Matrizen für Bakkalaureaten

Dr. M. Pfenniger

Studienjahr 17/18

Inhaltsverzeichnis

Kapitel 1

Abstraktion und Notation

Vier unabhängige Entwicklungen haben historisch zur Matrizenrechnung geführt.

- Die Untersuchung, wann beliebige lineare Gleichungssysteme der Form

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad (1 \leq i \leq m) \qquad \text{bzw. kurz} \quad A \cdot \vec{x} = \vec{b}$$

  lösbar sind und wie man sie effizient löst.

- Die Transformation autonomer linearer dynamischer Systeme in der Form

$$\begin{aligned} \vec{y}(k+1) &= A \cdot \vec{y}(k), && \text{(diskret)} \\ \vec{y}\,'(t) &= A \cdot \vec{y}(t), && \text{(kontinuierlich)} \end{aligned}$$

  in eine geeignete Basis mit Hilfe der Spektraltheorie von A. Damit lassen sich dann solche Systeme formelmässig lösen. Aus der expliziten Lösung $\vec{y}(k) = A^k \cdot \vec{y}(0)$ bzw. $\vec{y}(t) = e^{At} \cdot \vec{y}(0)$ kann beispielsweise das Langzeit- und das Stabilitätsverhalten des betreffenden Systems abgelesen werden.

- Die Transformation allgemeiner Gleichungen zweiten Grades

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + b_1 x + b_2 y + c = 0$$

  in ihr System von Hauptachsen und Klassifikation der entstehenden geometrischen Objekte — der sogn Kegelschnitte — als Ellipsen, Hyperbel, Parabel, Geradenpaare und Einsiedlerpunkte. Mit Hilfe der Hauptachsentransformation gelingt es, durch Drehen den Koeffizienten des gemischten Terms a_{12} und durch Verschieben die linearen Koeffizienten b_1, b_2 weitgehend zum Verschwinden zu bringen. Kegelschnitte spielen in der Astronomie und in der Optik eine zentrale Rolle.

  Entsprechendes gelingt für die höherdimensionale Version der Gleichung

$$\sum_{i,j=1}^{3} a_{ij} x_i x_j + \sum_{j=1}^{3} b_j x_j + c = 0$$

  Die zugehörigen geometrischen Objekte entpuppen sich diesmal als (von links nach rechts) als

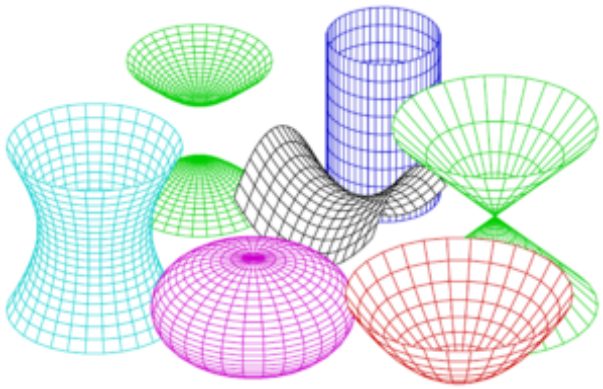


Abbildung 1.1: Graphen der Quadriken.

Ein- und zweischaliges Hyperboloid, Ellipsoid, hyperbolisches Paraboloid, Zylinder, elliptisches Paraboloid und Kegel, Zylinder oder Einsiedlerpunkt, d.h als höherdimensionale Versionen der Kegelschnitte.

In noch höheren Dimensionen führt dies zur Klassifikation der Quadriken

$$\vec{x}^T \cdot A \cdot \vec{x} + \langle \vec{b}, \vec{x} \rangle + c = 0$$

die durch die symmetrische Matrtix $A \in \mathbb{R}^{n,n}$ beschrieben werden.

Die Klassifikation symmetrischer Matrizen führt auch zur Spektralzerlegung von $A \in \mathbb{R}^{n,n}$. Dazu bestimmt man zunächst die m verschiedenen Eigenwerte $\lambda_1, \lambda_2, \ldots, \lambda_m$ im Spektrum $\sigma(A)$ von A. Es sind die Lösungen der Eigenwertgleichung

$$A \cdot \vec{v} = \lambda \vec{v}, \qquad (A - \lambda E_n) \cdot \vec{v} = \vec{0}$$

bzw. diejenigen reellen Zahlen $\lambda \in \mathbb{R}$, für die die quadratische Matrix

$$A - \lambda E$$

nicht invertierbar ist.

