlüssel hierzu ist das folgende Lemma.

Lemma 10.9. *Es seien $E \subseteq V$ ein endliches Erzeugendensystem und $U \subseteq V$ eine linear unabhängige Menge. Dann gilt für die Elementanzahlen:*

$$|U| \leq |E|.$$

Beweis. Als Teilmenge einer endlichen Menge ist auch $E \setminus U$ endlich. Wir benutzen Induktion nach $|E \setminus U|$. Wir schreiben $E = \{v_1, \ldots, v_n\}$ mit $v_1, \ldots, v_n$ paarweise verschieden.

1. Fall: $U \subseteq E$. Dann ist automatisch $|U| \leq |E|$, also nichts zu zeigen.
2. Fall: Es gibt ein $v \in U \setminus E$. Wir werden ein „Austauschargument“ benutzen und einen Vektor von E durch v ersetzen. Dies funktioniert folgendermaßen: Wegen $V = \langle E \rangle$ existieren $a_1, \ldots, a_n \in K$ mit

$$v = a_1 v_1 + \cdots + a_n v_n. \tag{10.1}$$

Wegen $v \notin E$ gilt $v \neq v_i$ für alle i. Es gibt ein i, so dass $v_i \notin U$ und $a_i \neq 0$, denn sonst ergäbe (10.1) die lineare Abhängigkeit von U. Nach Umnummerieren haben wir $v_1 \in E \setminus U$ und $a_1 \neq 0$. Dies zeigt auch, dass der Induktionsanfang ($|E \setminus U| = 0$) automatisch in den 1. Fall fällt. Mit $E' := \{v, v_2, \ldots, v_n\}$ ergibt sich aus (10.1):

$$v_1 = a_1^{-1} \cdot \left(v - \sum_{i=2}^{n} a_i v_i \right) \in \langle E' \rangle.$$

Hieraus folgt, dass auch E' ein Erzeugendensystem ist. Nach Definition von E' gilt $|E' \setminus U| = |E \setminus U| - 1$. Induktion liefert also $|U| \leq |E'|$. Wieder nach Definition gilt $|E'| = |E|$, und es folgt die Behauptung. □

Korollar 10.10. *Falls V ein endliches Erzeugendensystem hat, so sind alle Basen von V endlich und haben gleich viele Elemente.*

Beweis. B_1 und B_2 seien Basen von V. Da B_1 und B_2 linear unabhängig sind, liefert Lemma 10.9 $|B_1| < \infty$ und $|B_2| < \infty$. Weiter liefert Lemma 10.9 mit $U = B_1$ und $E = B_2$: $|B_1| \leq |B_2|$. Nach Rollenvertauschung erhalten wir ebenso $|B_2| \leq |B_1|$, also Gleichheit. □

Anmerkung. Es gilt die folgende, weitergehende Aussage: Je zwei Basen eines Vektorraums sind gleichmächtig. Der Beweis ist nicht schwierig, benutzt aber Methoden der *Kardinalzahlarithmetik*, die uns nicht zur Verfügung stehen. ◁

Nun können wir einen der wichtigsten Begriffe der linearen Algebra definieren.

Definition 10.11. *Falls V ein endliches Erzeugendensystem hat, so ist die* **Dimension** *von V die Elementanzahl einer (und damit jeder) Basis von V. Wir schreiben* $\dim(V)$ *für die Dimension von V. Falls V kein endliches Erzeugendensystem hat, schreiben wir* $\dim(V) = \infty$*, um diesen Sachverhalt auszudrücken. (Wir unterscheiden unendliche Basen also gewöhnlich nicht durch ihre Mächtigkeit.) Im ersten Fall heißt V* **endlich-dimensional**, *im zweiten* **unendlich-dimensional**.

Beispiel 10.12. (1) Der Standardraum K^n hat die Dimension n. Damit ist auch die Bezeichnung „n-dimensionaler Standardraum“ aufgeklärt.
(2) Der Lösungsraum des homogenenen LGS aus Beispiel 10.4(6) hat die Dimension 1.
(3) Der Nullraum $V = \{0\}$ hat die Dimension 0.
(4) Für $V = K[x]$ gilt $\dim(V) = \infty$. Hier können wir eine unendliche Basis angeben (siehe Beispiel 10.4(4)). Ist M eine unendliche Menge, so gilt auch $\dim\left(K^M\right) = \infty$. Wir können zwar keine Basis angeben, aber doch eine unendliche linear unabhängige Menge (siehe Beispiel 10.8), so dass K^M nach Lemma 10.9 nicht endlich erzeugt sein kann. ◁

Aus dem nach Beispiel 10.4 angegebenen Verfahren zum Finden einer Basis des Lösungsraums eines homogenen LGS gewinnen wir:

Proposition 10.13. *Gegeben sei ein homogenes LGS mit Koeffizientenmatrix $A \in K^{m\times n}$. Dann gilt für die Lösungsmenge L:*

$$\dim(L) = n - \operatorname{rg}(A).$$

Wie kann man eine Basis eines Unterraums $U \subseteq K^n$ finden? Wir nehmen an, U sei durch erzeugende Vektoren $v_1, \ldots, v_m$ gegeben. Dann bilden wir die Matrix $A \in K^{m\times n}$ mit den v_i als *Zeilen*. Nun bringen wir A mit dem Gauß-Algorithmus auf Zeilenstufenform. Dann bilden diejenigen Zeilen der Zeilenstufenform, die nicht komplett aus Nullen bestehen, eine Basis von U. *Begründung*: Nach Proposition 9.10 wird U von den Zeilen der Zeilenstufenform erzeugt, also auch durch die Zeilen $\neq 0$. Außerdem sieht man sofort, dass die Zeilen $\neq 0$ einer Matrix in Zeilenstufenform immer linear unabhängig sind.

Es folgt insbesondere: $\dim(U) = \operatorname{rg}(A)$. Damit haben wir bewiesen:

Proposition 10.14. *Der Rang einer Matrix $A \in K^{m\times n}$ ist die Dimension des von den Zeilen aufgespannten Unterraums von $K^{1\times n}$.*

Hiermit haben wir für den Rang eine nicht-prozedurale Charakterisierung gefunden. Hierdurch ist die Lücke, die sich durch Definition 9.8 ergeben hat, geschlossen. Eine weitere Charakterisierung des Rangs ist bereits in Proposition 10.13 enthalten. Auch diese zeigt die eindeutige Bestimmtheit des Rangs.

Wir ziehen noch ein paar weitere Folgerungen aus Lemma 10.9. Die erste ermöglicht in vielen Fällen, die Basiseigenschaft zu verifizieren oder zu falsifizieren.

Korollar 10.15. *Es seien $v_1, \ldots, v_n \in V$ paarweise verschieden und $S = \{v_1, \ldots, v_n\}$. Dann gelten:*

(a) S ist eine Basis von $V \iff \dim(V) = n$ und S ist linear unabhängig $\iff \dim(V) = n$ und $V = \langle S\rangle$.
(b) Falls $n < \dim(V)$, so folgt $V \neq \langle S\rangle$.
(c) Falls $n > \dim(V)$, so ist S linear abhängig.

Beweis. (a) Falls S eine Basis ist, so folgt aus Korollar 10.10 und Definition 10.3, dass $\dim(V) = n$, $V = \langle S\rangle$, und dass S linear unabhängig ist. Ist umgekehrt $\dim(V) = n$ und S linear unabhängig, so folgt aus Lemma 10.9, dass S maximal linear unabhängig ist, also ist S nach Satz 10.5 eine Basis. Falls $\dim(V) = n$ und $V = \langle S\rangle$, so folgt aus Lemma 10.9, dass S ein minimales Erzeugendensystem ist, also ist S nach Satz 10.5 eine Basis.

(b) Falls $n < \dim(V)$, so gibt es eine linear unabhängige Menge $U \subseteq V$ mit $|S| < |U|$. Nach Lemma 10.9 kann S kein Erzeugendensystem sein.

(c) Falls $n > \dim(V)$, so gibt es eine Basis $B \subseteq V$ mit $|B| < |S|$. Nach Lemma 10.9 kann S nicht linear unabhängig sein. □

An dieser Stelle lohnt es sich, auf einige formale Parallelen zwischen der Theorie der Basen von Vektorräumen und der Theorie der Spannbäume von Graphen hinzuweisen, auch wenn die tatsächlichen Inhalte der Theorien und die Beweise nichts miteinander zu tun haben. Hierbei entsprechen sich die Begriffe „Erzeugendensystem“ und „zusammenhängend“ sowie „linear unabhängig“ und „kreisfrei“. Genauer gibt es deutliche Parallelen zwischen Satz 10.6 und Satz 4.12, zwischen Korollar 10.15(a) und Satz 4.9(c), sowie zwischen Korollar 10.15(b),(c) und Satz 4.7(a),(b)

Korollar 10.16. *Es sei $U \subseteq V$ ein Unterraum. Dann gelten:*

(a) $\dim(U) \leq \dim(V)$.
(b) Falls $\dim(U) = \dim(V) < \infty$*, so folgt* $U = V$.

Beweis. Es sei A eine Basis von U. Wegen Satz 10.6 gibt es eine Basis B von V mit $A \subseteq B$. Hieraus folgt (a). Falls $\dim(U) = \dim(V) < \infty$, so folgt $A = B$, also $U = V$. □

11 Lineare Abbildungen

Auch in diesem Abschnitt sei K ein Körper. Weiter seien V und W zwei K-Vektorräume (über demselben Körper K!).

Definition 11.1. *Eine Abbildung $\varphi\colon V \to W$ heißt* **linear***, falls gelten:*

(1) Für alle $v, v' \in V$: $\varphi(v+v') = \varphi(v) + \varphi(v')$. (Hierbei ist das „+“ auf der linken Seite das von V, das auf der rechten das von W; φ ist also ein Homomorphismus von Gruppen.)
(2) Für alle $v \in V$ und $a \in K$: $\varphi(a \cdot v) = a \cdot \varphi(v)$.

Insbesondere bildet wegen Proposition 6.14(a) eine lineare Abbildung den Nullvektor von V auf den Nullvektor von W ab.

Beispiel 11.2. (1) Die folgenden geometrisch definierten Abbildungen $\mathbb{R}^2 \to \mathbb{R}^2$ sind linear: Drehungen um den Nullpunkt, Streckungen mit dem Nullpunkt als Zentrum, Spiegelungen an einer durch den Nullpunkt gehenden Geraden, Projektionen auf eine durch den Nullpunkt gehende Gerade. Drehungen um Punkte $\neq 0$ und Verschiebungen sind *nicht* linear.
(2) Die *Nullabbildung* $V \to W$, $v \mapsto 0$ ist linear.
(3) Sei $A = (a_{i,j}) \in K^{m\times n}$. Dann ist

$$\varphi_A\colon K^n \to K^m,\ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \quad \text{mit} \quad y_i = \sum_{j=1}^{n} a_{i,j}x_j$$

eine lineare Abbildung. Dies ist einer der wichtigsten Typen von linearen Abbildungen. Die Bezeichnung φ_A werden wir in Zukunft weiter benutzen.

(4) Für $V = \mathbb{R}[x]$ ist

$$\varphi\colon V \to V,\ f \mapsto f' \quad \text{(Ableitung)}$$

linear. Ebenso ist $\psi\colon V \to \mathbb{R},\ f \mapsto f(1)$ linear.

(5) Für $V = K^n$ und $i \in \{1, \dots, n\}$ ist

$$\pi_i\colon V \to K,\ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto x_i$$

linear. Man bezeichnet π_i als das i-te *Koordinatenfunktional*.

(6) Es sei M eine Menge und $x_1, \dots, x_n \in M$ irgendwelche (fest gewählten) Elemente. Dann ist

$$\varphi\colon V := K^M \to K^n,\ f \mapsto \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

linear. ◁

Sind $\varphi, \psi\colon V \to W$ linear, so gilt dies auch für

$$\varphi + \psi\colon V \to W,\ v \mapsto \varphi(v) + \psi(v).$$

Außerdem ist für ein $a \in K$ auch

$$a \cdot \varphi\colon V \to W,\ v \mapsto a \cdot \varphi(v)$$

linear. Dies bedeutet, dass die Menge $\mathrm{Hom}(V, W)$ aller linearer Abbildungen $V \to W$ einen K-Vektorraum bildet.

Weiter gilt: Sind $\varphi\colon V \to W$ und $\psi\colon W \to U$ (mit U ein weiterer K-Vektorraum) linear, so gilt dies auch für die Komposition $\psi \circ \varphi\colon V \to U$. Damit wird $\mathrm{Hom}(V, V)$ sogar zu einem Ring. (Wir werden sehen, dass dieser für $\dim(V) \geq 2$ nicht-kommutativ ist.)

Definition 11.3. *Es sei $\varphi\colon V \to W$ linear. Der* **Kern** *von φ ist die Menge*

$$\mathrm{Kern}(\varphi) := \{v \in V \mid \varphi(v) = 0\} \subseteq V.$$

Das **Bild** *von φ ist*

$$\mathrm{Bild}(\varphi) := \varphi(V) = \{\varphi(v) \mid v \in V\} \subseteq W.$$

Satz 11.4. *Es sei $\varphi\colon V \to W$ eine lineare Abbildung.*

(a) $\mathrm{Kern}(\varphi) \subseteq V$ *ist ein Unterraum.*
(b) $\mathrm{Bild}(\varphi) \subseteq W$ *ist ein Unterraum.*
(c) Es gilt die Äquivalenz:

$$\varphi \text{ ist injektiv} \iff \mathrm{Kern}(\varphi) = \{0\}.$$

Beweis. (a) Der Nullvektor von V ist in $\mathrm{Kern}(\varphi)$ enthalten. Für $v, v' \in \mathrm{Kern}(\varphi)$ gilt $\varphi(v + v') = \varphi(v) + \varphi(v') = 0$, also $v + v' \in \mathrm{Kern}(\varphi)$. Weiter gilt für $v \in \mathrm{Kern}(\varphi)$ und $a \in K$: $\varphi(a \cdot v) = a \cdot \varphi(v) = a \cdot 0 = 0$, also $a \cdot v \in \mathrm{Kern}(\varphi)$. Insgesamt folgt (a).
(b) folgt durch einfaches Nachrechnen.
(c) Dies folgt aus Proposition 6.14(e). □

Beispiel 11.5. (1) Sei $A \in K^{m\times n}$. Dann ist $\mathrm{Kern}(\varphi_A)$ die Lösungsmenge des homogenen LGS mit Koeffizientenmatrix A. Es folgt: φ_A ist injektiv $\iff \mathrm{rg}(A) = n$.
(2) Sei $V = \mathbb{R}[x]$ und $\varphi\colon V \to V,\ f \mapsto f'$ (Ableitung). $\mathrm{Kern}(\varphi)$ ist die Menge aller konstanter Polynome. (Wie wir wissen) ist φ nicht injektiv. Es gilt $\mathrm{Bild}(\varphi) = V$. ◁

Definition 11.6. *Eine lineare Abbildung $\varphi\colon V \to W$ heißt* **Isomorphismus***, falls φ bijektiv ist. Dann ist auch die Umkehrabbildung $\varphi^{-1}\colon W \to V$ ein Isomorphismus. V und W heißen* **isomorph***, falls es einen Isomorphismus $V \to W$ gibt. Notation: $V \cong W$.*

Wir betrachten einen K-Vektorraum V mit $n = \dim(V) < \infty$. Nachdem wir eine Basis $B = \{v_1, \ldots, v_n\}$ von V gewählt haben, können wir die lineare Abbildung

$$\varphi\colon K^n \to V,\ \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \mapsto \sum_{i=1}^{n} a_i v_i$$

definieren. Die lineare Unabhängigkeit von B liefert $\mathrm{Kern}(\varphi) = \{0\}$, also ist φ nach Satz 11.4(c) injektiv. Da B ein Erzeugendensystem ist, folgt die Surjektivität von φ. Also ist φ ein Isomorphismus. Die Umkehrabbildung ist dadurch gegeben, dass jedem $v \in V$ sein **Koordinatenvektor** bezüglich B zugewiesen wird, also der eindeutig bestimmte Vektor $\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in K^n$ mit $v = \sum_{i=1}^{n} a_i v_i$. Wir haben bewiesen:

Satz 11.7. *Es sei $n := \dim(V) < \infty$. Dann gilt*

$$V \cong K^n.$$

Beispiel 11.8. $V = \{f \in K[x] \mid \deg(f) < 3\} \cong K^3$. Ein Isomorphismus wird gegeben durch

$$\varphi\colon K^3 \to V,\ \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \mapsto a_1 + a_2 x + a_3 x^2.$$

◁

Der Isomorphismus aus Satz 11.7 kann immer erst nach Wahl einer Basis angegeben werden. Man spricht auch von einem *nicht kanonischen* Isomorphismus. Satz 11.7 besagt, dass man sich beim Studium von endlichdimensionalen Vektorräumen immer auf den Fall $V = K^n$ zurückziehen kann.

Satz 11.9 (Dimensionssatz für lineare Abbildungen)**.** *Sei $\varphi\colon V \to W$ linear. Dann gilt:*

$$\dim(V) = \dim\left(\mathrm{Kern}(\varphi)\right) + \dim\left(\mathrm{Bild}(\varphi)\right).$$

Beweis. Wir betrachten nur den Fall, dass $\mathrm{Kern}(\varphi)$ und $\mathrm{Bild}(\varphi)$ endlichdimensional sind. (Der allgemeine Fall geht genauso, benötigt aber aufwändigere Notation.) Es seien $\{w_1, \ldots, w_n\}$ eine Basis von $\mathrm{Bild}(\varphi)$ und $\{v_1, \ldots, v_m\}$ eine Basis von $\mathrm{Kern}(\varphi)$. Wir können $v_1', \ldots, v_n' \in V$ wählen mit $\varphi(v_i') = w_i$. Behauptung: $B := \{v_1, \ldots, v_m, v_1', \ldots, v_n'\}$ ist eine Basis von V.

Zum Nachweis der linearen Unabhängigkeit sei

$$a_1 v_1 + \cdots + a_m v_m + b_1 v_1' + \cdots + b_n v_n' = 0 \tag{11.1}$$

mit $a_i, b_i \in K$. Anwendung von φ auf (11.1) liefert:

$$0 = \varphi(0) = \sum_{i=1}^{m} a_i \varphi(v_i) + \sum_{i=1}^{n} b_i \varphi(v_i') = \sum_{i=1}^{n} b_i w_i.$$

Wegen der linearen Unabhängigkeit der w_i liefert dies $b_1 = \cdots = b_n = 0$. Nun folgt aus (11.1)

$$a_1 v_1 + \cdots + a_m v_m,$$

also auch $a_1 = \cdots = a_m = 0$.

Für den Nachweis, dass B ein Erzeugendensystem ist, sei $v \in V$ beliebig. Wegen $\varphi(v) \in \mathrm{Bild}(\varphi)$ können wir v schreiben als $\varphi(v) = \sum_{i=1}^{n} b_i w_i$ mit $b_i \in K$. Mit $\widetilde{v} := v - \sum_{i=1}^{n} b_i v_i'$ folgt

$$\varphi(\widetilde{v}) = \varphi(v) - \sum_{i=1}^{n} b_i \varphi(v_i') = \varphi(v) - \sum_{i=1}^{n} b_i w_i = 0,$$

also $\widetilde{v} \in \mathrm{Kern}(\varphi)$. Damit gibt es $a_1, \ldots, a_m \in K$, so dass

$$\widetilde{v} = a_1 v_1 + \cdots + a_m v_m.$$

Insgesamt erhalten wir

$$v = \widetilde{v} + \sum_{i=1}^{n} b_i v_i' = \sum_{i=1}^{m} a_i v_i + \sum_{i=1}^{n} b_i v_i',$$

also $v \in \langle B \rangle$.

Wir haben nachgewiesen, das B eine Basis von V ist, also $\dim(V) = |B| = m + n = \dim(\mathrm{Kern}(\varphi)) + \dim(\mathrm{Bild}(\varphi))$. □

Wir betrachten jetzt eine durch eine Matrix $A \in K^{m\times n}$ gegebene lineare Abbildung $\varphi_A\colon K^n \to K^m$ (siehe Beispiel 11.2(3)). Nach Proposition 10.13 hat $\mathrm{Kern}(\varphi_A)$ die Dimension $n - \mathrm{rg}(A)$. Satz 11.9 liefert $n = \dim(\mathrm{Kern}(\varphi_A)) + \dim(\mathrm{Bild}(\varphi_A))$, also folgt $\dim(\mathrm{Bild}(\varphi_A)) = \mathrm{rg}(A)$. Was ist $\mathrm{Bild}(\varphi_A)$? Das Bild besteht genau aus allen Linearkombinationen der Spalten von A. Damit haben wir bewiesen:

Korollar 11.10. *Der Rang einer Matrix $A \in K^{m\times n}$ ist die Dimension des von den Spalten aufgespannten Unterraums von K^m.*

Der Vergleich mit Proposition 10.14 ist besonders interessant! Die durch Proposition 10.14 und Korollar 11.10 gegebenen Interpretationen des Rangs laufen unter der Merkregel

„Zeilenrang“ = „Spaltenrang“.

Korollar 11.11. *Es gelte $\dim(V) = \dim(W) < \infty$, und $\varphi\colon V \to W$ sei eine lineare Abbildung. Dann sind äquivalent:*

(a) φ ist ein Isomorphismus.
(b) φ ist injektiv.
(c) φ ist surjektiv.

Beweis. Es wird behauptet, dass in der betrachteten Situation Injektivität und Surjektivität von φ äquivalent sind. Nach Satz 11.4(c) ist Injektivität gleichbedeutend mit $\mathrm{Kern}(\varphi) = \{0\}$, also mit $\dim(\mathrm{Kern}(\varphi)) = 0$. Wegen Satz 11.9 ist

$$\dim(\mathrm{Bild}(\varphi)) = \dim(V) - \dim(\mathrm{Kern}(\varphi)) = \dim(W) - \dim(\mathrm{Kern}(\varphi)).$$

Also ist φ genau dann injektiv, wenn $\dim(\mathrm{Bild}(\varphi)) = \dim(W)$. Dies ist wegen Korollar 10.16(b) gleichbedeutend mit $\mathrm{Bild}(\varphi) = W$, also mit der Surjektivität von φ. □

Zum Abschluss des Abschnitts beweisen wir einen Satz, der im folgenden Abschnitt eine wichtige Rolle spielen wird.

Satz 11.12 (lineare Fortsetzung)**.** *Es sei $B = \{v_1, \ldots, v_n\}$ eine Basis von V.*

(a) *Eine lineare Abbildung $\varphi\colon V \to W$ ist durch die Bilder der Basisvektoren v_i eindeutig bestimmt. Mit anderen Worten: Ist $\psi\colon V \to W$ eine weitere lineare Abbildung mit $\varphi(v_i) = \psi(v_i)$ für alle i, so folgt $\varphi = \psi$.*

(b) *Seien $w_1, \dots, w_n \in W$ beliebig. Dann gibt es eine lineare Abbildung $\varphi\colon V \to W$ mit $\varphi(v_i) = w_i$ für alle i.*

Zusammengefasst: Man kann lineare Abbildungen eindeutig definieren, indem man die Bilder der Basisvektoren angibt. Dies nennt man das Prinzip der linearen Fortsetzung.

Beweis. (a) Es gelte $\varphi(v_i) = \psi(v_i)$ für alle i. Sei $v \in V$. Dann gibt es $a_1, \dots, a_n \in K$ mit $v = \sum_{i=1}^{n} a_i v_i$, also

$$\varphi(v) = \varphi\left(\sum_{i=1}^{n} a_i v_i\right) = \sum_{i=1}^{n} a_i \varphi(v_i) = \sum_{i=1}^{n} a_i \psi(v_i) = \psi\left(\sum_{i=1}^{n} a_i v_i\right) = \psi(v).$$

Dies bedeutet $\varphi = \psi$.

(b) Wir definieren $\varphi\colon V \to W$ folgendermaßen: Für $v \in V$ sei $v = \sum_{i=1}^{n} a_i v_i$ mit $a_i \in K$. Dann setzen wir

$$\varphi(v) := \sum_{i=1}^{n} a_i w_i.$$

Die eindeutige Darstellungseigenschaft von B liefert die Wohldefiniertheit von φ. Die Linearität ergibt sich durch einfaches Nachprüfen. Außerdem gilt nach Konstruktion $\varphi(v_i) = w_i$. □

12 Darstellungsmatrizen und Matrixprodukt

In diesem Abschnitt seien K ein Körper, V und W endlich-dimensionale K-Vektorräume und $B = \{v_1, \dots, v_n\}$ bzw. $C = \{w_1, \dots, w_m\}$ Basen von V bzw. von W. Für das Folgende ist die *Reihenfolge* der Basisvektoren wichtig. Wir könnten dies zum Ausdruck bringen, indem wir als neues mathematisches Objekt eine *geordnete Basis* einführen, etwa als ein Element des n-fachen kartesischen Produkts $V \times \cdots \times V$ (mit den entsprechenden Zusatzeigenschaften einer Basis). Wir werden aber davon absehen, solchen begrifflichen und notationstechnischen Aufwand zu betreiben.

Nun sei $\varphi\colon V \to W$ eine lineare Abbildung. Für $j \in \{1, \dots, n\}$ können wir schreiben:

$$\varphi(v_j) = \sum_{i=1}^{m} a_{i,j} w_i$$

mit $a_{i,j} \in K$. Nun bilden wir die Matrix

$$A = (a_{i,j}) = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix} \in K^{m\times n}.$$

Die Spalten von A sind also die Koordinatenvektoren der $\varphi(v_i)$.

Definition 12.1. *Die oben definierte Matrix A heißt die* **Darstellungsmatrix** *von φ (bezüglich der Basen B und C). Schreibweise:*

$$A = D_{B,C}(\varphi).$$

Falls $V = W$ gilt, so verwendet man dieselbe Basis $B = C$ und schreibt $D_B(\varphi) \in K^{n\times n}$.

Als Merkregel halten wir fest:

Spalten der Darstellungsmatrix $\longleftrightarrow$ Bilder der Basisvektoren

Beispiel 12.2. (1) Es sei $V = W = \mathbb{R}^2$ mit Basis $B = \{e_1, e_2\}$, und $\varphi\colon V \to V$ sei eine Drehung um 60° nach links. Wir haben

$$\varphi(e_1) = \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix} = \frac{1}{2}e_1 + \frac{\sqrt{3}}{2}e_2,$$
$$\varphi(e_2) = \begin{pmatrix} -\sqrt{3}/2 \\ 1/2 \end{pmatrix} = -\frac{\sqrt{3}}{2}e_1 + \frac{1}{2}e_2,$$

also

$$D_B(\varphi) = \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix}.$$

(2) Es sei $V = \{f \in \mathbb{R}[x] \mid \deg(f) < 3\}$ mit Basis $B = \{1, x, x^2\}$. Für $\varphi\colon V \to V,\ f \mapsto f'$ (Ableitung) erhalten wir

$$\varphi(1) = 0,\ \varphi(x) = 1 \quad \text{und} \quad \varphi(x^2) = 2x,$$

also

$$D_B(\varphi) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

◁

Wir machen die Menge $K^{m\times n}$ aller $m\times n$-Matrizen zu einem K-Vektorraum, indem wir zwei Matrizen $A := (a_{i,j})$ und $B = (b_{i,j}) \in K^{m\times n}$ komponentenweise addieren, also

$$A + B = (a_{i,j} + b_{i,j})_{i,j},$$

und das Produkt mit einem Skalar $c \in K$ definieren als

$$c \cdot A = (c \cdot a_{i,j})_{i,j}.$$

Nun können wir formulieren:

Satz 12.3. *Es gilt*

$$\mathrm{Hom}(V, W) \cong K^{m \times n}.$$

Ein Isomorphismus wird gegeben durch

$$\Delta\colon \mathrm{Hom}(V, W) \to K^{m \times n},\ \varphi \mapsto D_{B,C}(\varphi).$$

Beweis. Die Linearität von Δ folgt direkt aus den Definitionen. Zum Beweis der Injektivität sei $\Delta(\varphi) = 0$. Dann folgt $\varphi = 0$ (die Nullabbildung) aus Satz 11.12(a). Für den Beweis der Surjektivität sei $A = (a_{i,j}) \in K^{m \times n}$. Wegen Satz 11.12(b) gibt es $\varphi \in \mathrm{Hom}(V, W)$ mit $\varphi(v_j) = \sum_{i=1}^{m} a_{i,j} w_i$. Es folgt $\Delta(\varphi) = A$. □

In Beispiel 11.2(3) haben wir mit Hilfe einer Matrix eine lineare Abbildung $K^n \to K^m$ definiert, also bereits eine Zuordnung zwischen Matrizen und linearen Abbildungen hergestellt. Besteht zwischen dieser Zuordnung und Definition 12.1 ein Zusammenhang?

Satz 12.4. *Gegeben seien $V = K^n$ und $W = K^m$ mit den Standardbasen B und C, und eine lineare Abbildung $\varphi\colon V \to W$. Mit $A := D_{B,C}(\varphi)$ gilt dann*

$$\varphi = \varphi_A.$$

Insbesondere sind alle linearen Abbildungen $V \to W$ von der Form φ_A mit $A \in K^{m \times n}$, und A ist die Darstellungsmatrix von φ_A bezüglich der Standardbasen.

Beweis. Wir schreiben $A = (a_{i,j})$. Für den Standardbasisvektor e_j gilt

$$\varphi(e_j) = \sum_{i=1}^{m} a_{i,j} e_i = \begin{pmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{pmatrix} = \varphi_A(e_j).$$

Aus Satz 11.12(a) folgt nun die Behauptung. □

Anmerkung. Aus der Wahl der Basen B und C erhalten wir Isomorphismen $\psi_B\colon K^n \to V$ und $\psi_C\colon K^m \to W$. Für die Darstellungsmatrix $A = D_{B,C}(\varphi)$ einer linearen Abbildung $\varphi\colon V \to W$ gilt dann:

$$\varphi_A = \psi_C^{-1} \circ \varphi \circ \psi_B.$$

Dies ist eine (leicht zu beweisende) Verallgemeinerung von Satz 12.4. ◁

Wir wissen, dass die Komposition von linearen Abbildungen wieder linear ist. Damit ergibt sich die Frage: Was passiert mit den Darstellungsmatrizen

bei Bildung der Komposition? Zur Beantwortung dieser Frage brauchen wir das Matrixprodukt.

Definition 12.5. *Für $A = (a_{i,j}) \in K^{m\times n}$ und $B = (b_{i,j}) \in K^{n\times l}$ ist das Produkt $A \cdot B \in K^{m\times l}$ definiert durch $A \cdot B = (c_{i,j})$ mit*

$$c_{i,j} := \sum_{k=1}^{n} a_{i,k} b_{k,j}.$$

Das Produkt ist also nicht *komponentenweise definiert. Es ist nur definiert, wenn die Spaltenzahl von A mit der Zeilenzahl von B übereinstimmt. Ein wichtiger Spezialfall ist das Produkt einer Matrix $A = (a_{i,j}) \in K^{m\times n}$ mit einem Spaltenvektor $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in K^n$:*

$$A \cdot v = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in K^m \quad \textit{mit} \quad y_i = \sum_{j=1}^{n} a_{i,j} x_j.$$

Beispiel 12.6.

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1\cdot 1 + 0\cdot 1 + 1\cdot 0 & 1\cdot 1 + 0\cdot 2 + 1\cdot 1 \\ 0\cdot 1 + 1\cdot 1 + 2\cdot 0 & 0\cdot 1 + 1\cdot 2 + 2\cdot 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 4 \end{pmatrix}.$$

◁

Zu $A \in K^{m\times n}$ kann man nun die lineare Abbildung $\varphi_A\colon K^n \to K^m$ durch $\varphi_A(v) := A \cdot v$ definieren. Außerdem können wir ein LGS mit erweiterter Koeffizientenmatrix $(A \mid b)$ schreiben als $A \cdot x = b$.

Satz 12.7. *Es seien U, V und W endlich-dimensionale K-Vektorräume mit Basen A, B bzw. C, und es seien $\varphi\colon U \to V$ und $\psi\colon V \to W$ lineare Abbildungen. Dann gilt*

$$D_{A,C}(\psi \circ \varphi) = D_{B,C}(\psi) \cdot D_{A,B}(\varphi).$$

Als Merkregel halten wir fest:

Komposition von linearen Abbildungen $\longleftrightarrow$ Matrixprodukt

Beweis. Wir müssen zunächst Bezeichnungen einführen. Wir schreiben $A = \{u_1, \ldots, u_n\}$, $B = \{v_1, \ldots, v_m\}$, $C = \{w_1, \ldots, w_l\}$ und

$$D_{B,C}(\psi) = (a_{i,j}) \in K^{l\times m},\ D_{A,B}(\varphi) = (b_{i,j}) \in K^{m\times n}.$$

Für $j \in \{1, \dots, n\}$ gilt:

$$(\psi \circ \varphi)(u_j) = \psi\left(\sum_{k=1}^{m} b_{k,j} v_k\right) = \sum_{k=1}^{m} b_{k,j}\psi(v_k) = \sum_{k=1}^{m}\left(b_{k,j}\sum_{i=1}^{l} a_{i,k} w_i\right) = \sum_{i=1}^{l}\left(\sum_{k=1}^{m} a_{i,k} b_{k,j}\right) w_i.$$

Aus der Beobachtung, dass im letzten Ausdruck der Koeffizient von w_i genau der (i,j)-te Eintrag des Produkts $D_{B,C}(\psi) \cdot D_{A,B}(\varphi)$ ist, folgt die Behauptung. □

Man könnte sagen, dass das Matrixprodukt so definiert ist, dass Satz 12.7 richtig wird. Da für drei lineare Abbildungen $\varphi_1\colon V_1 \to V_2$, $\varphi_2\colon V_2 \to V_3$ und $\varphi_3\colon V_3 \to V_4$ das „Assoziativitätsgesetz“ $\varphi_3 \circ (\varphi_2 \circ \varphi_1) = (\varphi_3 \circ \varphi_2) \circ \varphi_1$ gilt, folgt für Matrizen $A \in K^{m\times n}$, $B \in K^{n\times l}$ und $C \in K^{l\times r}$:

$$(A \cdot B) \cdot C = A \cdot (B \cdot C). \tag{12.1}$$

Wir haben schon gesehen, dass $\mathrm{Hom}(V,V)$ ein Ring wird. Aus Satz 12.7 folgt, dass $K^{n\times n}$ mit der Addition und Multiplikation von Matrizen ein Ring ist, der isomorph zu der $\mathrm{Hom}(V,V)$ ist. Das Einselement von $K^{n\times n}$ ist die **Einheitsmatrix**

$$I_n := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ & \ddots & & \\ \vdots & & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} = (\delta_{i,j})_{i,j} \in K^{n\times n}.$$

Für $n \geq 2$ ist $K^{n\times n}$ nicht kommutativ, wie das Beispiel

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \text{aber} \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

das sich auf beliebige $n \times n$-Matrizen mit $n \geq 2$ ausweiten lässt, zeigt. Damit ist auch $\mathrm{Hom}(V,V)$ für $\dim(V) \geq 2$ nicht kommutativ. Das wäre auch nicht zu erwarten gewesen, denn die Komposition von Abbildungen ist „selten“ kommutativ (siehe Anmerkung 2.5(b)).

Aus (12.1) folgt für $A \in K^{m\times n}$, $B \in K^{n\times l}$ und $v \in K^l$:

$$\varphi_{A\cdot B}(v) = (A \cdot B) \cdot v = A \cdot (B \cdot v) = \varphi_A\left(\varphi_B(v)\right),$$

also

$$\varphi_{A\cdot B} = \varphi_A \circ \varphi_B. \tag{12.2}$$

Wann ist eine Matrix $A \in K^{n\times n}$ **invertierbar**, d.h. wann gibt es eine **inverse Matrix** $A^{-1} \in K^{n\times n}$ mit $A \cdot A^{-1} = I_n$? Dies gilt wegen (12.2) genau dann, wenn die zugehörige lineare Abbildung $\varphi_A\colon K^n \to K^n$ surjektiv ist. Nach Korollar 11.11 ist dies gleichbedeutend mit der Injektivität von φ_A, also nach Beispiel 11.5(1) damit, dass $\operatorname{rg}(A) = n$. Wir halten fest:

$$A \in K^{n\times n} \text{ ist invertierbar} \quad \Longleftrightarrow \quad \operatorname{rg}(A) = n.$$

Für die Bedingung $\operatorname{rg}(A) = n$ haben wir auch die Sprechweise eingeführt, dass A regulär ist.

Da aus der Invertierbarkeit von A die Bijektivität von φ_A folgt, gilt auch $\varphi_A^{-1} \circ \varphi_A = \mathrm{id}$. Hieraus folgt mit (12.2), dass auch $A^{-1}A = I_n$ gilt.

Für das Berechnen einer inversen Matrix zu $A \in K^{n\times n}$ haben wir das folgende Verfahren.

(1) Bilde die „erweiterte“ Matrix $(A|I_n) \in K^{n\times(2n)}$ durch Anhängen einer Einheitsmatrix.
(2) Führe diese (mit dem Gauß-Algorithmus) über in strenge Zeilenstufenform, so dass zusätzlich in jeder Zeile $\neq 0$ der erste Eintrag $\neq 0$ eine 1 ist.
(3) 1. Fall: Die Zeilenstufenform hat die Gestalt $(I_n|B)$ mit $B \in K^{n\times n}$: Dann gilt $B = A^{-1}$, und wir sind fertig.
2. Fall: Die Zeilenstufenform hat eine andere Gestalt: Dann ist $\operatorname{rg}(A) < n$, A ist also nicht invertierbar.

Die Korrektheit des Algorithmus begründen wir wie folgt: Es werden simultan die LGSe $A \cdot x = e_i$ (i-ter Standardbasisvektor) gelöst. Der erste Fall ist der Fall eindeutiger Lösbarkeit. Dann sind die Spalten von B jeweils die Lösungsvektoren, und es folgt $A \cdot B = I_n$.

Beispiel 12.8. Wir möchten die Matrix $A = \begin{pmatrix} 1 & -2 & 0 \\ -1 & 3 & -2 \\ -1 & 2 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$ invertieren. Obiges Verfahren läuft wie folgt ab:

$$\left(\begin{array}{ccc|ccc} 1 & -2 & 0 & 1 & 0 & 0 \\ -1 & 3 & -2 & 0 & 1 & 0 \\ -1 & 2 & -1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & -2 & 0 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{array}\right) \longrightarrow$$

$$\left(\begin{array}{ccc|ccc} 1 & -2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & -2 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 2 & -4 \\ 0 & 1 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -1 & 0 & -1 \end{array}\right),$$

also $A^{-1} = \begin{pmatrix} -1 & 2 & -4 \\ -1 & 1 & -2 \\ -1 & 0 & -1 \end{pmatrix}$. Per Probe-Multiplikation prüft man leicht $A \cdot A^{-1} = A^{-1} \cdot A = I_3$ nach. ◁

Für zwei invertierbare Matrizen $A, B \in K^{n\times n}$ ist auch $A \cdot B$ invertierbar, die Inverse ist

$$(A \cdot B)^{-1} = B^{-1}A^{-1}.$$

Außerdem ist A^{-1} invertierbar. Es folgt, dass die Menge

$$\mathrm{GL}_n(K) := \left\{A \in K^{n\times n} \mid A \text{ invertierbar}\right\}$$

eine Gruppe bildet. Sie heißt die **allgemeine lineare Gruppe**. Für $n \geq 2$ ist $\mathrm{GL}_n(K)$ nicht abelsch.

Für den Rest des Abschnitts beschäftigen wir uns mit dem Thema Basiswechsel.

Wir wissen, dass Vektorräume verschiedene Basen haben. Was passiert mit der Darstellungsmatrix einer linearen Abbildung $V \to V$, wenn man die Basis von V wechselt?

Es sei $B = \{v_1, \ldots, v_n\}$ eine Basis von V, und $B' = \{v'_1, \ldots, v'_n\}$ sei eine weitere Basis. Wir können die „neuen“ Basisvektoren v'_j mit Hilfe der alten ausdrücken:

$$v'_j = \sum_{i=1}^{n} a_{i,j} v_i \tag{12.3}$$

mit $a_{i,j} \in K$. Hieraus können wir die Matrix $S := (a_{i,j}) \in K^{n\times n}$ bilden. S heißt die **Basiswechselmatrix**. Sie beschreibt den Übergang von B zu B'. Man schreibt bisweilen $S =: S_{B,B'}$. Die Basiswechselmatrix wird nach folgender Merkregel gebildet:

Spalten von S = Koordinatenvektoren der „neuen“ Basisvektoren

Man kann auch umgekehrt die v_j mit Hilfe der v'_i ausdrücken: $v_j = \sum_{i=1}^{n} b_{i,j} v'_i$ mit $b_{i,j} \in K$. Wir setzen $T := (b_{i,j}) \in K^{n\times n}$. Für alle $j \in \{1, \ldots, n\}$ folgt:

$$v_j = \sum_{i=1}^{n} b_{i,j} \left(\sum_{k=1}^{n} a_{k,i} v_k \right) = \sum_{k=1}^{n} \left(\sum_{i=1}^{n} a_{k,i} b_{i,j} \right) v_k.$$

Den in der rechten Klammer stehenden Ausdruck erkennen wir als den (k, j)-te Eintrag des Matrixprodukts $S \cdot T$. Aus der Gleichung folgt (wegen der linearen Unabhängigkeit von B), dass $S \cdot T = I_n$ gelten muss, also $T = S^{-1}$.

Wir bemerken noch, dass jede invertierbare Matrix $S = (a_{i,j}) \in \mathrm{GL}_n(K)$ einen Basiswechel beschreibt, indem man die neue Basis einfach durch (12.3) definiert.

Wir kehren zurück zu unserer Ausgangsfrage und betrachten eine lineare Abbildung $\varphi\colon V \to V$. Wir schreiben $D_B(\varphi) = (d_{i,j}) \in K^{n\times n}$ und möchten nun $D_{B'}(\varphi)$ bestimmen. Dazu rechnen wir

$$\varphi(v'_j) = \varphi\left(\sum_{i=1}^{n} a_{i,j}v_i\right) = \sum_{i=1}^{n} a_{i,j}\varphi(v_i) = \sum_{i=1}^{n} a_{i,j}\left(\sum_{k=1}^{n} d_{k,i}v_k\right) =$$
$$\sum_{i,k=1}^{n} d_{k,i}a_{i,j}\left(\sum_{l=1}^{n} b_{l,k}v'_l\right) = \sum_{l=1}^{n}\left(\sum_{i,k=1}^{n} b_{l,k}d_{k,i}a_{i,j}\right)v'_l.$$

Den in der rechten Klammer stehenden Ausdruck erkennen wir als den (l,j)-te Eintrag des Matrixprodukts $T\cdot D_B(\varphi)\cdot S$. Aus der Gleichung folgt, dass dieser Ausdruck andererseits der (l,j)-te Eintrag der Darstellungsmatrix $D_{B'}(\varphi)$ sein muss. Damit haben wir gezeigt:

Satz 12.9. *Es seien B und B' Basen eines endlich-dimensionalen K-Vektorraums V und $S := S_{B,B'}$ die Basiswechselmatrix. Dann gilt für eine lineare Abbildung $\varphi\colon V \to V$:*

$$\boxed{D_{B'}(\varphi) = S^{-1}\cdot D_B(\varphi)\cdot S.}$$

Dieser Satz beantwortet die Frage, was bei Wechsel der Basis mit der Darstellungsmatrix passiert. Für lineare Abbildungen zwischen verschiedenen Vektorräumen erhalten wir durch ein ganz entsprechendes (aber notationstechnisch aufwändigeres) Argument:

Satz 12.10. *Es seien B, B' endliche Basen von V und C, C' endliche Basen von W. Dann gilt für eine lineare Abbildung $\varphi\colon V \to W$:*

$$\boxed{D_{B',C'}(\varphi) = S^{-1}_{C,C'}\cdot D_{B,C}(\varphi)\cdot S_{B,B'}.}$$

Wir nehmen diese beiden Sätze (und die Bemerkung, dass jede invertierbare Matrix einen Basiswechsel vermittelt) zum Anlass für folgende Definition:

Definition 12.11. *(a) Zwei quadratische Matrizen $A, B \in K^{n\times n}$ heißen* **ähnlich**, *falls es $S \in \mathrm{GL}_n(K)$ gibt mit*

$$B = S^{-1}AS.$$

(b) Zwei Matrizen $A, B \in K^{m\times n}$ heißen **äquivalent**, *falls es $S \in \mathrm{GL}_n(K)$ und $T \in \mathrm{GL}_m(K)$ gibt mit*

$$B = T^{-1}AS.$$

Wie man sich leicht überlegt, sind Ähnlichkeit und Äquivalenz Äquivalenzrelationen. Von diesen beiden Begriffen ist die Ähnlichkeit der wichtigere.

Das folgende Beispiel soll einen Hinweis darauf geben, weshalb ein Basiswechsel nützlich sein kann.

Beispiel 12.12. Es seien $V = \mathbb{R}^2$ und $\varphi\colon V \to V$, $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} y \\ x \end{pmatrix}$. Mit der Standardbasis $B = \{e_1, e_2\}$ haben wir

$$D_B(\varphi) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Als neue Basis wählen wir $B' = \{\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}\}$. Die Basiswechselmatrix und ihre Inverse sind

$$S = S_{B,B'} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{und} \quad S^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Es ergibt sich

$$D_{B'}(\varphi) = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Die Darstellungsmatrix $D_{B'}(\varphi)$ beschreibt φ in einfacherer Weise: Der erste Basisvektor wird durch φ festgehalten, der zweite wird „umgeklappt“. ◁

Der Abschnitt sei mit einer eindringlichen Warnung abgeschlossen: Man neigt dazu, Basiswechselmatrizen eine Interpretation als eine Art von Darstellungsmatrizen zu geben, etwa als Matrizen, die eine Basis auf eine andere abbilden. Dies führt immer wieder zu großer Verwirrung. Man sollte sich vor solchen Interpretationen hüten, und den Formalismus des Basiswechsels stattdessen als Rezept, dessen Korrektheit bewiesen wurde, wortwörtlich anwenden.