Die Lösungen der Eigenwertgleichung zum Eigenwert λ_k

$$A \cdot \vec{v} = \lambda_k \vec{v}$$

bilden den Eigenraum V_k von A zum Eigenwert λ_k (invarianter Teilraum). Damit erhält man eine vollständige, orthogonale Zerlegung

$$\mathbb{R}^n = \bigoplus_{k=1}^{m} V_k$$

in die orthogonalen Eigenräume von A. Zwei solche Eigenräume zu verschiedenen Eigenwerten sind orthogonal, d.h. es ist $V_k \perp V_l$, falls $k \neq l$ ist.

Bezeichnet P_k den Projektor[1] auf den Eigenraum V_k, so gilt dank der Orthogonalität und der Vollständigkeit

$$P_k \cdot P_l = \delta_{kl} P_k, \qquad \sum_{k=1}^{m} P_k = E_n$$

Für A erhält man so die *Spektralzerlegung*

$$A = \sum_{k=1}^{m} \lambda_k P_k$$

Diese Zerlegung ist sogar eindeutig, falls $\lambda_k \neq \lambda_l$ ist, ausser für $k = l$.

Ordnet man die verschiedenen Eigenwerte in aufsteigender Reihenfolge

$$-\infty =: \lambda_0 < \lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_m < \lambda_{m+1} := \infty$$

an, erhält man das projektorwertige Mass $P(x)$ von A, indem man für jede reelle Zahl $\lambda_k \leq x < \lambda_{k+1}$ die Matrix

$$P(x) = \sum_{k=0}^{m} P_k, \qquad \text{für } \lambda_k \leq x < \lambda_{k+1} \quad (P_0 := 0)$$

bildet. Die zugehörigen Summen

$$\begin{array}{ll}
P(x) = P_0 = 0 & x < \lambda_1 \\
P(x) = P_1 & \lambda_1 \leq x < \lambda_2 \\
P(x) = P_1 + P_2 & \lambda_2 \leq x < \lambda_3 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
P(x) = P_1 + \cdots + P_{m-1} & \lambda_{m-1} \leq x < \lambda_m \\
P(x) = P_1 + \cdots + P_{m-1} + P_m = E_n & \lambda_m \leq x
\end{array}$$

sind wiederum Projektoren, weil die Räume V_k, auf die P_k projiziert, alle orthogonal sind. Bewegen wir uns also entlang von $\mathbb{R}$ von $-\infty$ nach ∞, so nimmt $P(x)$ bei den Eigenwerten von A in m Schritten zu und jeder Schritt ist einer der Projektoren P_k. Der Teilraum, auf den $P(x)$ unmittelbar nach einem Schritt projiziert, enthält den Teilraum, auf den $P(x)$ unmittelbar von dem Schritt projiziert hat.

Das projektorwertige Mass $P(x)$ lässt sich auf alle Borel-messbaren Mengen erweitern, indem man zunächst für das linksseitig offene Intervall $(a, b] = \{x \mid a < x \leq b\}$ das Spektralmass

$$E((a, b]) = P(b) - P(a)$$

[1] Unter einem Projektor versteht man ein quadratische Matrix P, die symmetrisch und idempontet ist, d.h. für die

$$P = P^T, \qquad P^2 = P$$

gilt. Man kann zeigen, dass die Projektoren von $\mathbb{R}^n$ genau den Orthogonalprojektionen auf einen Teilraum V von $\mathbb{R}^n$ entsprechen. Daher entspricht die Menge der Projektoren genau den Teilräumen von $\mathbb{R}^n$. Wir sagen, dass $P_1 \leq P_2$ gilt, falls für die zugehörigen Teilräume $V_1 \subseteq V_2$ ist, was äquivalent dazu ist, dass $P_1 \cdot P_2 = P_1$ gilt. Es ist dann $\langle \vec{v}, P \cdot \vec{v} \rangle = |P \cdot \vec{v}|^2$.