13 Diskrete Strukturen: Lineare Codes

In diesem Abschnitt werden die bisher erarbeiteten Konzepte auf die Datenübertragung über einen nicht perfekten Kanal angewandt. Wir stellen uns vor, dass nacheinander Bits $x_1, x_2, x_3, \ldots$ über einen Kanal gesendet (oder auf einem Datenträger gespeichert) werden. Hierbei sind Fehler möglich: Mit einer gewissen Wahrscheinlichkeit (etwa $p = 10^{-6}$) wird ein Bit fehlerhaft übertragen bzw. gespeichert. Um trotzdem die korrekten Daten rekonstruieren zu können, oder um zumindest mit großer Wahrscheinlichkeit auf einen Fehler aufmerksam zu werden, schickt man die Daten mit einer gewissen Redundanz.

Die naivste Idee ist hierbei das Wiederholen: Alle Daten werden zweimal gesendet (oder 3,4, ... mal). Bei Einteilung in Viererblocks wird also statt (x_1, x_2, x_3, x_4) das „Wort“ $(x_1, x_2, x_3, x_4, x_1, x_2, x_3, x_4)$ gesendet.

Als allgemeinen Rahmen wollen wir die folgende Situation betrachten: Ein Bit wird als ein Element des Körpers $K = \mathbb{F}_2 (= \mathbb{Z}/2\mathbb{Z})$ modelliert. Wir können jedoch auch Elemente eines anderen (endlichen) Körpers K betrachten. Der zu sendende Bit-Strom wird in Blocks der Länge k zerlegt, z.B. $k = 4$. Statt $(x_1, \ldots, x_k) \in K^k$ wird $(c_1, \ldots, c_n) \in K^n$ gesendet (bzw. gespeichert). Hierbei gibt es eine Zuordnung $(x_1, \ldots, x_k) \mapsto (c_1, \ldots, c_n)$. Diese ist häufig linear, d.h. gegeben durch eine Matrix $G \in K^{n \times k}$, also:

$$\begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = G \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}.$$

(Man beachte, dass wir hier je nach Bequemlichkeit Zeilen- und Spaltenvektoren schreiben.) Der gesendete Vektor $(c_1, \ldots, c_n)$ heißt **Codewort**, und $(x_1, \ldots, x_k)$ heißt **Informationswort**. G heißt **Generatormatrix**. Die Menge

$$C := \left\{ G \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \;\middle|\; \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \in K^k \right\}$$

aller Codewörter bildet einen Unterraum des K^n. Eine solche Datenübertragung ist nur sinnvoll, wenn die Zuordnung des Codeworts zu einem Datenwort injektiv ist. Das inhomogene LGS $G \cdot x = c$ muss also für alle $c \in C$ eindeutig lösbar sein, also $\operatorname{rg}(G) = k$. Aus unserem Test auf lineare Unabhängigkeit auf Seite 81 folgt, dass die Spalten von G linear unabhängig sind. Diese Spalten erzeugen C, also folgt

$$\dim(C) = k.$$

Ausgehend von dieser Situation machen wir folgende Definition:

Definition 13.1. *Ein* **linearer Code** *ist ein Unterraum $C \subseteq K^n$. Mit $k := \dim(C)$ bezeichnen wir C auch als einen (n, k)-Code. Die* **Länge** *von C ist n. Die* **Informationsrate** *ist k/n, die* **Redundanz** *ist $n - k$.*

Bei der Definition fällt auf, dass die Abbildung $K^k \to K^n$ nicht in die Definition des Codes aufgenommen wird. Für die meisten Fragestellungen der Codierungstheorie ist diese nämlich unerheblich. Als Generatormatrix eines Codes C kann man jede Matrix nehmen, deren Spalten eine Basis von C bilden. Wir bemerken noch, dass bisweilen auch nicht-lineare Codes betrachtet werden.

Beispiel 13.2. (1) Die Generatormatrix

$$G := \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix}$$

liefert den *Wiederholungscode*, bei dem alles einmal wiederholt wird. Dies ist ein (8,4)-Code, die Informationsrate ist also 1/2. Falls bei der Übertragung höchstens ein Fehler auftritt, wird dies beim Empfang festgestellt. Der Fehler kann jedoch nicht korrigiert werden. Man spricht von einem *1-fehlererkennenden Code*.

(2) Der sogenannte *Parity-Check-Code* ist gegeben durch die Generatormatrix

$$G := \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,1\,1\,1 \end{pmatrix}.$$

Als Abbildung kann man ihn als $(x_1, \ldots, x_4) \mapsto (x_1, \ldots, x_4, x_1+x_2+x_3+x_4)$ definieren. Dies ist ein (5,4)-Code. Falls einer oder 3 Fehler auftreten, wird dies erkannt. Also ist auch dieser Code *1-fehlererkennend*. Aber seine Informationsrate ist mit 4/5 höher als die des Wiederholungscodes. Der Parity-Check-Code ist wohl eine der ältesten Ideen der Informatik.

(3) Es ist auch möglich, jedes Informationswort dreimal zu senden. Der entsprechende Code hat die Generatormatrix

$$G = \begin{pmatrix} I_4 \\ I_4 \\ I_4 \end{pmatrix} \in K^{12\times 4}.$$

Dies ist ein (12,4)-Code. Falls höchstens ein Fehler auftritt, kann man diesen nach Empfang korrigieren. Man spricht von einem *1-fehlerkorrigierenden Code*. ◁

Das *Dekodieren* läuft folgendermaßen ab: Das empfangene Wort $c' = (c'_1, \ldots, c'_n)$ kann sich von dem gesendeten Wort c durch Übertragungsfehler unterscheiden. Falls c' ein Codewort ist, also $c' \in C$, so wird $c = c'$ angenommen, denn dann ist der wahrscheinlichste Fall, dass kein Fehler auftrat. In diesem Fall wird durch das Auflösen des LGS $G \cdot x = c'$ das (wahrscheinliche) Informationswort $x \in K^k$ ermittelt. Interessanter ist der Fall $c' \notin C$. Es wird (wieder) mit der Annahme gearbeitet, dass die Anzahl der Fehlerbits mit hoher Wahrscheinlichkeit klein ist. Also sucht man ein Codewort $c'' \in C$, das sich von c' an möglichst wenig Koordinaten unterscheidet. Falls es genau ein

solches c'' gibt, wird $c = c''$ angenommen und $x \in K^k$ mit $G \cdot x = c''$ ausgegeben. Andernfalls wird eine Fehlermeldung ausgegeben: dann ist sinnvolles Dekodieren nicht möglich. Die Güte eines Codes entscheidet sich darin, dass dieser Fall möglichst vermieden wird, und dass korrektes Dekodieren ($c'' = c$) mit möglichst hoher Wahrscheinlichkeit passiert.

Definition 13.3. *Für $c = (c_1, \ldots, c_n) \in K^n$ ist*

$$w(c) := \left| \left\{ i \in \{1, \ldots, n\} \,\middle|\, c_i \neq 0 \right\} \right|$$

das **Hamming-Gewicht** *von c. Für $c, c' \in K^n$ ist*

$$d(c, c') := w(c - c') = \left| \left\{ i \in \{1, \ldots, n\} \,\middle|\, c_i \neq c'_i \right\} \right|$$

der **Hamming-Abstand** *von c und c'. (Nebenbei: Dies ist eine* Metrik *auf K^n.) Für eine Teilmenge $C \subseteq K^n$ ist*

$$d(C) := \min \left\{ d(c, c') \,\middle|\, c, c' \in C,\ c \neq c' \right\}$$

der **Hamming-Abstand** *von C. (Falls $|C| \leq 1$, so setzen wir $d(C) := n+1$.) Falls C ein Unterraum ist, ergibt sich*

$$d(C) = \min \left\{ w(c) \,\middle|\, c \in C \setminus \{0\} \right\}.$$

Beispiel 13.4. (1) Der (8,4)-Wiederholungscode (Beispiel 13.2(1)) hat $d(C) = 2$.
(2) Der (5,4)-Parity-Check-Code (Beispiel 13.2(2)) hat ebenfalls $d(C) = 2$.
(3) Der (12,4)-Wiederholungscode (Beispiel 13.2(3)) hat $d(C) = 3$. ◁

Folgende Überlegung zeigt, dass der Hamming-Abstand entscheidend ist für die Güte eines Codes.

Es sei zunächst $d(C) = 2e + 1$ ungerade. Das (durch Übertragungsfehler bedingte) Ändern von höchstens e Bits in einem Codewort ergibt ein $c' \in K^n$ mit $d(c, c') \leq e$. Dann ist c das eindeutig bestimmte Codewort $c'' \in C$ mit $d(c'', c') \leq e$. Aus $d(c'', c') \leq e$ und $c'' \in C$ folgt nämlich $d(c'', c) \leq 2e$, also $c'' = c$ wegen der Annahme. Dies bedeutet, dass korrekt dekodiert wird, falls höchstens e Übertragungsfehler auftreten. Der Code ist also e-fehlerkorrigierend. (Bei mehr als e Fehlern ist allerdings eine misslungene oder gar falsche Dekodierung möglich.)

Nun sei $d(C) = 2e + 2$ gerade. Nach obigem Argument ist C auch e-fehlerkorrigierend. Zusätzlich gilt: Bei $e + 1$ Fehlern gibt es kein Codewort $c'' \in C$ mit $d(c'', c') \leq e$ (denn dann wäre $c'' \neq c$ und $d(c, c'') \leq d(c, c') + d(c', c'') \leq e + 1 + e < d(C)$, ein Widerspruch). Falls es nun ein eindeutig bestimmtes Codewort mit minimalem Abstand zu c' gibt, so ist dieses gleich c, und das Dekodieren liefert das korrekte Wort. Es ist aber möglich, dass c' „genau zwischen“ c und einem weiteren Codewort c'' liegt, d.h. $d(c, c') =$

$d(c'', c') = e + 1$. Dann wird eine Fehlermeldung ausgegeben. Dies bedeutet, dass $e+1$ Fehler zumindest erkannt werden. Ein Code mit Hamming-Abstand $2e + 2$ ist also in diesem Sinne $(e + 1)$-fehlererkennend.

Wir fassen zusammen:

Satz 13.5. *Sei $C \subseteq K^n$ ein Code.*

(a) Falls $d(C) = 2e + 1$, so ist C e-fehlerkorrigierend.
(b) Falls $d(C) = 2e+2$, so ist C e-fehlerkorrigierend und $(e+1)$-fehlererkennend.

Alles, was wir über das Dekodieren und den Hamming-Abstand gesagt haben, gilt auch für nicht-lineare Codes. Nun erinnern wir uns, dass wir lineare Codes betrachten wollen, also Unterräume $C \subseteq K^n$, die von den (linear unabhängigen) Spalten einer Matrix G erzeugt werden. Wegen $\operatorname{rg}(G) = k$ ist es gemäß Proposition 10.14 möglich, k linear unabhängige Zeilen von G auszusuchen. Durch Vertauschungen der Zeilen kann man also annehmen, dass die ersten k Zeilen von G linear unabhängig sind. Dies bedeutet, dass wir auch bei den Codewörtern $c \in C$ die Reihenfolge der Koordinaten c_i ändern, eine unwesentliche Änderung. Nun können wir auf G elementare *Spalten*operationen anwenden und G auf strenge *Spaltenstufenform* bringen; dies entspricht den gewohnten Zeilenoperationen auf der transponierten Matrix G^T. Wegen Proposition 9.10 ändern die Spaltenoperationen den Code C nicht. Wir ersetzen G durch die in strenge Spaltenstufenform gebrachte Matrix. Wegen der linearen Unabhängigkeit der ersten k Zeilen ergibt sich (nach Normieren der Diagonaleinträge)

$$G = \begin{pmatrix} I_k \\ A \end{pmatrix} \tag{13.1}$$

mit $A \in K^{(n-k)\times k}$. Bei unserern bisherigen Beispielen lag G jeweils schon zu Beginn in dieser Form vor. Nun bilden wir die Matrix

$$P := \begin{pmatrix} -A & I_{n-k} \end{pmatrix} \in K^{(n-k)\times n}.$$

P hat den Rang $n - k$, und es gilt

$$P \cdot G = \begin{pmatrix} -A & I_{n-k} \end{pmatrix} \cdot \begin{pmatrix} I_k \\ A \end{pmatrix} = 0.$$

Hieraus folgt $P \cdot c = 0$ für alle $c \in C$. Andererseits hat die Lösungsmenge L des homogenen LGS $P \cdot x = 0$ nach Proposition 10.13 die Dimension $n - (n - k) = k = \dim(C)$. Wegen Korollar 10.16(b) folgt $L = C$. Wir halten fest, dass für $c \in K^n$ gilt:

$$c \in C \quad \Longleftrightarrow \quad P \cdot c = 0.$$

P heißt die **Parity-Check-Matrix**. Nebenbei sei erwähnt, dass für lineare Codes auch ohne die Voraussetzung (13.1) eine Parity-Check-Matrix existiert.

Beispiel 13.6. (1) Der (8,4)-Wiederholungscode (Beispiel 13.2(1)) hat die Parity-Check-Matrix

$$P = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix} \in K^{4\times 8}.$$

(2) Der (5,4)-Parity-Check-Code (Beispiel 13.2(2)) hat die Parity-Check-Matrix

$$P = \begin{pmatrix} -1 & -1 & -1 & -1 & 1 \end{pmatrix} \in K^{1\times 5}.$$

Mit Hilfe der Parity-Check-Matrix kann man das Dekodierungsverfahren verbessern. Es sei $c' \in K^n$ das empfangene Wort. Den Unterschied von c und c' quantifizieren wir durch den (dem Empfänger nicht bekannten) *Fehlervektor* $f := c' - c \in K^n$. Es ergibt sich

$$P \cdot c' = P \cdot (c + f) = 0 + P \cdot f = P \cdot f.$$

Der Vektor $P \cdot c' \in K^{n-k}$ heißt das **Syndrom** von c'. Es misst, wie weit c' von einem Codewort abweicht. Nach obiger Gleichung haben empfangenes Wort und Fehlervektor das gleiche Syndrom. Das Dekodieren kann nun so geschehen: Man berechnet das Syndrom $P \cdot c'$. Nun sucht man ein $f \in K^n$, welches unter allen $f' \in K^n$ mit $P \cdot f' = P \cdot c'$ minimales Hamming-Gewicht hat. Falls $c' \in C$, so ergibt sich automatisch $f = 0$. Falls es ein eindeutig bestimmtes solches f gibt, setzt man $c'' := c' - f \in C$ und gibt $x \in K^k$ mit $G \cdot x = c''$ aus. Falls es kein eindeutiges f gibt, gibt man eine Fehlermeldung aus. Dies entspricht genau dem oben beschriebenen Dekodierungsverfahren. Da es nur $|K|^{n-k}$ mögliche Syndrome gibt, kann man das f (oder Fehlermeldung) zu jedem Syndrom in einer Tabelle speichern. Oft gibt es noch bessere Methoden zur Ermittlung von f. Dies ist in folgendem Beispiel der Fall.

Der (7,4)-Hamming-Code

Wir definieren nun den sogenannten (7,4)-Hamming-Code. Dieser zeigt, dass Codierungstheorie zu mehr in der Lage ist, als die bisherigen, relativ offensichtlichen Beispiele von Codes zu analysieren. Der Hamming-Code $C \subset \mathbb{F}_2^7$ wird durch die Generatormatrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} \in \mathbb{F}_2^{7\times 4}$$

definiert, als Abbildung $\mathbb{F}_2^4 \to \mathbb{F}_2^7$ also $(x_1, \ldots, x_4) \mapsto (x_1, x_2, x_3, x_4, x_2 + x_3 + x_4, x_1 + x_3 + x_4, x_1 + x_2 + x_4)$. C ist ein (7,4)-Code, hat also höhere Informationsrate als der (8,4)-Wiederholungscode aus Beispiel 13.2(1). Die Parity-Check-Matrix ist

$$P = \begin{pmatrix} 0\,1\,1\,1\,1\,0\,0 \\ 1\,0\,1\,1\,0\,1\,0 \\ 1\,1\,0\,1\,0\,0\,1 \end{pmatrix}.$$

Welchen Hamming-Abstand hat C? Dazu müssen wir $w(c)$ für $c \in C \setminus \{0\}$ ermitteln. Die Bedingung $c \in C$ ist gleichbedeutend mit $P \cdot c = 0$. Gibt es ein solches c mit $w(c) = 1$? Dies würde bedeuten, dass (mindestens) eine der Spalten von P eine Nullspalte ist, was nicht der Fall ist. Gibt es ein $c \in \mathbb{F}_2^7$ mit $P \cdot c = 0$ und $w(c) = 2$? Dies würde bedeuten, dass es in P zwei Spalten gibt, die linear abhängig sind. Auch dies ist nicht der Fall! Es folgt also $d(C) > 2$. In diesem Argument zeigt sich die eigentliche Idee des Hamming-Codes: Man beginnt mit der Parity-Check-Matrix und stellt sie so auf, dass sie keine zwei linear abhängigen Spalten enthält. Hieraus folgt dann $d(C) > 2$. Die Generatormatrix G leitet man dann aus der Parity-Check-Matrix her. Da G selbst (sogar mehr als) einen Vektor von Gewicht 3 enthält, folgt

$$d(C) = 3.$$

Der (7,4)-Hamming Code ist also 1-fehlerkorrigierend. Damit hat er einerseits eine höhere Informationsrate, andererseits bessere Fehlerkorrektureigenschaften als der (8,4)-Wiederholungscode!

Das Dekodieren ist hier ganz besonders einfach: Es gibt nur acht mögliche Syndrome, nämlich alle Vektoren von $\mathbb{F}_2^3$. Wir können diese schreiben als $v_0 = 0, v_1, \ldots, v_7$, wobei v_i die i-te Spalte von P ist ($i > 0$). Für v_0 ist der Nullvektor das Codewort kleinsten Gewichtes mit Syndrom v_0. Für v_i ($i > 0$) ist dies der i-te Standardbasisvektor e_i, denn $P \cdot e_i = v_i$. Der vollständige Dekodieralgorithmus läuft also so ab: Man ermittelt das Syndrom $s := P \cdot c'$ des empfangenen Wortes $c' = (c'_1, \ldots, c'_7)$. Falls $s = v_i$ mit $1 \leq i \leq 4$, so gibt man $(x_1, \ldots, x_4) = (c'_1, \ldots, c'_4) + e_i$ aus (d.h. das i-te Bit wird geändert). Andernfalls gibt man $(x_1, \ldots, x_4) = (c'_1, \ldots, c'_4)$ aus. (Falls das Syndrom einer der Vektoren v_5, v_6, v_7 ist, so wird e_i mit $i > 4$ zu c' hinzuaddiert, aber dies ändert $(x_1, \ldots, x_4)$ nicht.) In dem wahrscheinlichen Fall, dass bei der Übertragung höchstens ein Fehler auftritt, wird so das korrekte Informationswort ausgegeben.

Der Bauer-Code

Einen weiteren interessanten Code erhalten wir durch folgende Erweiterung des (7,4)-Hamming Codes: Wir hängen einfach zusätzlich noch ein Parity-Bit $c_8 = c_1 + \cdots + c_7$ an, d.h. wir benutzen die Abbildung

$$(x_1, \ldots, x_4) \mapsto (x_1, x_2, x_3, x_4, x_2{+}x_3{+}x_4, x_1{+}x_3{+}x_4, x_1{+}x_2{+}x_4, x_1{+}x_2{+}x_3).$$

Der hierdurch definierte Code C wird *Bauer-Code* (nach F. L. Bauer, Informatiker an der TU München) genannt. Es ist ein (8,4)-Code. Was ist der Hamming-Abstand $d(C)$? Auf jeden Fall mindestens 3, denn die ersten 7 Bits sind ja identisch mit dem Hamming-Code. Aber falls ein Wort $(c_1, \ldots, c_7)$ des Hamming-Codes das Gewicht 3 hat, so ist $c_1 + \cdots + c_7 = 1$, also hat das entsprechende Wort in C Gewicht 4. Wir erhalten $d(C) = 4$. Der Bauer-Code ist also 1-fehlerkorrigierend und 2-fehlererkennend. Er hat damit wesentlich bessere Eigenschaften als der (8,4)-Wiederholungscode.

14 Faktorräume

In diesem Abschnitt übertragen wir das Prinzip von Restklassenringen (siehe Satz 7.4) auf Vektorräume. Der folgende Satz ist zugleich auch eine Definition.

Satz 14.1. *Es seien V ein K-Vektorraum und $U \subseteq V$ ein Unterraum.*

(a) Auf V wird eine Äquivalenzrelation definiert durch

$$v \sim w \quad :\Longleftrightarrow \quad v - w \in U.$$

(b) Die Äquivalenzklasse eines $v \in V$ ist

$$[v]_\sim = \{v + u \mid u \in U\} =: v + U \subseteq V.$$

Teilmengen von V von der Gestalt $v + U$ nennt man auch **affine Unterräume**

(c) Die Faktormenge

$$V/U := \{v + U \mid v \in V\}$$

wird durch folgende Definitionen zu einem K-Vektorraum: Für $C_1, C_2 \in V/U$ und $a \in K$ wählen wir $v \in C_1$ und $w \in C_2$ und setzen

$$C_1 + C_2 := (v + w) + U \quad \text{und} \quad a \cdot C_1 = av + U.$$

Mit dieser Vektorraumstruktur heißt V/U der **Faktorraum** *von V nach U.*

(d) Die Abbildung

$$\pi\colon V \to V/U,\ v \mapsto v + U$$

ist linear und surjektiv. Der Kern ist $\mathrm{Kern}(\pi) = U$.

(e) Es gilt

$$\dim(U) + \dim(V/U) = \dim(V).$$

Beweis. (a) Die Reflexivität von $\sim$ folgt wegen $0 \in U$. Für $v, w \in V$ mit $v \sim w$ gilt $w - v = -(v - w) \in U$, also ist $\sim$ symmetrisch. Für $u, v, w \in V$

mit $u \sim v$ und $v \sim w$ folgt

$$u - w = u - v + v - w \in U,$$

also $u \sim w$. Damit ist $\sim$ auch transitiv.

(b) Für $w \in V$ gilt die Äquivalenz

$$w \in [v]_\sim \quad \Longleftrightarrow \quad \exists\, u \in U\colon w - v = u \quad \Longleftrightarrow \quad w \in v + U.$$

(c) Der wichtigste Schritt ist der Nachweis der Wohldefiniertheit, d.h. der Unabhängigkeit der Definitionen von der Wahl der Vertreter v und w. Es seien also $v', w' \in V$ mit $v' \sim v$ und $w' \sim w$. Dann folgt

$$(v'+w')-(v+w) = (v'-v)+(w'-w) \in U \quad \text{und} \quad av'-av = a(v'-v) \in U,$$

also $[v' + w']_\sim = [v + w]_\sim$ und $[av']_\sim = [av]_\sim$. Nachdem die Wohldefiniertheit geklärt ist, ist klar, dass sich die Vektorraumaxiome von V auf V/U vererben. Der Nullvektor von V/U ist $[0]_\sim = 0 + U = U$.

(d) Für $v, w \in V$ gilt $\pi(v + w) = v + w + U = (v + U) + (w + U)$, und für $a \in K$ gilt $\pi(av) = av+U = a(v+U)$. Also ist π linear. Die Surjektivität von π ist klar. Für $v \in V$ gilt

$$v \in \operatorname{Kern}(\varphi) \quad \Longleftrightarrow \quad v + U = 0 + U \quad \Longleftrightarrow \quad v \in U,$$

also $\operatorname{Kern}(\varphi) = U$.

(e) Dies folgt aus (d) und Satz 11.9 □

Beispiel 14.2. (1) In $V = \mathbb{R}^2$ sei $U \subseteq V$ eine Gerade durch den Nullpunkt. Dann ist V/U die Menge aller Geraden, die parallel zu U sind (aber nicht durch den Nullpunkt laufen müssen).

(2) Für $U = \{0\}$ ist $V/U = \{\{v\} \mid v \in V\}$. In diesem Fall ist π ein Isomorphismus, also $V/\{0\} \cong V$.

(3) Für $U = V$ ist $V/U = \{V\}$ der Nullraum.

◁

Als Anwendung des Faktorraums beweisen wir den folgenden Satz.

Satz 14.3 (Dimensionssatz für Unterräume)**.** *Es seien $U, W \subseteq V$ Unterräume eines K-Vektorraums. Dann gilt*

$$\dim(U \cap W) + \dim(U + W) = \dim(U) + \dim(W).$$

Beweis. Wir betrachten die Abbildung

$$\varphi\colon W \to V/U,\ w \mapsto w + U.$$

Es ist klar, dass φ linear ist. Außerdem gilt

$$\operatorname{Kern}(\varphi) = U \cap W \quad \text{und} \quad \operatorname{Bild}(\varphi) = (U + W)/U.$$

Mit Satz 11.9 folgt

$$\dim(W) = \dim(U \cap W) + \dim\left((U+W)/U\right).$$

Durch Addition von $\dim(U)$ auf beiden Seiten der Gleichung und Anwendung von Satz 14.1(e) ergibt sich die Behauptung. □

15 Direkte Summen

In diesem Abschnitt ist V immer ein Vektorraum über einem Körper K.

Wir erinnern uns an den Begriff des Summenraums. Sind $U_1, \dots, U_n \subseteq V$ Unterräume, so ist

$$\sum_{i=1}^{n} U_i = U_1 + \cdots + U_n = \{v_1 + \cdots + v_n \mid v_1 \in U_1, \dots, v_n \in U_n\} \subseteq V$$

der Summenraum der U_i. Dies ist ein Unterraum von V.

Definition 15.1. *(a) Es seien $U_1, \dots, U_n \subseteq V$ Unterräume. Die Summe $\sum_{i=1}^{n} U_i$ heißt* **direkt**, *falls für alle $v_1 \in U_1, \dots, v_n \in U_n$ gilt:*

$$v_1 + \cdots + v_n = 0 \quad \Longrightarrow \quad v_1 = \cdots = v_n = 0.$$

Wir schreiben dann

$$U_1 \oplus \cdots \oplus U_n = \bigoplus_{i=1}^{n} U_i$$

für $\sum_{i=1}^{n} U_i$.

(b) Sei $U \subseteq V$ ein Unterraum. Ein Unterraum $W \subseteq V$ heißt ein **Komplement** *von U, falls*

$$V = U \oplus W.$$

Proposition 15.2. *Für Unterräume $U_1, \dots, U_n \subseteq V$ sind äquivalent:*

(a) Die Summe $W := U_1 + \cdots + U_n$ ist direkt.

(b) Für alle $w \in W$ gibt es eindeutig bestimmte $v_1 \in U_1, \dots, v_n \in U_n$ mit $w = v_1 + \cdots + v_n$.

(c) Für alle $i \in \{1, \dots, n\}$ gilt

$$U_i \cap \left(\sum_{j \in \{1,\dots,n\} \setminus \{i\}} U_j \right) = \{0\}.$$

Für $n = 2$ lautet die Bedingung (c): $U_1 \cap U_2 = \{0\}$.

Beweis. Wir setzen (a) voraus und zeigen (b). Behauptet wird die Eindeutigkeit der v_i. Es seien also $v_1' \in U_1, \dots, v_n' \in U_n$ mit $w = v_1' + \dots + v_n'$. Dann gilt

$$(v_1 - v_1') + \dots + (v_n - v_n') = w - w = 0,$$

und wegen $v_i - v_i' \in U_i$ und (a) folgt $v_i = v_i'$ für alle i.

Nun zeigen wir, dass aus (b) die Bedingung (c) folgt. Es sei also $i \in \{1, \dots, n\}$ und $v_i \in U_i \cap \left(\sum_{j \neq i} U_j\right)$. Dann gilt

$$v_i = \sum_{j \neq i} v_j \quad \text{mit} \quad v_j \in U_j,$$

und wegen (b) folgt $v_i = 0$. Die Bedingung (c) gilt also.

Nun setzen wir (c) voraus und zeigen (a). Es sei also $v_1 + \dots + v_n = 0$ mit $v_i \in U_i$. Für $i \in \{1, \dots, n\}$ folgt

$$v_i = \sum_{j \neq i} (-v_j) \in \sum_{j \neq i} U_j,$$

also $v_i \in U_i \cap \sum_{j \neq i} U_j$. Wegen (c) folgt $v_i = 0$, also ist (a) gezeigt. □

Beispiel 15.3. (1) In $V = \mathbb{R}^3$ seien $U_1, U_2 \subseteq V$ Unterräume mit $\dim(U_1) = \dim(U_2) = 2$ und $U_1 \neq U_2$. Dann gilt $U_1 + U_2 = V$, aber nach Satz 14.3 folgt

$$\dim(U_1 \cap U_2) = \dim(U_1) + \dim(U_2) - \dim(V) = 1.$$

Also ist $U_1 \cap U_2 \neq \{0\}$. Die Summe $U_1 + U_2$ ist also nicht direkt.

(2) In $V = \mathbb{R}^3$ seien $U_1, U_2 \subseteq V$ Unterräume mit $\dim(U_1) = 1$, $\dim(U_2) = 2$ und $U_1 \not\subseteq U_2$. Dann gilt $U_1 + U_2 = V$ und nach Satz 14.3 folgt

$$\dim(U_1 \cap U_2) = \dim(U_1) + \dim(U_2) - \dim(V) = 0.$$

Die Summe $U_1 + U_2$ ist also direkt, und wir können sie als $U_1 \oplus U_2$ schreiben.

(3) Ist $\{v_1, \dots, v_n\}$ eine Basis von V, so folgt

$$V = \langle v_1 \rangle \oplus \dots \oplus \langle v_n \rangle.$$

(4) $U = V$ hat das Komplement $\{0\}$. ◁

Falls $W \subseteq V$ ein Komplement eines Unterraums $U \subseteq V$ ist, so ist die lineare Abbildung

$$\varphi \colon W \to V/U, \; w \mapsto w + U$$

ein Isomorphismus, denn $\text{Bild}(\varphi) = (W + U)/U = V/U$ und $\text{Kern}(\varphi) = W \cap U = \{0\}$. Also gilt $W \cong V/U$.

Satz 15.4. *Für eine direkte Summe $W := \bigoplus_{i=1}^n U_i$ von Unterräumen $U_i \subseteq V$ gilt*

$$\dim(W) = \sum_{i=1}^{n} \dim(U_i).$$

Beweis. Wir benutzen Induktion nach n. Für $n = 1$ ist nichts zu zeigen. Für $n > 1$ setzen wir $W' = \bigoplus_{i=2}^{n} U_i$. Wegen Proposition 15.2(c) folgt $U_1 \cap W' = \{0\}$, also $\dim(U_1 \cap W') = 0$. Es gilt $W = U_1 + W'$, und mit Satz 14.3 folgt

$$\dim(W) = \dim(U_1 \cap W') + \dim(U_1 + W') = \dim(U_1) + \dim(W').$$

Nach Induktion gilt $\dim(W') = \sum_{i=2}^{n} \dim(U_i)$, und der Satz ist bewiesen.

Alternativ lässt sich der Beweis auch führen, indem man Basen der U_i wählt und zeigt, dass deren Vereinigung eine Basis von W bildet. □

In Beispiel 15.3(1) sieht man, dass die Direktheit der Summe für die Gültigkeit von Satz 15.4 erforderlich ist.

Satz 15.5. *Jeder Unterraum $U \subseteq V$ besitzt ein Komplement.*

Beweis. Es sei A eine Basis von U. Nach dem Basisergänzungssatz (Satz 10.6) gibt es eine Basis B von V mit $A \subseteq B$. Wir setzen $C := B \setminus A$, $W = \langle C \rangle$ und behaupten, dass W ein Komplement von U ist.

Für den Nachweis von $U + W = V$ sei $v \in V$. Dann gibt es $v_1, \ldots, v_n \in A$, $w_1, \ldots, w_m \in C$ und $a_i, b_i \in K$, so dass

$$v = \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{m} b_i w_i \in U + W.$$

Weiter sei $v \in U \cap W$. Dann gibt es paarweise verschiedene $v_1, \ldots, v_n \in A$, paarweise verschiedene $w_1, \ldots, w_m \in C$ und $a_i, b_i \in K$, so dass

$$v = \sum_{i=1}^{n} a_i v_i \quad \text{und} \quad v = \sum_{i=1}^{m} b_i w_i.$$

Wegen $A \cap C = \emptyset$ sind die $v_1, \ldots, v_n, w_1, \ldots, w_m$ paarweise verschieden, und aus der Gleichung

$$\sum_{i=1}^{n} a_i v_i - \sum_{i=1}^{m} b_i w_i = 0$$

und der linearen Unabhängigkeit von B folgt $a_1 = \cdots = a_n = b_1 = \cdots = b_m = 0$, also $v = 0$. Damit ist $U \cap W = \{0\}$ gezeigt, und der Beweis ist abgeschlossen. □

Anmerkung. Man kann den Beweis von Satz 15.5 auch direkt mit dem Zornschen Lemma führen, indem man die Menge aller Unterräume $W \subseteq V$ mit $U \cap W = \{0\}$ betrachtet. ◁

Lineare Algebra: Normalformen

Das übergreifende Thema dieses Kapitels ist, für eine gegebene lineare Abbildung $\varphi\colon V \to V$ eines endlich-dimensionalen Vektorraums eine Basis B zu finden, so dass die Darstellungsmatrix $D_B(\varphi)$ möglichst übersichtlich wird. Wegen Satz 12.9 ist dies gleichbedeutend damit, zu einer gegebenen Matrix $A \in K^{n\times n}$ eine zu A ähnliche Matrix B zu finden (siehe Definition 12.11), die eine einfache Gestalt hat. In jeder Ähnlichkeitsklasse werden wir einen solch einfachen Vertreter B finden und diesen dann eine Normalform von A nennen.

Wir beginnen mit dem Begriff der Determinante, der bei weitem nicht nur für die Thematik der Normalformen von Bedeutung ist. Danach kommen wir zu den Eigenwerten und dem Begriff der Diagonalisierbarkeit. Den eigentlichen Normalformen werden wir uns nähern, indem wir zunächst Matrizen über $\mathbb{Z}$ und über dem Polynomring $K[x]$ behandeln.

16 Determinanten

Bevor wir die Determinante definieren, müssen wir uns mit der symmetrischen Gruppe beschäftigen. Zur Erinnerung: Für $n \in \mathbb{N}_{>0}$ ist die **symmetrische Gruppe** definiert als

$$S_n := \{\sigma\colon \{1,\dots,n\} \to \{1,\dots,n\} \mid \sigma \text{ ist bijektiv}\}\,.$$

Die Elemente von S_n heißen *Permutationen*, und die Verknüpfung ist durch die Komposition gegeben.

Definition 16.1. *Für $\sigma \in S_n$ definieren wir*

- $w(\sigma)$ *als die Anzahl der Paare* $(i,j) \in \mathbb{N} \times \mathbb{N}$ *mit* $1 \leq i < j \leq n$ *aber* $\sigma(i) > \sigma(j)$ *(solche Paare nennt man auch* Fehlstellen*);*
- $\operatorname{sgn}(\sigma) := (-1)^{w(\sigma)}$*, das* **Vorzeichen** *von* σ*.*

Beispiel 16.2. (1) Die Identität $\mathrm{id} \in S_n$ hat keine Fehlstellen, also $\operatorname{sgn}(\mathrm{id}) = 1$.

(2) Es sei $\sigma = (1,2) \in S_n$ (also $\sigma(1) = 2$, $\sigma(2) = 1$ und $\sigma(i) = i$ für $i > 2$). Offenbar ist $(1,2)$ die einzige Fehlstelle von σ, also $\operatorname{sgn}(\sigma) = -1$.

(3) Es seien $1 \leq i < j \leq n$, und $\sigma = (i,j) \in S_n$ (d.h. σ vertauscht i und j und lässt alle anderen Elemente von $\{1,\dots,n\}$ fest). Eine solche Permutation nennt man auch eine *Transposition*. Wir zählen Fehlstellen und kommen auf $w(\sigma) = 2(j-i)-1$, also $\operatorname{sgn}(\sigma) = -1$. ◁

Die wichtigste Eigenschaft des Vorzeichens ist seine Multiplikativität:

Satz 16.3. *Für* $\sigma, \tau \in S_n$ *gilt*

$$\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau).$$

Die Abbildung $\operatorname{sgn} : S_n \to \{1,-1\}$ *ist also ein Gruppen-Homomorphismus.*

Beweis. Es seien $x_1,\dots,x_n \in \mathbb{Q}$ paarweise verschiedene rationale Zahlen. Wir behaupten, dass für alle $\sigma \in S_n$ gilt:

$$\operatorname{sgn}(\sigma) = \prod_{1 \leq i < j \leq n} \frac{x_{\sigma(i)} - x_{\sigma(j)}}{x_i - x_j}. \tag{16.1}$$

Um dies einzusehen bemerken wir, dass Zähler und Nenner des Produkts bis auf das Vorzeichen übereinstimmen. Im Zähler tritt aber genau $w(\sigma)$ mal ein $x_k - x_l$ mit $k > l$ auf, während dies im Nenner nie vorkommt. Hieraus ergibt sich (16.1).

Nun setzen wir $y_i := x_{\sigma(i)}$. Ebenso wie die x_i sind auch die y_i paarweise verschieden, also gilt wegen (16.1) für alle $\tau \in S_n$

$$\operatorname{sgn}(\tau) = \prod_{1 \leq i < j \leq n} \frac{y_{\tau(i)} - y_{\tau(j)}}{y_i - y_j} = \prod_{1 \leq i < j \leq n} \frac{x_{\sigma\tau(i)} - x_{\sigma\tau(j)}}{x_{\sigma(i)} - x_{\sigma(j)}}. \tag{16.2}$$

Wir erhalten

$$\operatorname{sgn}(\sigma\tau) = \prod_{1 \leq i < j \leq n} \frac{x_{\sigma\tau(i)} - x_{\sigma\tau(j)}}{x_i - x_j} = \prod_{1 \leq i < j \leq n} \frac{x_{\sigma\tau(i)} - x_{\sigma\tau(j)}}{x_{\sigma(i)} - x_{\sigma(j)}} \cdot \prod_{1 \leq i < j \leq n} \frac{x_{\sigma(i)} - x_{\sigma(j)}}{x_i - x_j} \underset{(16.2)}{=} \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma).$$

□

Nun können wir die Determinante einer quadratischen Matrix definieren. Ab jetzt sei K ein Körper.

Definition 16.4. *Es sei $A = (a_{i,j}) \in K^{n\times n}$ eine quadratische Matrix. Die* **Determinante** *von A ist*

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

Die Definition lässt sich erweitern für den Fall, dass A Einträge in einem kommutativen Ring hat.

Beispiel 16.5. Für $n \leq 3$ machen wir Definition 16.4 explizit.

(1) Für $n = 1$ ist $A = (a)$ und

$$\det(A) = a.$$

(2) Für $n = 2$ ist $S_n = \{\mathrm{id}, \sigma\}$ mit $\sigma = (1, 2)$. Wir erhalten

$$\det \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}.$$

(3) Für $n = 3$ hat die S_n sechs Elemente: die Identität, die drei Transpositionen $(1, 2)$, $(1, 3)$ und $(2, 3)$, sowie die „zyklischen“ Permutationen $(1, 2, 3)$ und $(3, 2, 1)$ (siehe Beispiel 6.5(2)). Die zyklischen Permutationen haben Vorzeichen 1. Wir erhalten

$$\begin{aligned}\det \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} &= a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} \\ &\quad - a_{1,2}a_{2,1}a_{3,3} - a_{1,3}a_{2,2}a_{3,1} - a_{1,1}a_{2,3}a_{3,2}.\end{aligned}$$

Es gibt eine graphische Merkeregel für die Determinante einer 3×3-Matrix, die sogenannte *Sarrus-Regel*:

$$\begin{array}{cccccc} - & - & - & + & + & + \\ \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,1} & a_{3,2} \end{pmatrix} & & & & & \end{array}$$

Der Zusammenhang zwischen der obigen Formel und der Graphik dürfte selbsterklärend sein.

(4) Für die Einheitsmatrix I_n gilt: $\det(I_n) = 1$. ◁

Nun entwickeln wir die Theorie der Determinante.

Lemma 16.6. *Sei $A = (a_{i,j}) \in K^{n\times n}$.*

(a) $det(A^T) = \det(A)$ (transponierte Matrix).
(b) Es sei $\sigma \in S_n$. Wir definieren $b_{i,j} := a_{i,\sigma(j)}$ und $B := (b_{i,j}) \in K^{n\times n}$ (d.h. B geht aus A durch Permutation der Spalten gemäß σ hervor). Dann gilt

$$\det(B) = \operatorname{sgn}(\sigma) \cdot \det(A).$$

Entsprechendes gilt für Permutationen der Zeilen.

(c) *Falls in A zwei Zeilen oder zwei Spalten übereinstimmen, so folgt*

$$\det(A) = 0.$$

Beweis. (a) Wir rechnen

$$\begin{aligned}\det(A^T) &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot \prod_{i=1}^{n} a_{\sigma(i),i} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot \prod_{j=1}^{n} a_{j,\sigma^{-1}(j)} \\ &= \sum_{\tau \in S_n} \operatorname{sgn}(\tau^{-1}) \cdot \prod_{j=1}^{n} a_{j,\tau(j)} = \det(A).\end{aligned}$$

(b) Wir rechnen

$$\begin{aligned}\det(B) &= \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \cdot \prod_{i=1}^{n} b_{i,\tau(i)} = \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \cdot \prod_{i=1}^{n} a_{i,\sigma\tau(i)} \\ &= \sum_{\rho \in S_n} \operatorname{sgn}(\sigma^{-1}\rho) \cdot \prod_{i=1}^{n} a_{i,\rho(i)} = \operatorname{sgn}(\sigma^{-1}) \cdot \det(A),\end{aligned}$$

wobei Satz 16.3 für die letzte Gleichheit benutzt wurde. Satz 16.3 liefert auch $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$, also folgt die Behauptung.
Die entsprechende Aussage für Zeilenpermutationen lässt sich durch (a) auf die für Spaltenpermutationen zurückführen.

(c) Wegen (a) ist $\det(A) = 0$ nur für den Fall zweier gleicher Spalten nachzuweisen. Wir nehmen also an, dass es $1 \leq j < k \leq n$ gibt, so dass $a_{i,j} = a_{i,k}$ für alle i gilt. Es sei $\tau = (j,k) \in S_n$ die Transposition, die j und k vertauscht (siehe Beispiel 16.2(3)). Für alle $i, l \in \{1, \ldots, n\}$ gilt dann

$$a_{i,l} = a_{i,\tau(l)}. \tag{16.3}$$

Aus (b) folgt $\det(A) = \operatorname{sgn}(\tau)\det(A) = -\det(A)$. Im Fall $\operatorname{char}(K) \neq 2$ liefert dies die Behauptung $\det(A) = 0$. Da wir aber auch den Fall $\operatorname{char}(K) = 2$ mitnehmen möchten, müssen wir etwas mehr Aufwand betreiben. Wir definieren

$$A_n := \{\sigma \in S_n \mid \operatorname{sgn}(\sigma) = 1\}.$$

(Nebenbei gesagt folgt aus Satz 16.3, dass A_n eine Untergruppe der S_n ist; sie heißt die *alternierende Gruppe*.) Wegen $\operatorname{sgn}(\tau) = -1$ folgt aus Satz 16.3, dass S_n die *disjunkte Vereinigung* von A_n und $\tau A_n := \{\tau\sigma \mid \sigma \in A_n\}$ ist:

$$S_n = A_n \dot{\cup} \tau A_n.$$

(Hiermit ist die Vereinigungsmenge gemeint, wobei der Schnitt der beiden vereinigten Mengen leer ist; dies wird durch den Punkt ausgedrückt.) Nun folgt

$$\begin{aligned}\det(A) &= \sum_{\sigma\in A_n}\left(\operatorname{sgn}(\sigma)\cdot\prod_{i=1}^{n}a_{i,\sigma(i)} + \operatorname{sgn}(\tau\sigma)\cdot\prod_{i=1}^{n}a_{i,\tau\sigma(i)}\right)\\ &= \sum_{\sigma\in A_n}\operatorname{sgn}(\sigma)\cdot\left(\prod_{i=1}^{n}a_{i,\sigma(i)} - \prod_{i=1}^{n}a_{i,\tau(\sigma(i))}\right) = 0,\end{aligned}$$

wobei (16.3) für die letzte Gleichheit verwendet wurde. □

Der wohl wichtigste Satz über die Determinante ist der folgende.

Satz 16.7 (Determinantenmultiplikationssatz)**.** *Für $A, B \in K^{n\times n}$ gilt*

$$\boxed{\det(A\cdot B) = \det(A)\cdot\det(B).}$$

Beweis. Wie immer schreiben wir $A = (a_{i,j})$ und $B = (b_{i,j})$. Der (i,j)-te Eintrag von $A\cdot B$ ist $\sum_{k=1}^{n} a_{i,k}b_{k,j}$, also

$$\det(A\cdot B) = \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\cdot\prod_{i=1}^{n}\left(\sum_{k=1}^{n}a_{i,k}b_{k,\sigma(i)}\right).$$

Ausmultiplizieren des Produkts und Vertauschung der Summation liefern

$$\begin{aligned}\det(A\cdot B) &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\cdot\sum_{k_1,\dots,k_n=1}^{n}\prod_{i=1}^{n}\left(a_{i,k_i}b_{k_i,\sigma(i)}\right)\\ &= \sum_{k_1,\dots,k_n=1}^{n}\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\cdot\prod_{i=1}^{n}a_{i,k_i}\cdot\prod_{i=1}^{n}b_{k_i,\sigma(i)} =\\ &\qquad\sum_{k_1,\dots,k_n=1}^{n}\prod_{i=1}^{n}a_{i,k_i}\cdot\det(b_{k_j,l})_{j,l=1,\dots,n}. \qquad (16.4)\end{aligned}$$

Wegen Lemma 16.6(c) ist $\det(b_{k_j,l})_{j,l=1,\dots,n}$ nur dann $\neq 0$, wenn die k_j paarweise verschieden sind, d.h. wenn die Abbildung $\{1,\dots,n\}\to\{1,\dots,n\}$, $j\mapsto k_j$ eine Permutation ist. Statt über die $k_1,\dots,k_n$ zu summieren, können wir also auch über die Permutationen $\tau\in S_n$ summieren und erhalten

$$\det(A \cdot B) = \sum_{\tau \in S_n} \prod_{i=1}^{n} a_{i,\tau(i)} \cdot \det(b_{\tau(j),l})_{j,l=1,\dots,n}$$
$$= \sum_{\tau \in S_n} \prod_{i=1}^{n} a_{i,\tau(i)} \cdot \operatorname{sgn}(\tau) \cdot \det(B) = \det(A) \cdot \det(B),$$

wobei für die zweite Gleichheit Lemma 16.6(b) verwendet wurde. □

Die Determinante ist also multiplikativ. Als Warnung sei hier angemerkt, dass sie nicht additiv ist (außer im Fall $n = 1$)!

Der folgende Satz enthält zwei rekursive Formeln zur Berechnung der Determinante.

Satz 16.8. *Es sei $A = (a_{i,j}) \in K^{n \times n}$ mit $n \geq 2$. Für $i, j \in \{1, \dots, n\}$ sei $A_{i,j} \in K^{(n-1)\times(n-1)}$ die Matrix, die aus A durch Weglassen der i-ten Zeile und der j-ten Spalte entsteht. Für alle $i \in \{1, \dots, n\}$ gilt*

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \cdot \det(A_{i,j}), \tag{16.5}$$

und für alle $j \in \{1, \dots, n\}$ gilt

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{i,j} \cdot \det(A_{i,j}). \tag{16.6}$$

Die Berechnung der Determinante gemäß Formel (16.5) wird als *Entwicklung nach der i-ten Zeile* bezeichnet, und gemäß (16.6) als *Entwicklung nach der j-ten Spalte*. Man kann eine dieser Formeln anwenden und dabei i bzw. j nach Opportunitätsgesichtspunkten auswählen.

Beispiel 16.9. Wir möchten die Determinante von

$$A = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{pmatrix}$$

berechnen und entscheiden uns für Entwicklung nach der ersten Zeile. Es ergibt sich

$$\det(A) = 0 \cdot \det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} - 1 \cdot \det \begin{pmatrix} 3 & 5 \\ 6 & 8 \end{pmatrix} + 2 \cdot \det \begin{pmatrix} 3 & 4 \\ 6 & 7 \end{pmatrix}$$
$$= -(3 \cdot 8 - 6 \cdot 5) + 2 \cdot (3 \cdot 7 - 6 \cdot 4) = 6 - 6 = 0.$$

◁

Beweis von Satz 16.8. Wegen Lemma 16.6(a) genügt es, die Gleichung (16.5) nachzuweisen. Für $i \in \{1, \dots, n\}$ gilt

$$\begin{aligned}\det(A) &= \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\cdot\prod_{k=1}^{n} a_{k,\sigma(k)}\\ &= \sum_{j=1}^{n}\sum_{\substack{\sigma\in S_n\\ \text{mit } \sigma(i)=j}} \operatorname{sgn}(\sigma)\cdot\prod_{\substack{k\in\{1,\dots,n\}\\ \text{mit } k\neq i}} a_{k,\sigma(k)}\cdot a_{i,j}.\end{aligned}$$

Mit

$$c_{i,j} := \sum_{\substack{\sigma\in S_n\\ \text{mit } \sigma(i)=j}} \operatorname{sgn}(\sigma)\cdot\prod_{\substack{k\in\{1,\dots,n\}\\ \text{mit } k\neq i}} a_{k,\sigma(k)}$$

ist also $c_{i,j} = (-1)^{i+j}\det(A_{i,j})$ zu zeigen. Wir benutzen die beiden speziellen Permutationen

$$\eta = (i, i+1, \dots, n-1, n) \quad \text{und} \quad \rho = (j, j+1, \dots, n-1, n) \in S_n.$$

Es gelten $\operatorname{sgn}(\eta) = (-1)^{n-i}$ und $\operatorname{sgn}(\rho) = (-1)^{n-j}$. Mit

$$b_{k,l} := a_{\eta(k),\rho(l)}$$

gilt

$$A_{i,j} = (b_{k,l})_{k,l=1,\dots,n-1}.$$

Außerdem gilt für $\sigma \in S_n$ die Äquivalenz

$$\sigma(i) = j \quad \Longleftrightarrow \quad (\rho^{-1}\sigma\eta)(n) = n.$$

Mit $\tau := \rho^{-1}\sigma\eta$ als neue Summationsvariable erhalten wir

$$c_{i,j} = \sum_{\substack{\tau\in S_n\\ \text{mit } \tau(n)=n}} \operatorname{sgn}(\rho\tau\eta^{-1})\cdot\prod_{\substack{k\in\{1,\dots,n\}\\ \text{mit } k\neq i}} a_{k,(\rho\tau\eta^{-1})(k)},$$

und weiter mit $l := \eta^{-1}(k)$ (welches zwischen 1 und $n-1$ läuft)

$$c_{i,j} = \operatorname{sgn}(\rho)\operatorname{sgn}(\eta^{-1})\cdot\sum_{\tau\in S_{n-1}} \operatorname{sgn}(\tau)\cdot\prod_{l=1}^{n-1}\underbrace{a_{\eta(l),(\rho\tau)(l)}}_{=b_{l,\tau(l)}} = (-1)^{i+j}\det(A_{i,j}).$$

Dies schließt den Beweis ab. □

Wir nehmen Satz 16.8 zum Anlass für folgende Definition:

Definition 16.10. *Es sei $A \in K^{n\times n}$ mit $n \geq 2$. Für $i, j \in \{1, \dots, n\}$ sei $A_{i,j} \in K^{(n-1)\times(n-1)}$ die Matrix, die aus A durch Weglassen der i-ten Zeile und der j-ten Spalte entsteht. Mit*

$$c_{i,j} := (-1)^{i+j}\det(A_{j,i})$$

heißt $C := (c_{i,j}) \in K^{n\times n}$ *die* **adjunkte Matrix** *von* A.

Man beachte den kleinen Unterschied zwischen der Definition der $c_{i,j}$ im Beweis von Satz 16.8 und Definition 16.10.

Satz 16.11. *Es sei* $A \in K^{n\times n}$ *mit* $n \geq 2$. *Dann gilt für die adjunkte Matrix* $C \in K^{n\times n}$ *von* A*:*

$$A \cdot C = C \cdot A = \det(A) \cdot I_n.$$

Beweis. Wir schreiben $A = (a_{i,j})$. Der (i,i)-te Eintrag von $A \cdot C$ ist

$$\sum_{j=1}^{n} a_{i,j}c_{j,i} = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det(A_{i,j}) = \det(A),$$

wobei für die letzte Gleichheit (16.5) verwendet wurde. Nun sei $k \in \{1,\dots,n\}$ mit $k \neq i$, und $A' \in K^{n\times n}$ sei die Matrix, die aus A durch Weglassen der k-ten Zeile und durch Verdoppeln (zweimal untereinander schreiben) der i-ten Zeile entsteht. Der (i,k)-te Eintrag von $A \cdot C$ ist

$$\sum_{j=1}^{n} a_{i,j}c_{j,k} = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det(A_{k,j}) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det(A'_{i,j}) = \det(A').$$

Wegen Lemma 16.6(c) gilt aber $\det(A') = 0$. Insgesamt haben wir $A \cdot C = \det(A) \cdot I_n$ nachgewiesen, und der Beweis von $C \cdot A = \det(A) \cdot I_n$ läuft ebenso. □

Wir ziehen eine wichtige Folgerung.

Satz 16.12. *Für* $A \in K^{n\times n}$ *gilt die Äquivalenz*

$$\boxed{A \text{ ist invertierbar} \iff \det(A) \neq 0.}$$

Falls A *invertierbar ist, so gelten*

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

und

$$A^{-1} = \frac{1}{\det(A)} \cdot C, \tag{16.7}$$

wobei C *für die adjunkte Matrix steht.*

Beweis. Falls A invertierbar ist, folgt nach Satz 16.7 und Beispiel 16.5(4)

$$\det(A^{-1}) \cdot \det(A) = \det(A^{-1} \cdot A) = \det(I_n) = 1,$$

also $\det(A) \neq 0$ und $\det(A^{-1}) = 1/\det(A)$.

Ist umgekehrt $\det(A) \neq 0$, so liefert Satz 16.11 die Gleichung

$$\frac{1}{\det(A)} \cdot C \cdot A = I_n,$$

und es folgen (16.7) und die Invertierbarkeit von A. □

Anmerkung 16.13. Das Berechnen der Inversen nach der Formel (16.7) ist aufwändiger als durch das in Abschnitt 12 angegebene Verfahren. Die Formel kann jedoch nützlich sein, wenn in A Parameter vorkommen, oder um die auftretenden Nenner zu kontrollieren. Außerdem merken wir an, dass alles bisher gesagte auch gilt, wenn K durch einen kommutativen Ring ersetzt wird, wobei die Bedingung „$\det(A) \neq 0$“ in Satz 16.12 durch „$\det(A)$ ist (als Element von K) invertierbar“ zu ersetzen ist. ◁

Beispiel 16.14. (1) Für invertierbare 2×2-Matrizen liest sich (16.7) als

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \cdot \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Dies lässt sich auch direkt verifizieren.

(2) Für welche $a \in \mathbb{R}$ ist die Matrix $A = \left(\begin{smallmatrix} 1 & a \\ a & 1 \end{smallmatrix}\right)$ invertierbar? Die Bedingung hierfür ist nach Satz 16.12 $\det(A) \neq 0$, also $1 - a^2 \neq 0$. A ist also nur für $a = \pm 1$ nicht invertierbar. ◁

Wir haben inzwischen eine ganze Reihe Eigenschaften kennengelernt, die alle für eine quadratische Matrix $A \in K^{n \times n}$ äquivalent sind. Diese äquivalenten Eigenschaften sind:

- A ist regulär;
- A ist invertierbar (anders gesagt: $A \in GL_n(K)$);
- die Zeilen von A sind linear unabhängig;
- die Spalten von A sind linear unabhängig;
- die Abbildung φ_A ist injektiv;
- die Abbildung φ_A ist surjektiv;
- das LGS $A \cdot x = 0$ ist eindeutig lösbar.
- für alle $b \in K^n$ ist das LGS $A \cdot x = b$ eindeutig lösbar.
- $\det(A) \neq 0$.

Wir ziehen eine weitere Folgerung aus Satz 16.7.

Korollar 16.15. *Zwei Matrizen $A, B \in K^{n \times n}$ seien ähnlich. Dann gilt*

$$\det(A) = \det(B).$$

Beweis. Wir haben $B = S^{-1}AS$ mit $S \in \mathrm{GL}_n(K)$. Wegen der Sätze 16.7 und 16.12 folgt

$$\det(B) = \det(S)^{-1} \det(A) \det(S) = \det(A).$$

□

Korollar 16.15 hat eine interessante konzeptionelle Interpretation: Ist $\varphi\colon V \to V$ eine lineare Selbstabbildung eines endlich-dimensionalen Vektorraums V, so lässt sich $\det(\varphi)$ nach Wahl einer Basis B von V durch

$$\det(\varphi) := \det\left(D_B(\varphi)\right)$$

definieren. Denn bei einer anderen Basiswahl geht $D_B(\varphi)$ nach Satz 12.9 über in eine ähnliche Matrix.

Definition 16.16. *Die Menge*

$$\mathrm{SL}_n(K) := \left\{A \in K^{n\times n} \mid \det(A) = 1\right\}$$

heißt die **spezielle lineare Gruppe**. *Aus Satz 16.7 folgt, dass* $\mathrm{SL}_n(K)$ *eine Untergruppe der* $\mathrm{GL}_n(K)$ *ist, womit* $\mathrm{SL}_n(K)$ *selbst eine Gruppe ist.*

Nur quadratische Matrizen haben Determinanten. Bei beliebigen Matrizen $A \in K^{m\times n}$ kann man sogenannte **Minoren** (auch: *Unterdeterminanten*) betrachten. Für $r \leq \min\{m, n\}$ wird ein $r \times r$-Minor von A durch eine Auswahl von r Zeilen und r Spalten von A gebildet, wodurch eine $r\times r$-Matrix entsteht. Der Minor ist die Determinante dieser Matrix. Es gibt also im Allgemeinen eine ganze Menge Minoren. Beispielsweise ist die Anzahl der 2×2-Minoren einer 3×4-Matrix $3 \cdot 6 = 18$. Die 1×1-Minoren sind einfach die Einträge einer Matrix. Mit Hilfe von Korollar 11.10 und Satz 16.12 kann man zeigen, dass das maximale r, für dass es einen $r \times r$-Minor $\neq 0$ gibt, der Rang der Matrix ist.

Nun beschäftigen wir uns mit dem effizienten Berechnen der Determinante. Die Definition 16.4 ist explizit, so dass eine direkte Berechnung möglich ist. Sie erfordert jedoch wegen $|S_n| = n!$ etwa $n \cdot n!$ Körperoperationen, ein für große n nicht hinnehmbarer Aufwand. Wir werden ein besseres Verfahren entwickeln.

Wir können schon jetzt die Determinante einiger spezieller Matrizen im „Eilverfahren“ berechnen. Wir führen drei Fälle an. Begründen kann man die Ergebnisse jeweils entweder durch Entwicklung nach einer Zeile oder Spalte, oder indem man direkt mit Definition 16.4 arbeitet.

(1) Für eine *Diagonalmatrix*

$$A = \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{pmatrix}$$

gilt

$$\det(A) = a_1 \cdots a_n.$$

Man schreibt Diagonalmatrizen wie oben auch als

$$A = \mathrm{diag}(a_1, \ldots, a_n).$$

(2) Für eine *obere Dreiecksmatrix*

$$A = \begin{pmatrix} a_1 & & * \\ & \ddots & \\ 0 & & a_n \end{pmatrix} \tag{16.8}$$

gilt

$$\det(A) = a_1 \cdots a_n. \tag{16.9}$$

Zur Erklärung: (16.8) soll andeuten, dass oberhalb der Diagonalen irgendwelche Einträge stehen können, unterhalb aber lauter Nullen. Man könnte eine obere Dreiecksmatrix $A = (a_{i,j}) \in K^{n\times n}$ auch formaler durch die Bedingung $a_{i,j} = 0$ für $i > j$ definieren.
Dasselbe Ergebnis (16.9) gilt auch für untere Dreiecksmatrizen.

(3) Für eine Matrix

$$A = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$$

mit $B \in K^{l\times l}$, $D \in K^{(n-l)\times(n-l)}$ und $C \in K^{(n-l)\times l}$ gilt

$$\det(A) = \det(B) \cdot \det(D).$$

Man sagt auch, dass A *Block-Dreiecksgestalt* hat. Dies lässt sich erweitern auf Matrizen mit mehr als zwei Diagonal-Blöcken.

Nun wenden wir uns dem Berechnen der Determinante einer Matrix, die keine spezielle Gestalt hat, zu. Ziel ist es, auch hierfür den Gauß-Algorithmus einzusetzen. Wir müssen uns also überlegen, welche Auswirkungen elementare Zeilenoperationen auf die Determinante haben. Bei Operationen von Typ I (Vertauschen zweier Zeilen) geht die Antwort aus Lemma 16.6(b) hervor: Die Determinante ändert das Vorzeichen. Für Operationen vom Typ II und (wichtiger!) vom Typ III ist es zweckdienlich, diese als Links-Multiplikation mit gewissen Matrizen zu interpretieren: Multiplikation der i-ten Zeile von A mit einem Skalar $a \neq 0$ entspricht der Multiplikation von A mit der Matrix

$$S = \operatorname{diag}(1, \ldots, 1, a, 1, \ldots, 1),$$

wobei a der i-te Eintrag ist; also $A \to S\cdot A$. Wegen Satz 16.7 und der Regel (1) ergibt sich, dass sich bei einer Operation von Typ II die Determinante mit a multipliziert.

Um Operationen von Typ III zu behandeln, betrachten wir Matrizen $E_{i,j} \in K^{n\times n}$, die per Definition überall Nullen haben außer im (i,j)-ten Eintrag, der 1 ist. Nun sieht man leicht, dass Addition des a-fachen der j-ten Zeile zu der i-ten Zeile einer Multiplikation mit $I_n + a \cdot E_{i,j}$ von links entspricht: $A \to (I_n + a \cdot E_{i,j}) \cdot A$. Da $I_n + a \cdot E_{i,j}$ eine Dreiecksmatrix ist, folgt aus der Regel (2), dass $\det(I_n + a \cdot E_{i,j}) = 1$ ist, also ändert sich nach Satz 16.7 die Determinante bei Operationen von Typ III nicht. Wir fassen zusammen:

Typ I (Vertauschen zweier Zeilen): Die Determinante ändert das Vorzeichen.

Typ II (Multiplikation einer Zeile mit einem Skalar $a \in K \setminus \{0\}$): Die Determinante multipliziert sich mit a. Als Formel ausgedrückt:

$$\det(\text{neue Matrix}) = a \cdot \det(\text{alte Matrix}).$$

Typ III (Addition des a-fachen einer Zeile zu einer anderen): Die Determinante ändert sich nicht.

Wir bemerken noch, dass Entsprechendes auch für *elementare Spaltenoperationen* gilt.

Nun kann man den Gauß-Algorithmus zum Berechnen von Determinanten verwenden. Die Strategie ist, jeweils eine Spalte (oder Zeile) so weit auszuräumen, dass eine Entwicklung nach dieser Spalte (Zeile) sehr einfach wird. Man kann dabei den Gauß-Algorithmus variieren, denn es kommt nicht darauf an, welche Spalte bzw. Zeile jeweils ausgeräumt wird.

Beispiel 16.17. Wir berechnen (mit nachfolgenden Kommentaren zu den Rechenschritten)

$$\begin{aligned}\det\begin{pmatrix}1&3&4&2\\1&4&2&0\\0&2&1&3\\1&-5&0&-1\end{pmatrix} &\underset{(1)}{=} \det\begin{pmatrix}1&3&4&2\\0&1&-2&-2\\0&2&1&3\\0&-8&-4&-3\end{pmatrix} \underset{(2)}{=} 1\cdot\det\begin{pmatrix}1&-2&-2\\2&1&3\\-8&-4&-3\end{pmatrix}\\ &\underset{(3)}{=} \det\begin{pmatrix}5&0&4\\2&1&3\\0&0&9\end{pmatrix} \underset{(4)}{=} 1\cdot\det\begin{pmatrix}5&4\\0&9\end{pmatrix} \underset{(5)}{=} 5\cdot 9 = 45.\end{aligned}$$

Hierbei wurden folgende Schritte durchgeführt:

(1) Ausräumen der ersten Spalte durch Addition des (-1)-fachen der ersten Zeile zur zweiten und zur vierten Zeile;
(2) Entwicklung nach der ersten Spalte;
(3) Ausräumen der zweiten Spalte durch Addition des 2-fachen der zweiten Zeile auf die erste und Addition des 4-fachen der zweiten Zeile auf die dritte (Ausräumen der ersten Spalte wäre ein etwas größerer arithmetischer Aufwand gewesen: Wer möchte schon mit 8 multiplizieren?);
(4) Entwicklung nach der zweiten Spalte;
(5) die Formel für Dreiecksmatrizen (oder die Formel für 2×2-Determinanten).

◁

Zum Abschluss des Abschnitts geben wir noch eine geometrische Interpretation der Determinante. Für $v_1, v_2 \in \mathbb{R}^2$ ist $|\det(v_1 v_2)|$ der *Flächeninhalt* des Parallelogramms mit den Seiten v_1 und v_2. Dies lässt sich auf n-dimensionale Volumina verallgemeinern. Diese Interpretation ist solange nicht beweisbar, wie wir keinen mathematisch definierten Begriff von Flächeninhalt haben.

Flächeninhalte von Parallelogrammen (bzw. deren höher-dimensionalen Verallgemeinerungen) sind besonders wichtig, weil Parallelogramme bei Flächen-Integralen als „infinitessimale“ Flächenelemente auftreten.

17 Eigenwerte

Auch in diesem Abschnitt sei K ein Körper.

Definition 17.1. *Sei $A \in K^{n\times n}$ eine quadratische Matrix. Ein $\lambda \in K$ heißt* **Eigenwert** *von A, falls es $v \in K^n \setminus \{0\}$ gibt mit $A \cdot v = \lambda \cdot v$. Ein solcher Vektor v heißt dann ein* **Eigenvektor** *von A (zum Eigenwert λ).*

$$E_\lambda := \{v \in K^n \mid A \cdot v = \lambda \cdot v\}$$

heißt der **Eigenraum** *zum Eigenwert λ. Er besteht aus allen Eigenvektoren und dem Nullvektor. E_λ ist auch definiert, wenn $\lambda \in K$* kein *Eigenwert ist.*

Für eine lineare Abbildung $\varphi\colon V \to V$ eines K-Vektorraums V werden Eigenwerte, Eigenvektoren und Eigenräume durch die Eigenschaft

$$\varphi(v) = \lambda \cdot v$$

definiert.

Beispiel 17.2. (1) Für $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ gilt

$$A \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

also ist 1 ein Eigenwert von A und $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ein zugehöriger Eigenvektor. Ein weiterer Eigenwert ist -1, denn

$$A \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Der Eigenraum zu $\lambda = 1$ ist

$$E_1 = \left\{v \in K^2 \mid A \cdot v = v\right\} = \left\{v \in K^2 \mid (A - I_2) \cdot v = 0\right\},$$

also der Lösungsraum des homogenen LGS $(A - I_2) \cdot x = 0$. Die Matrix $A - I_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ hat den Rang 1, also folgt $\dim(E_1) = 1$ nach Proposition 10.13. Wir erhalten also

$$E_1 = \langle \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rangle,$$

und mit den gleichen Argumenten

$$E_{-1} = \langle \begin{pmatrix} 1 \\ -1 \end{pmatrix} \rangle.$$

Insgesamt stellen wir fest, dass $\{\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}\}$ eine Basis aus Eigenvektoren bildet. Die Frage, ob A außer ± 1 noch weitere Eigenwerte hat, werden wir bald beantworten können.

(2) Auf dem Vektorraum $V = C^\infty(\mathbb{R})$ der unendlich oft differenzierbaren Funktionen $\mathbb{R} \to \mathbb{R}$ sei $\varphi\colon V \to V,\ f \mapsto f'$ gegeben. Für $\lambda \in \mathbb{R}$ ist die Exponentialfunktion $f_\lambda\colon \mathbb{R} \to \mathbb{R},\ x \mapsto \exp(\lambda x)$ ein Eigenvektor (man spricht in diesem Zusammenhang auch von einer *Eigenfunktion*) zum Eigenwert λ. Die Theorie der gewöhnlichen Differenzialgleichungen liefert, dass der Eigenraum E_λ von f_λ erzeugt wird, er ist also eindimensional. Alle $\lambda \in \mathbb{R}$ sind in diesem Beispiel Eigenwerte.

(3) Für eine lineare Abbildung $\varphi\colon V \to V$ ist genau dann 0 ein Eigenwert, wenn φ nicht injektiv ist. Der Eigenraum ist $E_0 = \mathrm{Kern}(\varphi)$. ◁

Im obigen Beispiel haben wir bereits gesehen, dass Eigenräume Unterräume sind. Dies gilt allgemein, wie man leicht nachrechnet. Wir halten fest:

Proposition 17.3. *Für eine Matrix $A \in K^{n\times n}$ bzw. eine lineare Abbildung $\varphi\colon V \to V$ und $\lambda \in K$ ist E_λ ein Unterraum von K^n bzw. von V.*

Wie kann man Eigenwerte einer Matrix $A \in K^{n\times n}$ berechnen? Nach Definition ist $\lambda \in K$ genau dann ein Eigenwert, wenn $E_\lambda \neq \{0\}$, d.h. wenn das homogene LGS

$$(A - \lambda I_n) \cdot x = 0$$

nicht eindeutig lösbar ist. Dies ist nach den Ergebnissen von Abschnitt 16 äquivalent zu $\det(A - \lambda I_n) = 0$. Diese Überlegungen nehmen wir zum Anlass für eine Definition.

Definition 17.4. *Sei $A \in K^{n\times n}$ eine quadratische Matrix. Die* **charakteristische Matrix** *von A ist die Matrix*

$$x \cdot I_n - A \in K[x]^{n\times n}$$

mit Einträgen im Polynomring $K[x]$. Weiter heißt

$$\chi_A := \det(x \cdot I_n - A) \in K[x]$$

das **charakteristische Polynom** *von A.*

Den folgenden Satz haben wir bereits gezeigt.

Satz 17.5. *Die Eigenwerte einer quadratischen Matrix A sind die Nullstellen des charakteristischen Polynoms χ_A.*

Beispiel 17.6. (1) Für $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ gilt

$$\chi_A = \det \begin{pmatrix} x & -1 \\ -1 & x \end{pmatrix} = x^2 - 1,$$

also sind 1 und -1 die (einzigen) Eigenwerte.

(2) Für $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ gilt

$$\chi_A = \det \begin{pmatrix} x & -1 \\ 1 & x \end{pmatrix} = x^2 + 1,$$

also hat A keine Eigenwerte (in $\mathbb{R}$). ◁

Anmerkung 17.7. (a) Das charakteristische Polynom χ_A einer Matrix $A \in K^{n\times n}$ hat den Grad n und es ist **normiert**, d.h. der Koeffizient von x^n ist 1. Mit $A = (a_{i,j})$ gilt genauer

$$\chi_A = x^n - \Big(\sum_{i=1}^{n} a_{i,i}\Big) \cdot x^{n-1} + \cdots + (-1)^n \det(A).$$

Die in er Klammer stehende Summe über die Diagonaleinträge nennt man auch die *Spur* von A.

(b) Zwei ähnliche Matrizen $A, B \in K^{n\times n}$ haben gleiche charakteristische Polynome, denn aus $A = S^{-1}BS$ mit $S \in \mathrm{GL}_n(K)$ folgt

$$\chi_A = \det(xI_n - S^{-1}BS) = \det\left(S^{-1}(xI_n - B)S\right) = \chi_B$$

wegen Korollar 16.15. ◁

Aus Korollar 7.17 ergibt sich, dass eine $n \times n$-Matrix höchstens n Eigenwerte hat. Falls K algebraisch abgeschlossen ist, so hat jede quadratische Matrix über K Eigenwerte.

Im Lichte der bisherigen Überlegungen erscheinen die folgenden zwei Definitionen für die Vielfachheit eines Eigenwertes als natürlich.

Definition 17.8. *Es sei $\lambda \in K$ ein Eigenwert einer Matrix $A \in K^{n\times n}$.*

(a) Die **algebraische Vielfachheit** *$m_a(\lambda)$ von λ ist die Vielfachheit der Nullstelle λ im charakteristischen Polynom χ_A.*

(b) Die **geometrische Vielfachheit** *von λ ist*

$$m_g(\lambda) := \dim\left(E_\lambda\right).$$

Beispiel 17.9. (1) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ hat die Eigenwerte 1 und -1 (siehe Beispiel 17.2). Für beide Eigenwerte sind algebraische- und geometrische Vielfachheit gleich 1.

(2) Für $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ gilt

$$\chi_A = \det \begin{pmatrix} x-1 & -1 \\ 0 & x-1 \end{pmatrix} = (x-1)^2$$

(obere Dreiecksmatrix), also ist $\lambda = 1$ der einzige Eigenwert mit algebraische Vielfachheit $m_a(\lambda) = 2$. Zur Ermittlung der geometrischen Vielfachheit bemerken wir, dass

$$A - I_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

den Rang 1 hat, also $m_g(\lambda) = 1$. ◁

Satz 17.10. *Ist $\lambda \in K$ ein Eigenwert einer Matrix $A \in K^{n\times n}$, so gilt*

$$1 \le m_g(\lambda) \le m_a(\lambda).$$

Beweis. Die erste Ungleichung ist klar, denn für einen Eigenwert gilt $E_\lambda \neq \{0\}$, also $\dim(E_\lambda) \ge 1$.

Zur Beweis der zweiten Ungleichung setzen wir $m := m_g(\lambda)$ und wählen eine Basis $\{v_1, \ldots, v_m\}$ von E_λ. Diese können wir zu einer Basis $B = \{v_1, \ldots, v_n\}$ von K^n ergänzen. Für $1 \le i \le m$ gilt

$$\varphi_A(v_i) = A \cdot v_i = \lambda \cdot v_i,$$

also hat die Darstellungsmatrix von φ_A bzgl. B die Form

$$D_B(\varphi_A) = \left(\begin{array}{ccc|c} \lambda & & 0 & \\ & \ddots & & * \\ 0 & & \lambda & \\ \hline & 0 & & C \end{array}\right) =: D$$

mit $C \in K^{(n-m)\times(n-m)}$. Mit $S := (v_1 \ldots v_n) \in \mathrm{GL}_n(K)$ (die Matrix mit den v_i als Spalten) gilt $S^{-1}AS = D$ (wegen Satz 12.9), wegen Anmerkung 17.7(b) also

$$\chi_A = \chi_D.$$

Die Matrix $xI_n - D$ ist jedoch (ebenso wie D selbst) eine obere Block-Dreiecksmatrix. Damit können wir die Determinante ablesen und erhalten

$$\chi_A = (x-\lambda)^m \cdot \chi_C.$$

Also ist χ_A durch $(x-\lambda)^m$ teilbar, und wir schließen $m_a(\lambda) \geq m$, wie behauptet. □

Definition 17.11. *Eine quadratische Matrix $A \in K^{n\times n}$ heißt* **diagonalisierbar**, *falls es eine Basis von K^n bestehend aus Eigenvektoren von A gibt. Gleichbedeutend: A ist ähnlich zu einer Diagonalmatrix.*

Ebenso kann man von der Diagonalisierbarkeit einer linearen Abbildung $\varphi\colon V \to V$ eines K-Vektorraums V sprechen.

Beispiel 17.12. (1) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ ist diagonalisierbar (siehe Beispiel 17.2).

(2) $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ ist nicht diagonalisierbar. Es fehlen Eigenwerte (siehe Beispiel 17.6(2)).

(3) $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ ist nicht diagonalisierbar. Es fehlen Eigenvektoren (siehe Beispiel 17.9(2)). ◁

Wir werden folgendes Kriterium für Diagonalisierbarkeit beweisen. Es besagt, dass die in Beispiel 17.12(2) und (3) aufgetretenen Hindernisse für die Diagonalisierbarkeit tatsächlich die einzig möglichen Hindernisse sind.

Satz 17.13. *Eine Matrix $A \in K^{n\times n}$ ist genau dann diagonalisierbar, wenn beide der folgenden Bedingungen erfüllt sind:*

(a) Das charakteristische Polynom χ_A zerfällt in Linearfaktoren, also

$$\chi_A = \prod_{i=1}^{r} (x-\lambda_i)^{e_i}$$

mit $e_i = m_a(\lambda_i)$.

(b) Für alle Eigenwerte λ_i gilt

$$m_g(\lambda_i) = m_a(\lambda_i).$$

Das folgende Lemma benötigen wir für den Beweis.

Lemma 17.14. *Es seien $\lambda_1, \ldots, \lambda_r \in K$ paarweise verschiedene Eigenwerte einer Matrix $A \in K^{n\times n}$. Dann ist die Summe $\sum_{i=1}^{r} E_{\lambda_i}$ der Eigenräume direkt.*

Beweis. Wir benutzen Induktion nach r. Für $r = 1$ ist nichts zu zeigen. Wir können also ab jetzt $r \geq 2$ voraussetzen. Zum Nachweis der Direktheit der Summe seien $v_i \in E_{\lambda_i}$ $(i = 1, \ldots, r)$ mit $v_1 + \cdots + v_r = 0$. Wir rechnen:

$$\sum_{i=1}^{r} \lambda_i v_i = \sum_{i=1}^{r} A \cdot v_i = A \cdot \left(\sum_{i=1}^{r} v_i \right) = A \cdot 0 = 0.$$

Andererseits gilt

$$\sum_{i=1}^{r} \lambda_1 v_i = \lambda_1 \cdot \left(\sum_{i=1}^{r} v_i \right) = 0.$$

Wir subtrahieren beide Gleichungen und erhalten

$$\sum_{i=2}^{r} (\lambda_i - \lambda_1) v_i = 0.$$

Da $(\lambda_i - \lambda_1)v_i$ in E_{λ_i} liegt, liefert die Induktionsvoraussetzung $(\lambda_i - \lambda_1)v_i = 0$ für $i \in \{2, \ldots, r\}$. Wegen $\lambda_i \neq \lambda_1$ folgt $v_i = 0$ für $i \in \{2, \ldots, r\}$. Nun folgt auch $v_1 = -(v_2 + \cdots + v_r) = 0$. □

Beweis von Satz 17.13. Zunächst nehmen wir an, dass A diagonalisierbar ist, es gibt also eine Basis B von K^n aus Eigenvektoren. Sind $\lambda_1, \ldots, \lambda_r$ die Eigenwerte von A, so folgt mit $B_i := B \cap E_{\lambda_i}$:

$$n = |B| = \sum_{i=1}^{r} |B_i| \leq \sum_{i=1}^{r} m_g(\lambda_i) \leq \sum_{i=1}^{r} m_a(\lambda_i) \leq \deg(\chi_A) = n,$$

wobei die mittlere Ungleichung aus Satz 17.10 folgt und die letzte aus der Definition der $m_a(\lambda_i)$ als Vielfachheiten der Nullstellen von χ_A folgt. Es muss also überall Gleichheit gelten, und es folgen (a) und (b).

Nun nehmen wir umgekehrt an, dass (a) und (b) gelten. Für $i \in \{1, \ldots, r\}$ sei B_i eine Basis des Eigenraums E_{λ_i}. Wir setzen $B := B_1 \cup \cdots \cup B_r$. Es ist klar, dass B aus Eigenvektoren besteht. Aus Lemma 17.14 folgt, dass B linear unabhängig ist. Außerdem gilt

$$|B| = \sum_{i=1}^{r} |B_i| = \sum_{i=1}^{r} m_g(\lambda_i) \underset{\text{(b)}}{=} \sum_{i=1}^{r} m_a(\lambda_i) \underset{\text{(a)}}{=} \deg(\chi_A) = n.$$

Insgesamt folgt mit Korollar 10.15(a), dass B eine Basis von K^n ist. □

Aus Satz 17.13 und Satz 17.10 erhalten wir ein Kriterium, das in vielen Fällen bereits die Diagonalisierbarkeit einer Matrix garantiert.

Korollar 17.15. *Es sei $A \in K^{n \times n}$. Falls χ_A in Linearfaktoren zerfällt und nur Nullstellen der Vielfachheit 1 hat, so ist A diagonalisierbar.*

Als Anwendung betrachten wir ein physikalisches Beispiel. Wir stellen uns vor, dass zwei gleichschwere Massen mit identischen, masselosen Federn an gegenüberliegenden Wänden verbunden sind, und dass zwischen den Massepunkten eine weitere, andersartige Feder befestigt ist. Man spricht auch von *gekoppelten Schwingern*. Wenn $x_1(t)$ und $x_2(t)$ die Auslenkungen der Massepunkte (gemessen ab er Ruhelage) zur Zeit t bezeichnen, so gelten die Differentialgleichungen

$$\ddot{x}_1(t) = -ax_1(t) - b\left((x_1(t) - x_2(t)\right),$$
$$\ddot{x}_2(t) = -ax_2(t) - b\left((x_2(t) - x_1(t)\right),$$

wobei die Doppelpunkte wie üblich für die zweite Ableitung nach t stehen und die positiven Konstanten a und b von den Federeigenschaften und dem Gewicht der Massepunkte anhängen. In Matrixschreibweise:

$$\begin{pmatrix}\ddot{x}_1\\ \ddot{x}_2\end{pmatrix} = \underbrace{\begin{pmatrix}-a-b & b\\ b & -a-b\end{pmatrix}}_{=:A} \cdot \begin{pmatrix}x_1\\ x_2\end{pmatrix}.$$

Das charakeristische Polynom von A ist

$$\chi_A = \det\begin{pmatrix}x+a+b & -b\\ -b & x+a+b\end{pmatrix} = (x+a+b)^2 - b^2 = (x+a)(x+a+2b).$$

Korollar 17.15 garantiert, dass A diagonalisierbar ist. Die Eigenräume berechnen wir durch Auflösen von homogenen LGS (oder hinschauen):

$$E_{-a} = \langle\begin{pmatrix}1\\1\end{pmatrix}\rangle \quad \text{und} \quad E_{-a-2b} = \langle\begin{pmatrix}1\\-1\end{pmatrix}\rangle.$$

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

$$S^{-1}AS = \begin{pmatrix}-a & 0\\ 0 & -a-2b\end{pmatrix}.$$

Wir setzen $\begin{pmatrix}y_1\\ y_2\end{pmatrix} := S^{-1}\begin{pmatrix}x_1\\ x_2\end{pmatrix}$ und erhalten die Differentialgleichung

$$\begin{pmatrix}\ddot{y}_1\\ \ddot{y}_2\end{pmatrix} = S^{-1}\begin{pmatrix}\ddot{x}_1\\ \ddot{x}_2\end{pmatrix} = S^{-1}A\begin{pmatrix}x_1\\ x_2\end{pmatrix} = S^{-1}AS\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \begin{pmatrix}-a & 0\\ 0 & -a-2b\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}.$$

Die Diagonalisierung der Matrix hat also dazu geführt, dass wir zwei getrennte Differentialgleichungen für y_1 und y_2 bekommen haben. Diese können wir leicht lösen. Mit $\omega := \sqrt{a}$ und $\widetilde{\omega} := \sqrt{a+2b}$ lautet die allgemeine Lösung

$$\begin{pmatrix}y_1(t)\\ y_2(t)\end{pmatrix} = \begin{pmatrix}c_1\cos(\omega t) + c_2\sin(\omega t)\\ c_3\cos(\widetilde{\omega} t) + c_4\sin(\widetilde{\omega} t)\end{pmatrix}$$

mit Konstanten c_i. Durch Multiplikation mit S erhalten wir

$$\begin{pmatrix}x_1(t)\\ x_2(t)\end{pmatrix} = c_1\begin{pmatrix}\cos(\omega t)\\ \cos(\omega t)\end{pmatrix} + c_2\begin{pmatrix}\sin(\omega t)\\ \sin(\omega t)\end{pmatrix} + c_3\begin{pmatrix}\cos(\widetilde{\omega} t)\\ -\cos(\widetilde{\omega} t)\end{pmatrix} + c_4\begin{pmatrix}\sin(\widetilde{\omega} t)\\ -\sin(\widetilde{\omega} t)\end{pmatrix}.$$

Interessant ist die Lösung mit $c_1 = c_3 = 0$ und $c_2 = c_4 = 1$, die (nach ein paar Umformungen)

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = 2 \begin{pmatrix} \cos\left(\frac{\tilde{\omega}-\omega}{2} \cdot t\right) \cdot \sin\left(\frac{\tilde{\omega}+\omega}{2} \cdot t\right) \\ -\sin\left(\frac{\tilde{\omega}-\omega}{2} \cdot t\right) \cdot \cos\left(\frac{\tilde{\omega}+\omega}{2} \cdot t\right) \end{pmatrix}$$

lautet. Diese beschreibt ein periodisches Übertragen der Schwingung von der einen Masse zur anderen und zurück.

Bei der Definition von Polynomen war uns wichtig, Elemente eines größeren Rings in Polynome einsetzen zu können. Nun werden wir Matrizen in Polynome einsetzen.

Beispiel 17.16. Für $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ und $f = x^2 + 1$ gilt

$$f(A) = A^2 + I_2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

◁

Im obigen Beispiel haben wir eine Matrix in ihr eigenes charakteristische Polynom eingesetzt, und heraus kam die Nullmatrix. Der folgende Satz sagt, dass das kein Zufall war.

Satz 17.17 (Satz von Cayley-Hamilton)**.** *Für ein quadratische Matrix $A \in K^{n\times n}$ gilt*

$$\chi_A(A) = 0.$$

Beweis. Wir schreiben $A = (a_{i,j})$ und setzen $B := xI_n - A^T$, also die Transponierte der charakteristischen Matrix. Von B können wir die adjunkte Matrix $C \in K[x]^{n\times n}$ bilden. Satz 16.11 liefert

$$C \cdot B = \det(B) \cdot I_n = \chi_A \cdot I_n.$$

Für $j, k \in \{1, \ldots, n\}$ gilt also (mit $B = (b_{i,j})$ und $C = (c_{i,j})$)

$$\sum_{i=1}^{n} c_{k,i} b_{i,j} = \delta_{j,k} \cdot \chi_A.$$

In diese Gleichungen von Polynomen können wir $x = A$ einsetzen und erhalten

$$\sum_{i=1}^{n} c_{k,i}(A) b_{i,j}(A) = \delta_{j,k} \cdot \chi_A(A). \tag{17.1}$$

Nach Definition von B gilt $b_{i,j}(A) = \delta_{i,j} \cdot A - a_{j,i} \cdot I_n$. Wir schreiben e_j für den j-ten Standardbasisvektor und erhalten

$$\sum_{j=1}^{n} b_{i,j}(A) e_j = A \cdot e_i - \sum_{j=1}^{n} a_{j,i} e_j = 0. \tag{17.2}$$

Für $k \in \{1, \dots, n\}$ folgt

$$\chi_A(A) \cdot e_k = \sum_{j=1}^{n} \delta_{j,k} \cdot \chi_A(A) \cdot e_j \underset{(17.1)}{=} \sum_{i,j=1}^{n} c_{k,i}(A) b_{i,j}(A) e_j \underset{(17.2)}{=} 0,$$

woraus die Behauptung $\chi_A(A) = 0$ folgt. □

18 Die Smith-Normalform

Da es in diesem Abschnitt um ganzzahlige Lösungen geht, könnte man ihn mit einem gewissen Recht der diskreten Mathematik zurechnen. Der Ausgangspunkt der Überlegungen dieses Abschnitts sind ganzzahlige lineare Gleichungssysteme.

Beispiel 18.1. Für welche $b = \binom{b_1}{b_2} \in \mathbb{Z}^2$ ist das ganzzahlige LGS

$$\begin{aligned} 2x_1 + 3x_2 + 4x_3 &= b_1, \\ 5x_1 + 6x_2 + 7x_3 &= b_2, \end{aligned}$$

mit $x_i \in \mathbb{Z}$ lösbar? Wie sieht die Lösungsmenge aus? Was ist die Lösungsmenge für den Fall $b = 0$? Man kann das LGS in Matrixform als $A \cdot x = b$ schreiben mit $A \in \mathbb{Z}^{2\times 3}$. ◁

Die Fragestellungen als diesem Beispiel lassen sich mit der Smith-Normalform der Matrix A beantworten. Um diese zu definieren, werden wir an den Begriff der Äquivalenz von Matrizen (siehe Definition 12.11(b)) auf Matrizen über beliebigen Ringen ausweiten.

Definition 18.2. *Es sei R ein kommutativer Ring.*

(a) Eine quadratische Matrix $A \in R^{n\times n}$ heißt **invertierbar**, *falls $A^{-1} \in R^{n\times n}$ existiert mit $A^{-1} \cdot A = I_n$. Wegen Anmerkung 16.13 ist A genau dann invertierbar, wenn $\det(A) \in R$ ein invertierbares Element von R ist.*
Wir schreiben

$$\mathrm{GL}_n(R) := \{A \in R^{n\times n} \mid A \text{ ist invertierbar}\}$$

für die allgemeine lineare Gruppe über R, die mit dem Matrixprodukt eine Gruppe bildet.

(b) Zwei Matrizen $A, B \in R^{m\times n}$ heißen **äquivalent**, *falls es $S \in \mathrm{GL}_m(R)$ und $T \in \mathrm{GL}_n(R)$ gibt mit*

$$B = SAT.$$

Um dies auszudrücken, benutzen wir die (ad hoc) Schreibweise

$$A \approx B.$$

Beispiel 18.3. (1) Eine Matrix $A \in \mathbb{Z}^{n\times n}$ ist genau dann invertierbar, wenn $\det(A) \in \{1, -1\}$.

(2) Die Matrizen

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix} \quad \text{und} \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \in \mathbb{Z}^{2\times 3}$$

sind äquivalent, denn

$$\underbrace{\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}}_{=:S} \cdot \begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix} \cdot \underbrace{\begin{pmatrix} -1 & 3 & 1 \\ 1 & -2 & -2 \\ 0 & 0 & 1 \end{pmatrix}}_{=:T} = \begin{pmatrix} 2 & 3 & 4 \\ 3 & 3 & 3 \end{pmatrix} \cdot \begin{pmatrix} -1 & 3 & 1 \\ 1 & -2 & -2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix},$$

und man verifiziert anhand der Determinanten, dass S und T über $\mathbb{Z}$ invertierbar sind. ◁

Wir betrachten nun den Fall $R = \mathbb{Z}$. Später werden wir sämtliche Schritte auf den Fall $R = K[x]$ (Polynomring über einem Körper) übertragen. Wir kennzeichnen durch Fußnoten, welche Änderungen für den Übergang von $\mathbb{Z}$ nach $K[x]$ gemacht werden müssen. Diese Fußnoten können beim ersten Lesen des Skripts übergangen werden. Wir erinnern an die Schreibweise $a \mid b$ („a teilt b“).