definiert und dieses dann auf die anderen messbaren Teilmengen $\Delta \in \mathcal{B}(\mathbb{R})$ erweitert. Das so entstehende *Spektralmass*

$$E\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}^{n,n}, \qquad \Delta \mapsto E(\Delta)$$

ordnet der Borelmenge Δ den Projektor

$$E(\Delta) = \sum_{\{k|\lambda_k \in \Delta\}} P_k, \qquad \Delta \in \mathcal{B}(\mathbb{R})$$

zu und hat die folgenden charakteristischen Eigenschaften:

1. $E(\emptyset) = 0, E(\mathbb{R}) = E_n$.
2. $E(\Delta)$ ist ein Projektor.
3. Für $\Delta_1 \subseteq \Delta_2$ ist $E(\Delta_1) \leq E(\Delta_2)$.
4. Es ist $E(\Delta_1 \cap \Delta_2) = E(\Delta_1) \cdot E(\Delta_2)$.
5. Für $\Delta_1 \cap \Delta_2 = \emptyset$ projizieren $E(\Delta_1)$ und $E(\Delta_2)$ auf orthogonale Teilräume und es ist $E(\Delta_1 \cup \Delta_2) = E(\Delta_1) + E(\Delta_2)$.
6. Für alle Vektoren $\vec{v} \in \mathbb{R}$ liefert die Funktion

$$E_{\vec{v}}\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}, \qquad \Delta \mapsto \langle \vec{v}, E(\Delta) \cdot \vec{v} \rangle$$

ein reelles Mass auf der Borel σ-Algebra $\mathcal{B}(\mathbb{R})$.

Spektralzerlegung und Spektralmass symmetrischer Matrizen bilden das wichtigste Werkzeug für die Quantenmechanik, weil sie den Ja-Nein-Fragen über ein physikalisches System bzw. seinen Eigenschaften entsprechen.

- Der Versuch, das Konzept der Zahlen auf höhere Dimensionen zu verallgemeinern. Dabei stösst man auf die komplexen Zahlen $\mathbb{C}$, die Quaternionen $\mathbb{H}$ oder allgemeiner der Clifford-Algebren Cl_n und auf die endlichen Körper, die alle in gewissen Anwendungen eine zentrale Rolle spielen.

Obwohl wir zur gegebenen Zeit auf jeden dieser, auch für die Praxis wichtigen, Aspekte eingehen werden, stellen wir den ersten Punkt als Motiviation an den Anfang. Eines der wichtigsten Themen der linearen Algebra ist nämlich das Studium linearer Gleichungssysteme und ihrer Lösungsmengen. Eine provisorische Definition der linearen Algebra besagt, dass es sich dabei um die Theorie der Lösungen linearer Gleichungssysteme handelt.

Den Anwendern zuliebe werden wir also nicht den heute unter Mathematikern üblichen abstrakten[2] Zugang zur linearen Algebra via endlichdimensionale Vektorräume (über einem Körper) und strukturverträgliche lineare Abbildungen benutzen, sondern eine kanonische Basis wählen und dann in konkreter Weise via Matrizen vorgehen. Es gibt also zwei verschieden Arten, um lineare Algebra zu betreiben. In der abstrakten Art der Mathematiker interessiert man

[2] Abtrakt bedeutet konzeptionell und nicht weltfremd, wie gewisse Leute meinen. Beim Abstrahieren wird der wesentliche Kern einer Situation herausgeschält und studiert und weniger auf Beispiele geachtet. Dadurch lassen sich im vorliegenden Fall Algebra und Geometrie zusammenbringen, Quotientenstrukturen und Dualität besser verstehen, neue Anwendungen der selben Theorie und damit auch neue Beziehungen zwischen den Anwendungen erschliessen.

sich für Strukturen und Beweise. Dabei spielen Konzepte wie Basis, Dimension, lineare Abbildung, Kern, Bild, Rang, etc. die Hauptrolle und unnatürliche Wahlen werden möglichst lange vermieden. In der rechnerischen Art der Anwender sucht man Algorithmen, mit denen diese Konzepte berechnet werden können. Wir werden vorwiegend rechnerisch vorgehen. Wir suchen also effiziente Verfahren, um die gestellten Probleme zu lösen und werden zu unserer Überraschung feststellen, dass alle Probleme der linearen Algebra darauf hinauslaufen, ein gewisses lineares Gleichungssystem lösen zu müssen. Gelegentlich werden wir versuchen, etwas hinter diese Verfahren zu blicken und dabei die abstrakten Zusammenhänge zu erkennen beginnen. Wer sich aber wirklich in der abstrakten Seite der linearen Algebra auskennen will, was man aus heutiger Sicht auch als Anwender unbedingt sollte, wenn man diese Disziplin wirklich beherrschen will, muss das andernorts machen.