Definition 18.4. *Es sei $A = (a_{i,j}) \in \mathbb{Z}^{m\times n}$.*

(a) A heißt in **Smith-Normalform**, *falls*

$$A = \begin{pmatrix} d_1 & 0 & \cdots & 0 \cdots 0 \\ 0 & d_2 & & \\ \vdots & & \ddots & \vdots \quad \vdots \\ & & d_{r-1} & \\ 0 & \cdots & 0 \quad d_r & 0 \cdots 0 \end{pmatrix},$$

d.h. $a_{i,j} = \delta_{i,j} \cdot d_i$ mit $d_i \in \mathbb{Z}$ ($i = 1, \ldots, r := \min\{m, n\}$), und falls zusätzlich gelten:

$$d_i \geq 0^{1} \ (i = 1, \ldots, r) \quad \textit{und} \quad d_i \mid d_{i+1} \ (i = 1, \ldots, r-1).$$

(b) Eine Matrix $B \in \mathbb{Z}^{m\times n}$ heißt eine **Smith-Normalform** *von A, falls B in Smith-Normalform und äquivalent zu A ist.*

Beispiel 18.5. In Beispiel 18.3(2) ist B eine Smith-Normalform von A. Wir können damit das LGS aus Beispiel 18.1 behandeln. Wegen $SAT = B$ gilt

$$A \cdot x = b \quad \Longleftrightarrow \quad BT^{-1}x = S \cdot b,$$

[1] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ lautet die Bedingung: d_i ist normiert oder 0.

mit $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} := T^{-1}x$ ergibt sich also in diesem Beispiel das LGS

$$\begin{pmatrix} y_1 \\ 3y_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 - b_1 \end{pmatrix}.$$

Also ist das LGS genau dann lösbar, wenn $b_2 - b_1$ durch 3 teilbar ist, also wenn $b_1 \equiv b_2 \mod 3$. In diesem Fall liefert $y_1 = b_1$ und $y_2 = \frac{b_2 - b_1}{3}$ eine Lösung, also

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = T \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -1 & 3 & 1 \\ 1 & -2 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \frac{b_2 - b_1}{3} \\ c \end{pmatrix} = \begin{pmatrix} b_2 - 2b_1 \\ \frac{5b_1 - 2b_2}{3} \\ 0 \end{pmatrix} + c \cdot \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

mit $c \in \mathbb{Z}$ beliebig. Die Smith-Normalform (zusammen mit den transformierenden Matrizen S und T) liefert also ein Kriterium für die Lösbarkeit und die allgemeine Lösung. Insbesondere ergibt sich für $b = 0$ die Lösungsmenge $\mathbb{Z} \cdot \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$.

Es ist klar, dass dies für beliebige ganzzahlige LGS funktioniert. ◁

Unser nächstes Ziel ist der Nachweis, dass jede ganzzahlige Matrix eine Smith-Normalform besitzt. Danach werden wir zeigen, dass diese eindeutig bestimmt ist. Den Existenzbeweis führen wir, indem wir einen Algorithmus angeben, der eine Matrix in Smith-Normalform bringt. Das entscheidende Hilfsmittel im Algorithmus ist Division mit Rest.

Algorithmus 18.6 (Smith-Normalform)**.**

Eingabe: Eine Matrix $A \in \mathbb{Z}^{m \times n}$.
Ausgabe: Eine Smith-Normalform B von A.

(1) Setze $B := A$, schreibe $B = (b_{i,j})$.
(2) Falls $B = 0$, so ist B in Smith-Normalform und wird ausgegeben.
(3) Wähle $i \in \{1, \ldots, m\}$ und $j \in \{1, \ldots, n\}$ mit $b_{i,j} \neq 0$, so dass der Betrag $|b_{i,j}|$ minimal wird[2].
(4) Vertausche die i-te und die erste Zeile und die j-te und die erste Spalte von B, so dass das Element $\neq 0$ mit minimalem Betrag nun $b_{1,1}$ ist.
(5) Falls $b_{1,1} < 0$, multipliziere die erste Zeile von B mit -1. Danach ist $b_{1,1}$ positiv[3].
(6) Für $j = 2, \ldots, n$ durchlaufe die Schritte 7 bis 9.
(7) Führe Division mit Rest durch:

$$b_{1,j} = b_{1,1} \cdot q + r$$

[2] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ ist der Grad $\deg(b_{i,j})$ zu minimieren.

[3] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ wird mit dem Inversen des höchsten Koeffizienten von $b_{1,1}$ multipliziert, so dass $b_{1,1}$ normiert wird.

mit $q, r \in \mathbb{Z}$, so dass $|r| < |b_{1,1}|$ gilt[4].

(8) Subtrahiere das q-fache der ersten Spalte von der j-ten Spalte. Nun gilt $b_{1,j} = r$.

(9) Falls $b_{1,j} \neq 0$, gehe zu Schritt 3.

(10) Führe die Schritte 6 bis 9 analog für die Zeilen von B durch.

(11) Wenn dieser Schritt erreicht wird, sind außer $b_{1,1}$ alle Einträge der ersten Zeile und Spalte 0.
Falls $m = 1$ oder $n = 1$, so ist B in Smith-Normalform und wird ausgegeben.

(12) Falls $i, j > 1$ existieren, so dass $b_{1,1}$ *kein* Teiler von $b_{i,j}$ ist, addiere die i-te Zeile zur ersten und gehe zu Schritt 6. Eine der Divisionen mit Rest wird nun *nicht* aufgehen.

(13) Berechne durch einen rekursiven Aufruf eine Smith-Normalform D' von $B' = (b_{i,j})_{i,j \geq 2} \in \mathbb{Z}^{(m-1)\times(n-1)}$.

(14) Die Matrix

$$\left(\begin{array}{c|c} b_{1,1} & 0 \cdots 0 \\ \hline 0 & \\ \vdots & D' \\ 0 & \end{array}\right) \in \mathbb{Z}^{m\times n}$$

ist in Smith-Normalform und wird ausgegeben.

Das folgende Lemma brauchen wir für den Nachweis, dass Algorithmus 18.6 tatsächlich eine Smith-Normalform berechnet.

Lemma 18.7. *Die Operationen aus Algorithmus 18.6 lassen sich durch Multiplikation von links bzw. von rechts mit folgenden Matrizen realisieren (mit $k = m$ bzw. $k = n$):*

- $I_k + aE_{i,j}$ *mit* $a \in \mathbb{Z}$, $i, j \in \{1, \ldots, k\}$ *und* $i \neq j$ *(wobei* $E_{i,j} \in \mathbb{Z}^{k\times k}$ *die Matrix mit einer* 1 *als* (i, j)*-ten Eintrag und sonst lauter Nullen ist, siehe auf Seite 125);*
- *die Diagonalmatrix* $\operatorname{diag}(-1, 1, \ldots, 1) \in \mathbb{Z}^{k\times k}$.[5]

Beweis. Dies ist korrekt für die Schritte, bei denen ein Vielfaches einer Zeile oder Spalte zu einer anderen addiert wird. (Dies haben wir auf Seite 125 schon für Zeilen überlegt.) Schritt 4 lässt sich folgendemaßen realisieren: Addition der ersten Zeile zur i-ten, Subtraktion der i-ten Zeile von der ersten, Addition der ersten Zeile zur i-ten, Multiplikation der ersten Zeile mit -1, danach die entsprechenden Operationen mit der ersten und j-ten Spalte. Die Multiplikation der ersten Zeile bzw. Spalte mit -1 entspricht einer Multiplikation mit $\operatorname{diag}(-1, 1, \ldots, 1)$ von links bzw. rechts. Schritt 5 ist damit auch abgedeckt. □

[4] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ wird der Betrag durch den Grad ersetzt.

[5] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ muss man statt der -1 alle konstanten Polynome $\neq 0$ zulassen.

Satz 18.8. *Algorithmus 18.6 terminiert nach endlich vielen Schritten und liefert eine Smith-Normalform von A. Insbesondere besitzt jede Matrix in $\mathbb{Z}^{m\times n}$ eine Smith-Normalform.*

Beweis. Aus Lemma 18.7 folgt, dass die Matrix B zu jeder Zeit während des Algorithmus äquivalent zu A ist.

Jedesmal, wenn die Division durch $b_{1,1}$ einen Rest $r \neq 0$ lässt, wird das minimale $|b_{i,j}|$ mit $b_{i,j} \neq 0$ kleiner. Deshalb wird Schritt 13 irgendwann erreicht. Per Induktion nach $\min\{m,n\}$ folgt, dass der rekursive Aufruf eine Smith-Normalform D' von B' liefert. Wegen der Äquivalenz von B' und D' sind alle Einträge von D' Linearkombinationen der Einträge von B' mit Koeffizienten aus $\mathbb{Z}$. Da die Einträge von B' beim Erreichen von Schritt 13 Vielfache von $b_{1,1}$ sind, folgt dies also auch für die Einträge von D'. Also ist die Matrix in Schritt 14 tatsächlich in Smith-Normalform. □

Man kann Algorithmus 18.6 so variieren, dass die transformierenden Matrizen S und T mitberechnet werden, indem man, ähnlich wie beim Verfahren zur Berechnung einer inversen Matrix aus Seite 99, eine $m \times m$- und eine $n \times n$-Einheitsmatrix mitführt, auf die man alle Zeilen- bzw. Spaltenoperationen ausübt. Wegen Lemma 18.7 erhält man aus diesen am Schluss des Algorithmus die Matrizen S und T. Wir werden dies im Beispiel 18.9(2) durchführen.

Beispiel 18.9. (1) Wir beginnen mit einer relativ großen Matrix. An diesem Beispiel kann man lernen, dass es entscheidend ist, die Matrix-Einträge im Verlauf der Rechnung möglichst klein zu halten. Stures Vorgehen nach Algorithmus 18.6 ließe die Einträge explodieren. Wir betrachten

$$A = \begin{pmatrix} 8 & 2 & 9 & -2 \\ 22 & 2 & 28 & -8 \\ 20 & -6 & 31 & -12 \end{pmatrix} \in \mathbb{Z}^{3\times 4}$$

und rechnen

$$\begin{pmatrix} 8 & 2 & 9 & -2 \\ 22 & 2 & 28 & -8 \\ 20 & -6 & 31 & -12 \end{pmatrix} \xrightarrow[(1)]{} \begin{pmatrix} 8 & 2 & 9 & -2 \\ -2 & -4 & 1 & -2 \\ -4 & -12 & 4 & -6 \end{pmatrix} \xrightarrow[(2)]{} \begin{pmatrix} 1 & -4 & -2 & -2 \\ 9 & 2 & 8 & -2 \\ 4 & -12 & -4 & -6 \end{pmatrix} \xrightarrow[(3)]{}$$

$$\begin{pmatrix} 1 & -4 & -2 & -2 \\ 0 & 38 & 26 & 16 \\ 0 & 4 & 4 & 2 \end{pmatrix} \xrightarrow[(4)]{} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 38 & 26 & 16 \\ 0 & 4 & 4 & 2 \end{pmatrix} \xrightarrow[(5)]{} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 4 & 4 \\ 0 & 16 & 26 & 38 \end{pmatrix} \xrightarrow[(6)]{}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 4 & 4 \\ 0 & 0 & -6 & 6 \end{pmatrix} \xrightarrow[(7)]{} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -6 & 6 \end{pmatrix} \xrightarrow[(8)]{} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 6 & 0 \end{pmatrix}.$$

Die Schritte waren: (1) Subtraktion des 3-fachen der ersten Zeile von der zweiten und dritten, (2) Vertauschung der ersten und zweiten Zeile sowie

der ersten und dritten Spalte, (3) Subtraktion der 9- bzw. 4-fachen der ersten Zeile von der zweiten bzw. dritten, (4) Addition des 4-, 2- bzw. 2-fachen der ersten Spalte zu der zweiten, dritten bzw. vierten, (5) Vertauschung der zweiten und vierten Spalte sowie der zweiten und dritten Zeile, (6) Subtraktion des 8-fachen der zweiten Zeile von der dritten, (7) Subtraktion des 2-fachen der zweiten Spalte von der dritten und vierten und (8) Addition der dritten Spalte zur vierten und Multiplikation der dritten Spalte mit -1. Normalerweise kennzeichnet man diese Schritte direkt an den Matrizen wie in Beispiel 9.4.

(2) Wir betrachten wie in Beispiel 18.3(2) die Matrix

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix} \in \mathbb{Z}^{2\times 3}$$

Bei der Rechnung führen wir eine Einheitmatrix rechts von A und eine weitere unterhalb von A mit, und wenden alle Zeilenoperationen auf die erste und alle Spaltenoperationen auf die zweite mit an.

$$\left(\begin{array}{ccc|cc} 2 & 3 & 4 & 1 & 0 \\ 5 & 6 & 7 & 0 & 1 \\ \hline 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \end{array}\right) \longrightarrow \left(\begin{array}{ccc|cc} 2 & 3 & 1 & 1 & 0 \\ 5 & 6 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & & \\ 0 & 1 & -1 & & \\ 0 & 0 & 1 & & \end{array}\right) \longrightarrow \left(\begin{array}{ccc|cc} 2 & 1 & 1 & 1 & 0 \\ 5 & 1 & 1 & 0 & 1 \\ \hline 1 & -1 & 0 & & \\ 0 & 1 & -1 & & \\ 0 & 0 & 1 & & \end{array}\right) \longrightarrow$$

$$\left(\begin{array}{ccc|cc} 1 & 2 & 0 & 1 & 0 \\ 1 & 5 & 0 & 0 & 1 \\ \hline -1 & 1 & 1 & & \\ 1 & 0 & -2 & & \\ 0 & 0 & 1 & & \end{array}\right) \longrightarrow \left(\begin{array}{ccc|cc} 1 & 2 & 0 & 1 & 0 \\ 0 & 3 & 0 & -1 & 1 \\ \hline -1 & 1 & 1 & & \\ 1 & 0 & -2 & & \\ 0 & 0 & 1 & & \end{array}\right) \longrightarrow \left(\begin{array}{ccc|cc} 1 & 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & -1 & 1 \\ \hline -1 & 3 & 1 & & \\ 1 & -2 & -2 & & \\ 0 & 0 & 1 & & \end{array}\right).$$

Auf die Beschreibung der einzelnen Schritte soll hier verzichtet werden. Wir erhalten die Smith-Normalform $B = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \end{smallmatrix}\right)$ und die transformierenden Matrizen $S = \left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$ und $T = \left(\begin{smallmatrix} -1 & 3 & 1 \\ 1 & -2 & -2 \\ 0 & 0 & 1 \end{smallmatrix}\right)$. In Beispiel 18.3(2) haben wir $SAT = B$ schon nachgerechnet. ◁

Die Bezeichnung „Smith-Normalform“ suggeriert, dass diese eindeutig bestimmt ist. Wir zeigen dies, indem wir die Diagonaleinträge einer Smith-Normalform mit größten gemeinsamen Teilern von Minoren in Verbindung bringen. Den Begriff „größter gemeinsamer Teiler“ (ggT) erläutern wir kurz: Sind $a_1, \ldots, a_n \in \mathbb{Z}$ ganze Zahlen, so heißt eine ganze Zahl $a \geq 0$ ein **größter gemeinsamer Teiler** (**ggT**) von $a_1, \ldots, a_n$, wenn a ein gemeinsamer Teiler der a_i und gleichzeitig ein Vielfaches von jedem anderen gemeinsamen Teiler ist.[6] Nach dieser Definition ist es zunächst gar nicht klar, dass es immer einen ggT gibt. Wenn es aber einen gibt, so ist dieser eindeutig bestimmt,

[6] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ wird statt „$a \geq 0$“ gefordert, dass a normiert oder 0 ist.

denn zwei ggT's von $a_1, \dots, a_n$ müssten sich gegenseitig teilen, sind also wegen der Bedingung „$a \geq 0$“ gleich.

Satz 18.10. *Für $A \in \mathbb{Z}^{m\times n}$ sei $B \in \mathbb{Z}^{m\times n}$ eine Smith-Normalform mit Diagonaleinträgen $d_1, \dots, d_r$ (wobei $r = \min\{m, n\}$). Dann gilt für $k = 1, \dots, r$: Das Produkt $d_1 \cdots d_k$ ist der ggT aller $k \times k$-Minoren von A.*

Insbesondere ist die Smith-Normalform von A eindeutig bestimmt.

Beweis. Wir schreiben $A = (a_{i,j})$ und nehmen ein $k \in \{1, \dots, r\}$. Zunächst zeigen wir, dass sich die Menge der gemeinsamen Teiler der $k \times k$-Minoren von A nicht ändert, wenn A von links mit einer Matrix $S = (s_{i,j}) \in \mathrm{GL}_m(\mathbb{Z})$ multipliziert wird. Wir betrachten zunächst den mit den ersten k Zeilen und Spalten von $S \cdot A$ gebildeten Minor M und erhalten durch dieselbe Rechung wie in (16.4)

$$M = \sum_{\sigma\in S_k} \operatorname{sgn}(\sigma) \cdot \prod_{i=1}^{k} \left(\sum_{j=1}^{m} s_{i,j} a_{j,\sigma(i)} \right) = \\ = \sum_{j_1,\dots,j_k=1}^{m} \left(\prod_{i=1}^{k} s_{i,j_i} \right) \cdot \det(a_{j_t,l})_{t,l=1,\dots,k}.$$

Die $\det(a_{j_t,l})_{t,l=1,\dots,k}$ sind gewisse $k \times k$-Minoren von A, die Gleichung zeigt also, dass jeder gemeinsame Teiler der $k \times k$-Minoren von A auch ein Teiler von M ist. Aus Symmetriegründen (und durch die selbe Rechnung) sehen wir, dass dies auch gilt, wenn M *irgendein* $k \times k$-Minor von $C := S \cdot A$ ist. Jeder gemeinsame Teiler der $k \times k$-Minoren von A ist also auch ein gemeinsamer Teiler der $k \times k$-Minoren von C. Wegen $A = S^{-1}C$ gilt die Umkehrung, also bleibt die Menge der gemeinsamen Teiler aller $k \times k$-Minoren unverändert, wenn man A durch $S \cdot A$ ersetzt. Ebenso bleibt diese Menge unverändert, wenn man A durch $A \cdot S$ mit $S \in \mathrm{GL}_n(\mathbb{Z})$ ersetzt, denn $AS = (S^T A^T)^T$ (transponierte Matrizen), und die Minoren ändern sich beim Transponieren nicht. Es folgt insbesondere, dass die Menge der gemeinsamen Teiler der $k \times k$-Minoren beim Übergang von A zur Smith-Normalform B unverändert bleibt.

Die $k \times k$-Minoren von B sind gleich 0 oder Produkte von k der d_i. Wegen $d_i \mid d_{i+1}$ für $i < r$ folgt: Eine ganze Zahl ist genau dann Teiler aller $k \times k$-Minoren, wenn sie Teiler des Produkts $d_1 \cdots d_k$ ist. Die Menge der gemeinsamen Teiler der $k \times k$-Minoren von B ist also identisch mit der Menge der Teiler von $d_1 \cdots d_k$. Andererseits haben wir gesehen, dass diese Menge identisch ist mit der Menge der gemeinsamen Teiler der $k \times k$-Minoren von A. Also ist $d_1 \cdots d_k$ tatsächlich der ggT der $k \times k$-Minoren von A.

Hieraus folgt sofort die eindeutige Bestimmtheit der Diagonaleinträge bis zu dem kleinsten k, bei dem $d_k = 0$ gilt. Dieses k ist auch eindeutig bestimmt, und wegen $d_k \mid d_i$ für $i > k$ sind alle d_i mit $i > k$ auch 0 und damit ebenso eindeutig bestimmt. □

Nach Satz 18.10 sind die Diagonaleinträge d_i in der Smith-Normalform einer Matrix $A \in \mathbb{Z}^{m\times n}$ eindeutig bestimmt. Man nennt die d_i die **Elementarteiler** (manchmal auch *invariante Faktoren*) von A.

Korollar 18.11. *Zwei Matrizen $A, B \in \mathbb{Z}^{m\times n}$ sind genau dann äquivalent, wenn ihre Elementarteiler übereinstimmen.*

Beweis. Falls $A \approx B$, so ist die Smith-Normalform von A auch eine Smith-Normalform von B, also sind die Smith-Normalformen von A und B identisch. Falls umgekehrt A und B die gleiche Smith-Normalform haben, so sind A und B zu ein und derselben Matrix äquivalent, also $A \approx B$. □

Man kann das Korollar auch so ausdrücken, dass die Äquivalenzklassen von Matrizen in $\mathbb{Z}^{m\times n}$ durch die Elementarteiler klassifiziert werden. Das wichtigste über die Smith-Normalform haben wir nun erarbeitet.

Als Anwendung werden wir nun die Existenz von ggT's nachweisen und den Satz über eindeutige Primzerlegung in $\mathbb{Z}$ herleiten. Wir wenden Satz 18.10 auf ganz bestimmte Matrizen an. Es seien $a_1, \dots, a_n \in \mathbb{Z}$ und $A := (a_1, \dots, a_n) \in \mathbb{Z}^{1\times n}$. Die Smith-Normalform von A hat dann die Form $B = (d, 0, \dots, 0)$, und wegen Satz 18.10 ist d der ggT von $a_1, \dots, a_n$. Wir erhalten also die Existenz von ggT's. Wir schreiben

$$d := \mathrm{ggT}(a_1, \dots, a_n).$$

Da A und B äquivalent sind, folgt insbesondere, dass sich d als $d = x_1a_1 + \cdots + x_na_n$ mit $x_i \in \mathbb{Z}$ darstellen lässt, wobei die x_i aus den transformierenden Matrizen S und T gewonnen werden. Wir haben damit die folgende wichtige Aussage über ganze Zahlen bewiesen.

Proposition 18.12. *Zu $a_1, \dots, a_n \in \mathbb{Z}$ gibt es $x_1, \dots, x_n \in \mathbb{Z}$, so dass*

$$\mathrm{ggT}(a_1, \dots, a_n) = \sum_{i=1}^{n} x_i a_i.$$

Beispiel 18.13. Der ggT von 15 und 21 ist 3, und es gilt $3 = 3 \cdot 15 - 2 \cdot 21$. ◁

Aus Proposition 18.12 können wir den Fundamentalsatz der Arithmetik, d.h. den Satz über die eindeutige Primzerlegung in $\mathbb{Z}$ herleiten. Wir erinnern daran, dass eine ganze Zahl $p > 1$ eine **Primzahl** heißt, wenn 1 und p die einzigen positiven ganzzahligen Teiler von p sind[7].

Satz 18.14 (Fundamentalsatz der Arithmetik)**.** *Jede ganze Zahl $a > 1$ ist Produkt von (nicht notwendig verschiedenen) Primzahlen:*

$$a = p_1 \cdots p_r.$$

[7] Ein normiertes, nicht konstantes Polynom $p \in K[x]$ heißt **Primpolynom**, falls 1 und p die einzigen normierten Teiler von p sind.

Hierbei sind die Primzahlen p_i bis auf die Reihenfolge eindeutig bestimmt.[8]

Beweis. In Satz 3.16 haben wir bereits gezeigt, dass jedes $a > 1$ Produkt von Primzahlen ist.

Für den Beweis der Eindeutigkeit betrachten wir zunächst eine Primzahl p und $b, c \in \mathbb{Z}$ mit $p \mid (b \cdot c)$. Falls p kein Teiler von b ist, so ist 1 der ggT von p und b, also gibt es nach Proposition 18.12 ganze Zahlen x und y mit $1 = xb + yp$. Es folgt

$$c = xbc + ypc,$$

also ist p ein Teiler von c. Wir haben gesehen: Falls eine Primzahl ein Produkt ganzer Zahlen teilt, so teilt sie mindestens einen der Faktoren.

Nun seien $a = p_1 \cdots p_r$ und $a = q_1 \cdots q_s$ zwei Darstellungen von a als Produkte von Primzahlen. Falls $r = 1$ ist, ist a eine Primzahl, also $s = 1$ und $q_1 = p_1$. Wir können also $r > 1$ annehmen. Wegen der obigen Aussage gibt es ein $i \in \{1, \ldots, s\}$ mit $p_1 \mid q_i$, also $p_1 = q_i$, da q_i eine Primzahl ist. Nun folgt $p_2 \cdots p_r = q_1 \cdots q_{i-1} q_{i+1} \cdots q_s$, und der Rest folgt per Induktion nach r. □

Natürlich können wir die Zerlegung einer ganzen Zahl $a > 1$ auch so anordnen, dass gleiche Primzahlen in eine Potenz zusammengefasst werden, also

$$a = \prod_{i=1}^{r} p_i^{e_i} =: \prod_{i=1}^{r} q_i \tag{18.1}$$

mit p_i paarweise verschiedene Primzahlen und $e_i \in \mathbb{N}$. Wir nennen dies eine *Zerlegung von a in Primzahlpotenzen*. Nun ergibt sich auch die Existenz von **kleinsten gemeinsamen Vielfachen** (**kgV**).

In der folgenden Proposition, die (in ihrer Version für Polynome in $K[x]$) in Abschnitt 19 gebraucht wird, geht es um die Elementarteiler von Diagonalmatrizen mit Primzahlpotenzen als Einträgen. Es ist praktisch, einen Elementarteiler einer Matrix als **wesentlich** zu bezeichnen, falls er $\neq 1$ ist. Es ist klar, dass Korollar 18.11 auch gilt, wenn nur die wesentlichen Elementarteiler betrachtet werden.

Proposition 18.15. *Seien $d_1, \ldots, d_r \in \mathbb{Z}$ mit $d_i > 1$ für alle i und $d_i \mid d_{i+1}$ für $i < r$. Sei A die Diagonalmatrix mit den Primzahlpotenzen aus den Zerlegungen der d_i in Primzahlpotenzen als Einträge. Dann sind die d_i die wesentlichen Elementarteiler von A.*

Beweis. Wir betrachten zunächst den Fall $r = 1$, also $d_1 = q_1, \ldots, q_s$ mit q_i paarweise teilerfremde Primzahlpotenzen. Die $(s-1) \times (s-1)$-Minoren von $A = \operatorname{diag}(q_1, \ldots, q_s)$ sind Null oder bis auf Vorzeichen Produkte der $q_1, \ldots, q_s$, bei denen ein Faktor q_i fehlt. Aus Satz 18.14 folgt, dass der ggT dieser Minoren 1 ist. Aus Satz 18.10 folgt, dass die ersten $s-1$ Elementarteiler von A

[8] Beim Ersetzen von $\mathbb{Z}$ durch $K[x]$ lautet der Satz: Jedes nicht konstante, normierte Polynom lässt sich eindeutig (bis auf Reihenfolge) als Produkt von Primpolynomen darstellen.

gleich 1 sind. Das Produkt der Elementarteiler ist aber gleich $\det(A) = d_1$, also muss der letzte (und einzig wesentliche) Elementarteiler d_1 sein.

Nun betrachten wir den Fall $r > 1$. Es seien $d_i = \prod_{j=1}^{s_i} q_{i,j}$ die Zerlegungen in Primzahlpotenzen. Mit $A_i := \operatorname{diag}(q_{i,1}, \ldots, q_{i,s_i})$ folgt die Äquivalenz

$$A_i \approx \operatorname{diag}(1, \ldots, 1, d_i)$$

aus dem Fall $r = 1$, also

$$A = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_r \end{pmatrix} \approx \operatorname{diag}(1, \ldots, 1, d_1, \ldots, d_r).$$

Da die rechte Matrix in Smith-Normalform ist, folgt die Behauptung. □

Bereits zu Beginn des Abschnitts haben wir angekündigt, dass sich die gesamte in diesem Abschnitt entwickelte Mathematik von $\mathbb{Z}$ auf den Polynomring $K[x]$ über einem Körper K überträgt. Was haben diese beiden Ringe gemeinsam? Beides sind kommutative Ringe, in denen es eine Division mit Rest gibt (siehe Satz 7.14). Division mit Rest ist die entscheidende Technik, die den Algorithmus 18.6 zum Laufen bringt. Wir haben durch Fußnoten gekennzeichnet, welche Änderungen beim Übergang von $\mathbb{Z}$ zu $K[x]$ zu machen sind. Statt des Betrags einer ganzen Zahl wird der Grad eines Polynoms betrachtet. Den positiven ganzen Zahlen entsprechen die normierten Polynome. Mit diesen Änderungen zieht sich die gesamte Theorie durch. Matrizen in $K[x]^{m\times n}$ haben also eindeutig bestimmte Smith-Normalformen. Die Elementarteiler sind normierte Polynome oder 0. Auch die Existenz von ggT's und der Satz über eindeutige Primzerlegung übertragen sich.

Beispiel 18.16. Wir betrachten die charakteristische Matrix $xI_3 - A$ von

$$A = \begin{pmatrix} -3 & -1 & 2 \\ 4 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

und bringen sie mit folgenden Schritten in Smith-Normalform:

$$\begin{pmatrix} x+3 & 1 & -2 \\ -4 & x-1 & 4 \\ 0 & 0 & x+1 \end{pmatrix} \xrightarrow[(1)]{} \begin{pmatrix} 1 & x+3 & -2 \\ x-1 & -4 & 4 \\ 0 & 0 & x+1 \end{pmatrix} \xrightarrow[(2)]{}$$

$$\begin{pmatrix} 1 & x+3 & -2 \\ 0 & -x^2-2x-1 & 2x+2 \\ 0 & 0 & x+1 \end{pmatrix} \xrightarrow[(3)]{} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -(x+1)^2 & 2(x+1) \\ 0 & 0 & x+1 \end{pmatrix} \xrightarrow[(4)]{}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & x+1 & 0 \\ 0 & 2(x+1) & -(x+1)^2 \end{pmatrix} \xrightarrow[(5)]{} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x+1 & 0 \\ 0 & 0 & (x+1)^2 \end{pmatrix}.$$

Die Schritte waren: (1) Vertauschung der ersten und zweiten Spalte, (2) Addition des $-(x-1)$-fachen der ersten Zeile zur zweiten, (3) Addition des $-(x+3)$- bzw. 2-fachen der ersten Spalte zur zweiten bzw. dritten, (4) Vertauschung der zweiten und dritten Spalte und der zweiten und dritten Zeile, (5) Addition des -2-fachen der zweiten Zeile zur dritten und Multiplikation der dritten Spalte mit -1.

Die wesentlichen Elementarteiler der charakteristischen Matrix $xI_3 - A$ sind also $x+1$ und $(x+1)^2$. ◁

Wir haben gesehen, dass die Mathematik dieses Abschnitts für die Ringe $\mathbb{Z}$ und $K[x]$ entwickelbar ist. Der gemeinsame Oberbegriff dieser beiden Ringe ist der Begriff eines **euklidischen Rings**. Euklidische Ringe werden (etwas grob gesagt) definiert als kommutative Ringe, bei denen Division mit Rest möglich ist. Der Rest muss dabei bezüglich einer geeigneten Bewertung (in unseren Beispielen Betrag einer ganzen Zahl bzw. Grad eines Polynoms) kleiner sein als der Divisor. Weitere Beispiele für euklidische Ringe sind:

- Der Ring $R = \left\{a + b\sqrt{-1} \mid a, b \in \mathbb{Z}\right\} \subseteq \mathbb{C}$ der *Gaußschen ganzen Zahlen* mit
$$R \to \mathbb{N},\ a + b\sqrt{-1} \mapsto a^2 + b^2$$
als Bewertungsfunktion.
- Jeder Körper K mit
$$K \to \mathbb{N},\ a \mapsto \begin{cases} 1 & \text{falls } a \neq 0, \\ 0 & \text{sonst} \end{cases}$$
als Bewertungsfunktion.

Ein Beispiel für einen nicht euklidischen Ring ist der Polynomring $\mathbb{Z}[x]$ über $\mathbb{Z}$. Dies kann man beispielsweise daran sehen, dass die Matrix $(2, x) \in \mathbb{Z}[x]^{1\times 2}$ keine Smith-Normalform besitzt.

19 Die Jordansche Normalform und allgemeine Normalform

In diesem Abschnitt geht es um die Frage, wie man eine quadratische Matrix umformen kann in eine ähnliche Matrix, die eine möglichst übersichtliche Gestalt hat. Dies ist gleichbedeutend zu der Frage, wie man zu einer linearen Abbildung $\varphi\colon V \to V$ eines endlich-dimensionalen Vektorraums V eine Basis B von V finden kann, so dass die Darstellungsmatrix $D_B(\varphi)$ übersichtlich wird. Dies Thema wurde schon im Abschnitt 17 unter dem Stichwort „Diagonalisierbarkeit“ angeschnitten. Wir werden in jeder Ähnlichkeitsklasse von Matrizen in $K^{n\times n}$ einen „Standardvertreter“ finden und somit die Ähnlichkeitsklassen klassifizieren. Dieser Standardvertreter wird die allgemeine Normalform oder, falls das charakteristische Polynom in Linearfaktoren zerfällt, die Jordansche Normalform genannt. Im Falle einer diagonalisierbaren Matrix wird die Jordansche Normalform eine Diagonalmatrix sein.

Die Ergebnisse des vorherigen Abschnitts werden eine zentrale Rolle spielen. Dort ging es um Äquivalenz von Matrizen, nicht um Ähnlichkeit. Die Brücke zwischen beiden Begriffen wird durch den folgenden, erstaunlichen Satz gebildet. Wie zuvor steht in diesem Abschnitt K immer für einen Körper.

Satz 19.1. *Zwei quadratische Matrizen über K sind genau dann ähnlich, wenn ihre charakteristischen Matrizen äquivalent sind.*

Beweis. Es seien $A, B \in K^{n\times n}$. Zunächst setzen wir voraus, dass A und B ähnlich sind, und leiten daraus die Äquivalenz der charakteristischen Matrizen $xI_n - A$ und $xI_n - B$ her. Es gibt $S \in \mathrm{GL}_n(K)$ mit $S^{-1}AS = B$, also

$$S^{-1}(xI_n - A)S = S^{-1}xI_nS - S^{-1}AS = xI_n - B,$$

also in der Tat $xI_n - A \approx xI_n - B$.

Umgekehrt setzen wir nun die Äquivalenz von $xI_n - A$ und $xI_n - B$ voraus und zeigen die Ähnlichkeit von A und B. Dies ist der schwierigere Teil des Beweises. Wir haben also $S, T \in \mathrm{GL}_n(K[x])$, so dass

$$xI_n - A = S \cdot (xI_n - B) \cdot T. \tag{19.1}$$

Ist $C \in K[x]^{n\times n}$ irgendeine Matrix mit Einträgen in $K[x]$, so können wir schreiben $C = \sum_{i=0}^{m} x^i C_i$ mit $C_i \in K^{n\times n}$ und definieren

$$C(A) := \sum_{i=0}^{m} A^i C_i \in K^{n\times n}. \tag{19.2}$$

Für jede weitere Matrix $D \in K[x]^{n\times n}$ mit $D = \sum_{j=0}^{k} x^j D_j$ (wobei $D_j \in K^{n\times n}$) gelten dann die Regeln

$$(C + D)(A) = C(A) + D(A), \tag{19.3}$$

$$\begin{aligned}(C \cdot D)(A) &= \left(\sum_{i=0}^{m}\sum_{j=0}^{k} x^{i+j} C_i D_j\right)(A) = \sum_{i,j} A^{i+j} C_i D_j \\ &= \sum_{j=0}^{k} A^j \left(\sum_{i=0}^{m} A^i C_i\right) \cdot D_j = (C(A) \cdot D)(A)\end{aligned} \tag{19.4}$$

und

$$C \in K^{n\times n} \quad \Longrightarrow \quad C(A) = C. \tag{19.5}$$

Es gilt

$$(xI_n - A)(A) = AI_n - A = 0,$$

wegen (19.4) also

$$0 = \left((xI_n - A)\cdot T^{-1}\right)(A) \underset{(19.1)}{=} (S\cdot(xI_n - B))(A) \underset{(19.3)}{=} (xS)(A) - (SB)(A)$$
$$\underset{(19.4)}{=} A\cdot S(A) - (S(A)\cdot B)(A) \underset{(19.5)}{=} A\cdot S(A) - S(A)\cdot B$$

und damit $A \cdot S(A) = S(A) \cdot B$. Per Induktion ergibt sich hieraus

$$A^i \cdot S(A) = S(A) \cdot B^i \tag{19.6}$$

für alle $i \in \mathbb{N}$. Wir zeigen nun, dass $S(A)$ invertierbar ist. Wegen $S \in \mathrm{GL}_n(K[x])$ gibt es $C \in K[x]^{n\times n}$ mit $S\cdot C = I_n$. Wir schreiben $C = \sum_{i=0}^{m} x^i C_i$ mit $C_i \in K^{n\times n}$ und erhalten

$$\begin{aligned}I_n &\underset{(19.5)}{=} I_n(A) = (S\cdot C)(A) \underset{(19.4)}{=} (S(A)\cdot C)(A) \\ &= \sum_{i=0}^{m} A^i S(A) C_i \underset{(19.6)}{=} S(A)\cdot \sum_{i=0}^{m} B^i C_i = S(A)\cdot C(B).\end{aligned}$$

Wie behauptet folgt also $S(A) \in \mathrm{GL}_n(K)$, und aus (19.6) erhalten wir

$$S(A)^{-1}\cdot A\cdot S(A) = B.$$

Also sind A und B in der Tat ähnlich. □

Aus dem Beweis sieht man, wie man aus Matrizen $S, T \in \mathrm{GL}_n(K[x])$ mit (19.1) eine Matrix gewinnt, die die Ähnlichkeit von A und B „realisiert“: Mit $R := S(A)$ (gebildet gemäß (19.2)) gilt nämlich

$$R^{-1}AR = B.$$

Mit Korollar 18.11 (übertragen auf den Fall von Matrizen mit Einträgen in $K[x]$) erhalten wir:

Korollar 19.2. *Zwei quadratische Matrizen über K sind genau dann ähnlich, wenn ihre charakteristischen Matrizen dieselben (wesentlichen) Elementarteiler haben.*

Man kann die Ähnlichkeitsklasse einer quadratischen Matrix also an den Elementarteilern der charakteristischen Matrix ablesen. Die Aufgabe, in derer Ähnlichkeitsklasse einen „übersichtlichen“ Vertreter zu finden, reduziert sich nun darauf, zu einer gegebenen Folge von Elementarteilern eine „übersichtliche“ Matrix zu finden, deren charakteristische Matrix genau diese Elementarteiler hat.

Beispiel 19.3. Wir betrachten

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Wir könnten die Elementarteiler der charakteristischen Matrix

$$xI_3 - A = \begin{pmatrix} x+1 & 0 & 0 \\ 0 & x+1 & 0 \\ 0 & -1 & x+1 \end{pmatrix}$$

berechnen, indem wir sie auf Smith-Normalform bringen. Alternativ wählen wir den Weg, die ggT's der Minoren zu berechnen und daraus die Elementarteiler gemäß Satz 18.10 zu gewinnen. Wegen des Eintrags -1 haben die 1×1-Minoren den ggT 1. Man sieht außerdem, dass der ggT der 2×2-Minoren $x+1$ ist. Die Determinante ist $(x+1)^3$, und wir erhalten die wesentlichen Elementarteiler $x+1$ und $(x+1)^2$. Ein Vergleich mit Beispiel 18.16 zeigt, dass die charakteristische Matrix der dort betrachteten Matrix dieselben wesentlichen Elementarteiler hat. Nach Korollar 19.2 sind also

$$\begin{pmatrix} -3 & -1 & 2 \\ 4 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

ähnlich. Die zweite Matrix ist hierbei übersichtlicher. Sie ist ein Beispiel für ein Matrix in Jordanscher-Normalform, die wir in Kürze definieren werden. ◁

Die folgende Definition ist Bestandteil unseres Projekts, übersichtliche Matrizen zu finden, deren charakteristische Matrizen vorgegebenen Elementarteiler haben. Wir erinnern daran, dass ein Primpolynom ein normiertes, nicht konstantes Polynom $f \in K[x]$ ist, dessen einzige normierten Teiler 1 und f selbst sind. Beispielsweise ist jedes Polynom der Form $x-a$ ein Primpolynom, und $x^2+1 \in \mathbb{R}[x]$ ist ein Primpolynom.

Definition 19.4. *(a) Sei $f = x^n - a_{n-1}x^{n-1} - \cdots - a_1x - a_0 \in K[x]$ ein nicht konstantes, normiertes Polynom. Dann heißt*

$$B_f := \begin{pmatrix} 0 & & & 0 & a_0 \\ 1 & \ddots & & & a_1 \\ & & \ddots & 0 & \vdots \\ & & & 1 & a_{n-1} \end{pmatrix} \in K^{n\times n}$$

die **Begleitmatrix** *von f. Besonders wichtig ist der Fall $f = x - a$, in dem B_f nichts weiter als eine 1×1-Matrix mit dem Eintrag a ist.*

(b) Ist $f \in K[x]$ wie in (a) und $e \in \mathbb{N}_{>0}$ eine positive ganze Zahl, so setzen wir

$$B_f^{(e)} := \begin{pmatrix} B_f & & & & 0 \\ 1 & B_f & & & \\ & 1 & \ddots & & \\ & & \ddots & B_f & \\ 0 & & & 1 & B_f \end{pmatrix} \in K^{en\times en}.$$

$B_f^{(e)}$ ist also eine Block-Diagonalmatrix mit e identischen Blöcken B_f und zusätzlich Einsern an den Positionen links unterhalb der Berührpunkte der Blöcke. Für $e = 1$ ist $B_f^{(1)} = B_f$. In dem wichtigen Spezialfall $f = x - a$ heißt

$$B_{x-a}^{(e)} = \begin{pmatrix} a & & & 0 \\ 1 & a & & \\ & \ddots & & \\ & & a & \\ 0 & & 1 & a \end{pmatrix} \in K^{e\times e}$$

ein **Jordan-Kästchen**. *Es hat a als Diagonaleinträge und Einsen in der unteren Nebendiagonalen. (Manchmal werden Jordan-Kästchen auch mit Einsen auf der oberen Nebendiagonalen definiert; dies ist eine Frage der Konvention.)*

(c) Eine quadratische Matrix $A \in K^{n\times n}$ heißt in **allgemeiner Normalform**, *falls*

$$A = \begin{pmatrix} \boxed{B_{f_1}^{(e_1)}} & & & 0 \\ & \boxed{B_{f_2}^{(e_2)}} & & \\ & & \ddots & \\ 0 & & & \boxed{B_{f_s}^{(e_s)}} \end{pmatrix} =: \operatorname{diag}\left(B_{f_1}^{(e_1)}, \ldots, B_{f_s}^{(e_s)}\right)$$

eine Block-Diagonalmatrix ist mit Matrizen $B_{f_i}^{(e_i)}$ als Blöcke, wobei die $f_i \in K[x]$ Primpolynome sind. Falls alle f_i den Grad 1 haben (falls also die $B_{f_i}^{(e_i)}$ Jordan-Kästchen sind), so heißt A in **Jordanscher Normalform**.

(d) Sei $A \in K^{n\times n}$ eine quadratische Matrix. Eine Matrix $B \in K^{n\times n}$ heißt eine **allgemeine Normalform** *von A, falls B in allgemeiner Normalform und ähnlich zu A ist. Falls B sogar in Jordanscher Normalform ist, so heißt sie eine* **Jordansche Normalform** *von A.*

Beispiel 19.5. (1) Die Begleitmatrix eines normierten Polynoms $f = x^2 - ax - b$ von Grad 2 ist

$$B_f = \begin{pmatrix} 0 & b \\ 1 & a \end{pmatrix}$$

(2) Die Matrizen

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

sind in Jordanscher Normalform, die Matrix

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

aber nicht.

(3) Wegen Beispiel 19.3 hat

$$A = \begin{pmatrix} -3 & -1 & 2 \\ 4 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

die Matrix

$$B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

als Jordansche Normalform.

(4) Über $K = \mathbb{R}$ ist $x^2 + x + 1$ ein Primpolynom, also sind die Matrizen

$$\begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

in allgemeiner Normalform. ◁

Lemma 19.6. *Es sei $f \in K[x]$ ein nicht konstantes, normiertes Polynom und $e \in \mathbb{N}_{>0}$.*

(a) Das charakteristische Polynom der Begleitmatrix B_f ist $\chi_{B_f} = f$.

(b) Die charakteristische Matrix von $B_f^{(e)}$ hat den einzigen wesentlichen Elementarteiler f^e.

Beweis. (a) Wir schreiben $f = x^n - a_{n-1}x^{n-1} - \cdots - a_1 x - a_0$ und $A := B_f$. Für die Standardbasisvektoren e_i mit $1 \le i \le n-1$ gilt $A \cdot e_i = e_{i+1}$, also

$$A^i \cdot e_1 = e_{1+i} \quad (i = 0, \ldots, n-1). \tag{19.7}$$

Weiter gilt

$$A^n \cdot e_1 \underset{(19.7)}{=} A \cdot e_n = \sum_{i=0}^{n-1} a_i e_{i+1} \underset{(19.7)}{=} \sum_{i=0}^{n-1} a_i A^i \cdot e_1.$$

Es folgt

$$f(A) \cdot e_1 = 0.$$

Andererseits folgt aus dem Satz von Cayley-Hamilton (Satz 17.17) mit $g := \chi_A$ die Beziehung $g(A) \cdot e_1 = 0$. Da f und g normiert vom Grad n sind, können wir $f - g = \sum_{i=0}^{n-1} b_i x^i$ mit $b_i \in K$ schreiben, und es folgt

$$0 = (f-g)(A) \cdot e_1 = \sum_{i=0}^{n-1} b_i A^i \cdot e_1 \underset{(19.7)}{=} \sum_{i=0}^{n-1} b_i e_{1+i},$$

also $b_i = 0$ für alle i und damit $g = f$. Dies war zu zeigen.

(b) Wenn wir in der charakteristischen Matrix $xI_m - B_f^{(e)}$ (mit $m := en$) die erste Zeile und die letzte Spalte streichen, erhalten wir eine untere Dreiecksmatrix mit dem Eintrag -1 überall auf der Diagonalen. Also tritt $(-1)^{m-1}$ als einer der $(m-1) \times (m-1)$-Minoren auf. Es folgt, dass 1 der ggT der $(m-1) \times (m-1)$-Minoren ist. Wegen Satz 18.10 (in der Version für Matrizen über $K[x]$) folgt, dass die ersten $m-1$ Elementarteiler 1 sind.