In diesem Abschnitt wollen wir an Hand einiger typischer Fragestellungen andeuten, wo lineare Gleichungssysteme bzw. Matrizen in aussermathematischen Anwendungen eine zentrale Rolle spielen. Dabei wird schnell klar werden, dass die in der Mittelschule gepflegten Umgangsformen mit linearen Gleichungssystemen und Vektorgeometrie zwar weitgehend unbrauchbar sind, aber zur Illustration nützlich sein können.

Matrizenrechnung und der Spezialfall der Vektorrechnung spielt in fast allen Anwendungen der Mathematik — etwa in der Elektrotechnik, Mechanik, Hydraulik, Elastizitätstheorie, Informatik, Thermodynamik, Optik, Stöchiometrie usw. die zentrale Rolle. Zur Beschreibung von Gleichgewichtszuständen von Netzwerkproblemen — etwa bei der Verwendung von Gleichströmen oder in der Theorie der Fachwerke bzw. beim Vorhandensein von konstanten Kräften in Masse-Federn-Systeme — treten lineare Gleichungssysteme der Form

$$A \cdot \vec{x} = \vec{b}$$

direkt auf, wenn man die involvierten Grössen bilanziert. Die Ausgleichsrechnung liefert für das Extremum des skalaren quadratischen Fehlers

$$E(\vec{x}) = |A \cdot \vec{x} - \vec{b}|^2 = \langle A \cdot \vec{x} - \vec{b}, A \cdot \vec{x} - \vec{b} \rangle = \langle A \cdot \vec{x}, A \cdot \vec{x} \rangle - 2\langle A \cdot \vec{x}, \vec{b} \rangle + \langle \vec{b}, \vec{b} \rangle$$

das lineare Gleichungssystem

$$(A^T \cdot A) \cdot \vec{x} = A^T \cdot \vec{b}$$

der sogn. Normalengleichungen. Sie dienen etwa in der Geodäsie dazu, trotz der unvermeidlichen Messfehlern das Optimum an Information aus Messungen herauszuholen. In den erwähnten Anwendungen als Netzwerkprobleme haben die entstehenden linearen Gleichungssysteme die durch die Diagonalmatrix C mit strikt positiven Elementen gewichtete spezielle Form

$$(A^T \cdot C \cdot A) \cdot \vec{x} = A^T \cdot C \cdot \vec{b} - \vec{f}$$

für gegebene Quellenterme $\vec{b}$ und $\vec{f}$, deren Koeffizientenmatrix $A^T \cdot C \cdot A$ symmetrisch und positiv definit ist. Sie extremalisieren den Skalar

$$E(\vec{x}) = \frac{1}{2}\langle A \cdot \vec{x} - \vec{b}, C \cdot (A \cdot \vec{x} - \vec{b}) \rangle - \langle \vec{x}, \vec{f} \rangle$$

der dort die Rolle der Energie spielt. Im Spezialfall $C = E$ und $\vec{f} = \vec{0}$ wird aus dem Netzwerkproblem offenbar das Problem der Ausgleichsrechnung. Allgemeiner führt Optimieren auf sogn. lineare Programme, bei denen neben linearen Gleichungen noch lineare Ungleichungen vorkommen dürfen. Mit der einer Matrix A zugehörigen linearen Abbildungen lassen sich in der linearen Kodierungstheorie Kodes effektiv behandelt, mit denen Fehler erkannt und korrigiert und Informationen chiffriert werden können.