Der letzte Elementarteiler muss daher gleich der Determinante von $xI_m - B_f^{(e)}$ sein. Dies ist eine untere Block-Dreiecksmatrix mit Diagonalblöcken $xI_n - B_f$. Wegen (a) ist der gesuchte letzte Elementarteiler also f^e. □

Wir kommen nun zum Hauptergebnis dieses Abschnitts, dass jede quadratische Matrix eine allgemeine Normalform besitzt. Der Satz 18.14 über eindeutige Primzerlegung überträgt sich auf Polynome. Insbesondere kann man bei der Primzerlegung eines nicht konstanten, normierten Polynoms $f \in K[x]$ jeweils gleiche Primpolynome f_i zu Potenzen zusammenfassen und erhält so eine Zerlegung

$$f = \prod_{i=1}^{s} f_i^{e_i}$$

in Primpolynompotenzen.

Satz 19.7. *Sei $A \in K^{n\times n}$ eine quadratische Matrix.*

(a) A hat eine allgemeine Normalform. Anders gesagt: A ist ähnlich zu einer Matrix B in allgemeiner Normalform.

(b) A hat genau dann eine Jordansche Normalform, wenn das charakteristische Polynom χ_A in Linearfaktoren zerfällt. Falls K algebraisch abgeschlossen ist (z.B. $K = \mathbb{C}$), so hat also jede quadratische Matrix eine Jordansche Normalform. Die Diagonaleinträge der Jordanschen Normalform sind die Eigenwerte von A.

Beweis. (a) Es seien $d_1, \dots, d_r \in K[x]$ die wesentlichen Elementarteiler von $xI_n - A$, und $f_1^{e_1}, \dots, f_s^{e_s}$ seien die Primpolynompotenzen aus der Zerlegung der d_i. Wir bilden die Block-Diagonalmatrix

$$B = \operatorname{diag}\left(B_{f_1}^{(e_1)}, \dots, B_{f_s}^{(e_s)}\right)$$

also eine Matrix in allgemeiner Normalform. Jedes $B_{f_1}^{(e_i)}$ hat $e_i \cdot \deg(f_i)$ Zeilen und Spalten, wegen

$$\sum_{i=1}^{s} e_i \deg(f_i) = \deg\left(\prod_{i=1}^{s} f_i^{e_i}\right) = \deg\left(\prod_{i=1}^{r} d_i\right) = \deg(\chi_A) = n$$

gilt $B \in K^{n\times n}$. Wegen Lemma 19.6(b) gilt die Äquivalenz

$$xI_n - B \approx \operatorname{diag}\left(1, \dots, 1, f_1^{e_1}, \dots, f_s^{e_s}\right).$$

Wegen Proposition 18.15 (in der Version für Polynome in $K[x]$) gilt weiter

$$\operatorname{diag}\left(f_1^{e_1}, \dots, f_s^{e_s}\right) \approx \operatorname{diag}\left(1, \dots, 1, d_1, \dots, d_r\right),$$

insgesamt also $xI_n - B \approx \operatorname{diag}(1, \ldots, 1, d_1, \ldots, d_r)$. Dies bedeutet, dass $d_1, \ldots, d_r$ die wesentlichen Elementarteiler von $xI_n - B$ sind. Aus Korollar 19.2 folgt, dass A ähnlich zu B ist.

(b) Falls χ_A in Linearfaktoren zerfällt, so gilt dies wegen $d_1 \cdots d_r = \chi_A$ auch für die Elementarteiler d_i. Die f_i aus dem Beweis von (a) haben also den Grad 1, also ist B in Jordanscher Normalform, und die Diagonaleinträge sind die Nullstellen von χ_A, also die Eigenwerte.
Falls umgekehrt A ähnlich ist zu einer Matrix B in Jordanscher Normalform, so folgt $\chi_A = \chi_B$ (siehe Anmerkung 17.7(b)), und χ_B zerfällt in Linearfaktoren, denn die charakteristische Matrix $xI_n - B$ ist eine untere Dreiecksmatrix mit normierten Polynomen vom Grad 1 auf der Diagonalen. □

Wir können mit Hilfe der Elementarteiler auch die Eindeutigkeit der allgemeinen Normalform beweisen.

Satz 19.8. *Die allgemeine Normalform einer quadratischen Matrix $A \in K^{n\times n}$ ist bis auf die Reihenfolge der Blöcke eindeutig bestimmt.*

Genauer gilt: Die Blöcke $B_{f_i}^{(e_i)}$ der allgemeinen Normalform gehören zu den Primpolynompotenzen $f_i^{e_i}$, die in der Zerlegung der wesentlichen Elementarteiler der charakteristischen Matrix $xI_n - A$ auftreten.

Beweis. Es sei $B = \operatorname{diag}\left(B_{f_1}^{(e_1)}, \ldots, B_{f_s}^{(e_s)}\right)$ eine Matrix in allgemeiner Normalform, die zu A ähnlich ist. Wegen Satz 19.1 und Lemma 19.6 folgt

$$xI_n - A \approx xI_n - B \approx \operatorname{diag}(1, \ldots, 1, f_1^{e_1}, \ldots, f_s^{e_s}).$$

Aus der Liste von Primpolynompotenzen $f_i^{e_i}$ bilden wir nun wie folgt eine Sequenz $d_1, \ldots, d_r$ von Polynomen: Zunächst sei d_1 das kleinste gemeinsame Vielfache der $f_i^{e_i}$. Die Zerlegung von d_1 in Primpolynompotenzen besteht aus einigen der $f_i^{e_i}$, die wir aus nun der Liste streichen. Von den verbleibenden $f_i^{e_i}$ bilden wir erneut das kgV und setzen es d_2. So fahren wir fort, bis alle $f_i^{e_i}$ abgearbeitet sind. Die $f_i^{e_i}$ sind nun genau die Primpolynompotenzen, die in der Zerlegung der d_j auftreten. Außerdem ist jedes d_j ein Vielfaches des nachfolgenden. Indem wir die Reihenfloge der d_j umdrehen, erreichen wir also $d_j \mid d_{j+1}$ für $j < r$. Wegen Proposition 18.15 (in der Version für Polynome in $K[x]$) folgt

$$\operatorname{diag}(f_1^{e_1}, \ldots, f_s^{e_s}) \approx \operatorname{diag}(1, \ldots, 1, d_1, \ldots, d_r).$$

Zusammen mit der obigen Äquivalenz ergibt sich, dass $xI_n - A$ die Smith-Normalform $\operatorname{diag}(1, \ldots, 1, d_1, \ldots, d_r)$ hat, also sind die d_j die wesentlichen Elementarteiler von $xI_n - A$. Damit ist der Satz bewiesen. □

Zum Berechnen der allgemeinen Normalform kann man also Algorithmus 18.6 auf die charakteristische Matrix anwenden und erhält die Elementarteiler. Aus deren Zerlegung in Primpolynompotenzen geht dann die allgemei-

ne Normalform hervor. Dies Berechnungsverfahren ist allerdings aufwändig. Wesentlich schneller geht es mit gewissen Rang-Formeln, die wir hier für den Fall der Jordanschen Normalform besprechen möchten. Da wir die Eindeutigkeit der allgemeinen Normalform nachgewiesen haben, werden wir bei dem nächsten Satz auf einen Beweis verzichten.

Satz 19.9. *Es sei $A \in K^{n\times n}$ eine quadratische Matrix, für die es eine Jordansche Normalform gibt. Für jeden Eigenwert λ von A gelten dann:*

(a) Für $e \in \mathbb{N}_{>0}$ ist

$$c_e(\lambda, A) := \operatorname{rg}\left((A-\lambda I_n)^{e-1}\right) - 2\operatorname{rg}\left((A-\lambda I_n)^{e}\right) + \operatorname{rg}\left((A-\lambda I_n)^{e+1}\right)$$

die Anzahl der Jordan-Kästchen der Länge e zum Eingenwert λ.

(b) Die Gesamtlänge der Jordan-Kästchen zum Eigenwert λ ist gleich der algebraischen Vielfachheit des Eigenwerts λ.

(c) Die Anzahl der Jordan-Kästchen zum Eigenwert λ ist gleich der geometrischen Vielfachheit des Eigenwerts λ.

Wir fassen die Methode zur Berechnung der Jordanschen Normalform, die sich aus Satz 19.9 ergibt, zusammen.

Der erste Schritt ist die Berechnung des charakteristischen Polynoms χ_A und das Auffinden der Nullstellen. Wir setzen voraus, dass χ_A in Linearfaktoren zerfällt. Damit sind die Eigenwerte und deren algebraische Vielfachheiten bekannt. Hat ein Eigenwert λ die algebraische Vielfachheit 1, so gibt es zu λ genau ein Jordan-Kästchen der Länge 1, also einen Diagonaleintrag λ in der Jordanschen Normalform ohne Einsen in der Nebendiagonalen. Bei algebraischer Vielfachheit > 1 berechnet man die geometrische Vielfachheit, also $n - \operatorname{rg}(A-\lambda I_n)$. Damit kennt man die Anzahl der Jordan-Kästchen zum Eigenwert λ, womit man zusammen mit der Kenntnis der Gesamtlänge (= algebraische Vielfachheit) häufig schon deren Längen bestimmen kann. Falls das nicht geht, muss man die Ränge der Matrizen $(A-\lambda I_n)^k$ berechnen und daraus die $c_e(\lambda, A)$ gemäß Satz 19.9(a). Das macht man solange, bis man aufgrund der Kenntnis der Gesamtlänge die Längen aller Jordan-Kästchen zum Eigenwert λ bestimmt hat. Auf diese Art arbeitet man alle Eigenwerte λ ab.

Beispiel 19.10. (1) Wir betrachten nochmals die Matrix

$$A = \begin{pmatrix} -3 & -1 & 2 \\ 4 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3},$$

deren Jordansche Normalform wir eigentlich schon kennen (siehe Beispiel 19.5(3)). Das charakteristische Polynom ist

$$\chi_A = \det \begin{pmatrix} x+3 & 1 & -2 \\ -4 & x-1 & 4 \\ 0 & 0 & x+1 \end{pmatrix} = (x+1) \cdot \det \begin{pmatrix} x+3 & 1 \\ -4 & x-1 \end{pmatrix}$$
$$= (x+1) \cdot (x^2+2x+1) = (x+1)^3,$$

wobei wir im ersten Schritt nach der dritten Zeile entwickelt haben. Der einzige Eigenwert ist also $\lambda = -1$ mit algebraischer Vielfachheit 3. Der Rang von

$$A + I_3 = \begin{pmatrix} -2 & -1 & 2 \\ 4 & 2 & -4 \\ 0 & 0 & 0 \end{pmatrix}$$

ist 1, also gibt es zwei Jordan-Kästchen. Da die Gesamtlänge 3 ist, müssen sie die Länge 1 und 2 haben, die Jordansche Normalform ist also

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

(2) Wir betrachten die Matrix

$$A = \begin{pmatrix} -3 & -1 & 4 & -3 & -1 \\ 1 & 1 & -1 & 1 & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 4 & 1 & -4 & 5 & 1 \\ -2 & 0 & 2 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{5\times 5}.$$

Das Berechnen des charakteristischen Polynoms ist aufwändig:

$$\chi_A = \det \begin{pmatrix} x+3 & 1 & -4 & 3 & 1 \\ -1 & x-1 & 1 & -1 & 0 \\ 1 & 0 & x-2 & 0 & 0 \\ -4 & -1 & 4 & x-5 & -1 \\ 2 & 0 & -2 & 2 & x-1 \end{pmatrix}$$
$$\underset{(1)}{=} \det \begin{pmatrix} x+3 & 1 & -x^2-x+2 & 3 & 1 \\ -1 & x-1 & x-1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ -4 & -1 & 4x-4 & x-5 & -1 \\ 2 & 0 & -2x+2 & 2 & x-1 \end{pmatrix}$$
$$\underset{(2)}{=} \det \begin{pmatrix} 1 & -x^2-x+2 & 3 & 1 \\ x-1 & x-1 & -1 & 0 \\ -1 & 4x-4 & x-5 & -1 \\ 0 & -2x+2 & 2 & x-1 \end{pmatrix}$$

$$
\begin{aligned}
&\underset{(3)}{=} \det \begin{pmatrix} 0 & -x^2+3x-2 & x-2 & 0 \\ x-1 & x-1 & -1 & 0 \\ -1 & 4x-4 & x-5 & -1 \\ -x+1 & 4x^2-10x+6 & x^2-6x+7 & 0 \end{pmatrix} \\
&\underset{(4)}{=} \det \begin{pmatrix} 0 & -x^2+3x-2 & x-2 \\ x-1 & x-1 & -1 \\ -x+1 & 4x^2-10x+6 & x^2-6x+7 \end{pmatrix} \\
&\underset{(5)}{=} \det \begin{pmatrix} 0 & -x^2+3x-2 & x-2 \\ x-1 & x-1 & -1 \\ 0 & 4x^2-9x+5 & x^2-6x+6 \end{pmatrix} \\
&\underset{(6)}{=} -(x-1)\cdot \det \begin{pmatrix} -x^2+3x-2 & x-2 \\ 4x^2-9x+5 & x^2-6x+6 \end{pmatrix} \\
&\underset{(7)}{=} -(x-1)\cdot \det \begin{pmatrix} 0 & x-2 \\ x^3-3x^2+3x-1 & x^2-6x+6 \end{pmatrix} \\
&= (x-1)(x-2)(x^3-3x^2+3x-1) = (x-2)(x-1)^4.
\end{aligned}
$$

Die Schritte waren: (1) Addieren des $(-x+2)$-fachen der ersten Spalte zur dritten, (2) Entwickeln nach der dritten Zeile, (3) Addition der dritten Zeile zur ersten und des $(x-1)$-fachen der dritten Zeile zur letzten, (4) Entwickeln nach der letzten Spalte, (5) Addieren der zweiten Zeile zur dritten, (6) Entwickeln nach der ersten Spalte und (7) Addieren des $(x-1)$-fachen der zweiten Spalte zur ersten.
Der Eigenwert 2 ergibt ein Jordan-Kästchen der Länge 1. Der Eigenwert 1 hat algebraische Vielfachheit 4. Wir berechnen den Rang von $A - I_5$:

$$
\begin{aligned}
\operatorname{rg}(A-I_5) &= \operatorname{rg} \begin{pmatrix} -4 & -1 & 4 & -3 & -1 \\ 1 & 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 4 & 1 & -4 & 4 & 1 \\ -2 & 0 & 2 & -2 & 0 \end{pmatrix} = \operatorname{rg} \begin{pmatrix} 0 & -1 & 0 & 1 & -1 \\ 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\
&= \operatorname{rg} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = 3.
\end{aligned}
$$

Es gibt also $5-3=2$ Jordan-Kästchen zum Eigenwert 1. Dafür gibt es zwei Möglichkeiten (zwei Kästchen der Länge 2 oder je eines der Länge 1 und 3). Um die Anzahl $c_1(1,A)$ der Jordan-Kästchen der Länge 1 nach Satz 19.9(a) zu berechnen, brauchen wir den Rang von $(A-I_5)^2$:

$$\operatorname{rg}\left((A-I_5)^2\right) = \operatorname{rg}\begin{pmatrix} 1 & 1 & -1 & 1 & 1 \\ 1 & 0 & -1 & 1 & 0 \\ 3 & 1 & -3 & 3 & 1 \\ 3 & 0 & -3 & 3 & 0 \\ -2 & 0 & 2 & -2 & 0 \end{pmatrix} = \operatorname{rg}\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = 2.$$

Wir erhalten $c_1(1, A) = 5 - 2 \cdot 3 + 2 = 1$. Es gibt also ein Jordan-Kästchen der Länge 1, und A hat die Jordansche Normalform

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

◁

Oft ist es von Interesse, nicht nur die allgemeine bzw. Jordansche Normalform B einer Matrix $A \in K^{n\times n}$ zu bestimmen, sondern auch eine transformierende Matrix $S \in \mathrm{GL}_n(K)$ mit $B = S^{-1}AS$. Dies ist gleichbedeutend mit der Bestimmung einer Basis von K^n, bezüglich der φ_A die Darstellungsmatrix B hat. Bisweilen wird eine solche Basis (im Falle der Jordanschen Normalform) eine *Jordan-Basis* genannt.

Eine Methode zur Berechnung einer transformierenden Matrix wird aus der Bemerkung vor Korollar 19.2 klar: Aus der Kenntnis einer der transformierenden Matrizen für die Äquivalenz der charakteristischen Matrizen $xI_n - A$ und $xI_n - B$ erhält man eine transformierende Matrix für die Ähnlichkeit von A und B. Diese Methode ist jedoch meist zu aufwändig. Daher wird normalerweise eine wesentlich effizientere Methode verwendet, die wir nun (im Fall der Jordanschen Normalform) skizzieren.

Es wird vorausgesetzt, dass die Jordansche Normalform einer Matrix $A \in K^{n\times n}$ bekannt ist, und das Ziel ist die Bestimmung einer Jordan-Basis. Diese setzt man zusammen aus Vektoren, die durch Anwendung von A gemäß den einzelnen Jordan-Kästchen transformiert werden. Man behandelt die Eigenwerte λ nacheinander. Zu einem Eigenwert λ sucht man zunächst Basisvektoren, die zu den längsten Jordan-Kästchen zum Eigenwert λ gehören. Ist deren Länge e, so berechnet man den sogenannten *Hauptraum*

$$E_\lambda^{(e)} := \{v \in K^n \mid (A - \lambda I_n)^e \cdot v = 0\}.$$

Haupträume stellen eine Verallgemeinerung der Eigenräume dar. Man ergänzt nun eine Basis des Unterraums $E_\lambda^{(e-1)}$ zu einer Basis von $E_\lambda^{(e)}$. Die ergänzenden Vektoren bilden die „Keime“ der zu den Jordan-Kästchen gehörenden Basisvektoren. Ist $v \in E_\lambda^{(e)}$ ein solcher, so setzen wir nämlich

$$v_1 := v, \quad v_2 := Av_1 - \lambda v_1, \quad \ldots, \quad v_e := Av_{e-1} - \lambda v_{e-1}. \tag{19.8}$$

Für $i \leq e-1$ folgt $A \cdot v_i = \lambda \cdot v_i + v_{i+1}$, also genau das Verhalten, das durch ein Jordan-Kästchen beschrieben wird. Aus $v \in E_\lambda^{(e)}$ folgt weiter $Av_e = \lambda \cdot v_e$, was auch dem Jordan-Kästchen entspricht. Die Vektoren v_i fügt man zu der Jordan-Basis hinzu, und so verfährt man mit allen Vektoren, die eine Basis von $E_\lambda^{(e-1)}$ zu einer von $E_\lambda^{(e)}$ ergänzen. Nun hat man Basisvektoren, die zu den Jordan-Kästchen zum Eigenwert λ mit der maximalen Länge e gehören.

Es geht weiter mit den Basisvektoren zu den Jordan-Kästchen der Länge $e-1$ (falls vorhanden). Um lineare Abhängigkeit mit den schon in der Jordan-Basis befindlichen Vektoren zu vermeiden, muss man Basen von $E_\lambda^{(e-2)}$ und von $(A - \lambda I_n) \cdot E_\lambda^{(e)}$ zu einer Basis von $E_\lambda^{(e-1)}$ ergänzen. Eine Basis von $(A-\lambda I_n) \cdot E_\lambda^{(e)}$ erhält man hierbei aus den „Abkömmlingen“ v_2 gemäß (19.8) der Vektoren aus der Basisergänzung von $E_\lambda^{(e-1)}$ zu $E_\lambda^{(e)}$. Auch hier bilden die ergänzenden Basisvektoren die „Keime“ der zu den Jordan-Kästchen der Länge $e-1$ gehörenden Basisvektoren.

Beispiel 19.11. Zur Illustration der Methode betrachten wir unsere Standardbeispiele.

(1) Wir betrachten wieder

$$A = \begin{pmatrix} -3 & -1 & 2 \\ 4 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Wir wissen, dass es zwei Jordan-Kästchen der Länge 1 und 2 zum Eigenwert -1 gibt (siehe Beispiel 19.5(3)). Der Eigenraum E_{-1} hat also die Dimension 2, der Hauptraum $E_{-1}^{(2)}$ muss also Dimension 3 haben. (Diese Dimensionen ergeben sich auch aus der Formel in Satz 19.9(a).) Wir können als „Keim“ einer Jordanbasis also mit einem beliebigen Vektor außerhalb E_{-1} beginnen. Wir wählen den ersten Standardbasisvektor $v_1 := e_1$. Weiter setzen wir

$$v_2 := Av_1 + v_1 = \begin{pmatrix} -2 \\ 4 \\ 0 \end{pmatrix}.$$

Diese beiden Vektoren gehören zum Jordan-Kästchen der Länge 2. Um einen Basisvektor zum Jordan-Kästchen der Länge 1 zu bekommen, ergänzen wir v_2 durch

$$v_3 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$$

zu einer Basis von E_{-1}. In der Reihenfolge v_3, v_1, v_2 bilden unsere Vektoren eine Jordan-Basis zu der Jordanschen Normalform mit der Reihenfolge der Kästchen wie in Beispiel 19.5(3). Eine transformierende Matrix ist

$$S = \begin{pmatrix} 0 & 1 & -2 \\ 2 & 0 & 4 \\ 1 & 0 & 0 \end{pmatrix}.$$

(2) Nun betrachten wir unser zweites Standardbeispiel, nämlich

$$A = \begin{pmatrix} -3 & -1 & 4 & -3 & -1 \\ 1 & 1 & -1 & 1 & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 4 & 1 & -4 & 5 & 1 \\ -2 & 0 & 2 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{5\times 5}$$

(siehe Beispiel 19.10(2)). Für den Eigenwert $\lambda = 2$ finden wir durch Lösen des entsprechenden homogenen LGS den Eigenvektor

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ -2 \end{pmatrix},$$

den wir als ersten Vektor in die Jordan-Basis aufnehmen. Nun behandeln wir den Eigenwert $\lambda = 1$ und suchen als erstes einen Vektor für das Jordan-Kästchen der Länge 3. Hierzu müssen wir $E_1^{(3)}$, also den Kern von $(A - I_5)^3$, berechnen. Wir kennen aus Beispiel 19.10(2) bereits die Ränge von $A - I_5$ und $(A - I_5)^2$ (nämlich 3 und 2), und erhalten $\operatorname{rg}\left((A - I_5)^3\right) = 1$ durch Auflösen der Formel aus Satz 19.9(a). Es genügt also, eine Zeile von $(A - I_5)^3$ zu berechnen, wobei wir $(A - I_5)^2$ schon aus Beispiel 19.10(2) kennen. Am einfachsten ist die dritte Zeile von $(A - I_5)^3$, die sich zu $(2, 0, -2, 2, 0)$ ergibt. Wir wählen

$$v_3 := \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \in E_1^{(3)} \setminus E_1^{(2)}.$$

und weiter

$$v_4 := (A - I_5) \cdot v_3 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad v_5 := (A - I_5) \cdot v_4 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Die Vektoren v_3, v_4, v_5 gehören zum Jordan-Kästchen der Länge 3, was wir durch Nachrechnen von

$$A \cdot v_3 = v_3 + v_4, \quad A \cdot v_4 = v_4 + v_5 \quad \text{und} \quad A \cdot v_5 = v_5$$

bestätigen können. Für das Jordan-Kästchen der Länge 1 brauchen wir einen Vektor aus $E_1^{(1)}$ (also einen Eigenvektor), der zusammen mit v_5 linear unabhängig ist. Wir haben $A - I_5$ in Beispiel 19.10(2) bereits mit Spaltenoperationen behandelt und sind auf die Matrix

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

gekommen, an der man die Basis

$$\begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

des Eigenraums $E_1^{(1)}$ abliest. Wir können also als letzten Basisvektor

$$v_2 = \begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

wählen. Die Nummerierung der v_i haben wir so gemacht, dass sie mit der gewählten Reihenfolge der Jordan-Kästchen in Beispiel 19.10(2) kompatibel ist. Als transformierende Matrix erhält man

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

◁

Dies ist eine geeignete Stelle, um den Begriff des Minimalpolynoms einer Matrix $A \in K^{n \times n}$ einzuführen. Nach dem Satz von Cayley-Hamilton (Satz 17.17) gilt für das charakteristische Polynom χ_A die Beziehung $\chi_A(A) = 0$, also existiert ein (normiertes) Polynom, das A als „Nullstelle“ hat. (Dies hätten wir auch daraus folgern können, dass wegen $\dim(K^{n \times n}) < \infty$ die Potenzen von A linear abhängig sein müssen.) Das **Minimalpolynom** von A

ist das normierte Polynom $g \in K[x]$ minimalen Grades, so dass $g(A) = 0$ gilt. Es ist nicht schwer zu sehen, dass g eindeutig bestimmt ist, und dass die Polynome $f \in K[x]$ mit $f(A) = 0$ genau die Vielfachen von g sind. Außerdem haben ähnliche Matrizen das gleiche Minimalpolynom.

Beispiel 19.12. Für die „Projektionsmatrix“

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

gilt $A^2 = A$, und A hat das Minimalpolynom $x^2 - x = x(x-1)$. Das charakteristische Polynom ist $\chi_A = x^2(x-1)^2$. ◁

Aus der Theorie der Jordanschen Normalform sieht man: Ist $\chi_A = \prod_{i=1}^{r}(x-\lambda_i)^{e_i}$ mit paarweise verschiedenen Eigenwerten λ_i, so ist

$$g = \prod_{i=1}^{r}(x-\lambda_i)^{l_i}$$

mit l_i die maximale Länge eines Jordan-Kästchens zum Eigenwert λ_i das Minimalpolynom. Entsprechend verhält es sich mit der allgemeinen Normalform. Äquivalent ist folgende Aussage: Das Minimalpolynom von A ist der letzte Elementarteiler d_n der charakteristischen Matrix $xI_n - A$.

20 Dualraum

Dieser Abschnitt passt nicht wirklich unter das Stichwort „Normalformen“.

Weiterhin steht K immer für einen Körper. Wir erinnern daran, dass für zwei K-Vektorräume V, W auch die Menge $\mathrm{Hom}(V, W)$ der linearen Abbildungen $V \to W$ ein Vektorraum wird, wobei die Operationen punktweise definiert sind.

Definition 20.1. *Es sei V ein K-Vektorraum. Eine* **Linearform** *(auf V) ist eine lineare Abbildung $V \to K$. Der Raum*

$$V^* := \mathrm{Hom}(V, K)$$

aller Linearformen heißt der **Dualraum** *von V.*

Beispiel 20.2. (1) Eine Linearform auf dem n-dimensionalen Standardraum $V = K^n$ hat eine Darstellungsmatrix (bzgl. der Standardbasen) aus $K^{1\times n}$. Umgekehrt liefert jeder Zeilenvektor aus $K^{1\times n}$ eine Linearform, und die Addition bzw. Multiplikation mit Skalaren von Zeilenvektoren entspricht den entsprechenden Operationen der Linearformen. Wir

können V^* also mit dem Vektorraum $K^{1\times n}$ der Zeilenvektoren identifizieren.

(2) Sei $V = K[x]$ der Polynomring. Zu jeder Linearform $\varphi\colon V \to K$ erhalten wir eine Folge $(b_0, b_1, \ldots)$ durch $b_i := \varphi(x^i) \in K$. Ist umgekehrt $(b_0, b_1, \ldots)$ eine Folge mit $b_i \in K$, so liefert

$$\varphi\colon V \to K,\ \sum_{i=0}^{n} a_i x^i \mapsto \sum_{i=0}^{n} a_i b_i$$

eine Linearform. Wir können V^* also mit dem Raum der K-wertigen Folgen identifizieren. ◁

Es sei nun V ein K-Vektorraum und B eine Basis. Jedes $v \in V$ lässt sich also eindeutig schreiben als

$$v = \sum_{w\in B} a_w \cdot w$$

mit $a_w \in K$, wobei nur endlich viele der a_w ungleich 0 sind. Wir fixieren jetzt einen Basisvektor $b \in B$ und definieren eine Abbildung

$$b^*\colon V \to K,\ v = \sum_{w\in B} a_w \cdot w \mapsto a_b.$$

Es ist klar, dass b^* eine Linearform ist, also $b^* \in V^*$. Die Menge

$$B^* := \{b^* \mid b \in B\}$$

heißt die **Dualbasis** zu B. Die Bezeichnung „Dualbasis“ ist etwas irreführend, wie der Teil (b) des folgenden Satzes zeigt.

Satz 20.3. *Es seien V ein K-Vektorraum und B eine Basis.*

(a) Die Dualbasis $B^* \subseteq V^*$ ist linear unabhängig.

(b) B^* ist genau dann eine Basis von V^*, falls $\dim(V) < \infty$. In diesem Fall gilt also

$$\dim(V) = \dim(V^*).$$

Beweis. (a) Es seien $b_1, \ldots, b_n \in B$ paarweise verschieden und $a_1, \ldots, a_n \in K$, so dass

$$f := \sum_{i=1}^{n} a_i b_i^* = 0.$$

Dann gilt für alle $j = 1, \ldots, n$

$$0 = f(b_j) = \sum_{i=1}^{n} a_i b_i^*(b_j) = a_j.$$

Also sind $b_1^*, \ldots, b_n^*$ linear unabhängig.

(b) Es sei $\dim(V) < \infty$ und $B = \{b_1, \ldots, b_n\}$. Für $f \in V^*$ setzen wir $a_i := f(b_i) \in K$ und $g := \sum_{i=1}^n a_i b_i^*$. Dann gilt für $j \in \{1, \ldots, n\}$

$$g(b_j) = \sum_{i=1}^n a_i b_i^*(b_j) = a_j = f(b_j),$$

f und g stimmen also auf der Basis B überein. Wegen Satz 11.12(a) folgt $f = g$. Wegen $g \in \langle B^* \rangle$ erhalten wir $V^* = \langle B^* \rangle$, also ist B^* eine Basis.
Nun sei B unendlich. Jede Linearkombination von B^* ist eine Linearform, die nur auf endlich vielen Basisvektoren einen Wert $\neq 0$ annimmt. Also liegt die Linearform

$$f\colon V \to K,\ v = \sum_{w \in B} a_w \cdot w \mapsto \sum_{w \in B} a_w$$

nicht in $\langle B^* \rangle$, B^* ist also keine Basis. □

Das Wesen des Dualraums wird klarer, wenn man sich sogenannte duale Abbildungen anschaut. Diese werden wie folgt gebildet. Ist $\varphi\colon V \to W$ eine lineare Abbildung zwischen zwei K-Vektorräumen, so definieren wir die **duale Abbildung**

$$\varphi^*\colon W^* \to V^*,\ f \mapsto f \circ \varphi.$$

Offenbar ist φ^* auch linear. Die duale Abbildung φ^* geht in umgekehrter Richtung wie φ.

Man kann auch den Dualraum des Dualraums bilden, also

$$V^{**} := (V^*)^*.$$

Man nennt V^{**} den **Bidualraum**. Für $v \in V$ können wir ein ganz spezielles Element $\varphi_v \in V^{**}$ wie folgt definieren:

$$\varphi_v\colon V^* \to K,\ f \mapsto f(v).$$

In der Tat gelten für $f, g \in V^*$ und $a \in K$:

$$\varphi_v(f+g) = (f+g)(v) = f(v) + g(v) = \varphi_v(f) + \varphi_v(g)$$

und

$$\varphi_v(a \cdot f) = (a \cdot f)(v) = a \cdot f(v) = a \cdot \varphi_v(f).$$

Satz 20.4. *Es sei V ein K-Vektorraum.*

(a) Die Abbildung

$$\Phi\colon V \to V^{**},\ v \mapsto \varphi_v$$

ist linear und injektiv.

(b) Genau dann ist Φ ein Isomorphismus, wenn $\dim(V) < \infty$.

Beweis. (a) Für $v, w \in V$, $a \in K$ und $f \in V^*$ gelten

$$\varphi_{v+w}(f) = f(v+w) = f(v) + f(w) = \varphi_v(f) + \varphi_w(f)$$

und

$$\varphi_{av}(f) = f(av) = af(v) = a\varphi_v(f).$$

also

$$\Phi(v+w) = \varphi_{v+w} = \Phi(v) + \Phi(w) \quad \text{und} \quad \Phi(av) = \varphi_{av} = a\Phi(v).$$

Damit ist Φ linear. Für den Nachweis von $\operatorname{Kern}(\Phi) = \{0\}$ nehmen wir ein $v \in V$ mit $v \neq 0$. Wir können $\{v\}$ zu einer Basis B von V ergänzen. Für $f := v^* \in B^*$ gilt dann $f(v) = 1$, also $\varphi_v(f) \neq 0$. Es folgt $v \notin \operatorname{Kern}(\Phi)$. Damit ist auch die Injektivität von Φ gezeigt.

(b) Falls $\dim(V) < \infty$, so liefert zweimaliges Anwenden von Satz 20.3(b)

$$\dim(V) = \dim(V^*) = \dim(V^{**}).$$

Aus (a) und Korollar 11.11 folgt, dass Φ ein Isomorphismus ist.
Nun sei V unendlich-dimensional und B eine Basis. Die Dualbasis B^* ist nach Satz 20.3(a) linear unabhängig, also lässt sie sich zu einer Basis C^* von V^* ergänzen. Wir definieren $\varphi \in V^{**}$ durch

$$\varphi\colon V^* \to K,\ f = \sum_{c \in C^*} a_c \cdot c \mapsto \sum_{c \in C^*} a_c$$

und behaupten, dass $\varphi \neq \varphi_v$ für alle $v \in V$ gilt, also $\varphi \notin \Phi(V)$. Es sei also

$$v = \sum_{b \in B} a_b \cdot b \in V.$$

Da a_b nur für endlich viele $b \in B$ ungleich 0 ist, gibt es $b \in B$ mit $a_b = 0$, also

$$\varphi_v(b^*) = b^*(v) = a_b = 0 \neq 1 = \varphi(b^*).$$

Dies schließt den Beweis ab. □

Diskrete Strukturen: Zählen

21 Binomialkoeffizienten und Kombinatorik

Laut Wiktionary ist Kombinatorik die „mathematische Disziplin, die sich mit der Frage befasst, welche Möglichkeiten (Kombinationen) es gibt, eine bestimmte Anzahl von Dingen miteinander zu kombinieren“. Wenn die „Dinge“ mathematische Objekte sind, läuft das Zählen von Kombinationen in der Regel auf das Bestimmen der Elementzahl einer endlichen Menge hinaus. Hiervon handelt der erster Satz des Abschnitts.

Satz 21.1. *(a) Sind A und B zwei gleichmächtige endliche Mengen, so gilt $|A| = |B|$.*

(b) Sind $A_1, \dots, A_n$ paarweise disjunkte endliche Mengen (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$), so gilt

$$|A_1 \cup \cdots \cup A_n| = |A_1| + \cdots + |A_n|.$$

Anmerkung*: Man nennt die Vereinigung paarweiser disjunkter Mengen auch* **disjunkte Vereinigung** *und schreibt sie als*

$$A_1 \dot\cup \cdots \dot\cup A_n.$$

(c) Sind $A_1, \dots, A_n$ endliche Mengen, so gilt für das kartesische Produkt

$$|A_1 \times \cdots \times A_n| = |A_1| \cdots |A_n|.$$

Zum Beweis des Satzes bemerken wir, dass der Teil (a) direkt aus Definition 2.13 folgt. Will man aber die Teile (b) und (c) beweisen, so braucht man eine Definition der Addition und der Multiplikation von natürlichen Zahlen. Da wir dies nicht gemacht, sondern nur grob angedeutet haben (siehe Seite 12), können wir die Beweise nicht führen. Verfügt man über Defini-

tionen für Addition und der Multiplikation von natürlichen Zahlen, so ist der Nachweis der Teile (b) und (c) jedoch nicht schwer. Da diese Teile intuitiv unmittelbar einsichtig sind, sollte uns das Fehlen formaler Beweise keine Kopfschmerzen bereiten.

Den schwierigeren Fall von nicht disjunkten Vereinigungen werden wir später behandeln (siehe Satz 21.8).

Beispiel 21.2. Sind A und B endliche Mengen mit $k := |A|$ und $n := |B|$, so können wir die Anzahl der injektiven Funktionen $g\colon A \to B$ bestimmen. Falls $A = \emptyset$, so gibt es genau eine solche Funktion. Andernfalls wählen wir $a_0 \in A$ und zerlegen die Menge F der injektiven Funktionen $A \to B$ disjunkt als

$$F = \dot{\bigcup_{b \in B}} \underbrace{\big\{g\colon A \to B \mid g \text{ injektiv},\ g(a_0) = b\big\}}_{=:F_b}.$$

Die Einschränkung auf $A \setminus \{a_0\}$ liefert eine Bijektion von F_b auf die Menge der Injektionen $A\setminus\{a_0\} \to B\setminus\{b\}$. Die gesuchte Elementanzahl $f(k,n) = |F|$ erfüllt also die Gleichung

$$f(k,n) = n \cdot f(k-1, n-1).$$

Nun liefert eine einfache Induktion nach k die Formel

$$|F| = n(n-1)\cdots(n-k+1) = \prod_{i=0}^{k-1}(n-i) =: n^{\underline{k}}.$$

Die Zahl $n^{\underline{k}}$ wird die k-te **fallende Faktorielle** (von n) genannt. Für $k \le n$ ist dies gleich $n!/(n-k)!$, und für $k > n$ ist dies gleich 0. Insbesondere erhalten wir die bekannte Formel $|S_n| = n!$ für die symmetrische Gruppe. ◁

Wir hätten das obige Beispiel auch weniger formal behandeln können und die Zahl $|F|$ interpretieren können als die Anzahl der Möglichkeiten, aus n Kugeln hintereinander k Stück ohne Zurücklegen zu ziehen, wobei es auf die Reihenfolge der Züge ankommt. Kommt es jedoch *nicht* auf die Reihenfolge der Züge an, so muss man die fallende Faktorielle durch $k!$ dividieren. Es gibt also $n^{\underline{k}}/k!$ Möglichkeiten, eine ungeordnete Menge von k Kugeln aus einer Urne mit n Kugeln ohne Zurücklegen zu ziehen. Wir werden dies später mathematischer formulieren und formaler beweisen (siehe Satz 21.6), nehmen die obige Formel aber zum Anlass für die folgende etwas allgemeinere Definition.

Definition 21.3. *Für eine komplexe Zahl $a \in \mathbb{C}$ und eine natürliche Zahl $k \in \mathbb{N}$ ist*

$$\binom{a}{k} = \frac{\prod_{i=0}^{k-1}(a-i)}{k!} = \frac{a}{1} \cdot \frac{a-1}{2} \cdots \frac{a-(k-1)}{k} = \frac{a^{\underline{k}}}{k!}$$

der (k-te) **Binomialkoeffizient** *(von a). Er wird häufig gelesen als „a über k“. Im Fall $k = 0$ interpretieren wir das leere Produkt als 1, also $\binom{a}{0} = 1$.*

Ebensogut wie für komplexe Zahlen a lässt sich $\binom{a}{k}$ für Elemente $a \in R$ eines kommutativen Ringes R, der die rationalen Zahlen $\mathbb{Q}$ enthält, definieren. Interessant und wichtig ist dabei der Fall $a = x \in \mathbb{Q}[x]$, in dem

$$\binom{x}{k} = \frac{\prod_{i=0}^{k-1}(x-i)}{k!} \in \mathbb{Q}[x]$$

ein Polynom vom Grad k ist. Einsetzen von $x = a$ in dieses Polynom ergibt $\binom{a}{k}$.

Die Binomialkoeffizienten stehen in einem Geflecht von Beziehungen zueinander und zu anderen Größen. Die vielleicht wichtigsten davon fassen wir in folgendem Satz zusammen.

Satz 21.4. *Für alle $k \in \mathbb{N}$, $n \in \mathbb{Z}$ und $a, b \in \mathbb{C}$ (oder allgemeiner a, b Elemente in einem kommutativen Ring, der $\mathbb{Q}$ enthält) gelten:*

(a)

$$\binom{n}{k} \in \mathbb{Z} \quad \textit{(Ganzzahligkeit).}$$

(b) Falls $n \geq k$, dann

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} > 0.$$

(c) Falls $n \geq k$, dann

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

(d) Falls $0 \leq n < k$, dann

$$\binom{n}{k} = 0 \quad \textit{(Nullstellen).}$$

(e)

$$\binom{a}{k} + \binom{a}{k+1} = \binom{a+1}{k+1} \quad \textit{(Formel für das Pascalsche Dreieck).}$$

(f)

$$\binom{a}{0} = 1 \quad \textit{und} \quad \binom{a}{1} = a \quad \textit{(spezielle Werte).}$$

(g)

$$\binom{-a}{k} = (-1)^k \binom{a+k-1}{k}.$$

(h) Falls $n \geq 0$, dann

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \quad \textit{(binomische Formel).}$$

(Dies gilt sogar dann, wenn a und b aus einem kommutativen Ring kommen, der nicht $\mathbb{Q}$ enthält.)

(i) Falls $n \geq 0$, dann

$$\sum_{i=0}^{n} \binom{i}{k} = \binom{n+1}{k+1} \quad \textit{(Summenformel).}$$

(j)

$$\sum_{j=0}^{k} \binom{a}{j} \binom{b}{k-j} = \binom{a+b}{k} \quad \textit{(Vandermondesche Identität).}$$

Beweis. Die Aussagen (b), (d), (f) und (g) folgen direkt aus der Definition, und (c) folgt aus (b). Teil (e) ergibt sich aus der Rechung

$$\begin{aligned}\binom{a}{k} + \binom{a}{k+1} &= \frac{(k+1)\prod_{i=0}^{k-1}(a-i) + \prod_{i=0}^{k}(a-i)}{(k+1)!} \\ &= \frac{(k+1+a-k)\prod_{j=1}^{k}(a-j+1)}{(k+1)!} = \binom{a+1}{k+1}.\end{aligned}$$

Wie wir sehen werdem, wird (e) für alle weiteren Nachweise entscheidend verwendet.

Die Ganzzahligkeit (a) gilt für $k = 0$ oder $n = 0$ nach (f) und (d). Für positive n und k folgt sie per Induktion nach n mit (e), und für negative n dann aus (g).

Die binomische Formel (h) zeigen wir per Induktion. Für $n = 0$ folgt sie aus (f). Weiter gilt

$$\begin{aligned}&(a+b)^{n+1} = \\ &(a+b)\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k} = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n+1-k} + \sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k} = \\ &1\cdot b^{n+1} + \sum_{k=1}^{n}\left(\binom{n}{k} + \binom{n}{k-1}\right)a^k b^{n+1-k} + 1\cdot a^{n+1} = \\ &\qquad\qquad \sum_{k=0}^{n+1}\binom{n+1}{k}a^k b^{n+1-k},\end{aligned}$$

wobei wir für die erste Gleichheit Induktion und für die letzte (e) benutzt haben.

Teil (i) lässt sich ebenfalls durch Induktion nach n zeigen. Für $n = 0$ lautet die Behauptung $\binom{0}{k} = \binom{1}{k+1}$, was wegen (d) und (f) stimmt. Weiter gilt

$$\sum_{i=0}^{n+1} \binom{i}{k} = \binom{n+1}{k+1} + \binom{n+1}{k} = \binom{n+2}{k+1},$$

wobei wir für die erste Gleichheit Induktion und für die zweite (e) benutzt haben.

Der Nachweis der Vandermondeschen Identität (j) ist der schwierigste und interessanteste. Wir beginnen mit dem Spezialfall, dass a und b im Polynomring $\mathbb{Q}[x]$ liegen, und zwar, noch spezieller, dass $a = x$ und $b = n \in \mathbb{N}$. Wir benutzen Induktion nach n. Zu zeigen ist also

$$\sum_{j=0}^{k} \binom{x}{j}\binom{n}{k-j} = \binom{x+n}{k}, \tag{21.1}$$

was für $n = 0$ zu der Gleichung $\binom{x}{k} = \binom{x}{k}$ wird, und für $k = 0$ zu $1 \cdot 1 = 1$. Wir setzen nun $k > 0$ voraus und rechnen

$$\sum_{j=0}^{k} \binom{x}{j}\binom{n+1}{k-j} = \binom{x}{k} + \sum_{j=0}^{k-1} \binom{x}{j}\left(\binom{n}{k-j} + \binom{n}{k-j-1}\right) = \binom{x+n}{k} + \binom{x+n}{k-1} = \binom{x+n+1}{k},$$

wobei wir für die erste und dritte Gleichheit (e) und für die zweite Induktion verwendet haben. Hiermit ist (21.1) nachgewiesen. Es mag erstaunen, dass dieser Spezialfall eigentlich schon den allgemeinen Fall beinhaltet. Denn (21.1) sagt, dass (für jedes $k \in \mathbb{N}$) das Polynom

$$\sum_{j=0}^{k} \binom{x}{j}\binom{y}{k-j} - \binom{x+y}{k}$$

in der Variablen y und mit Koeffizienten aus $\mathbb{Q}[x]$ alle natürlichen Zahlen als Nullstellen hat. Es muss sich also um das Nullpolynom handeln. Setzt man nun in dieses Polynom $x = a$ und $y = b$ ein, ergibt sich (j). □

Alle Eigenschaften aus 21.4 sind wichtig für Anwendungen. Für die Formel in (e) werden wir einige sehen. Nicht direkt klar ist, warum die Summenformel (i) und die Vandermondesche Identität (21.1) wichtig sind. Letztere wird im nächsten Abschnitt 22 eine interessante Rolle spielen. Die Summenformel wenden wir in folgendem Beispiel an.