Bemerkenswerterweise erlaubt die Matrizenrechnung neben der Beschreibung der stationären Zustände d.h. der Statik auch die Beschreibung der ganzen dynamischen Entwicklung dieser Systeme. Mit der Eigenwertgleichung

$$A \cdot \vec{x} = \lambda \vec{x}, \qquad \text{bzw.} \qquad (A - \lambda E) \cdot \vec{x} = \vec{0}$$

in der die charakteristische Matrix $A - \lambda E$ die Hauptrolle spielt, können Prozesse untersucht, gesteuert und geregelt werden, die wachsen, zerfallen schwingen oder komplex strömen. Zusammen mit etwas Graphentheorie, die eng mit der linearen Algebra verwoben ist und in der die Matrizenrechnung ebenfalls eine wichtige Rolle spielt, lassen sich mit Matrizen die Prototypen von Maschinen und ihre Dynamik beschreiben. Die jüngste erfolgreiche Anwendung der Eigenwertgleichung stochastischer Matrizen zur Berechnung der Dynamik diskreter Markov-Prozesse dient in modernen Suchmaschinen dazu, Dokumente im Internet zu gewichten, indem jedem Dokument ein um so grösseres Gewicht gegeben wird, je mehr Dokumente mit möglichst grossem Gewicht auf dieses Dokument verweisen. Damit wird das Verhalten eines Zufallssurfers simuliert. Allgemeiner wird durch einen diskreten Markov-Prozess die zeitliche Entwicklung einen Systems modelliert, in dem der künftige Zustand zufällig vom aktuellen Zustand und nicht von der ganzen Vorgeschichte abhängt. Damit können beispielsweise zufällige Pfade in Graphen und ihre statistischen Eigenschaften beschrieben werden. Diese enge Beziehung zwischen Stochastik und linearer Algebra gehört ins Gepäck jedes Informatikers, der sein Geld wert ist. Weitere Anwendungen der linearen Algebra betreffen Bildbearbeitung, Kryptologie, Maschinenlernen, Optimierung, Graphentheorie, Quantenberechnungen etc. Angewandte Mathematik ist also kein Zweig der Mathematik, sondern höchstens die Kunst, Probleme auf lineare Algebra zu reduzieren.

Lösungsmengen linearer Gleichungssysteme und Matrizen spielen beim Studium der Geometrie eine zentrale Rolle. Anschaulich kann man sich eine einzige lineare Gleichung als eine i.a. hochdimensionale Ebene — man redet von einer Hyperebene — vorstellen. Ein lineares Gleichungssystem beschreibt dann das Schnittgebilde mehrerer solcher Ebenen und lineare Abbildungen führen solche linearen Räume ineinander über. Die Grundaufgaben der analytischen Geometrie formuliert und löst man zweckmässigerweise mit Hilfe der Matrizenrechnung. Die Aufgabe, eine lineare Abbildung umzukehren oder sie zu iterieren bzw. interessante Information über sie zu erhalten, führt auf gewisse lineare Gleichungssysteme. Matrizen und die mit ihnen beschriebenen linearen Abbildungen spielen in Form von Symmetrien bzw. von Koordinatentransformationen oder beim Studium der Projektionen und der Perspektive in der Computergraphik oder in der speziellen Relativitätstheorie eine zentrale Rolle, weil es sich herausstellt, dass die meisten Lie-Gruppen als sogn. Matrizengruppen realisiert werden können. Die lineare Algebra hat also neben der algebraischen auch eine geometrische Seite und die algebraischen sind mit den geometrischen Aspek-

te so eng verwoben, so dass sie gemeinsam untersucht werden müssen, um ein vollständige mathematische Theorie zu liefern. Klar geworden scheint das erst so richtig Lagrange[3], der meinte:

> As long as algebra and geometry were separate subjects, their progress was slow and their uses limited, but when those two sciences united, they have lent each other their forces, and have since marched together to perfection.

Aus moderner mathematischer Sicht handelt lineare Algebra von linearen Abbildungen zwischen Vektorräumen. Wir werden uns hier mit dem äquivalenten Skelett dieser abstrakten Theorie befassen, jeweils eine Basis wählen und daher mit konkreten Matrizen rechnen. Bezeichnen wir also mit $\mathbf{Mat}_{\mathbb{R}}$ die kleine Kategorie, deren Objekte durch die natürlichen Zahlen n, m ... (Dimensionen) indiziert und deren Morphismen reelle $m \times n$-Matrizen aus

$$\mathbf{Mat}_{\mathbb{R}}(n, m) = \{(m \times n) - \text{Matrizen mit Elementen in } \mathbb{R}\}$$