Beispiel 21.5. Die berühmte Formel

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

ist der Fall $k = 1$ in Satz 21.4(i). Man bekommt auch Summenformeln für $\sum_{i=0}^{n} i^k$, indem man das Polynom x^k als (ganzzahlige) Linearkombination von $\binom{x}{1}, \ldots, \binom{x}{k}$ darstellt und dann Satz 21.4(i) für die Summation der Binomialkoeffizienten benutzt. Im Beispiel $k = 3$ läuft das so:

$$x^3 = 6\binom{x}{3} + 6\binom{x}{2} + \binom{x}{1},$$

also nach Satz 21.4(i)

$$\sum_{i=0}^{n} i^3 = 6\binom{n+1}{4} + 6\binom{n+1}{3} + \binom{n+1}{2} = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4} = \frac{n^2(n+1)^2}{4},$$

wobei wir die Details der Rechnung weggelassen haben. ◁

Nun können wir unsere frühere Behauptung über „ungeordnes Ziehen von k Kugeln ohne Zurücklegen“ beweisen. Dies bedeutet nichts anderes als die Auswahl einer k-elementigen Teilmenge, und entsprechend formulieren wir das Ergebnis.

Satz 21.6. *Es seien A eine endliche Menge mit n Elementen und $k \in \mathbb{N}$ eine natürliche Zahl. Dann gilt für die Menge*

$$M := \big\{T \subseteq A \mid |T| = k\big\}$$

aller k-elementigen Teilmengen von A die Formel

$$|M| = \binom{n}{k}.$$

Beweis. Im Falle $k = 0$ gilt $M = \{\emptyset\}$, also $|M| = 1 = \binom{n}{k}$ wegen Satz 21.4(f). Wir setzen ab jetzt $k > 0$ voraus und benutzen Induktion nach n. Für $n = 0$ gilt $|M| = 0 = \binom{n}{k}$ wegen Satz 21.4(d). Es bleibt der Fall $n > 0$, in dem wir ein Element $x \in A$ wählen können. Es gilt

$$M = \underbrace{\big\{T \subseteq A \mid |T| = k \text{ und } x \in T\big\}}_{=:M_x} \dot\cup \underbrace{\big\{T \subseteq A \mid |T| = k \text{ und } x \notin T\big\}}_{=:M_{\overline{x}}}.$$

Ordnet man einer $(k-1)$-elementigen Teilmenge $S \subseteq A \setminus \{x\}$ die Menge $S \cup \{x\}$ zu, so ergibt dies eine Bijektion $\big|\big\{S \subseteq A \setminus \{x\} \mid |S| = k-1\big\} \to M_x$. Außerdem gilt $M_{\overline{x}} = \big\{T \subseteq A \setminus \{x\} \mid |T| = k\big\}$. Wir erhalten

$$|M| \underset{\text{Satz } 21.1(b)}{=} |M_x| + |M_{\overline{x}}| \underset{\text{Satz } 21.1(a)}{=} \left|\left\{S \subseteq A \setminus \{x\} \mid |S| = k-1\right\}\right| + \left|\left\{T \subseteq A \setminus \{x\} \mid |T| = k\right\}\right| \underset{\text{Induktion}}{=} \binom{n-1}{k-1} + \binom{n-1}{k} \underset{\text{Satz } 21.4(e)}{=} \binom{n}{k}.$$

Damit ist der Satz bewiesen. □

Beispiel 21.7. Ein Beispiel für Lottospieler: Die Anzahl der Möglichkeiten, aus 49 Zahlen 6 auszuwählen, ist

$$\binom{49}{6} = 13983816.$$

Auf der Webseite lotto.de wird die Wahrscheinlichkeit für Gewinnklasse 2 (sechs richtige) mit 1 zu 15537573 angegeben. Warum? ◁

Durch Satz 21.6 motiviert ist die häufig benutzte Schreibweise

$$\binom{A}{k} := \left\{T \subseteq A \mid |T| = k\right\}$$

für eine Menge A. Mit dieser Schreibweise lautet der Satz

$$\left|\binom{A}{k}\right| = \binom{|A|}{k}.$$

Eine weitere gängige Schreibweise ist

$$[n] := \{1, 2, \ldots, n\}$$

für $n \in \mathbb{N}$.

Im folgenden Satz geht es um die Elementanzahl einer Vereinigung von endlich vielen endlichen Mengen, die nicht disjunkt sein müssen. Wir könnten solche Mengen als $A_1, \ldots, A_n$ aufzählen oder sie als endliches Mengensystem zu schreiben. Der Satz enthält beide Varianten und benutzt die obigen Schreibweisen.

Satz 21.8 (Inklusion-Exklusion)**.** *Für endliche Mengen $A_1, \ldots, A_n$ gilt*

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{I \in \binom{[n]}{k}} \left|\bigcap_{i \in I} A_i\right| \right). \tag{21.2}$$

Gleichbedeutend hierzu ist: Sei $\mathcal{M}$ ein Mengensystem bestehend aus endlichen Mengen. Dann gilt

$$\left|\bigcup \mathcal{M}\right| = \sum_{\emptyset \neq \mathcal{N} \subseteq \mathcal{M}} (-1)^{|\mathcal{N}|-1} \left|\bigcap \mathcal{N}\right|. \tag{21.3}$$

Beispiel 21.9. Vor dem Beweis des Satzes schauen wir die Fälle $n = 2$ und $n = 3$ an, die lauten (mit endlichen Mengen A, B, C):

$$|A \cup B| = |A| + |B| - |A \cap B| \tag{21.4}$$

und

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

Beide Formeln sind einigermaßen einsichtigund werden nun bewiesen. ◁

Beweis von Satz 21.8. Wir benutzen Induktion nach n. Für $n = 1$ lautet die Behauptung $|A_1| = |A_1|$ (und auch für $n = 0$ stimmt der Satz mit den leeren Summen interpretiert als 0). Für den Induktionsschritt brauchen wir den Fall $n = 2$. Es seien also A und B irgendwelche endliche Mengen. Dann gelten

$$A \cup B = A \,\dot\cup\, (B \setminus A) \quad \text{und} \quad B = (B \setminus A) \,\dot\cup\, (A \cap B),$$

woraus sich mit Satz 21.1(b) die Formel (21.4) ergibt. Für den Fall $n \geq 3$ ist es günstig, die behauptete Formel (21.2) umzuschreiben zu

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq I \subseteq [n]} (-1)^{|I|-1} \left|\bigcap_{i \in I} A_i\right|,$$

wodurch auch die Äquivalenz zu (21.3) klar wird.

Wir setzen $B := \bigcup_{i=1}^{n-1} A_i$. Wegen $\bigcup_{i=1}^{n} A_i = B \cup A_n$ ergibt sich aus (21.4)

$$\left|\bigcup_{i=1}^{n} A_i\right| = |B| + |A_n| - |B \cap A_n|.$$

Anwendung der Induktionsannahme auf B und auf $B \cap A_n = \bigcup_{i=1}^{n-1} (A_i \cap A_n)$ liefert

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq I \subseteq [n-1]} (-1)^{|I|-1} \left|\bigcap_{i \in I} A_i\right| + |A_n| + \sum_{\emptyset \neq I \subseteq [n-1]} (-1)^{|I|} \left|\bigcap_{i \in I \cup \{n\}} A_i\right| = \sum_{\emptyset \neq I \subseteq [n]} (-1)^{|I|-1} \left|\bigcap_{i \in I} A_i\right|,$$

was den Beweis abschließt. □

Beispiel 21.10. In diesem Beispiel geht es um sogenannte *fixpunktfreie Permutationen*. Damit ist folgendes gemeint: Eine Permutationen $\sigma \in S_A$ einer Menge A heißt **fixpunktfrei**, falls es kein $x \in A$ gibt mit $\sigma(x) = x$. Wir setzen voraus, dass A endlich ist und schreiben $D \subseteq S_A$ für die Menge aller

fixpunktfreien Permutationen. Das Komplement $S_A \setminus D$ ist die Vereinigungsmenge der Mengen

$$(S_A)_x := \{\sigma \in S_A \mid \sigma(x) = x\} \qquad (x \in A),$$

aber die Vereinigung ist nicht disjunkt. Mit $n := |A|$ liefert Satz 21.8

$$|S_A \setminus D| = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{I \in \binom{A}{k}} \left| \bigcap_{x \in I} (S_A)_x \right| \right).$$

Für eine nicht leere Teilmenge $I \subseteq A$ besteht die Schittmenge $\bigcap_{x\in I}(S_A)_x$ aus den Permutationen, die jedes Element von I fixieren. Die Abbildung

$$\bigcap_{x \in I} (S_A)_x \to S_{A\setminus I},\ \sigma \mapsto \sigma|_{A\setminus I} \quad \text{(Einschränkung)}$$

ist bijektiv, also $\left|\bigcap_{x\in I}(S_A)_x\right| = |S_{A\setminus I}| = (n - |I|)!$ wegen Satz 21.1(a). Wir erhalten

$$|D| = |S_A| - |S_A \setminus D| = n! - \sum_{k=1}^{n} \left((-1)^{k-1} \binom{n}{k} (n-k)! \right) \underset{\text{Satz } 21.4(b)}{=} \sum_{k=0}^{n} \frac{n!}{k!} (-1)^k.$$

Der Quotient $\frac{|D|}{n!} = \frac{D}{|S_A|}$ ist also $\sum_{k=0}^{n} \frac{(-1)^k}{k!}$, was für $n \to \infty$ sehr schnell gegen $\exp(-1) = \frac{1}{e}$ konvergiert. Als Fazit haben wir gelernt, dass ungefähr 37 Prozent aller Permutationen fixpunktfrei sind. ◁

22 Erzeugende Funktionen

Wir beginnen mit den berühmten Fibonacci-Zahlen, benannt nach Leonardo Fibonacci, der im frühen 13. Jahrhundert das Wachstun von Kaninchenpopulationen unter folgenden idealisierten Annahmen untersuchte:

- Ab dem Alter von zwei Monaten bekommen Kaninchen Nachwuchs.
- Jedes Paar von Kaninchen bekommt pro Monat zwei Nachkommen, die sich gleichmäßig in männliche und weibliche aufteilen.
- Sie hören nie auf, Nachkommen zu bekommen.

Aus diesen (sicher nur für eingeschränkte Zeiträume halbwegs realistischen) Annahmen ergibt sich für die Anzahl a_{n+2} der Kaninchenpaare im Monat $n+2$ die Gleichung

$$a_{n+2} = a_{n+1} + a_n,$$

wobei die Summanden den Bestand vom Vormonat und die Nachkommen der mindestens zwei Monate alten Kaninchen darstellen. Die Zahlen a_n sind die Fibonacci-Zahlen, wobei zusätzlich

$$a_0 = 0 \quad \text{und} \quad a_1 = 1$$

festgelegt wird. Dies ergibt die Folge $(a_n)_n = (0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots)$. Gesucht ist eine Formel für die Berechnung der a_n. Eine solche lässt sich finden, indem man die obige Rekursionsgleichung in Matrix-Schreibweise bringt und dann per Diagonalisierung die Potenzen der vorkommenden Matrix berechnet. Wir wählen hier einen anderen Weg, der auf eine größere Klasse von durch Rekursionsgleichungen definierten Zahlenfolgen anwendbar ist.

Wir stellen die sogenannte **erzeugende Funktion** auf, womit die Potenzreihe mit den a_n als Koeffizienten gemeint ist, also

$$f = \sum_{n=0}^{\infty} a_n x^n.$$

Dieser Ansatz ist zugleich gewagt und naiv. Gewagt, weil es zunächst unplausibel erscheint, dass beim Verwendung irgendeiner Zahlenfolge als Koeffizienten einer Potenzreihe eine „sinnvolle“ Funktion herauskommt. Und naiv deshalb, weil wir uns keine Gedanken über die Konvergenz der Potenzreihe gemacht haben. Diese Gedanken werden wir nachholen und der Konvergenzfrage auf eine vielleicht überraschende Art begegnen. Zunächst rechnen wir weiter und tun so, also sei alles in Ordnung. Aus den obigen Gleichungen ergibt sich

$$f = x + \sum_{n=0}^{\infty} a_{n+2} x^{n+2} = x + \sum_{n=0}^{\infty} (a_n + a_{n+1}) x^{n+2} = x + (x + x^2) f,$$

und durch Auflösen nach f:

$$f = \frac{x}{1 - x - x^2}.$$

Dies ist tatsächlich eine handhabbare Funktion, deren Potenzreihenentwicklung wir nun durch *Partialbruchzerlegung* bestimmen werden. Hiermit ist gemeint, dass wir den Ansatz

$$f = \frac{\beta_1}{1 - \gamma_1 x} + \frac{\beta_2}{1 - \gamma_2 x}$$

mit $\beta_1, \beta_2, \gamma_1, \gamma_2 \in \mathbb{C}$ machen, der äquivalent ist zu

$$x(1 - \gamma_1 x)(1 - \gamma_2 x) = \big(\beta_1(1 - \gamma_2 x) + \beta_2(1 - \gamma_1 x)\big)(1 - x - x^2).$$

Vergleich der konstanten Koeffizienten liefert $\beta_2 = -\beta_1$, nach Division mit x also

$$(1-\gamma_1 x)(1-\gamma_2 x) = \beta_1(\gamma_1-\gamma_2)(1-x-x^2),$$

also $\beta_1 = \frac{1}{\gamma_1-\gamma_2}$, $\gamma_1+\gamma_2 = 1$ und $\gamma_1\gamma_2 = -1$. Die γ_i müssen also Nullstellen des Polynoms $x^2 - x - 1$ sein, also etwa

$$\gamma_1 = \frac{1+\sqrt{5}}{2},\ \gamma_2 = \frac{1-\sqrt{5}}{2}, \beta_1 = \frac{1}{\sqrt{5}} \quad \text{und} \quad \beta_2 = \frac{-1}{\sqrt{5}}.$$

Was haben wir durch die Partialbruchzerlegung gewonnen? Wir können $\frac{1}{1-\gamma_i x}$ durch die geometrische Reihe $\sum_{n=0}^{\infty}(\gamma_i x)^n$ ausdrücken und erhalten

$$f = \sum_{n=0}^{\infty}(\beta_1\gamma_1^n + \beta_2\gamma_2^n)x^n = \sum_{n=0}^{\infty}\frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right)x^n.$$

Durch Koeffizientenvergleich erhalten wir nun die gewünschte Formel für die Fibonacci-Zahlen:

$$a_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right). \tag{22.1}$$

Wir haben unser Ziel erreicht, müssen allerdings unsere Rechnungen nun (nachträglich) auf solide Füße stellen.

Dies tun wir, indem wir Potenzreihen nicht als (konvergente) Reihen betrachten, sondern sie definieren als *formale* Potenzreihen, bei denen Konvergenzbetrachtungen keine Rolle spielen. Dazu erinnern wir uns an unsere Definition von Polynomen (Definition 7.10) und modifizieren diese auf scheinbar geringfügige Weise.

Definition 22.1. *Es sei R ein kommutativer Ring.*

(a) Eine **formale Potenzreihe** *über R ist eine Abbildung $f\colon \mathbb{N} \to R,\ n \mapsto a_n$ (d.h. ein R-wertige Folge). Die a_n heißen die* **Koeffizienten** *von f. Der Unterschied zwischen einem Polynom und einer formalen Potenzreihe ist also, dass bei einem Polynom nur endliche viele Koeffizienten ungleich 0 sein dürfen.*

(b) Für zwei formale Potenzreihen $f\colon \mathbb{N} \to R,\ n \mapsto a_n$ und $g\colon \mathbb{N} \to R,\ n \mapsto b_n$ definieren wir

$$f+g\colon \mathbb{N} \to R,\ n \mapsto a_n + b_n$$

und

$$f \cdot g\colon \mathbb{N} \to R,\ n \mapsto \sum_{j=0}^{n} a_j b_{n-j} = \sum_{\substack{j,k\in\mathbb{N} \\ mit\ j+k=n}} a_j \cdot b_k.$$

(c) Für eine formale Potenzreihe benutzen wir die Schreibweise

$$f = \sum_{n=0}^{\infty} a_n x^n,$$

wobei man statt x bisweilen andere Variablennamen verwendet. Mit dieser Schreibweise erkennen wir das oben definierte Produkt als das übliche Cauchy-Produkt von Potenzreihen.

(d) Die Menge aller formalen Potenzreihen über R heißt der **formale Potenzreihenring** *über R und wird mit $R[[x]]$ bezeichnet. Es gilt also $R[x] \subseteq R[[x]]$.*

Der formale Potenzreihenring ist tatsächlich ein kommutativer Ring, wobei sich der Beweis von Satz 7.11(a) wörtlich überträgt. Das Betrachten von formalen Potenzreihen hat gegenüber dem Betrachten von Potenzreihen in der Analysis einige Vorteile:

- Man braucht sich nicht um Konvergenz zu kümmern.
- Die Definition funktioniert über beliebigen Ringen R, auch über solchen, in denen überhaupt kein Konvergenzbegriff existiert.
- Die Aussage, dass zwei Potenzreihen genau dann übereinstimmen, wenn alle ihre Koeffizienten übereinstimmen („Koeffizientenvergleich“), ergibt sich unmittelbar aus der Definition.

Es gibt jedoch, verglichen mit Polynomen und Potenzreihen, auch Einschränkungen:

- Man kann formale Potenzreihen nicht auswerten, d.h. man kann keine Werte einsetzen. Deshalb kann man sie auch nicht als Funktionen $R \to R$ interpretieren.
- Formale Potenzreihen haben keinen Grad. Allerdings bildet das *minimale n* mit $a_n \neq 0$ in mancher Hinsicht einen Ersatz.

Im Polynomring $R[x]$ sind die einzigen invertierbaren Element die konstanten Polynome $a \in R$, bei denen a als Element von R invertierbar ist. Im formalen Potenzreihenring verhält sich dies ganz anders, wie der folgende Satz zeigt.

Satz 22.2. *Eine formale Potenzreihe $f = \sum_{n=0}^{\infty} a_n x^n$ über einem kommutativen Ring R ist genau dann invertierbar (als Element von $R[[x]]$), falls a_0 (als Element von R) invertierbar ist.*

Beweis. Zunächst sei f invertierbar, es gibt also eine formale Potenzreihe $g \in R[[x]]$ mit $fg = 1$. Dann muss das Produkt von a_0 und dem konstanten Koeffizienten von g geich 1 sein, also ist a_0 invertierbar.

Nun setzen wir umgekehrt voraus, dass a_0 invertierbar ist, also gibt es $b_0 \in R$ mit $a_0 b_0 = 1$. Wir definieren rekursiv eine Folge (b_n) durch

$$b_n := -b_0 \cdot \sum_{j=1}^{n} a_j b_{n-j} \quad \text{für} \quad n \geq 1.$$

Es folgt $\sum_{j=0}^{n} a_j b_{n-j} = 0$, für die formale Potenzreihe $g = \sum_{n=0}^{\infty} b_n x^n$ ergibt sich also $f \cdot g = 1$ direkt aus der Definition des Produkts. □

Beispiel 22.3. Das Polynom $1 - x$ hat in $R[[x]]$ die geometrische Reihe $\sum_{n=0}^{\infty} x^n$ als Inverse. Da das inverse Element eines Ringelements r gewöhnlich als r^{-1} oder $\frac{1}{r}$ schreibt, ist die Gleichung

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \tag{22.2}$$

korrekt. Für Ringelemente β und $\gamma \in R$ gilt weiter

$$\frac{\beta}{1-\gamma x} = \sum_{n=0}^{\infty} \beta\gamma^n x^n. \tag{22.3}$$

◁

Nun können wir unsere Rechnungen zu den Fibonacci-Zahlen vollständig rechtfertigen: Die Potenzreihe $f = \sum_{n=0}^{\infty} a_n x^n$ (mit a_n die Fibonacci-Zahlen) ist eine formale Potenzreihe in $\mathbb{C}[[x]]$, und für diese haben wir die Gleichung $f = x + (x + x^2)f$ hergeleitet. Da $1 - x - x^2$ in $\mathbb{C}[[x]]$ invertierbar ist, folgt $f = \frac{x}{1-x-x^2}$. Auch die Rechungen zur Partialbruchzerlegung spielen sich komplett im formalen Potenzreihenring ab, und die Formel (22.1) für die a_n ergibt sich aus (22.3).

In $R[[x]]$ gibt es auch eine ganze Menge Elemente mit Quadratwurzeln. Man kann beispielsweise ähnlich wie in Satz 22.2 zeigen, dass $f = \sum_{n=0}^{\infty} a_n x^n$ in $R[[x]]$ eine Quadratwurzel hat, falls $a_0 \neq 0$ in R eine Quadratwurzel hat und außerdem 2 in R invertierbar ist. Inverse bzw. Quadratwurzeln werden hierbei urch rekursive Formeln gegeben. Aus dem folgenden Satz werden wir dies für einige formale Potenzreihen explizit machen.

Satz 22.4. *Es seien $a, b \in R$ Elemente eines kommutativen Rings, der $\mathbb{Q}$ enthält. Mit*

$$F_a := \sum_{k=0}^{\infty} \binom{a}{k} x^k \in R[[x]]$$

gilt dann

$$F_a \cdot F_b = F_{a+b}.$$

Insbesondere gelten

$$F_a^{-1} = F_{-a} \quad \textit{und} \quad F_a^n = F_{na} \quad \textit{für } n \in \mathbb{N}.$$

Man nennt F_a eine Binomialreihe.

Beweis. Die erste Gleichung ist eine direkte Folgerung aus der Vandermondeschen Identität (Satz 21.4(j)). Die zweite folgt wegen $F_0 = 1$, und die dritte per Induktion. □

Beispiel 22.5. Für und $a = 1$ gilt $F_a = 1 + x$, also

$$\frac{1}{1+x} = \sum_{k=0}^{\infty} \binom{-1}{k} x^k \underset{\text{Satz 21.4(g)}}{=} \sum_{k=0}^{\infty} (-1)^k x^k,$$

woraus (22.2) folgt. Weiter gilt $F_{1/2}^2 = F_1$, was man salopp als

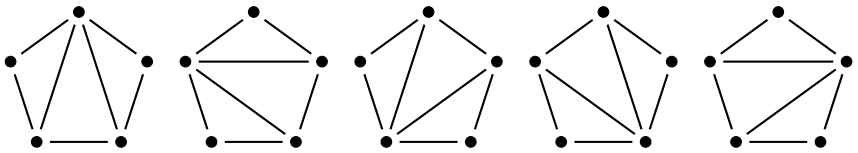
$$\sqrt{1+x} = \sum_{k=0}^{\infty} \binom{1/2}{k} x^k$$

schreiben kann. (Die Schreibweise ist deshalb salopp, weil die Quadratwurzel nicht eindeutig definiert ist.) ◁

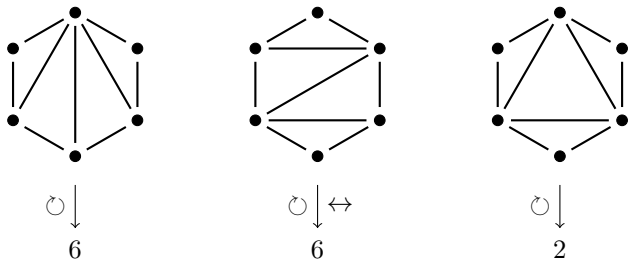
Als eine interessante Anwendung von erzeugenden Funktionen und des obigen Satzes werden wir nun die sogenannten Catalan-Zahlen behandeln. Wir werden zwei Zählprobleme anschauen und jeweils aus Rekursionsgleichungen mit Hilfe von erzeugenden Funktionen Ausdrücke für die gesuchten Zahlen herleiten.

Als erstes fragen wir, wieviele Möglichkeiten es gibt, ein regelmäßiges n-Eck (oder allgemeiner ein konvexes n-Eck) durch Verbindungslinien zwischen einigen der Eckpunkte in Dreiecke aufzuteilen. Eine solche Aufteilung in Dreiecke nennt man eine Triangulation. Triangulationen spielen in der Topologie eine wichtige Rolle, weil sie helfen können, Oberflächen und andere geometrische Objekte zu klassifizieren. Für Vierecke ($n = 4$) gibt es zwei Möglichkeiten:

Für $n = 5$ sieht man, dass es genau fünf Möglichkeiten gibt:



Alle fünf Triangulierungen gehen durch Symmetrie (genauer: Drehungen) auseinander hervor, und ebenso verhält es sich für $n = 4$. Komplizierter ist der Fall $n = 6$. Hier kommt man auf drei wesentlich verschiedene Triangulierungen:



Die ersten beiden Triangulierungen führen durch Anwendung von Drehungen und Spiegelungen zu jeweils fünf weiteren, und die dritte zu einer. Insgesamt erhalten wir so 14 verschiedene Triangulierungen. Durch angestrengtes Nachdenken findet man, dass es keine weiteren gibt. Beginnen wir mit der Anzahl der Triangulierungen für $n = 3$, also mit 1, so erhalten wir die Zahlenfolge

$$1, 2, 5, 14, \ldots$$

Das zweite Zählproblem, das in keinem offensichtlichen Zusammenhang mit dem ersten steht, fragt, wieviele Möglichkeiten es gibt, ein Produkt $A_1 \cdots A_n$ von n nicht kommutierenden quadratischen Matrizen auszurechnen, indem man nacheinander Multiplikationen von jeweils zwei Matrizen ausführt. Diese Möglichkeiten entsprechen *Klammerungen* des Produkts. Für $n = 3$ gibt es beispielsweise zwei Möglichkeiten $K_{3,1}$ und $K_{3,2}$:

$$\underbrace{A_1(A_2A_3)}_{K_{3,1}} \quad \text{und} \quad \underbrace{(A_1A_2)A_3}_{K_{3,2}}.$$

Für $n = 4$ erhalten wir genau fünf Klammerungen:

$$A_1(A_2(A_3A_4)),\ A_1((A_2A_3)A_4),\ (A_1A_2)(A_3A_4),\ (A_1(A_2A_3))A_4 \quad \text{und} \quad ((A_1A_2)A_3)A_4.$$

Bei der dritten Klammerung hat man die Möglichkeiten, zuerst A_1A_2 und dann A_3A_4 auszurechnen oder umgekehrt. Wir lassen jedoch die Reihenfolge der Berechnungen außer Acht, betrachten also tatsächlich nur die Klammerung. Wer in der Auflistung der Klammerungen für $n = 4$ eine Systematik entdeckt hat, kann nun in der selben Weise fortfahren mit dem Fall $n = 5$, bei dem die Sequenz der Klammerungen beginnt mit

$$A_1(A_2(A_3(A_4A_5))),\ A_1(A_2((A_3A_4)A_5)), A_1((A_2A_3)(A_4A_5)) \ldots$$

Am Ende erhält man 14 Klammerungen. Da es für $n = 2$ genau eine Klammerung gibt, erhalten wir die Zahlenfolge

$$1, 2, 5, 14, \ldots$$

Wir beobachten Übereinstimmung für beide Zählprobleme, und solche Übereinstimmungen sind selten Zufall!

Um der Sache auf den Grund zu gehen, erarbeiten wir nun eine Rekursionsformel für die Anzahl a_n der Möglichkeiten, ein Produkt $A_1 \cdots A_n$ zu klammern. Dazu teilen wir eine Klammerung in zwei Teilprodukte auf, im Beispiel $n = 4$ also

$$\underbrace{A_1}_{K_{1,1}}\underbrace{(A_2(A_3A_4))}_{\substack{K_{3,1}\\ j=1}},\ \underbrace{A_1}_{K_{1,1}}\underbrace{((A_2A_3)A_4)}_{\substack{K_{3,2}\\ j=1}},\ \underbrace{(A_1A_2)}_{K_{2,1}}\underbrace{(A_3A_4)}_{\substack{K_{2,1}\\ j=2}},$$

$$\underbrace{(A_1(A_2A_3))}_{\substack{K_{3,1}\\ j=3}}\underbrace{A_4}_{K_{1,1}} \quad \text{und} \quad \underbrace{((A_1A_2)A_3)}_{\substack{K_{3,2}\\ j=3}}\underbrace{A_4}_{K_{1,1}}.$$

Bei jeder Klammerung bezeichnet die darunterstehende Zahl j die Anzahl der Matrizen, die zum linken Teilprodukt gehören. Ist M die Menge aller Kammerungen des Produkts $A_1 \cdots A_n$, so erhalten wir eine disjunkte Zerlegung

$$M = M_1 \,\dot\cup \cdots \dot\cup\, M_{n-1},$$

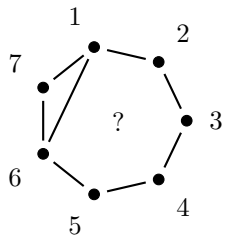
wobei M_j die Menge aller Klammerungen ist, bei der das linke Teilprodukt j Matrizen enthält. Da linkes und rechtes Teilprodukt beliebig geklammert sein können, ergibt sich mit Satz 21.1(c) $|M_j| = a_j \cdot a_{n-j}$, wobei wir (wie oben eingeführt) $a_n := |M|$ schreiben. Dies ergibt mit Satz 21.1(b) die gewünschte Rekursionsformel

$$a_n = \sum_{j=0}^{n} a_j a_{n-j} \qquad (n \geq 2), \tag{22.4}$$

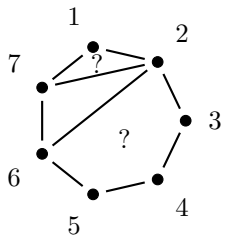
wobei wir $a_0 := 0$ gesetzt haben. Die Folge der a_n ist durch (22.4) und $a_0 = 0$, $a_1 = 1$ eindeutig bestimmt.

Bevor wir eine explizite Formel für die a_n herleiten, wollen wir uns vergewissern, dass unsere beiden Zählprobleme wirklich dieselbe Lösung haben. Dies bedeutet, dass a_n die Anzahl der Triangulierungen eines $(n+1)$-Ecks sein sollte. Wenn wir b_n für die Anzahl der Triangulierungen eines $(n+1)$-Ecks schreiben und $b_0 := 0$, $b_1 := 1$ setzen, so ist die Rekursionsformel (22.4) für die b_n nachzuweisen.

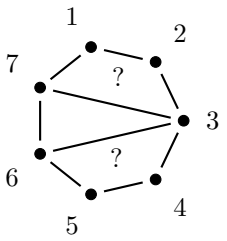
Wir nummerieren die Ecken unseres $(n+1)$-Ecks mit $1, 2, \ldots, n+1$. Bei jeder Triangulierung ist die Kante zwischen den Punkten n und $n+1$ Bestandteil von genau einem Dreieck. Die dritte Ecke dieses Dreiecks sei die Ecke j, also $j \in \{1, \ldots, n-1\}$. Indem wir die Triangulierungen nach dem Wert von j sortieren, erhalten wir eine disjunkte Zerlegung der Menge aller Triangulierungen. Für $n = 6$ sieht diese wie folgt aus:



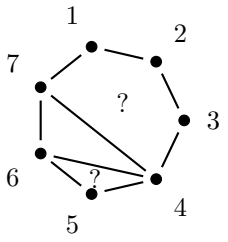
j = 1



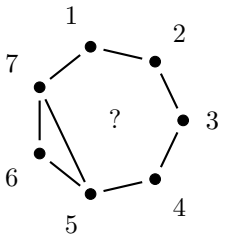
j = 2



j = 3



j = 4



j = 5

Die Fragezeichen deuten dabei an, dass in den m-Ecken ober- und unterhalb des gewählten Dreiecks beliebige Triangulationen vorgenommen werden können. Genauer haben wir oberhalb ein $(j+1)$-Eck und unterhalb ein $(n-j+1)$-Eck, wobei es für $j=1$ bzw. $j=n-1$ kein m-Eck ober- bzw. unterhalb gibt. Für die Anzahl b_n der Triangulationen des $(n+1)$-Ecks ergibt sich damit die Formel

$$b_n = b_{n-1} + \sum_{j=2}^{n-2} b_j b_{n-j} + b_{n-1} = \sum_{j=0}^{n} b_j b_{n-j} \qquad (n \geq 2),$$

wobei sich die letzte Gleichung aufgrund der Konventionen $b_0 = 0$ und $b_1 = 1$ ergibt. Da die Folge der a_n dieselbe Rekursionsgleichung und dieselben Anfangswerte hat, folgt in der Tat $a_n = b_n$ für alle n.

Nun wollen wir eine Formel für die Zahlen $a_n = b_n$ herleiten, und dazu benutzen wir die erzeugende Funktion

$$f := \sum_{n=0}^{\infty} a_n x^n \in \mathbb{C}[[x]].$$

Aus (22.4) und $a_0 = 0$, $a_1 = 1$ erhalten wir

$$f = x + \sum_{n=2}^{\infty} \left(\sum_{j=0}^{n} a_j a_{n-j} \right) x^n = x + \sum_{n=0}^{\infty} \left(\sum_{j=0}^{n} a_j a_{n-j} \right) x^n = x + f^2,$$

also $f^2 - f + x = 0$ und durch Auflösen nach f

$$f = \frac{1 \pm \sqrt{1-4x}}{2}.$$

Die Existenz und Bedeuting der Quadratwurzel wird hierbei durch Beispiel 22.5 gegeben, also

$$\sqrt{1-4x} = \sum_{n=0}^{\infty} \binom{1/2}{n} (-4)^n x^n \in \mathbb{C}[[x]].$$

Der konstante Koeffizient hiervon ist 1. Da f den konstanten Koeffizient 0 hat, muss von den obigen Lösungen für f diejenige mit „$-$“ die richtige sein. Wir erhalten die Formel

$$a_n = \frac{-1}{2}(-4)^n \binom{1/2}{n} \qquad (n \geq 1).$$

Diese Formel lässt sich vereinfachen, und das geht am besten, wenn man die Folge der a_n verschiebt, indem man $c_n := a_{n+1}$ setzt. Für $n \geq 0$ gilt

$$c_n = \frac{-1}{2}(-4)^{n+1}\binom{1/2}{n+1} = \frac{-(-4)^{n+1}\prod_{i=0}^{n}(1/2-i)}{2(n+1)!} =$$
$$\frac{-2^{n+1}\prod_{i=0}^{n}(2i-1)}{2(n+1)!} = \frac{2^n \prod_{i=1}^{n}(2i-1)}{(n+1)!} = \frac{\prod_{i=1}^{n}(2i)\prod_{i=1}^{n}(2i-1)}{n!(n+1)!}$$
$$= \frac{(2n)!}{(n+1)(n!)^2} = \frac{1}{n+1}\binom{2n}{n}.$$

Die Zahl c_n heißt die n-te **Catalan-Zahl**. Die Folge der Catalan-Zahlen beginnt mit

$$c_0 = 1,\ c_1 = 1,\ c_2 = 2,\ c_3 = 5,\ c_4 = 14,\ c_5 = 42, \ldots$$

Wir fassen zusammen:

Satz 22.6. *Für $n \geq 1$ gibt die n-te Catalan-Zahl*

$$c_n = \frac{1}{n+1}\binom{2n}{n} \in \mathbb{N}$$

an, auf wieviele Arten man ein Produkt von $n+1$ nicht kommutierenden Matrizen klammern kann, und auf wieviele Arten man ein regelmäßiges $(n+2)$-Eck triangulieren kann.

Die zwei Anwendungsbeispiele, anhand derer wir die Catalan-Zahlen eingeführt haben, reichen nicht aus, um deren Wichtigkeit erahnen zu lassen. Tatsächlich gehören sie zu den „Stars“ der Kombinatorik, wie man in den Lehrbüchern von Stanley (R.P. Stanley, *Enumerative Combinatorics*, Band 1

und 2, Cambridge University Press), in denen über 60 Anwendungen gegeben werden, nachlesen kann.

Lineare Algebra: Euklidische und unitäre Räume

Bis jetzt haben wir die gesamte Theorie über beliebigen Körpern entwickelt. Dabei hat jeglicher Begriff von „Abstand“ gefehlt. Die Einführung eines Abstandsbegriffs ist über allgemeinen Körpern auch nicht (in geometrisch sinnvoller Weise) möglich. Nun spezialisieren wir den Grundkörper zu $\mathbb{R}$ oder $\mathbb{C}$ und führen das Skalarprodukt ein. Mit diesem werden dann Längen, Abstände und auch Winkel definiert. Schließlich wenden wir uns nochmal der Diagonalisierbarkeit von Matrizen zu.

23 Skalarprodukte

Auf $\mathbb{R}^n$ ist das **Standard-Skalarprodukt** zweier Vektoren $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ und $w = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ durch

$$\langle v, w \rangle := \sum_{i=1}^{n} x_i y_i \quad (= v^T \cdot w) \in \mathbb{R}$$

definiert. Achtung: Die Notation ist anfällig für Verwechselungen mit dem Erzeugnis!

Es gelten die folgenden Regeln:

(a) Für alle $u, v, w \in \mathbb{R}^n$ und $a \in \mathbb{R}$ gelten:

$$\langle u, v + a \cdot w \rangle = \langle u, v \rangle + a \cdot \langle u, w \rangle$$

und

$$\langle u + a \cdot v, w\rangle = \langle u, w\rangle + a \cdot \langle v, w\rangle.$$

(Man sagt auch, dass das Skalarprodukt **bilinear** ist.)

(b) Für $v, w \in \mathbb{R}^n$ gilt

$$\langle v, w\rangle = \langle w, v\rangle.$$

(Man sagt auch, dass das Skalarprodukt **symmetrisch** ist.)

(c) Für $v \in \mathbb{R}^n$ mit $v \neq 0$ gilt

$$\langle v, v\rangle > 0.$$

(Man sagt auch, dass das Skalarprodukt **positiv definit** ist.)

Wir nehmen dies zum Anlass für folgende Definition:

Definition 23.1. *Es sei V ein* reeller Vektorraum *(d.h. ein Vektorraum über $\mathbb{R}$). Eine Abbildung*

$$V \times V \to \mathbb{R},\ (v, w) \mapsto \langle v, w\rangle$$

heißt eine **symmetrische Bilinearform***, falls sie symmetrisch und bilinear ist. Eine symmetrische Bilinearform heißt ein* **Skalarprodukt***, wenn sie zusätzlich positiv definit ist.*

Ein reeller Vektorraum zusammen mit einem Skalarprodukt heißt ein **euklidischer Raum***.*

Beispiel 23.2. (1) $V = \mathbb{R}^n$ ist zusammen mit dem Standardskalarprodukt ein euklidischer Raum.

(2) Für reelle Zahlen $a < b$ sei $V := C([a, b], \mathbb{R})$ der Vektorraum aller stetiger Funktionen $[a, b] \to \mathbb{R}$ auf dem abgeschlossenen Intervall $[a, b]$. Durch

$$\langle f, g\rangle := \int_a^b f(x)g(x)dx$$

wird ein Skalarprodukt auf V definiert.

(3) Auf $V = \mathbb{R}^2$ wird für $v = \left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)$ und $w = \left(\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right)$ ein Skalarprodukt erklärt durch

$$\langle v, w\rangle = 5x_1y_1 + 3x_1y_2 + 3x_2y_1 + 2x_2y_2.$$

Die Bilinearität und Symmetrie sind klar, und die positive Definitheit geht aus

$$\langle v, v\rangle = 5x_1^2 + 6x_1x_2 + 2x_2^2 = (2x_1 + x_2)^2 + (x_1 + x_2)^2$$

hervor.

(4) Ebenso wie oben kann man

$$\langle v, v\rangle = x_1y_1 - x_2y_2$$

definieren und erhält ein Beispiel für eine nicht positiv definite, symmetrische Bilinearform. ◁

Zu einer symmetrischen Bilinearform auf $\mathbb{R}^n$ erhält man durch Einsetzen der Standardbasisvektoren Zahlen $a_{i,j} := \langle e_i, e_j \rangle$, die man zu einer Matrix $A = (a_{i,j}) \in \mathbb{R}^{n \times n}$ zusammenfassen kann. A ist symmetrisch und wird die **Darstellungsmatrix** der symmetrischen Bilinearform genannt. Die Bilinearform wird durch A „codiert," denn für $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ und $w = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ gilt

$$\langle v, w \rangle = \langle \sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{n} y_j e_j \rangle = \sum_{i,j=1}^{n} x_i y_j a_{i,j} = v^T \cdot A \cdot w.$$

Die Darstellungsmatrix des Standard-Skalarprodukts ist die Einheitsmatrix.

Allgemeiner kann man auch Darstellungsmatrizen von symmetrischen Bilinearformen auf endlich-dimensionalen Vektorräumen betrachten, indem man eine Basis wählt und die Basisvektoren in die Form einsetzt. Nun kann man auch überlegen, wie sich ein Basiswechsel auf die Darstellungsmatrix auswirkt. Wir werden dieses Thema nicht weiter verfolgen, sondern uns nun mit komplexen Vektorräumen beschäftigen.

In einem komplexen Vektorraum V (d.h. einem Vektorraum über $\mathbb{C}$) kann es kein Skalarprodukt im Sinne von Definition 23.1 geben (es sei denn, $V = \{0\}$). Denn für $0 \neq v \in V$ müsste $\langle v, v \rangle > 0$ gelten, also

$$\langle iv, iv \rangle = i^2 \langle v, v \rangle = -\langle v, v \rangle < 0.$$

(Darüber hinaus wäre beispielsweise $\langle (i+1) \cdot v, (i+1) \cdot v \rangle = 2i \langle v, v \rangle$ nicht einmal reell.) Man behilft sich, indem man die *komplexe Konjugation* benutzt, die wir nun in Erinnerung rufen: Für $z = a + bi \in \mathbb{C}$ ist das **komplex konjugierte**

$$\overline{z} := a - bi \in \mathbb{C}.$$

Man rechnet nach, dass für $z, w \in \mathbb{C}$ die Regeln

$$\overline{z + w} = \overline{z} + \overline{w} \quad \text{und} \quad \overline{z \cdot w} = \overline{z} \cdot \overline{w}$$

gelten. Wir haben es also mit einem Ring-Homomorphismus zu tun. Außerdem gilt

$$\overline{z} \cdot z = a^2 + b^2 \in \mathbb{R}_{\geq 0},$$

was die Definition des Betrags $| z | := \sqrt{\overline{z} \cdot z}$ möglich macht. Nur die Null hat den Betrag Null. Es ist klar, dass z genau dann reell ist, wenn $z = \overline{z}$.

Das Standard-Skalarprodukt auf $\mathbb{R}^n$ wird nun ersetzt durch das Produkt

$$\langle v, w \rangle := \sum_{i=1}^{n} \overline{x}_i y_i \quad (= \overline{v}^T \cdot w) \in \mathbb{C} \tag{23.1}$$

für $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ und $w = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{C}^n$ ersetzt. Dies ist ein komplexes Skalarprodukt gemäß der folgenden Definition.

Definition 23.3. *Es sei V ein komplexer Vektorraum. Eine Abbildung*

$$V \times V \to \mathbb{C}, \ (v, w) \mapsto \langle v, w \rangle$$

heißt

(a) **sesquilinear**, *falls für $u, v, w \in V$ und $a \in \mathbb{C}$ die Regeln*

$$\langle u, v + a \cdot w \rangle = \langle u, v \rangle + a \cdot \langle u, w \rangle$$

und

$$\langle u + a \cdot v, w \rangle = \langle u, w \rangle + \overline{a} \cdot \langle v, w \rangle$$

gelten;

(b) **hermitesch**, *falls für $v, w \in V$ die Regel*

$$\langle v, w \rangle = \overline{\langle w, v \rangle}$$

gilt;

(c) **positiv definit**, *falls für $v \in V \setminus \{0\}$*

$$\langle v, v \rangle \in \mathbb{R} \quad \textit{und} \quad \langle v, v \rangle > 0$$

gilt.

Man spricht dann auch von einer **Sesquilinearform** *bzw. einer* **hermiteschen Form**. *Eine positiv definite, hermitesche Sesquilinearform heißt ein* **komplexes Skalarprodukt**.

Ein komplexer Vektorraum zusammen mit einem komplexen Skalarprodukt heißt ein **unitärer Raum**.

Anmerkung. Man drückt die Bedingung der Sesquilinearität auch aus, indem man sagt, dass die Form linear im zweiten und *semilinear* im ersten Argument ist. Einige Autoren treffen die umgekehrte Konvention, indem sie Linearität im ersten und Semilinearität im zweiten Argument fordern. ◁

Beispiel 23.4. (1) $V = \mathbb{C}^n$ mit dem *Standardprodukt* (23.1) ist ein unitärer Raum.

(2) Für reelle Zahlen $a < b$ sei $V := C([a, b], \mathbb{C})$ der Vektorraum aller stetiger Funktionen $[a, b] \to \mathbb{C}$ auf dem abgeschlossenen Intervall $[a, b] \subseteq \mathbb{R}$. Durch

$$\langle f, g \rangle := \int_a^b \overline{f(x)} g(x) dx$$

wird ein komplexes Skalarprodukt auf V definiert. ◁

Zu einer hermiteschen Sesquilinearform auf einem endlich-dimensionalen Vektorraum mit einer Basis $\{v_1, \ldots, v_n\}$ erhält man eine Matrix $A = (a_{i,j}) \in \mathbb{C}^{n\times n}$ durch $a_{i,j} := \langle v_i, v_j\rangle$. Es folgt $a_{i,j} = \overline{a_{j,i}}$ für alle $i, j \in \{1, \ldots, n\}$, also

$$A^T = \overline{A}.$$

Matrizen mit dieser Eigenschaft nennt man **hermitesch**. Die Darstellungsmatrizen von hermiteschen Sesquilinearformen sind also hermitesche Matrizen.

Von nun an sei V ein euklidischer oder unitärer Raum. Wir kommen nun zum Abstands- und Längenbegriff.

Definition 23.5. *Für $v \in V$ heißt*

$$|| \, v \, ||:= \sqrt{\langle v, v\rangle} \in \mathbb{R}_{\geq 0}$$

die **Länge** *(auch:* **Norm***) von v.*

Für $v, w \in V$ heißt

$$d(v, w) := ||v - w|| \in \mathbb{R}_{\geq 0}$$

der **Abstand** *von v und w.*

Proposition 23.6 (Schwarzsche Ungleichung)**.** *Für $v, w \in V$ gilt*

$$|\langle v, w\rangle| \leq ||v|| \cdot ||w||.$$

Hierbei gilt Gleichheit genau dann, wenn v und w linear abhängig sind.

Beweis. Wir können $w \neq 0$ annehmen, da für $w = 0$ die Ungleichung und die Zusatzbehauptung erfüllt sind.

Für $a \in \mathbb{R}$ oder (im Falle eines komplexen Vektorraums) $a \in \mathbb{C}$ gilt

$$0 \leq ||v - aw||^2 = \langle v - aw, v - aw\rangle = ||v||^2 - a\langle v, w\rangle - \overline{a}\langle w, v\rangle + \overline{a}a||w||^2.$$

Speziell für $a = \frac{\langle w,v\rangle}{||w||^2}$ ergibt dies

$$\begin{aligned} 0 &\leq ||v||^2 - \frac{\langle w, v\rangle\langle v, w\rangle}{||w||^2} - \frac{\overline{\langle w, v\rangle}\langle w, v\rangle}{||w||^2} + \frac{\overline{\langle w, v\rangle}\langle w, v\rangle}{||w||^2} \\ &= \frac{1}{||w||^2}\left(||v||^2||w||^2 - |\langle v, w\rangle|^2\right). \end{aligned}$$

Dies liefert die Ungleichung und zeigt, dass genau dann Gleichheit gilt, wenn $v = \frac{\langle w,v\rangle}{||w||^2} \cdot w$. Die lineare Abhängigkeit ist also notwendig für die Gleichheit. Ist umgekehrt $v = aw$ mit $a \in \mathbb{R}$ bzw. $a \in \mathbb{C}$, so folgt

$$\frac{\langle w, v\rangle}{||w||^2} = \frac{a||w||^2}{||w||^2} = a,$$

also Gleichheit. □

Nun können wir die wichtigsten Eigenschaften der Länge und des Abstands beweisen.

Satz 23.7. *Für alle $u, v, w \in V$ und $a \in \mathbb{R}$ bzw. $a \in \mathbb{C}$ gelten:*

(a) *Falls $v \neq 0$, so folgt $||v|| > 0$.*
(b) $||a \cdot v|| = |a| \cdot ||v||$.
(c) $||v + w|| \leq ||v|| + ||w||$ *(*Dreiecksungleichung*).*
(d) *Falls $v \neq w$, so folgt $d(v, w) > 0$.*
(e) $d(v, w) = d(w, v)$.
(f) $d(u, w) \leq d(u, v) + d(v, w)$ *(*Dreiecksungleichung*).*

Beweis. Die Teile (a), (b), (d) und (e) sind unmittelbar klar. Für den Nachweis von (c) rechnen wir:

$$\begin{aligned} ||v+w||^2 &= ||v||^2 + \langle v, w\rangle + \langle w, v\rangle + ||w||^2 = ||v||^2 + 2\,\mathrm{Re}\,(\langle v, w\rangle) + ||w||^2 \\ &\leq ||v||^2 + 2\,|\langle v, w\rangle| + ||w||^2 \underset{\text{Proposition 23.6}}{\leq} ||v||^2 + 2||v|| \cdot ||w|| + ||w||^2 \\ &= (||v|| + ||w||)^2\,, \end{aligned}$$

wobei $\mathrm{Re}(z) := a$ für $z = a + bi \in \mathbb{C}$ den Realteil bezeichnet. Der Nachweis von (f) wird durch

$$d(u, w) = ||u - w|| = ||u - v + v - w|| \underset{\text{(c)}}{\leq} ||u - v|| + ||v - w|| = d(u, v) + d(v, w)$$

erbracht. □

Wir nehmen diesen Satz zum Anlass, ein paar Begriffe zu erwähnen, die in dieser Vorlesung nicht weiter vorkommen werden.

Anmerkung 23.8. (a) Ein **normierter Vektorraum** ist ein reeller oder komplexer Vektorraum V mit einer Abbildung

$$V \to \mathbb{R}_{\geq 0},\ v \mapsto ||v||,$$

die (a)–(c) aus Satz 23.7 erfüllt.

(b) Ein **metrischer Raum** ist eine Menge V mit einer Abbildung

$$d\colon V \times V \to \mathbb{R}_{\geq 0},$$

die (d)–(f) aus Satz 23.7 erfüllt. Die Abbildung d heißt dann eine **Metrik** auf V.

(c) Sobald man einen Abstandsbegriff hat, kann man von konvergenten Folgen und von Cauchy-Folgen sprechen. Vollständigkeit bedeutet, dass jede Cauchy-Folge konvergent ist. Ein **Banachraum** ist ein vollständiger normierter Raum. Ein **Hilbertraum** ist ein vollständiger euklidischer oder unitärer Raum. ◁

Wir erhalten eine hierarchische Anordnung unserer Begriffe: Jeder euklidische oder unitäre Raum ist normiert, und jeder normierte Raum ist metrisch. Jeder Hilbertraum ist ein Banachraum.

Beispiel 23.9. (1) Beispiele für Normen, die nicht von einem Skalarprodukt kommen, sind die *Manhattan-Norm* auf $\mathbb{R}^n$, definiert durch

$$||v|| = \sum_{i=1}^{n} |v_i|$$

(wobei v_i die Komponenten von $v \in \mathbb{R}^n$ sind) und die *Maximum-Norm* auf $C([a,b],\mathbb{C})$, definiert durch

$$||f|| := \max\{|f(x)| \mid x \in \mathbb{R},\ a \le x \le b\}\,.$$

(2) Ein Beispiel für eine Metrik, die nicht von einer Norm kommt, ist die *Hamming-Metrik* auf $\mathbb{R}^n$ (oder K^n mit einem Körper K), definiert durch

$$d(v,w) := |\{i \in \{1,\dots,n\} \mid v_i \neq w_i\}|\,,$$

wobei v_i und w_i die Komponenten von $v, w \in \mathbb{R}^n$ sind.

(3) Es ist nicht schwer zu zeigen, dass jeder endlich-dimensionale euklidische oder unitäre Raum ein Hilbertraum ist. Ebenso ist jeder endlich-dimensionale normierte Raum ein Banachraum.

(4) Der euklidische Raum $C([a,b],\mathbb{R})$ (siehe Beispiel 23.2(2)) ist nicht vollständig, also kein Hilbertraum.

(5) Man kann zeigen, dass $C([a,b],\mathbb{R})$ und $C([a,b],\mathbb{C})$ zusammen mit der Maximum-Norm (siehe (1)) Banachräume sind. Der durch die Maximum-Norm gegebene Konvergenzbegriff ist die gleichmäßige Konvergenz.

(6) Das wohl einfachste Beispiel für einen unendlich-dimensionalen Hilbertraum ist der Raum ℓ^2 aller komplexer Folgen $\mathbf{a} = (a_n)$ mit der Eigenschaft, dass $\sum_{n=1}^{\infty} |a_n|^2$ konvergiert. Das Skalarprodukt wird durch

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{n=1}^{\infty} \overline{a}_n b_n$$

definiert. Der Nachweis der Vollständigkeit von ℓ^2 ist nicht ganz einfach. ◁

Die Schwarzsche Ungleichung (Proposition 23.6) ermöglicht es, für Vektoren $v, w \in V$ positiver Länge in einem *euklidischen* Raum den **Winkel** zwischen v und w als die eindeutig bestimmte Zahl α in dem abgeschlossenen Intervall $[0,\pi]$ mit

$$\cos(\alpha) = \frac{\langle v, w \rangle}{||v|| \cdot ||w||}$$

zu definieren. Diese Definition erscheint zunächst willkürlich, sie liefert aber genau das Erwartete.

Beispiel 23.10. Für $v = \binom{1}{0}$ und $w = \binom{1}{1} \in \mathbb{R}^2$ ist

$$\frac{\langle v, w\rangle}{||v|| \cdot ||w||} = \frac{1}{\sqrt{2}},$$

also beträgt der Winkel $\pi/4$. ◁

In unitären Räumen lässt sich kein sinnvoller Winkelbegriff definieren, man kann aber (ebenso wie in euklidischen Räumen) davon sprechen, dass zwei Vektoren senkrecht aufeinander stehen. Dies ist Inhalt der folgenden Definition.

Definition 23.11. *Es sei V ein euklidischer oder unitärer Raum.*

(a) Zwei Vektoren $v, w \in V$ heißen **orthogonal** *(gleichbedeutend:* **senkrecht***), falls*

$$\langle v, w\rangle = 0.$$

(b) Eine Menge $S \subseteq V$ heißt ein **Orthogonalsystem***, falls je zwei Vektoren $v, w \in S$ mit $v \neq w$ orthogonal sind.*

(c) Ein Orthogonalsystem $S \subseteq V$ heißt ein **Orthonormalsystem***, falls zusätzlich alle Vektoren $v \in S$ die Länge $||v|| = 1$ haben.*

(d) Ein Orthonormalsystem $S \subseteq V$ heißt **Orthonormalbasis***, falls es zusätzlich eine Basis ist.*

(e) Zu einem Unterraum $U \subseteq V$ heißt

$$U^\perp := \{v \in V \mid \langle v, u\rangle = 0 \text{ für alle } u \in U\}$$

das **orthogonale Komplement** *von U. Es ist klar, dass $U^\perp$ ein Unterraum von V ist.*

Beispiel 23.12. (1) Die Standardbasis ist eine Orthonormalbasis von $\mathbb{R}^n$ bzw. $\mathbb{C}^n$ mit dem Standard-Skalarprodukt.

(2) Die Vektoren

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\1\end{pmatrix} \quad \text{und} \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\-1\end{pmatrix}$$

bilden ein Orthonormalsystem im $\mathbb{R}^3$.

(3) Im Raum $C([0, 2\pi], \mathbb{C})$ der stetigen komplexen Funktionen auf den Intervall $[0, 2\pi]$ mit dem Skalarprodukt aus Beispiel 23.4 bilden die Funktionen

$$f_n(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{int} \quad (n \in \mathbb{Z})$$

ein Orthonormalsystem. Die Theorie der Fourierreihen basiert hierauf. ◁

Satz 23.13. *Jedes Orthogonalsystem $S \subseteq V$ in einem euklidischen oder unitären Raum, das nicht den Nullvektor enthält, ist linear unabhängig. Falls $|S| = \dim(V) < \infty$, so ist S eine Basis.*

Beweis. Seien $v_1, \ldots, v_n \in S$ paarweise verschieden. Weiter sei

$$a_1 v_1 + \cdots + a_n v_n = 0$$

mit $a_i \in \mathbb{R}$ bzw. $a_i \in \mathbb{C}$. Für alle $j \in \{1, \ldots, n\}$ folgt durch Bildung des Skalaprodukts mit v_j:

$$0 = \langle v_j, 0\rangle = \left\langle v_j, \sum_{i=1}^{n} a_i v_i \right\rangle = \sum_{i=1}^{n} a_i \langle v_j, v_i\rangle = a_j \langle v_j, v_j\rangle.$$

Wegen $v_j \neq 0$ sind also sind alle $a_j = 0$, und die lineare Unabhängigkeit ist bewiesen.

Die zweite Aussage folgt mit Korollar 10.15(a). □

Orthonormalbasen haben einige günstige Eigenschaften. Ist beispielsweise $S = \{v_1, \ldots, v_n\}$ eine Orthonormalbasis eines endlich-dimensionalen euklidischen oder unitären Raums und $v \in V$, so sind die Skalarprodukte $\langle v_i, v\rangle$ genau die Koordinaten von v bezüglich der Basis S. Gilt nämlich $v = a_1 v_1 + \cdots + a_n v_n$, so folgt

$$\langle v_i, v\rangle = \left\langle v_i, \sum_{j=1}^{n} a_j v_j \right\rangle = \sum_{j=1}^{n} a_j \langle v_i, v_j\rangle = a_i \langle v_i, v_i\rangle = a_i.$$

Mit Orthonormalbasen lassen sich also Koeffizienten „isolieren“. Es stellt sich die Frage, ob jeder endlich-dimensionale euklidische oder unitäre Raum eine Orthonormalbasis hat. Diese Frage werden wir konstruktiv durch das Schmidtsche Orthogonalisierungsverfahren beantworten.

Algorithmus 23.14 (Schmidtsches Orthogonalisierungsverfahren)**.**

Eingabe: Vektoren $v_1, \ldots, v_k$ eines euklidischen oder unitären Raums V.
Ausgabe: Eine Orthonormalbasis $\{u_1, \ldots, u_m\}$ des von den v_i erzeugten Unterraums von V.

(1) Setze $m := 0$.
(2) Für $i = 1, \ldots, k$ führe Schritte (3) und (4) aus.
(3) Setze

$$w_i := v_i - \sum_{j=1}^{m} \langle u_j, v_i\rangle \cdot u_j. \tag{23.2}$$

(Im Fall $m = 0$ bedeutet dies $w_i := v_i$.)
(4) Falls $w_i \neq 0$, setze $m := m + 1$ und

$$u_m := \frac{1}{||w_i||} \cdot w_i.$$

Satz 23.15. *Algorithmus 23.14 liefert eine Orthonormalbasis von* $\langle v_1, \ldots, v_k \rangle \subseteq V$.

Beweis. Wir benutzen Induktion nach der Anzahl k der Erzeuger von V und können $k \geq 1$ voraussetzen. Nach Induktion gelten nach Durchlaufen der Schleife für $i = 1, \ldots, k-1$:

$$\langle u_i, u_j \rangle = \delta_{i,j} \quad (1 \leq i, j \leq m) \tag{23.3}$$

und

$$\langle v_1, \ldots, v_{k-1} \rangle = \langle u_1, \ldots, u_m \rangle, \tag{23.4}$$

wobei m das „aktuelle“ m nach $k-1$ Schleifendurchläufen ist. Aus (23.2) folgt für $i \leq m$

$$\langle u_i, w_k \rangle = \langle u_i, v_k \rangle - \sum_{j=1}^{m} \langle u_j, v_k \rangle \cdot \langle u_i, u_j \rangle \underset{(23.3)}{=} \langle u_i, v_k \rangle - \langle u_i, v_k \rangle = 0.$$

Außerdem folgt aus (23.2)

$$\langle u_1, \ldots, u_m, w_k \rangle = \langle u_1, \ldots, u_m, v_k \rangle \underset{(23.4)}{=} \langle v_1, \ldots, v_k \rangle.$$

Falls $w_k = 0$, so folgt $\langle v_1, \ldots, v_k \rangle = \langle u_1, \ldots, u_m \rangle$. Falls $w_k \neq 0$, so wird $\{u_1, \ldots, u_{m+1}\}$ ein Orthonormalsystem und ein Erzeugendensystem von $\langle v_1, \ldots, v_k \rangle$, also nach Satz 23.13 eine Orthonormalbasis. □

Beispiel 23.16. Wir wollen Algorithmus 23.14 auf

$$V := \left\langle \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \right\rangle \subseteq \mathbb{R}^3$$

anwenden. Wir erhalten

$$w_1 = v_1 = \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix} \quad \text{und} \quad u_1 = \frac{1}{||w_1||} \cdot w_1 = \begin{pmatrix} 3/5 \\ 0 \\ 4/5 \end{pmatrix}.$$

Im zweiten Schritt erhalten wir

$$w_2 = v_2 - \langle u_1, v_2 \rangle \cdot u_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{3}{5} \cdot \begin{pmatrix} 3/5 \\ 0 \\ 4/5 \end{pmatrix} = \frac{1}{25} \cdot \begin{pmatrix} 16 \\ 0 \\ -12 \end{pmatrix}$$

und

$$u_2 = \frac{1}{||w_2||} \cdot w_2 = \begin{pmatrix} 4/5 \\ 0 \\ -3/5 \end{pmatrix}.$$

Der dritte Schritt liefert

$$w_3 = v_3 - \langle u_1, v_3\rangle \cdot u_1 - \langle u_2, v_3\rangle \cdot u_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} - \frac{11}{5} \cdot \begin{pmatrix} 3/5 \\ 0 \\ 4/5 \end{pmatrix} + \frac{2}{5} \cdot \begin{pmatrix} 4/5 \\ 0 \\ -3/5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Also ist $\{u_1, u_2\}$ eine Orthonormalbasis von V. ◁

Wenn man das Schmidtsche Orthogonalisierungsverfahren auf eine Basis $B = \{v_1, \ldots, v_k\}$ von V anwendet, bekommt man eine Orthonormalbasis $B' = \{u_1, \ldots, u_k\}$. Es ist interessant, dass die Basiswechselmatrix $S_{B,B'}$ automatisch eine obere Dreiecksmatrix wird. Dies folgt aus (23.4).

Aus der Korrektheit von Algorithmus 23.14 folgt:

Korollar 23.17. *Jeder endlich-dimensionale euklidische oder unitäre Raum hat eine Orthonormalbasis.*

Zwischen euklidischen bzw. unitären Räumen kann man „strukturerhaltende“ Abbildungen studieren.

Definition 23.18. *Es seien V und W zwei euklidische bzw. zwei unitäre Räume. Eine lineare Abbildung $\varphi\colon V \to W$ heißt* **orthogonal** *bzw.* **unitär**, *falls für alle $u, v \in V$ gilt:*

$$\langle \varphi(u), \varphi(v)\rangle = \langle u, v\rangle.$$

Eine unitäre oder orthogonale Abbildung φ ist injektiv, denn aus $\varphi(v) = 0$ für $v \in V$ folgt $\langle v, v\rangle = \langle \varphi(v), \varphi(v)\rangle = 0$, also $v = 0$. Weiter gilt

$$||\varphi(v)|| = ||v||$$

für alle $v \in V$ und damit auch

$$d(\varphi(u), \varphi(v)) = d(u, v)$$

für $u, v \in V$, φ ist also „abstandserhaltend“. Abbildungen zwischen metrischen Räumen mit dieser Eigenschaft nennt man auch *Isometrien*. Es ist nicht schwer zu zeigen, dass jede lineare Isometrie zwischen euklidischen oder unitären Räumen eine orthogonale bzw. unitäre Abbildung ist.

Beispiel 23.19. (1) Jede Drehung um den Nullpunkt definiert eine orthogonale Abbildung $\mathbb{R}^2 \to \mathbb{R}^2$.

(2) Auf dem Raum $V = C([a,b],\mathbb{C})$ der stetigen Funktionen eines Intervalls $[a,b]$ in $\mathbb{C}$ wird durch $\varphi\colon V \to V,\ f \mapsto \hat{f}$ mit $\hat{f}(x) = f(a+b-x)$ eine unitäre Abbildung gegeben. ◁

Was sind die orthogonalen bzw. unitären Abbildungen $V \to V$ für $V = K^n$ mit $K = \mathbb{R}$ bzw. $K = \mathbb{C}$? Ist φ eine solche, so muss φ jede Orthonormalbasis wieder auf eine Orthonormalbasis abbilden. Ist $A \in K^{n\times n}$ die Darstellungsmatrix von φ bezüglich der Standardbasis (also $\varphi = \varphi_A$), so folgt, dass die Spalten von A eine Orthonormalbasis von V bilden. Dies kann man ausdrücken durch die Bedingungen

$$A^T \cdot A = I_n \quad (\text{für } K = \mathbb{R}) \tag{23.5}$$

bzw.

$$\overline{A}^T \cdot A = I_n \quad (\text{für } K = \mathbb{C}), \tag{23.6}$$

wobei $\overline{A}$ durch komplexe Konjugation aller Einträge aus A hervorgeht. (Die zweite Bedingung umfasst eigentlich die erste, da $\overline{A} = A$ für $K = \mathbb{R}$.) Ist umgekehrt $A \in K^{n\times n}$ eine Matrix, die (23.5) bzw. (23.6) erfüllt, so folgt für $u, v \in V$

$$\langle \varphi_A(u), \varphi_A(v)\rangle = (\overline{Au})^T \cdot (Av) = \overline{u}^T \overline{A}^T A v = \langle u, v\rangle.$$

Dies bedeutet, dass genau die Matrizen mit (23.5) bzw. (23.6) orthogonale bzw. unitäre Abbildungen $V \to V$ definieren. Wir nehmen dies zum Anlass für die folgende Definition.

Definition 23.20. *(a) Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt* **orthogonal**, *falls sie (23.5) erfüllt. Dies ist gleichbedeutend damit, dass die Spalten von A eine Orthonormalbasis von $\mathbb{R}^n$ bilden, und wegen $A \cdot A^T = I_n$ auch damit, dass die Zeilen von A eine Orthonormalbasis von $\mathbb{R}^n$ bilden.*

(b) Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt **unitär**, *falls sie (23.6) erfüllt. Dies ist gleichbedeutend damit, dass die Spalten von A eine Orthonormalbasis von $\mathbb{C}^n$ bilden, und wegen $A \cdot \overline{A}^T = I_n$ auch damit, dass die Zeilen von A eine Orthonormalbasis von $\mathbb{C}^n$ bilden.*

(c) Die Untergruppe

$$\mathrm{O}_n := \left\{A \in \mathbb{R}^{n\times n} \mid A^T \cdot A = I_n\right\} \subseteq \mathrm{GL}_n(\mathbb{R})$$

heißt die **orthogonale Gruppe**, *und*

$$\mathrm{SO}_n := \mathrm{O}_n \cap \mathrm{SL}_n(\mathbb{R})$$

heißt die **spezielle orthogonale Gruppe**.

(d) Die Untergruppe

$$\mathrm{U}_n := \left\{A \in \mathbb{C}^{n\times n} \mid \overline{A}^T \cdot A = I_n\right\} \subseteq \mathrm{GL}_n(\mathbb{C})$$

heißt die **unitäre Gruppe**, *und*

$$\mathrm{SU}_n := \mathrm{U}_n \cap \mathrm{SL}_n(\mathbb{C})$$

heißt die **spezielle unitäre Gruppe**.

Besonders interessante orthogonale bzw. unitäre Abbildungen sind sogenannte Spiegelungen, die man folgendermaßen definieren kann. Ist $e \in V$ ein Vektor mit $||e|| = 1$, so heißt

$$\varphi_e \colon V \to V,\ v \mapsto v - 2\langle e, v\rangle \cdot e$$

die **Spiegelung** entlang e. Der folgende Satz sagt aus, dass die orthogonale Gruppe O_n durch Spiegelungen erzeugt werden.

Satz 23.21. *Es sei V ein euklidischer oder unitärer Raum.*

(a) Jede Spiegelung φ_e (mit $e \in V$, $||e|| = 1$) ist eine orthogonale bzw. unitäre Abbildung.

(b) Ist V euklidisch und $n = \dim(V) < \infty$, so lässt sich jede orthogonale Abbildung $\varphi\colon V \to V$ als Komposition von höchstens n Spiegelungen schreiben. Die orthogonale Gruppe wird also durch Spiegelungen erzeugt.

Beweis. (a) Es ist klar, dass φ_e linear ist. Für $v, w \in V$ gilt

$$\begin{aligned}\langle \varphi_e(v), \varphi_e(w)\rangle &= \Big\langle v - 2\langle e, v\rangle \cdot e, w - 2\langle e, w\rangle \cdot e \Big\rangle \\ &= \langle v, w\rangle - 2\langle e, w\rangle\langle v, e\rangle - 2\overline{\langle e, v\rangle}\langle e, w\rangle + 4\overline{\langle e, v\rangle}\langle e, w\rangle \\ &= \langle v, w\rangle,\end{aligned}$$

also ist φ_e orthogonal bzw. unitär.

(b) Wir führen den Beweis per Induktion nach n. Im Fall $\varphi = \mathrm{id}_V$ (der den Induktionsanfang $n = 0$ einschließt) ist nichts zu zeigen. Wir setzen also $\varphi \neq \mathrm{id}_V$ voraus und wählen $v \in V$ mit $\varphi(v) \neq v$. Mit

$$e := \frac{1}{||\varphi(v) - v||} \cdot (\varphi(v) - v)$$

folgt

$$\begin{aligned}\varphi_e(v) &= v - 2\frac{\langle \varphi(v) - v, v\rangle}{||\varphi(v) - v||^2} \cdot (\varphi(v) - v) \\ &= v - 2\frac{\langle \varphi(v), v\rangle - ||v||^2}{||\varphi(v)||^2 - 2\langle \varphi(v), v\rangle + ||v||^2} \cdot (\varphi(v) - v) \\ &= v + (\varphi(v) - v) = \varphi(v).\end{aligned}$$

Nun setzen wir

$$\varphi' := \varphi_e^{-1} \circ \varphi$$

und bemerken, dass auch φ' orthogonal ist. Es folgt $\varphi'(v) = v$. Für $u \in U := \langle v\rangle^{\perp}$ folgt

$$\langle v, \varphi'(u)\rangle = \langle \varphi'(v), \varphi'(u)\rangle = \langle v, u\rangle = 0,$$

also $\varphi'(u) \in U$. Damit ist die Einschränkung $\varphi'|_U$ eine orthogonale Abbildung auf U. Wegen $\dim(U) < n$ erhalten wir per Induktion die Existenz von $e_1, \ldots, e_k \in U$ mit $k < n$ und $||e_i|| = 1$, so dass

$$\varphi'|_U = \varphi_{e_1} \circ \cdots \circ \varphi_{e_k},$$

wobei die φ_{e_i} hier Spiegelungen auf U sind. Wenn wir die φ_{e_i} als Spiegelungen von V auffassen, gilt $\varphi_{e_i}(v) = v$ wegen $e_i \in U$. Es sei nun $w \in V$. Mit $a := \frac{\langle v,w\rangle}{\langle v,v\rangle}$ gilt dann $w - av \in U$, also

$$\begin{aligned}\varphi'(w) &= \varphi'(av) + \varphi'(w - av) = av + (\varphi_{e_1} \circ \cdots \circ \varphi_{e_k})(w - av)\\ &= (\varphi_{e_1} \circ \cdots \circ \varphi_{e_k})(w).\end{aligned}$$

Also gilt $\varphi' = \varphi_{e_1} \circ \cdots \circ \varphi_{e_k}$ und damit $\varphi = \varphi_e \circ \varphi_{e_1} \circ \cdots \circ \varphi_{e_k}$. □

24 Der Spektralsatz

In diesem Abschnitt steht V wieder für einen euklidischen oder unitären Raum.

Definition 24.1. *Sei $\varphi\colon V \to V$ eine lineare Abbildung. Eine lineare Abbildung $\psi\colon V \to V$ heißt zu φ* **adjungiert***, falls für alle $v, w \in V$ gilt:*

$$\langle v, \varphi(w)\rangle = \langle \psi(v), w\rangle.$$

In diesem Fall schreiben wir auch $\psi = \varphi^*$.

Es besteht Verwechselungsgefahr mit der dualen Abbildung! Das Zusammenfallen der Notationen ist Ausdruck eines Zusammenhangs zwischen dualer und adjungierter Abbildung. Bevor wir Beispiele betrachten, wollen wir uns überzeugen, dass die adjungierte Abbildung eindeutig bestimmt ist (wie die Notation φ^* ja schon andeutet).

Proposition 24.2. *Sei $\varphi\colon V \to V$ linear.*

(a) Falls φ eine adjungierte Abbildung hat, so ist diese eindeutig bestimmt.
(b) Falls φ eine adjungierte Abbildung φ^* hat, so ist deren adjungierte Abbildung φ, d.h.

$$\varphi^{**} = \varphi.$$

Beweis. (a) Es seien $\psi, \psi'\colon V \to V$ zwei adjungierte Abbildungen von φ. Für $v, w \in V$ gilt dann

$$\langle \psi(v) - \psi'(v), w\rangle = \langle \psi(v), w\rangle - \langle \psi'(v), w\rangle = \langle v, \varphi(w)\rangle - \langle v, \varphi(w)\rangle = 0.$$

Setzt man speziell $w = \psi(v) - \psi'(v)$ ein, so ergibt sich $\psi(v) = \psi'(v)$, also $\psi = \psi'$.

(b) Für $v, w \in V$ gilt

$$\langle v, \varphi^*(w)\rangle = \overline{\langle \varphi^*(w), v\rangle} = \overline{\langle w, \varphi(v)\rangle} = \langle \varphi(v), w\rangle,$$

also ist φ zu φ^* adjungiert. □

Beispiel 24.3. (1) Es sei $V = C([a,b], \mathbb{C})$ wie in Beispiel 23.4. Für ein fest gewähltes $h \in V$ betrachten wir $\varphi_h \colon V \to V,\ f \mapsto h \cdot f$. Für $f, g \in V$ gilt

$$\langle f, \varphi_h(g)\rangle = \int_a^b \overline{f(x)} h(x) g(x) dx = \int_a^b \overline{f(x)\overline{h(x)}} g(x) dx = \langle \overline{h} f, g\rangle,$$

also $\varphi_h^* = \varphi_{\overline{h}}$.

(2) Es sei V wie oben und $x_0 \in [a,b]$ fest gewählt. Wir betrachten $\varphi \colon V \to V,\ f \mapsto f(x_o)$, wobei $f(x_0)$ als konstante Funktion angesehen wird. Für $f, g \in V$ gilt

$$\langle f, \varphi(g)\rangle = \int_a^b \overline{f(x)} g(x_0) dx = g(x_0) \int_a^b \overline{f(x)} dx$$

Falls φ eine adjungierte Abbildung hätte, so würde mit $h := \varphi^*(f)$ für alle $g \in V$ gelten:

$$g(x_0) \int_a^b \overline{f(x)} dx = \langle h, g\rangle = \int_a^b \overline{h(x)} g(x) dx.$$

Eine solche Funktion h gibt es aber nur, falls $\int_a^b \overline{f(x)} dx = 0$, was nicht für alle f der Fall ist. Es folgt, dass φ keine adjungierte Abbildung hat. ◁

Die folgende Proposition klärt die Situation bei den Standard-Räumen $\mathbb{R}^n$ und $\mathbb{C}^n$.

Proposition 24.4. *(a) Es seien $V = \mathbb{R}^n$ mit dem Standardskalarprodukt und $A \in \mathbb{R}^{n\times n}$. Dann gilt*

$$\varphi_A^* = \varphi_{A^T}.$$

(b) Es seien $V = \mathbb{C}^n$ mit dem Standardskalarprodukt und $A \in \mathbb{C}^{n\times n}$. Dann gilt

$$\varphi_A^* = \varphi_{\overline{A}^T}.$$

Beweis. Wir führen nur den (etwas schwereren) Nachweis von (b). Für $v, w \in \mathbb{C}^n$ gilt

$$\langle v, \varphi_A(w)\rangle = \overline{v}^T A w = (A^T \overline{v})^T w = \overline{(\overline{A}^T v)}^T w = \langle \varphi_{\overline{A}^T}(v), w\rangle.$$

Dies liefert die Behauptung. □

Entsprechend verhält es sich bei linearen Abbildungen $\varphi\colon V \to V$ von endlich-dimensionalen euklidischen oder unitären Räumen: Ist S eine Orthonormalbasis von V, so wird die adjungierte Abbildung φ^* gegeben durch die Darstellungsmatrix

$$D_S(\varphi^*) = \overline{D_S(\varphi)}^T.$$

Definition 24.5. *(a) Eine lineare Abbildung $\varphi\colon V \to V$ heißt* **normal**, *falls die adjungierte Abbildung φ^* existiert und*

$$\varphi \circ \varphi^* = \varphi^* \circ \varphi$$

gilt.

(b) Eine Matrix $A \in \mathbb{R}^{n\times n}$ bzw. $A \in \mathbb{C}^{n\times n}$ heißt **normal**, *falls*

$$A \cdot \overline{A}^T = \overline{A}^T \cdot A$$

gilt. Im Fall $A \in \mathbb{R}^{n\times n}$ liest sich das als $A^T \cdot A = A \cdot A^T$.

Wir haben bereits eine Reihe normaler Abbildungen und Matrizen kennengelernt.

Beispiel 24.6. (1) Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch oder $A \in \mathbb{C}^{n\times n}$ hermitesch. Dann ist A normal.

(2) Sei $A \in \mathbb{R}^{n\times n}$ mit $A^T = -A$. (Solche Matrizen heißen *antisymmetrisch.*) Dann ist A normal. Ebenso sind antihermitesche Matrizen (mit der offensichtlichen Begriffsbildung) normal.

(3) Jede othogonale oder unitäre Matrix ist normal.

(4) Für die Matrix $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$ gilt

$$A^T \cdot A = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix} \quad \text{aber} \quad A \cdot A^T = \begin{pmatrix} 5 & 11 \\ 11 & 25 \end{pmatrix},$$

also ist A nicht normal.

(5) Sei $\varphi\colon V \to V$ eine surjektive orthogonale bzw. unitäre Abbildung. Dann ist φ bijektiv, und es gilt für $v, w \in V$:

$$\langle v, \varphi(w)\rangle = \langle \varphi^{-1}(v), \varphi^{-1}(\varphi(w))\rangle = \langle \varphi^{-1}(v), w\rangle.$$

Es folgt $\varphi^* = \varphi^{-1}$, also ist φ normal.

(6) Für die Abbildung φ_h aus Beispiel 24.3(2) gilt $\varphi_h^* = \varphi_{\overline{h}}$, also ist φ_h normal. Falls h nur reelle Werte annimmt, so gilt $\varphi_h^* = \varphi_h$. Lineare Abbildungen mit dieser Eigenschaft nennt man *selbstadjungiert.* ◁

Unser nächstes Ziel ist es zu zeigen, dass jede normale Abbildung eines endlich-dimensionalen unitären Raums diagonalisierbar ist. Für eine lineare Abbildung $\varphi\colon V \to V$ und $\lambda \in \mathbb{R}$ bzw. $\lambda \in \mathbb{C}$ betrachten wir den Eigenraum

$$E_\lambda(\varphi) := \{v \in V \mid \varphi(v) = \lambda v\}.$$

Das folgende Lemma ist entscheidend.

Lemma 24.7. *Es sei $\varphi\colon V \to V$ normal.*

(a) Für $\lambda \in \mathbb{R}$ bzw. $\lambda \in \mathbb{C}$. Dann gilt

$$E_\lambda(\varphi) = E_{\overline{\lambda}}(\varphi^*).$$

(b) Sind $v \in E_\lambda(\varphi)$ und $w \in E_\mu(\varphi)$ mit $\lambda, \mu \in \mathbb{R}$ bzw. $\lambda, \mu \in \mathbb{C}$ verschieden, so folgt $\langle v, w\rangle = 0$.

(c) Sei $L \subseteq V$ das Erzeugnis aller Eigenvektoren (zu allen Eigenwerten) von φ. Dann gilt

$$\varphi\left(L^\perp\right) \subseteq L^\perp,$$

und $L^\perp$ enthält keine Eigenvektoren von φ.

Beweis. (a) Für $v \in E_\lambda(\varphi)$ gelten

$$||\varphi^*(v)||^2 = \langle v, \varphi\left(\varphi^*(v)\right)\rangle = \langle v, \varphi^*\left(\varphi(v)\right)\rangle = \langle v, \varphi^*(\lambda v)\rangle = \lambda\langle v, \varphi^*(v)\rangle$$

und

$$\langle \varphi^*(v), v\rangle = \langle v, \varphi(v)\rangle = \langle v, \lambda v\rangle = \lambda \cdot ||v||^2,$$

also

$$||\varphi^*(v) - \overline{\lambda}v||^2 = ||\varphi^*(v)||^2 - \overline{\lambda}\langle\varphi^*(v), v\rangle - \lambda\langle v, \varphi^*(v)\rangle + |\lambda|^2||v||^2 = 0.$$

Es folgt $v \in E_{\overline{\lambda}}(\varphi^*)$, also $E_\lambda(\varphi) \subseteq E_{\overline{\lambda}}(\varphi^*)$. Durch Anwenden auf φ^* und $\overline{\lambda}$ ergibt sich

$$E_{\overline{\lambda}}(\varphi^*) \subseteq E_\lambda(\varphi^{**}) = E_\lambda(\varphi),$$

also Gleichheit.

(b) Die Behauptung ergibt sich aus

$$(\lambda - \mu)\langle v, w\rangle = \langle\overline{\lambda}v, w\rangle - \langle v, \mu w\rangle = \underbrace{\langle\varphi^*(v), w\rangle}_{=\langle v, \varphi(w)\rangle} - \langle v, \varphi(w)\rangle = 0,$$

wobei die zweite Gleichheit aus (a) folgt.

(c) Ist v ein Eigenvektor, so gilt $v \in L \setminus \{0\}$ und $\langle v, v\rangle \neq 0$, also $v \notin L^\perp$. Nun sei $v \in L^\perp$. Für den Nachweis von $\varphi(v) \in L^\perp$ genügt es zu zeigen, dass $\varphi(v)$ zu allen Eigenvektoren $w \in V$ orthogonal ist. Es sei also $\varphi(w) = \lambda w$ mit $\lambda \in \mathbb{R}$ bzw. $\lambda \in \mathbb{C}$. Dann gilt

$$\langle w, \varphi(v)\rangle = \langle\varphi^*(w), v\rangle = \langle\overline{\lambda}w, v\rangle = \lambda\langle w, v\rangle = 0,$$

wobei die zweite Gleichheit aus (a) folgt. Dies schließt den Beweis ab. □

Satz 24.8 (Spektralsatz für unitäre Räume)**.** *Es seien V ein endlich-dimensionaler unitärer Raum und $\varphi\colon V \to V$ eine normale Abbildung. Dann besitzt V eine Orthonormalbasis B, die aus Eigenvektoren von φ besteht. Genauer:*

Jede Vereinigungsmenge von Orthonormalbasen der Eigenräume von φ bildet eine solche Basis B. Insbesondere ist φ diagonalisierbar.

Beweis. Es seien $\lambda_1, \dots, \lambda_r$ die Eigenwerte von φ. Wegen Korollar 23.17 gibt es für jeden Eigenraum E_{λ_i} eine Orthonormalbasis B_i. Wegen Lemma 24.7(b) ist $B := B_1 \cup \dots \cup B_r$ ein Orthonormalsystem. Wegen Satz 23.13 ist B also eine Orthonormalbasis des Unterraums $L \subseteq V$, der von allen Eigenvektoren von φ erzeugt wird. Es ist klar, dass B aus Eigenvektoren von φ besteht. Also ist nur noch $L = V$ zu zeigen.

Wir schreiben $B = \{v_1, \dots, v_n\}$. Dann ist $L^\perp$ der Kern der linearen Abbildung

$$\psi\colon V \to \mathbb{C}^n,\ v \mapsto (\langle v_1, v\rangle, \dots, \langle v_n, v\rangle)\,,$$

wegen Satz 11.9 also

$$\dim(V) = \dim(L^\perp) + \dim\left(\mathrm{Bild}(\psi)\right) \leq \dim(L^\perp) + \dim(L).$$

(In Wirklichkeit gilt Gleichheit, aber das wird hier nicht gebraucht.) Wäre $L^\perp \neq \{0\}$, so enthielte $L^\perp$ wegen der algebraischen Abgeschlossenheit von $\mathbb{C}$ und der ersten Aussage von Lemma 24.7(c) einen Eigenvektor von φ, was der zweiten Aussage von Lemma 24.7(c) widerspräche. Es folgt $L^\perp = \{0\}$, also liefert die obige Dimensionsungleichung $L = V$. □

Korollar 24.9 (Spektralsatz für komplexe normale Matrizen)**.** *Sei $A \in \mathbb{C}^{n\times n}$ normal. Dann gibt es eine unitäre Matrix $S \in \mathrm{U}_n$, so dass $S^{-1}AS$ eine Diagonalmatrix ist. Wegen $S \in \mathrm{U}_n$ gilt $S^{-1}AS = \overline{S}^T AS$.*

Anmerkung 24.10. Es gilt auch die Umkehrung von Korollar 24.9: Sei $A \in \mathbb{C}^{n\times n}$ eine Matrix, für die $S \in \mathrm{U}_n$ existiert, so dass $S^{-1}AS = D$ eine Diagonalmatrix ist. Dann folgen

$$A = SDS^{-1} = SD\overline{S}^T \quad \text{und} \quad \overline{A}^T = S\overline{D}\overline{S}^T,$$

also

$$A \cdot \overline{A}^T = SD\overline{S}^T S\overline{D}\overline{S}^T = SD\overline{D}\overline{S}^T = \overline{A}^T \cdot A.$$

Damit ist A normal. ◁

Nun wenden wir uns der Frage zu, was im reellen Fall passiert.

Lemma 24.11. *Es seien $A \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{C}$ und $v \in \mathbb{C}^n$ mit $A \cdot v = \lambda v$.*

(a) Für den Vektor $\overline{v} \in \mathbb{C}^n$, der aus v durch Konjugation aller Koordinaten entsteht, gilt

$$A \cdot \overline{v} = \overline{\lambda}\overline{v}.$$

(b) Für den Real- und Imaginärteil von v gelten

$$A \cdot \mathrm{Re}(v) = \mathrm{Re}(\lambda)\,\mathrm{Re}(v) - \mathrm{Im}(\lambda)\,\mathrm{Im}(v)$$

und

$$A \cdot \operatorname{Im}(v) = \operatorname{Im}(\lambda)\operatorname{Re}(v) + \operatorname{Re}(\lambda)\operatorname{Im}(v).$$

Beweis. (a) Dies ergibt sich aus

$$A \cdot \overline{v} = \overline{A} \cdot \overline{v} = \overline{A \cdot v} = \overline{\lambda v} = \overline{\lambda}\overline{v}.$$

(b) Es gilt

$$\begin{aligned} A \cdot \operatorname{Re}(v) + iA \cdot \operatorname{Im}(v) &= A \cdot v = \lambda v \\ &= \operatorname{Re}(\lambda)\operatorname{Re}(v) - \operatorname{Im}(\lambda)\operatorname{Im}(v) + i\left(\operatorname{Im}(\lambda)\operatorname{Re}(v) + \operatorname{Re}(\lambda)\operatorname{Im}(v)\right). \end{aligned}$$

Die Behauptung ergibt sich durch Vergleich von Real- und Imaginärteil. □

Korollar 24.12 (Spektralsatz für reelle normale Matrizen)**.** *Sei $A \in \mathbb{R}^{n\times n}$ normal. Dann gibt es eine orthogonale Matrix $S \in \mathrm{O}_n$, so dass*

$$S^{-1}AS = \begin{pmatrix} \lambda_1 & & & & & 0 \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & \boxed{\begin{smallmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{smallmatrix}} & & \\ & & & & \ddots & \\ 0 & & & & & \boxed{\begin{smallmatrix} a_s & -b_s \\ b_s & a_s \end{smallmatrix}} \end{pmatrix}$$

mit $\lambda_1, \ldots, \lambda_r, a_1, \ldots, a_s, b_1, \ldots, b_s \in \mathbb{R}$ und $b_i > 0$ für alle i.
Wegen $S \in \mathrm{O}_n$ gilt $S^{-1}AS = S^TAS$.

Beweis. Das charakteristische Polynom χ_A zerfällt über $\mathbb{C}$ in Linearfaktoren, wir können also schreiben

$$\chi_A = \prod_{i=1}^{r}(x - \lambda_i)\prod_{i=1}^{s}(x - \mu_i)\prod_{i=1}^{t}(x - \nu_i)$$

mit $\lambda_i \in \mathbb{R}$, $\mu_i.\nu_i \in \mathbb{C}$, so dass $\operatorname{Im}(\mu_i) < 0$ und $\operatorname{Im}(\nu_i) > 0$. Aus der eindeutigen Primzerlegung folgt durch komplexe Konjugation wegen $\overline{\chi_A} = \chi_A$

$$\prod_{i=1}^{t}(x - \nu_i) = \prod_{i=1}^{s}(x - \overline{\mu_i}),$$

also $s = t$ und

$$n = \deg(\chi_a) = r + s + t = r + 2s.$$

Wir wenden Satz 24.8 auf $\varphi_A\colon \mathbb{C}^n \to \mathbb{C}^n$ an und erhalten Vektoren $u_1, \ldots, u_r, v_1, \ldots, v_s \in \mathbb{C}^n$ mit

$$A \cdot u_i = \lambda_i u_i, \quad A \cdot v_i = \mu_i v_i, \quad \langle u_i, u_j \rangle = \delta_{i,j}, \quad \text{und} \quad \langle v_i, v_j \rangle = \delta_{i,j}$$

für alle i, j. (Satz 24.8 liefert auch Eigenvektoren für die Eigenwerte ν_i, aber die brauchen wir hier nicht.) Die u_i können aus beliebigen Orthonormalbasen der Eigenräume E_{λ_i} gewählt werden, also können wir $u_i \in \mathbb{R}^n$ annehmen. Für $i = 1, \ldots, s$ setzen wir

$$w_i := \sqrt{2}\,\mathrm{Re}(v_i), \quad w_i' := \sqrt{2}\,\mathrm{Im}(v_i), \quad a_i := \mathrm{Re}(\mu_i) \quad \text{und} \quad b_i := -\,\mathrm{Im}(\mu_i).$$

Falls

$$B := \{u_1, \ldots, u_r, w_1, w_1', \ldots, w_s, w_s'\}$$

eine Basis von $\mathbb{C}^n$ bildet, so folgt aus Lemma 24.11(b), dass $D_B(\varphi_A)$ genau die im Korollar angegebene Block-Diagonalmatrix ist, also folgt die Behauptung mit $S := (v_1, \ldots, v_r, w_1, w_1', \ldots, w_s, w_s') \in \mathrm{GL}_n(\mathbb{R})$. Wegen $n = |B|$ genügt es nach Satz 23.13 zu zeigen, dass B ein Orthonormalsystem ist, und dann folgt auch $S \in \mathrm{O}_n$. Für $j \in \{1, \ldots, r\}$ und $k \in \{1, \ldots, s\}$ gilt

$$\langle u_j, w_k \rangle + i\langle u_j, w_k' \rangle = \sqrt{2}\langle u_j, v_k \rangle = 0,$$

also $\langle u_j, w_k \rangle = \langle u_j, w_k' \rangle = 0$. Weiter gilt für $j, k \in \{1, \ldots, s\}$:

$$\begin{aligned} \langle w_j, w_k \rangle &= \left\langle \frac{1}{\sqrt{2}}(v_j + \overline{v_j}), \frac{1}{\sqrt{2}}(v_k + \overline{v_k}) \right\rangle \\ &= \frac{1}{2}\left(\langle v_j, v_k \rangle + \langle v_j, \overline{v_k} \rangle + \langle \overline{v_j}, v_k \rangle + \langle \overline{v_j}, \overline{v_k} \rangle \right) = \delta_{j,k}, \end{aligned}$$

wobei $\langle v_j, \overline{v_k} \rangle = \langle \overline{v_j}, v_k \rangle = 0$ aus Lemma 24.11(a) und Lemma 24.7(b) folgen. Entsprechende Rechnungen liefern

$$\langle w_j', w_k' \rangle = \delta_{j,k} \quad \text{und} \quad \langle w_j, w_k' \rangle = 0.$$

Dies schließt den Beweis ab. □

Wir spezialisieren dies Resultat nun für die beiden wichtigsten Klassen von normalen reellen Matrizen, die orthogonalen und die symmetrischen Matrizen. Wir beginnen mit dem orthogonalen Fall.