sind, so kann eine Matrix $A \in \mathbf{Mat}_{\mathbb{R}}(n, m)$ als Morphismus $n \to m$ bzw. als lineare Abbildung $A\colon \mathbb{R}^n \to \mathbb{R}^m$ aufgefasst werden. Komposition ist das übliche Matrizenprodukt. Diese Zuordnung liefert einen voll-treuen Funktor $F\colon \mathbf{Mat}_{\mathbb{R}} \to \mathbf{FinVect}_{\mathbb{R}}$ in die Kategorie der endlich dimensionalen Vektorräume, der auf den Objekten surjektiv ist und daher eine Äquivalenz zwischen diesen Kategorien liefert, aber nicht kanonisch ist, weil er von der Wahl einer Basis abhängt. So gesehen wird die etwas armseelige natürliche Zahl $n \in \mathbb{N}$ durch den reichhaltigeren Vektorraum $\mathbb{R}^n$ der Dimension n ersetzt; die Summe natürlicher Zahlen $n + m$ entspricht bei dieser Übersetzung von der Zahlentheorie in die Theorie der Vektorräume der direkten Summe $\mathbb{R}^n \oplus \mathbb{R}^m$ und das Produkt $n \cdot m$ entpricht dem Tensorprodukt $\mathbb{R}^n \otimes \mathbb{R}^m$. Fasst man einen Vektor $\vec{v} \in \mathbb{R}^n$ als Funktion $v\colon [1..n] \to \mathbb{R}$ auf, für die $v(i) = v_i$ gilt, so lassen sich die Elemente der direkten Summe $\mathbb{R}^n \oplus \mathbb{R}^m \cong \mathbb{R}^{n+m}$ als Funktionen $[1..n] + [1..m] \cong [1..n + m] \to \mathbb{R}$ und die Elemente des Tensorproduktes $\mathbb{R}^n \otimes \mathbb{R}^m \cong \mathbb{R}^{n \cdot m}$ als Funktionen $[1..n] \times [1..m] \cong [1..n \cdot m] \to \mathbb{R}$ auffassen. Statt dass man also zwischen zwei typischen Objekten der Zahlentheorie, d.h. zwischen zwei natürlichen Zahlen n und m kaum interessante Beziehungen findet, erhält man in der reichhaltigeren Theorie der endlich dimensionalen Vektorräume $\mathbf{FinVect}_{\mathbb{R}}$ zwischen den zugehörigen Vektorräumen $\mathbb{R}^n$ und $\mathbb{R}^m$ interessante Beziehungen, die sich effizient in Form von Matrizen $A \in \mathbb{R}^{m,n}$, d.h. als Morphismen $A\colon \mathbb{R}^n \to \mathbb{R}^m$ der Kategorie $\mathbf{Mat}_{\mathbb{R}}$ darstellen lassen. Beispielsweise hat die Zahl 3 keine interessante innere Struktur; im Gegensatz dazu lässt der zugehörige Vektorraum $\mathbb{R}^3$ interessante Symmetrien in Form der speziellen orthogonalen Gruppe $SO_3(\mathbb{R})$ zu, die sich geometrisch in Form von Raumdrehungen interpretieren lassen.

Der Vektorraum $\mathbb{R}^m$ hat weitere Struktur. Die assoziative und kommutative Vektoraddition

$$\mu\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m, \qquad (\vec{x}, \vec{y}) \mapsto \vec{x} + \vec{y}$$

liefert ein kommutatives Monoid mit der eindeutigen linearen Abbildung

$$\eta\colon \mathbb{R}^0 \to \mathbb{R}^m, \qquad \vec{0} \mapsto \vec{0}$$

[3] $1736 - 1813$.