Sei also $A \in \mathrm{O}_n$. Wegen Korollar 24.12 gibt es $S \in \mathrm{O}_n$, so dass $B := S^{-1}AS$ die im Korollar angegebene Form hat. Dann muss B selbst orthogonal sein, also gilt für die λ_i, a_i und b_i:

$$\lambda_i = \pm 1 \quad \text{und} \quad a_i^2 + b_i^2 = 1.$$

Wegen $b_i > 0$ folgt insbesondere $|a_i| < 1$, also $a_i = \cos(\alpha_i)$ mit $0 < \alpha_i < \pi$ und $b_i = \sin(\alpha_i)$. Für $\alpha \in \mathbb{R}$ schreiben wie

$$D(\alpha) := \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$

und nennen dies ein **Drehkästchen**. Es beschreibt eine Drehung der Ebene $\mathbb{R}^2$ um den Winkel α mit festgehaltenem Nullvektor. Wir formulieren unser Resultat geometrisch.

Korollar 24.13. *Es sei $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ eine orthogonale Abbildung. Dann gibt es eine Orthonormalbasis B, bezüglich der die Darstellungmatrix von φ die Block-Diagonalgestalt*

$$D_B(\varphi) = \begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & \boxed{D(\alpha_1)} & & \\ & & & & & & & \ddots & \\ 0 & & & & & & & & \boxed{D(\alpha_s)} \end{pmatrix}$$

mit $\alpha_i \in \mathbb{R}$, $0 < \alpha_i < \pi$ annimmt.

Beispiel 24.14. Jede orthogonale Abbildung $\varphi\colon \mathbb{R}^3 \to \mathbb{R}^3$ hat bezüglich einer geeigneten Orthonormalbasis B die Darstellungsmatrix

$$D_B(\varphi) = \begin{pmatrix} \pm 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

mit $0 \leq \alpha \leq \pi$. Genau dann liegt φ in der speziellen orthogonalen Gruppe, wenn der erste Eintrag der Matrix 1 ist. Die Elemente der SO_3 beschreiben also Drehungen um eine gewisse Achse. ◁

Wir behandeln nun die symmetrischen Matrizen und beweisen das wichtige Resultat, dass sie diagonalisierbar sind. Dies Ergebnis läuft manchmal unter der Bezeichnung *Hauptachsentransformation*. Außerdem beweisen wir, dass auch hermitesche Matrizen reelle Eigenwerte haben.

Korollar 24.15. *(a) Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Dann gibt es eine orthogonale Matrix $S \in \mathrm{O}_n$, so dass $S^{-1}AS$ eine Diagonalmatrix ist. Insbesondere sind alle Eigenwerte von A reell, und A ist diagonalisierbar.*
(b) Sei $A \in \mathbb{C}^{n\times n}$ hermitesch (so dass A nach Korollar 24.9 mit einer unitären Matrix diagonalisierbar ist). Dann sind alle Eigenwerte von A reell.

Beweis. (a) Nach Korollar 24.12 gibt es $S \in \mathrm{O}_n$, so dass $S^T AS =: D$ die im Korollar angegebene Gestalt hat. Es folgt

$$D^T = S^T A^T S = S^T AS = D,$$

d.h. D ist symmetrisch. Hieraus folgt, dass in D kein Block der Form $\left(\begin{smallmatrix} a_i & -b_i \\ b_i & a_i \end{smallmatrix}\right)$ auftritt, da ein solcher wegen $b_i > 0$ der Symmetrie widersprechen würde.

(b) Wegen Korollar 24.9 gibt es $S \in \mathrm{U}_n$ mit $\overline{S}^T AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) =: D$. Es folgt

$$\overline{D} = \overline{D}^T = \overline{S}^T \overline{A}^T S = \overline{S}^T AS = D$$

also $\lambda_i \in \mathbb{R}$ für alle i. □

Beispiel 24.16. (1) Wir betrachten die symmetrische Matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Um A zu diagonalisieren, berechnen wir das charakteristische Polynom und erhalten

$$\chi_A = \det \begin{pmatrix} x-2 & -1 & -1 \\ -1 & x-2 & -1 \\ -1 & -1 & x-2 \end{pmatrix} = (x-2)^3 - 2 - 3(x-2) =$$
$$x^3 - 6x^2 + 9x - 4 = (x-1)(x^2 - 5x + 4) = (x-1)^2(x-4).$$

Damit wissen wir schon, dass A zu $\mathrm{diag}(1, 1, 4)$ ähnlich ist. Wir wollen eine orthogonale Transformationsmatrix ausrechnen. Hierfür müssen wir die Eigenräume bestimmen. Der Eigenraum E_1 zum Eigenwert 1 ergibt sich als Lösungsraums des homogenen LGS mit Matrix $A - I_3$. Wir erhalten

$$E_1 = \left\langle \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\rangle.$$

Auf die Basis von E_1 wenden wir das Schmidtsche Orthogonalisierungsverfahren an. Der erste Schritt liefert

$$u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Weiter erhalten wir

$$w_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{2}} u_1 = \frac{1}{2} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix},$$

also

$$u_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Nun berechnen wir E_4 und erhalten durch Lösen des entsprechenden LGS (oder durch die Beobachtung, dass alle Zeilensummen von A gleich 4 sind)

$$E_4 = \left\langle \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\rangle.$$

Normieren liefert als letzten Vektor der Orthonormalbasis

$$u_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Damit gilt

$$S = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix} \in O_3(\mathbb{R})$$

und

$$S^{-1}AS = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

(2) Es stellt sich die Frage, ob Korollar 24.15(a) auch über anderen Körpern außer $\mathbb{R}$ gilt, z.B. über $\mathbb{C}$. Um diese zu beantworten, betrachten wir die symmetrische Matrix

$$A = \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix} \in \mathbb{C}^{2\times 2}.$$

Das charakteristische Polynom ist

$$\chi_A = \det \begin{pmatrix} x-1 & -i \\ -i & x+1 \end{pmatrix} = (x-1)(x+1) + 1 = x^2,$$

also haben wir 0 als einzigen Eigenwert. Die algebraische Vielfachheit ist 2, die geometrische aber 1, also ist A nicht diagonalisierbar. Mit $\mathbb{C}$ statt $\mathbb{R}$ wäre Korollar 24.15(a) also nicht korrekt. Ebenso verhält es sich mit $\mathbb{Q}$ statt $\mathbb{R}$. ◁

Anmerkung 24.17. Die Aussagen über reelle Eigenwerte in Korollar 24.15 stehen in einem breiteren Kontext. In der Tat sind die Eigenwerte einer selbstadjungierten Abbildung $\varphi\colon V \to V$ eines unitären Raums immer reell. Es seien nämlich $\lambda \in \mathbb{C}$ ein Eigenwert und $v \in V \setminus \{0\}$ ein zugehöriger Eigenvektor. Dann gilt

$$\lambda \cdot ||v||^2 = \langle v, \lambda v\rangle = \langle v, \varphi(v)\rangle = \langle \varphi(v), v\rangle = \langle \lambda v, v\rangle = \overline{\lambda} \cdot ||v||^2.$$

Hieraus folgt $\lambda \in \mathbb{R}$. ◁

Korollar 24.15(a) hat beispielsweise physikalische Anwendungen. Zu einem starren Körper betrachtet man den sogenannten *Trägheitstensor*. Dieser ist eine Matrix in $I \in \mathbb{R}^{3\times 3}$, die die Winkelgeschwindigkeit (als Vektor) mit dem Drehimpuls verbindet, ähnlich wie die Masse die Geschwindigkeit mit dem Impuls verbindet. Es stellt sich heraus, dass I symmetrisch ist. Also liefert Korollar 24.15, dass es für jeden starren Körper drei senkrecht zueinander stehende Achsen gibt, so dass bei einer Drehung um diese Achsen die Drehgeschwindigkeit und der Drehimpuls in dieselbe Richtung zeigen. Diese Achsen heißen *Hauptträgheitsachsen*. Wegen des Drehimpulserhaltungssatzes bedeutet dies, dass Drehungen um die Hauptträgheitsachsen „schlingerfrei" möglich sind. Bei konstantem Drehimpuls ist eine Drehung um die Achse mit dem größten Eigenwert (= *Hauptträgheitsmoment*) die energetisch günstigste und daher stabilste.

Wir haben bereits im Zusammenhang mit symmetrischen Bilinearformen und hermiteschen Sesquilinearformen von positiver Definitheit gesprochen. Nun übertragen wir dies auf Matrizen. Da alle Eigenwerte einer symmetrischen (reellen) oder hermiteschen Matrix oder reell sind, können wir fragen, ob sie positiv sind.

Definition 24.18. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch bzw. $A \in \mathbb{C}^{n\times n}$ hermitesch. A heißt*

- **positiv definit**, *falls alle Eigenwerte von A positiv sind;*
- **positiv semidefinit**, *falls alle Eigenwerte von A positiv oder Null sind;*
- **negativ definit**, *falls alle Eigenwerte von A negativ sind;*
- **negativ semidefinit**, *falls alle Eigenwerte von A negativ oder Null sind;*
- **indefinit**, *falls es sowohl positive als auch negative Eigenwerte gibt.*

Satz 24.19. *Eine symmetrische bzw. hermitesche Matrix $A \in \mathbb{R}^{n\times n}$ bzw. $A \in \mathbb{C}^{n\times n}$ ist genau dann positiv definit, wenn für alle $v \in \mathbb{R}^n \setminus \{0\}$ bzw. $v \in \mathbb{C}^n \setminus \{0\}$ gilt:*

$$\langle v, A \cdot v\rangle > 0.$$

Die Bedingung bedeutet, dass die durch A definierte Bilinearform bzw. Sesquilinearform positiv definit ist. A ist positiv semidefinit, wenn $\langle v, A \cdot v\rangle \geq 0$ gilt. Entsprechendes gilt für negativ (semi-)definit.

Beweis. Wegen Korollar 24.15 gibt es $S \in \mathrm{O}_n$ bzw. $S \in \mathrm{U}_n$ mit

$$\overline{S}^T A S = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} =: D,$$

wobei die $\lambda_i \in \mathbb{R}$ die Eigenwerte von A sind. Wegen der Invertierbarkeit von $\overline{S}^T$ ist für jeden Vektor $v \in \mathbb{R}^n \setminus \{0\}$ bzw. $v \in \mathbb{C}^n \setminus \{0\}$ auch $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} := \overline{S}^T \cdot v$ ungleich 0, und jeder Vektor aus $\mathbb{R}^n \setminus \{0\}$ bzw. $\mathbb{C}^n \setminus \{0\}$ tritt als ein solches $\overline{S}^T \cdot v$ auf. Es gilt

$$\langle v, A \cdot v \rangle = \overline{v}^T S D \overline{S}^T v = (\overline{x}_1, \ldots, \overline{x}_n) D \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} \lambda_i |x_i|^2.$$

Hieraus folgen alle Behauptungen. □

Beispiel 24.20. Wir betrachten

$$A = \begin{pmatrix} a & 0 & -a & 0 \\ 0 & b & 0 & -b \\ -a & 0 & a & 0 \\ 0 & -b & 0 & b \end{pmatrix} \quad \text{mit} \quad a, b \in \mathbb{R}.$$

Wir wenden Satz 24.19 zur Feststellung der Definitheitseigenschaften von A an. Für $v = \begin{pmatrix} x_1 \\ \vdots \\ x_4 \end{pmatrix} \in \mathbb{R}^4$ gilt

$$\langle v, A \cdot v \rangle = (x_1, x_2, x_3, x_4) \cdot \begin{pmatrix} a(x_1 - x_3) \\ b(x_2 - x_4) \\ -a(x_1 - x_3) \\ -b(x_2 - x_4) \end{pmatrix} = a(x_1 - x_3)^2 + b(x_2 - x_4)^2.$$

Damit ist A positiv semidefinit, falls $a, b \geq 0$, negativ semidefinit, falls $a, b \leq 0$, und sonst indefinit. ◁

25 Singulärwertzerlegung und Moore-Penrose-Inverse

Eine in der numerischen Mathematik wichtige Technik ist die sogenannte *Singulärwertzerlegung*, die durch den folgenden Satz gegeben wird. Wie wir im Beweis sehen werden, verdankt die Singulärwertzerlegung ihre Existenz dem Korollar 24.15.

Satz 25.1 (Singulärwertzerlegung)**.** *Sei $A \in \mathbb{C}^{m\times n}$ eine (nicht notwendig quadratische) Matrix. Dann gibt es unitäre Matrizen $U \in \mathrm{U}_m$ und $V \in \mathrm{U}_n$, so dass*

$$\overline{U}^T A V = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) =: \Sigma \in \mathbb{R}^{m\times n} \tag{25.1}$$

mit $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, wobei $r = \mathrm{rg}(A)$. Im Fall $A \in \mathbb{R}^{m\times n}$ können $U \in \mathrm{O}_m$ und $V \in \mathrm{O}_n$ gewählt werden. Die zur obigen Gleichung äquivalente Gleichung

$$A = U\Sigma\overline{V}^T$$

bezeichnet man als **Singulärwertzerlegung** *von A.*

Beweis. Die Matrix $\overline{A}^T A \in \mathbb{C}^{n\times n}$ ist wegen

$$\left(\overline{A}^T A\right)^T = A^T\overline{A} = \overline{\overline{A}^T A}$$

hermitesch. Außerdem ist sie gemäß Satz 24.19 positiv semidefinit, denn für $v \in \mathbb{C}^n$ gilt

$$\langle v, \overline{A}^T A v\rangle = \overline{v}^T\overline{A}^T A v = \overline{Av}^T Av = \langle Av, Av\rangle \geq 0.$$

Wegen Korollar 24.15 gibt es $V \in \mathrm{U}_n$ (wobei $V \in \mathrm{O}_n$ im reellen Fall), so dass

$$\overline{V}^T\overline{A}^T A V = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{25.2}$$

mit $\lambda_i \in \mathbb{R}_{\geq 0}$, die wir so anordnen können, dass $\lambda_1 \geq \cdots \geq \lambda_n$. Es sei r maximal mit $\lambda_r > 0$. (Später werden wir $r = \mathrm{rg}(A)$ sehen.) Für $i \in \{1, \ldots, r\}$ setzen wir

$$\sigma_i := \sqrt{\lambda_i}.$$

Wir schreiben $v_1, \ldots, v_n$ für die Spalten von V, und für $i \in \{1, \ldots, r\}$ setzen wir

$$u_i := \sigma_i^{-1} A v_i \in \mathbb{C}^m \tag{25.3}$$

Sind $i, j \in \{1, \ldots, r\}$, so folgt

$$\langle u_i, u_j\rangle = (\sigma_i\sigma_j)^{-1}\overline{Av_i}^T Av_j = (\sigma_i\sigma_j)^{-1}\overline{v_i}^T\overline{A}^T Av_j$$
$$\underset{(25.2)}{=} (\sigma_i\sigma_j)^{-1}\lambda_i\delta_{i,j} = \delta_{i,j},$$

also bilden $u_1, \dots, u_r$ ein Orthonormalsystem. Dies lässt sich, etwa mit dem Schmidtschen Orthogonalisierungsverfahren, zu einer Orthonormalbasis $u_1, \dots, u_m$ von $\mathbb{C}^m$ ergänzen. Wir setzen

$$U := (u_1, \dots, u_m) \in \mathrm{U}_m\,.$$

Sei $i \in \{1, \dots, m\}$ und $j \in \{1, \dots, n\}$. Falls $j \le r$, so gilt

$$\overline{u_i}^T A v_j \underset{(25.3)}{=} \overline{u_i}^T \sigma_j u_j = \delta_{i,j}\sigma_j.$$

Falls $j > r$, so folgt

$$||Av_j||^2 = \overline{v_j}^T \overline{A}^T A v_j \underset{(25.2)}{=} \lambda_j = 0,$$

also $Av_j = 0$ und daher auch

$$\overline{u_i}^T A v_j = 0.$$

damit ist (25.1) gezeigt. Es folgt nun auch $A = U\Sigma\overline{V}^T$. Da U und $\overline{V}^T$ reguläre Matrizen sind, folgt hieraus

$$\mathrm{rg}(A) = \mathrm{rg}(\Sigma) = r.$$

Schließlich bemerken wir, dass im Fall $A \in \mathbb{R}^{m\times n}$ alle vorkommenden Matrizen reell sind und insbesondere U und V orthogonal. □

Anmerkung 25.2. (a) Ist $A \in \mathbb{C}^{m\times n}$ mit Singulärwertzerlegung $A = U\Sigma\overline{V}^T$, so folgt

$$\overline{A}^T A = V\overline{\Sigma}^T\overline{U}^T U\Sigma\overline{V}^T = V\Sigma^T\Sigma\overline{V}^T,$$

also ist $\Sigma^T\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0) \in \mathbb{R}^{n\times n}$ ähnlich zu $\overline{A}^T A$. Die σ_i^2 sind also genau die Eigenwerte von $\overline{A}^T A$, die nicht Null sind. Damit sind die σ_i (wegen $\sigma_1 \ge \cdots \ge \sigma_r$) eindeutig bestimmt. Man nennt sie die **Singulärwerte** von A.
Die Matrizen U, V aus der Singulärwertzerlegung sind im Allgemeinen nicht eindeutig bestimmt.

(b) Die folgende Rechnung liefert eine Interpretation des größten Singlärwerts σ_1. Für $v \in \mathbb{C}^n \setminus \{0\}$ setzen wir $w := \overline{V}^T v$ und schreiben w_i für die Koordinaten von w. Dann gilt

$$||Av|| = ||U\Sigma w|| = ||\Sigma w|| = \sqrt{\sum_{i=1}^{r} \sigma_i^2 |w_i|^2} \le \sigma_1 \cdot ||w|| = \sigma_1 \cdot ||v||,$$

wobei Gleichheit gilt, wenn v die erste Spalte von V ist. Es folgt

$$\sigma_1 = \max\left\{ \frac{||Av||}{||v||} \,\middle|\, v \in \mathbb{C}^n \setminus \{0\} \right\} =:|| A ||_s .$$

Die mit $||A||_s$ bezeichnete Zahl nennt man die *Spektralnorm* von A. Wir haben also die Gleichheit von Spektralnorm und dem ersten Singulärwert gezeigt. Die Spektralnorm ist eine Norm auf $\mathbb{C}^{m\times n}$ im Sinne von Anmerkung 23.8(a), die zusätzlich *submultiplikativ* ist, d.h. es gilt die Regel $||AB||_s \leq ||A||_s \cdot ||B||_s$ für $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times l}$.

(c) Ist $A \in \mathbb{C}^{n\times n}$ quadratisch und $A = U\Sigma\overline{V}^T$ eine Singulärwertzerlegung, so folgt

$$A = U\overline{V}^T V\Sigma\overline{V}^T = B \cdot C$$

mit $B = U\overline{V}^T \in \mathrm{U}_n$ unitär und $C = V\Sigma\overline{V}^T$ hermitesch und positiv semidefinit (definit genau dann, wenn $A \in \mathrm{GL}_n(\mathbb{C})$). Man nennt eine Zerlegung $A = BC$ mit B unitär und C hermitesch und positiv semidefinit eine *Polarzerlegung* von A. ◁

Beispiel 25.3. Die Matrix

$$A = \begin{pmatrix} 1 & 2 \\ -2 & -3.99 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

hat den Rang 2, ist aber nahe an einer Matrix vom Rang 1. Dies wird widergespiegelt durch die Singulärwerte, die sich näherungsweise zu

$$\sigma_1 \approx 4.992 \quad \text{und} \quad \sigma_2 \approx 0.002$$

ergeben. Ersetzt man -3.99 in A durch -4, so sieht man, dass für $v = \left(\frac{1}{2}\right)$ (der kein Eigenvektor ist) die Spektralnorm 5 „erreicht“ wird. ◁

Die Singulärwertzerlegung spielt in der numerischen Mathematik eine große Rolle. Weitere Anwendungen gibt es beispielsweise in der Bildkompression. Ein (digitales) Bild mit $m \times n$ Pixeln lässt sich durch eine $m \times n$-Matrix A darstellen. Bei vielen Bildern weist die Folge der Singulärwerte (σ_i) einen dramatischen Abbruch auf, d.h. ab einem gewissen (kleinen) s sind die Werte der σ_i für $i > s$ extrem klein im Verhältnis zu den σ_i mit $i \leq s$. Setzt man in der Singulärwertzerlegung

$$A = U\Sigma\overline{V}^T$$

alle σ_i mit $i > s$ gleich Null, so erhält man eine neue Matrix Σ', so dass der Übergang von A zu $A' := U\Sigma'\overline{V}^T$ zwar einen Datenverlust darstellt, der aber im Bild nicht sichtbar ist. Der Gewinn ist, dass man für das Auswerten von $A' = U\Sigma'\overline{V}^T$ nur die ersten s Spalten von $U\Sigma'$ und von V speichern muss, insgesamt also

$$s \cdot (n + m) \quad \text{statt} \quad m \cdot n$$

Einträge. Dies kann zu einer erheblichen Datenkompression führen.

Eine weitere wichtige Anwendung der Singulärwertzerlegung ist die Berechnung (und der Existenznachweis) der Moore-Penrose-Inversen, die wir nun definieren. Die Moore-Penrose-Inverse ist wohl die wichtigste Vertreterin der *Pseudo-Inversen*, die das Ziel haben, für nicht invertierbare Matrizen einen für gewisse Zwecke tauglichen Ersatz für eine Inverse zur Verfügung zu stellen.

Definition 25.4. *Es sei $A \in \mathbb{C}^{m\times n}$ eine (nicht notwendig quadratische) komplexe Matrix. Eine Matrix $A^+ \in \mathbb{C}^{n\times m}$ heißt* **Moore-Penrose-Inverse** *von A, falls gelten:*

(1) $AA^+A = A$,
(2) $A^+AA^+ = A^+$ *und*
(3) AA^+ *und* A^+A *sind hermitesch.*

Wir werden nun die Existenz und Eindeutigkeit der Moore-Penrose-Inversen beweisen. Falls A invertierbar ist, erfüllt A^{-1} alle Eigenschaften (1)–(3), also liefert die Eindeutigkeit in diesem Fall $A^+ = A^{-1}$. Die Moore-Penrose-Inverse verallgemeinert also die Inverse.

Satz 25.5. *Es sei $A \in \mathbb{C}^{m\times n}$.*

(a) Ist

$$A = \left(\begin{array}{c|c} \sigma_1 & \\ \ddots & 0 \\ \hline & \sigma_r \\ \hline 0 & 0 \end{array}\right) \in \mathbb{R}^{m\times n}$$

eine Diagonalmatrix mit $r \leq \min\{m,n\}$ und $\sigma_i \neq 0$ für alle i, so ist

$$A^+ = \left(\begin{array}{c|c} \sigma_1^{-1} & \\ \ddots & 0 \\ \hline & \sigma_r^{-1} \\ \hline 0 & 0 \end{array}\right) \in \mathbb{R}^{n\times m}$$

eine Moore-Penrose-Inverse von A.

(b) Ist $A = U\Sigma\overline{V}^T$ eine Singulärwertzerlegung von A, so ist

$$A^+ = V\Sigma^+\overline{U}^T$$

eine Moore-Penrose-Inverse von A. Dabei kann Σ^+ aus (a) verwendet werden.

(c) Die Moore-Penrose-Inverse von A ist eindeutig bestimmt.

Beweis. Der Nachweis von (a) und (b) geschieht durch direktes Nachprüfen der Eigenschaften (1)–(3) in Definition 25.4. Für den Nachweis von (c) machen wir folgende Vorbemerkung. Für eine Matrix $B \in \mathbb{C}^{m\times n}$ mit $\overline{B}^T \cdot B = 0$ folgt

$$||Bv||^2 = \overline{v}^T \overline{B}^T Bv = 0 \quad \text{für alle} \quad v \in \mathbb{C}^n,$$

also $B = 0$. Es seien nun $A^+, \widetilde{A} \in \mathbb{C}^{n\times m}$ zwei Moore-Penrose-Inverse von A. Dann gelten

$$\overline{(A^+A - \widetilde{A}A)}^T (A^+A - \widetilde{A}A) \underset{(3)}{=} (A^+A - \widetilde{A}A)^2 \underset{(1)}{=} A^+A - A^+A - \widetilde{A}A + \widetilde{A}A = 0$$

und

$$\overline{(AA^+ - A\widetilde{A})}^T (AA^+ - A\widetilde{A}) \underset{(3)}{=} (AA^+ - A\widetilde{A})^2 \underset{(1)}{=} AA^+ - A\widetilde{A} - AA^+ + A\widetilde{A} = 0,$$

also gemäß unserer Vorbemerkung $A^+A = \widetilde{A}A$ und $AA^+ = A\widetilde{A}$. Hieraus folgt

$$A^+ \underset{(2)}{=} A^+AA^+ = \widetilde{A}AA^+ = \widetilde{A}A\widetilde{A} \underset{(2)}{=} \widetilde{A},$$

die Eindeutigkeit ist also bewiesen. □

Die Moore-Penrose-Inverse hat viele interessante Eigenschaften. Um die wichtigsten zu beweisen, werden wir uns mit dem Begriff einer orthogonalen Projektion beschäftigen, der von unabhängigen Interesse ist.

Satz 25.6. *Sei $\varphi\colon V \to V$ eine lineare Abbildung eines euklidischen oder unitären Raums V, für die $\varphi^2 = \varphi$ (mit $\varphi^2 := \varphi \circ \varphi$) gilt. Wir schreiben $U := \mathrm{Bild}(\varphi)$*

(a) Genau dann ist φ selbstadjungiert, wenn für alle $u \in U$ und $w \in \mathrm{Kern}(\varphi)$ gilt: $\langle u, w\rangle = 0$ (d.h. Bild und Kern von φ stehen senkrecht aufeinander). In diesem Fall heißt φ eine **orthogonale Projektion** *(auf U).*

(b) Falls φ eine orthogonale Projektion ist, so gilt für alle $v \in V$: $\varphi(v)$ ist der eindeutig bestimmte Vektor aus U, der zu v minimalen Abstand hat.

(c) Falls φ eine orthogonale Projektion ist, so gilt dies auch für $\psi := \mathrm{id}_V - \varphi$.

Beweis. (a) Zunächst sei φ selbstadjungiert und $u \in U$ und $w \in \mathrm{Kern}(\varphi)$, also $u = \varphi(v)$ mit $v \in V$. Es folgt

$$\langle u, w\rangle = \langle \varphi(v), w\rangle = \langle v, \varphi(w)\rangle = \langle v, 0\rangle = 0.$$

Umgekehrt nehmen wir an, dass Bild und Kern von φ senkrecht aufeinander stehen. Für $v, w \in V$ folgt

$$\langle v, \varphi(w)\rangle = \langle \underbrace{v - \varphi(v)}_{\in \mathrm{Kern}(\varphi)} + \varphi(v), \varphi(w)\rangle = \langle \varphi(v), \varphi(w)\rangle,$$

und ebenso $\langle \varphi(v), w\rangle = \langle \varphi(v), \varphi(w)\rangle$. Also ist φ selbstadjungiert.

(b) Es sei $u \in U$, also auch $u - \varphi(v) \in U$. Wegen $\varphi^2 = \varphi$ gilt $\varphi(v) - v \in \mathrm{Kern}(\varphi)$, also $\langle u - \varphi(v), \varphi(v) - v\rangle = 0$. Es folgt

$$\begin{aligned} ||u - v||^2 &= \langle u - \varphi(v) + \varphi(v) - v, u - \varphi(v) + \varphi(v) - v\rangle \\ &= ||u - \varphi(v)||^2 + ||\varphi(v) - v||^2. \end{aligned}$$

Also wird $||u - v||$ genau für $u = \varphi(v)$ minimal.

(c) Dies folgt aus

$$\psi^2 = \mathrm{id}_V^2 - 2\varphi + \varphi^2 = \mathrm{id}_V - \varphi = \psi$$

und $\psi^* = \mathrm{id}_V^* - \varphi^* = \mathrm{id}_V - \varphi = \psi$. □

Aus dem nächsten Satz geht hervor, dass die Moore-Penrose-Inverse sich in Bezug auf das Lösen von linearen Gleichungssystemen so verhält, wie man dies von einer Pseudo-Inversen erwarten würde. Interessant ist, dass hierbei Aussagen über nicht lösbare sowie über nicht eindeutig lösbare lineare Gleichungssysteme gemacht werden können.

Satz 25.7. *Zu $A \in \mathbb{C}^{m\times n}$ und $b \in \mathbb{C}^m$ betrachten wir das lineare Gleichungssystem $Ax = b$.*

(a) Ist das lineare Gleichungssystem lösbar, so ist $x = A^+b \in \mathbb{C}^n$ eine Lösung, und A^+b hat unter allen Lösungen die minimale Länge.

(b) Für alle $x \in \mathbb{C}^n$ gilt:

$$||Ax - b|| \geq ||AA^+b - b||.$$

A^+b liefert also eine bestmögliche näherungsweise Lösung. Unter allen Vektoren, die eine bestmögliche näherungsweise Lösung liefern, ist A^+b der kürzeste.

(c) Im Falle $b = 0$ (homogenes lineares Gleichungssystem) wird der Lösungsraum L durch die Spalten von $I_n - A^+A$ erzeugt. Genauer: $I_n - A^+A$ definiert eine orthogonale Projektion auf L.

Beweis. (c) Wegen $A^+AA^+A = A^+A$ und weil A^+A hermitesch ist, wird durch A^+A gemäß Satz 25.6(a) eine orthogonale Projektion gegeben, also nach Satz 25.6(c) auch durch $I_n - A^+A$. Wegen

$$A \cdot (I_n - A^+A) = A - AA^+A = A - A = 0$$

liegt deren Bild im Lösungsraum L, und umgekehrt gilt für $x \in L$:

$$(I_n - A^+A)x = I_n x = x,$$

also ist L im Bild der Projektion enthalten.

(b) Wegen $AA^+AA^+ = AA^+$ und weil AA^+ hermitesch ist, wird durch AA^+ gemäß Satz 25.6(a) eine orthogonale Projektion $\varphi\colon \mathbb{C}^m \to \mathbb{C}^m$ gegeben. Es gilt

$$\text{Bild}(\varphi) \subseteq \{Ax \mid x \in \mathbb{C}^n\} =: U,$$

und umgekehrt gilt für $Ax \in U$

$$Ax = AA^+Ax = \varphi(Ax) \in \text{Bild}(\varphi).$$

Also ist φ eine orthogonale Projektion auf U. Damit folgt aus Satz 25.6(b) die behauptete Ungleichung.
Für den Beweis der zweiten Behauptung in (b) sei $x \in \mathbb{C}^n$ mit $||Ax-b|| = ||AA^+b - b||$. Aus der Eindeutigkeit des Vektors aus U mit minimalem Abstand zu b folgt $Ax = AA^+b$, also $A^+b - x \in L$. Weiter gilt:

$$\begin{aligned} ||x|| \text{ minimal } &\Leftrightarrow \underbrace{A^+b - x}_{\in L} \text{ hat minimalen Abstand zu } A^+b \\ &\Leftrightarrow\ A^+b - x = (I_n - A^+A)A^+b \ \Leftrightarrow\ A^+b - x = 0 \ \Leftrightarrow\ x = A^+b, \end{aligned}$$

wobei die zweite Äquivalent aus (c) und Satz 25.6(b) folgt. Dies liefert die zweite Behauptung.

(a) Ist das lineare Gleichungssystem lösbar, so gibt es $x \in \mathbb{C}^n$ mit $||Ax-b|| = 0$. Aus (b) folgt $AA^+b = b$ und die Minimalität der Länge von A^+b unter den Lösungen. □

Satz 25.5(b) enthält eine Methode zur Bestimmung der Moore-Penrose-Inversen über die Singulärwertzerlegung, deren Berechnung aus dem Beweis von Satz 25.1 hervorgeht. Diese Methode ist numerisch stabil, aber aufwändig. Eine einfachere Methode funktioniert wie folgt: Ist $A \in \mathbb{C}^{m\times n}$ mit $r = \text{rg}(A)$, so lässt sich A zerlegen als

$$A = B \cdot C$$

mit $B \in \mathbb{C}^{m\times r}$ und $C \in \mathbb{C}^{r\times n}$, beide vom Rang r. Beispielsweise kann man r linear unabhängige Spalten von A aussuchen und diese in B schreiben und dann in C „hineinkodieren“, wie sich die Spalten von A als Linearkombinationen der Spalten von B ausdrücken. Aus Anmerkung 25.2(a) folgt die Beziehung

$$\text{rg}(A) = \text{rg}(\overline{A}^T A) = \text{rg}(A\overline{A}^T),$$

angewandt auf B und C ergibt dies also die Invertierbarkeit der Produkte $\overline{B}^T B$ und von $C\overline{C}^T$. Nun verifiziert man durch Überprüfung der Eigenschaften aus Definition 25.4, dass

$$A^+ = \overline{C}^T \left(C\overline{C}^T\right)^{-1} \left(\overline{B}^T B\right)^{-1} \overline{B}^T \tag{25.4}$$

gilt.

Beispiel 25.8. Bei

$$A := \begin{pmatrix} 2 & 3 & -2 \\ 3 & 5 & -3 \\ -2 & -3 & 2 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

ist die dritte Spalte gleich dem Negativen der ersten, also

$$A = \begin{pmatrix} 2 & 3 \\ 3 & 5 \\ -2 & -3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} =: B \cdot C.$$

Auswerten von (25.4) liefert

$$A^+ = \frac{1}{4} \begin{pmatrix} 5 & -6 & -5 \\ -6 & 8 & 6 \\ -5 & 6 & 5 \end{pmatrix}.$$

Für das lineare Gleichungssystem

$$Ax = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} =: b$$

liefert

$$x = A^+ b = \begin{pmatrix} -3 \\ 4 \\ 3 \end{pmatrix}$$

nach Satz 25.7(b) den kürzesten Vektor, dessen Produkt mit A möglichst nah an b liegt. ◁

26 Diskrete Strukturen: Spektren von Graphen

In diesem Abschnitt greifen wir nochmals die Graphentheorie auf und verbinden sie mit den Methoden der linearen Algebra, insbesondere aus Abschnitt Abschnitt 24. Der Einfachheit halber betrachten wir wieder einfache Graphen gemäß Definition 4.1. Ein Graph ist also ein Paar $G = (V, E)$ mit V einer endlichen, nicht leeren Menge von „Knoten“ und E einer Menge

$$E \subseteq \{\{x, y\} \mid x, y \in V,\ x \neq y\}$$

von „Kanten“.

Definition 26.1. *Zwei Graphen $G = (V, E)$ und $G' = (V', E')$ heißen* **isomorph**, *falls es eine Bijektion $f\colon V \to V'$ gibt, so dass*

$$\big\{\{f(x), f(y)\} \mid \{x, y\} \in E\big\} = E'.$$

Gewissenmaßen sind isomorphe Graphen bis auf die Bezichnung oder Nummerierung ihrer Knoten identisch. Es ist ein schwieriges Problem, zu zwei gegebenen (großen) Graphen festzustellen, ob sie isomorph sind. Eine Methode, um das Problem anzugehen, ist das Vergleichen der *Spektren* der Graphen, die wir nun einführen.

Es sei $G = (V, E)$ ein Graph mit $V = \{x_1, \ldots, x_n\}$. Wir setzen

$$g_{i,j} := \begin{cases} 1 & \text{falls } \{x_i, x_j\} \in E \\ 0 & \text{sonst} \end{cases} \quad \text{und} \quad A := (g_{i,j}) \in \mathbb{R}^{n \times n}.$$

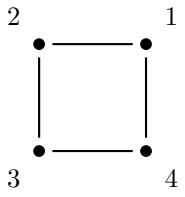
A heißt die **Adjazenzmatrix** von G. Die Menge der Eigenwerte von A (gezählt mit Vielfachheiten) ist das **Spektrum** von G.

Aus der Definition ist klar, dass die Adjazenzmatrix symmetrisch ist. Daher sind wegen Korollar 24.15(a) alle Eigenwerte reell, und die algebraischen und geometrischen Vielfachheiten stimmen überein. Da das Spektrum eine Menge mit Vielfachheiten ist, ist es zweckmäßig, die Eigenwerte als der Größe nach geordnete Liste anzugeben.

Beispiel 26.2. Der Graph G mit

$$V = \{1, 2, 3, 4\} \quad \text{und} \quad E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{1, 4\}\}$$

wird wie folgt gezeichnet:



Die Adjazenzmatrix ist

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

Deren charakteristisches Polynom ergibt sich nach kurzer Rechnung zu

$$\chi_A = \det \begin{pmatrix} x&-1&0&-1 \\ -1&x&-1&0 \\ 0&-1&x&-1 \\ -1&0&-1&x \end{pmatrix} = x^4 - 4x^2.$$

Als Spektrum bekommen wir $-2, 0, 0, 2$. ◁

Das Interesse am Spektrum eines Graphen ist durch fogenden Satz begründet.

Satz 26.3. *Die Spektren isomorpher Graphen stimmen überein.*

Beweis. Es seien $A = (g_{i,j})$ und $A' = (g'_{i,j}) \in \mathbb{R}^{n\times n}$ die Adjazenzmatrizen zweier isomorpher Graphen. Die Isomorphie bedeutet, dass es $\sigma \in S_n$ gibt mit

$$g'_{i,j} = g_{\sigma(i),\sigma(j)}.$$

Also geht A' aus A hervor, indem die Permuation σ auf die Zeilen und auf die Spalten angewandt wird. Ebenso geht die Matrix $(x \cdot I_n - A') \in \mathbb{R}[x]^{n\times n}$ aus $x \cdot I_n - A$ durch Permutation der Zeilen und Spalten mit σ hervor. Aus Lemma 16.6(b) folgt $\chi_{A'} = \chi_A$, also stimmen die Spektren überein. □

Man drückt Satz 26.3 auch aus, indem man sagt, dass das Spektrum eine Graph-Invariante ist. In analoger Sprechweise könnte man auch sagen, dass die Dimension eine Invariante eines Vektorraums ist, oder die Ordnung eine Invariante einer Gruppe. Eine weitere Graphinvariante ist die Anzahl der Zusammenhangskomponenten. Die Adjazenzmatrix selbst ist aber keine Graph-Invariante.

Gilt auch die Umkehrung von Satz 26.3? Werden also Graphen bis auf Isomorphie durch ihr Spektrum bestimmt? Wie das folgende Beispiel zeigt, ist dies leider nicht der Fall.

Beispiel 26.4. Die Graphen G und G', gegeben durch

(bei G ist der in der Mitte gezeichnete Punkt mit keinem verbunden), haben beide das Spektrum $-2, 0, 0, 0, 2$. Sie sind aber nicht isomorph. Dies kann man z.B. daran sehen, dass G' zusammenhängend ist, G aber nicht. ◁

Zwei Graphen mit demselben Spektrum nennt man *isospektral*. Wir führen nun eine Variante des Spektrums ein.

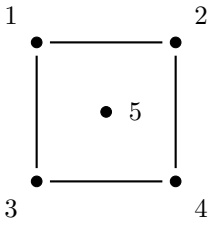
Definition 26.5. *Es sei G ein Graph mit Knoten $\{x_1, \ldots, x_n\}$ und Adjazenzmatrix $A = (g_{i,j}) \in \mathbb{R}^{n\times n}$. Für $i = 1, \ldots, n$ setzen wir $d_i := \deg(x_i) = \sum_{j=1}^{n} g_{i,j}$, den Grad des Knotens x_i. Wir bilden die Matrix*

$$L = (l_{i,j}) \in \mathbb{R}^{n\times n} \quad \textit{mit} \quad l_{i,j} = \begin{cases} -g_{i,j} & \textit{falls } i \neq j \\ d_i & \textit{falls } i = j \end{cases}.$$

L heißt die **Laplace-Matrix** *von G. Die Menge der Eigenwerte von L (gezählt mit Vielfachheiten) ist das* **Laplace-Spektrum** *von G.*

Da auch L symmetrisch ist, sind die Eigenwerte reell. Außerdem haben isomorphe Graphen identische Laplace-Spektren. Dies beweist man genau so wie Satz 26.3.

Beispiel 26.6. (a) Wenn wir die Knoten des Graphen G aus Beispiel 26.4 wie folgt nummerieren,

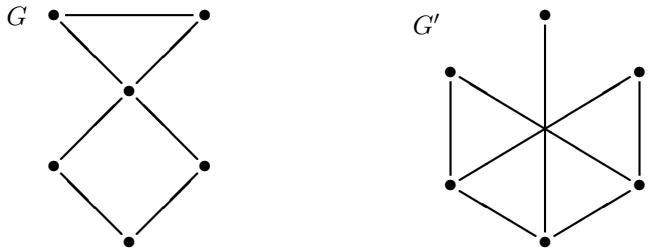


so ergibt sich die Laplace-Matrix

$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 \\ -1 & 0 & 2 & -1 & 0 \\ 0 & -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Rechnung liefert, dass das Laplace-Spektrum 0,0,2,2,4 ist. Der Graph G' aus Beispiel 26.4 hat im Gegensatz dazu das Laplace-Spektrum 0,1,1,1,5. Diese beiden Graphen lassen sich also durch ihre Laplace-Spektren trennen! Wir sehen also, dass das Laplace-Spektrum eine neue Invariante ist, die weitere Informationen liefert.

(b) Nun betrachten wir die folgenden Graphen G und G':



Aufstellen der Laplace-Matrizen und Berechnen der Eigenwerte ergibt, dass G und G' *beide* das Laplace-Spektrum

$$0, 3 - \sqrt{5}, 2, 3, 3, 3 + \sqrt{5}$$

haben. G und G' sind aber nicht isomorph. Dies kann man z.B. daran sehen, dass G' einen Knoten von Grad 1 enthält, G aber nicht. ◁

Man kann auch Beispiele nicht isomorpher Graphen finden, bei denen das Spektrum und das Laplace-Spektrum übereinstimmen.

Satz 26.7. *Die Laplace-Matrix eines Graphen ist positiv semidefinit. Das Laplace-Spektrum besteht also aus lauter nicht-negativen Zahlen.*

Beweis. Es sei $A = (g_{i,j}) \in \mathbb{R}^{n\times n}$ die Adjazenzmatrix eines Graphen. Wir benutzen Satz 24.19. Für $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ gilt

$$\langle v, L\cdot v\rangle = \sum_{i,j=1}^{n} x_i l_{i,j} x_j = \sum_{i=1}^{n} d_i x_i^2 - \sum_{i\neq j} g_{i,j}x_i x_j =$$
$$\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} g_{i,j}x_i^2 - \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} g_{i,j}x_i x_j = \sum_{1\leq i<j\leq n} g_{i,j}\left(x_i^2 + x_j^2 - 2x_i x_j\right) =$$
$$\sum_{1\leq i<j\leq n} g_{i,j}(x_i - x_j)^2 \geq 0. \quad (26.1)$$

□

Indem wir den obigen Beweis nochmal anschauen und analysieren, für welche Vektoren $v \in \mathbb{R}^n$ die Gleichung $\langle v, L\cdot v\rangle = 0$ gilt, erhalten wir einen interessanten Zusatz.

Satz 26.8. *Die Anzahl der Zusammenhangskomponenten eines Graphen G ist die Vielfachheit des Eigenwertes 0 im Laplace-Spektrum.*

Beweis. Für welche Vektoren $v = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ gilt $\langle v, L\cdot v\rangle = 0$? Wegen (26.1) muss $x_i = x_j$ für alle i, j mit $g_{i,j} = 1$ gelten. Wegen der Transitivität der Gleichheitsbeziehung gilt dann auch automatisch $x_i = x_j$, wenn i und j in derselben Zusammenhangskomponente von G liegen. Umgekehrt kann man für jede Zusammenhangskomponente Z_k eine Zahl $\alpha_k \in \mathbb{R}$ wählen und dann für alle Knoten $i \in Z_k$ $x_i := \alpha_k$ setzen. So erhält man einen Vektor v mit $\langle v, L\cdot v\rangle = 0$. Wir fassen zusammen: Mit

$$E_0 := \{v \in \mathbb{R}^n \mid \langle v, L\cdot v\rangle = 0\}$$

gilt

$$\dim(E_0) = \text{Anzahl der Zusammenhangskomponenten.} \quad (26.2)$$

Warum ist $\dim(E_0)$ die Vielfachheit des Eigenwertes 0 von L? Wegen Korollar 24.15(a) gibt es eine Orthonormalbasis $\{v_1, \ldots, v_n\}$ aus Eigenvektoren. Also $L\cdot v_i = \lambda_i v_i$ mit $\lambda_i \geq 0$ wegen Satz 26.7. Durch Umordnen können wir $\lambda_1 = \cdots = \lambda_l = 0$ und $\lambda_i > 0$ für $i > l$ erreichen. Für $v = \sum_{i=1}^{n} y_i v_i \in \mathbb{R}^n$ folgt

$$\langle v, L \cdot v \rangle = \sum_{i,j=1}^{n} y_i \lambda_j y_j \langle v_i, v_j \rangle = \sum_{i=1}^{n} \lambda_i y_i^2,$$

also $v \in E_0$ genau dann wenn $y_{l+1} = \cdots = y_n = 0$. Dies ergibt

$$\dim(E_0) = l = \text{Vielfachheit des Eigenwertes } 0 \text{ von } L.$$

Mit (26.2) folgt die Behauptung. □

Notation

Index