als Einheit. Dual liefert die koassoziative und kokommutative Diagonalenabbildung (Duplikation)

$$\hat{\mu}\colon \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^m, \qquad \vec{x} \mapsto (\vec{x}, \vec{x})$$

ein kommutativen Komonoid mit der eindeutigen linearen Abbildung (Löschen)

$$\hat{\eta}\colon \mathbb{R}^m \to \mathbb{R}^0, \qquad \vec{x} \mapsto \vec{0}$$

als Koeinheit. Wie diese Strukturen miteinander verträglich sind, beschreibt die Algebra eines bikommutatives Bimonoides. Weil die Multiplikation mit einem Skalar $r \in \mathbb{R}$ eine weitere verträgliche Struktur

$$s\colon \mathbb{R} \to \mathrm{End}(\mathbb{R}^m), \qquad r \mapsto s_r, \quad [s_r(\vec{x}) := r\vec{x}]$$

liefert, hat der Vektorraum $\mathbb{R}^m$ die Struktur eines bikommutativen Bimonoids über $\mathbb{R}$. Die Kategorie $\mathbf{Mat}_{\mathbb{R}}$ entpuppt sich als *freie* symmetrische Monoidalkategorie über einem bikommutativen Monoid über $\mathbb{R}$ und die erwähnten Verträglichkeiten charakterisieren die lineare Algebra vollständig.

Aus postmodernern Sicht hat die Matrizenrechnung folgende Charakterisierung.

Satz. Für einen kommutativen Rig R ist $\mathbf{Mat}_R$ der PROP für die bikommutativen Bimonoide über R.

Dabei versteht man unter einem kommutativen Rig einen kommutativen Ring ohne Negative, d.h. auf die Bedingung, dass jedes Element ein additives Inverses hat, wird verzichtet. Beispiele sind neben dem Körper $(\mathbb{R}, +, 0, \times, 1)$ und dem Ring der ganzen Zahlen $(\mathbb{Z}, +, 0, \times, 1)$ die natürlichen Zahlen $(\mathbb{N}, +, 0, \times, 1)$ oder die Boolesche Algebra $(\mathbb{B}, \vee, 0, \wedge, 1)$ mit der Disjunktion $\vee$ und der Konjunktion $\wedge$. Ein PROP ist eine strikt symmetrische Monoidalkategorie, deren Objekte die natürlichen Zahlen sind und deren Tensorprodukt auf den Objekten die Addition ist. Ein Morphismus $[n] \to [m]$ in einem PROP kann man sich als Blackbox mit m Inputs und n Outputs vorstellen, die beliebig vertauscht werden dürfen. Die $(m \times n)$-Matrizen mit Elementen aus R bilden die Morphismen $[n] \to [m]$ eines PROP $\mathbf{Mat}_R$ und im Satz wird behauptet, dass ihre Algebren, d.h. die Kategorie der symmetrischen Monoidalfunktoren $\mathbf{Mat}_R \to \mathbf{C}$ in eine beliebige symmetrische Monoidalkategorie $\mathbf{C}$ äquivalent zur Kategorie $\mathbf{CBMon}_R(\mathbf{C})$ der bikommutativen Bimonoide über R in der Kategorie $\mathbf{C}$ ist, d.h. es gilt

$$\mathrm{SymMon}(\mathbf{Mat}_R, \mathbf{C}) \cong \mathbf{CBMon}_R(\mathbf{C})$$

Die Situation ist analog zur Kategorie **FinSet** der endlichen Mengen und Funktionen. Deren Skelett $\mathbf{\Phi}$ hat als Objekte die durch die natürlichen Zahlen indizierten endlichen Mengen

$$\langle n \rangle := \{0, 1, \ldots, n-1\}, \quad n \in \mathbb{N}$$

und die Einbettung $F\colon \mathbf{\Phi} \to \mathbf{FinSet}$. Die Addition liefert eine symmetrische Monoidalstruktur mit dem Neutralelement $\langle 0 \rangle$. Weil $(\mathbf{\Phi}, +, \langle 0 \rangle)$ die freie symmetrische Monoidalkategorie mit dem kommutativen Monoid $\langle 1 \rangle$ ist, beschreiben die Funktionen endlicher Mengen einen PROP, dessen Algebren die kommutativen Monoide sind. Es ist in der Tat das universelle kommutative Monoid (freie

symmetrische Monoidalkategorie mit einem kommutativen Monoid $\langle 1\rangle$), d.h. für jede symmetrische Monoidalkategorie $\mathbf{C}$ erhält man eine Äquivalenz

$$\mathrm{SymMon}(\mathbf{\Phi}, \mathbf{C}) \cong \mathbf{CMon}(\mathbf{C})$$

Beschränkt man sich auf die monoton wachsenden Funktionen $\langle n\rangle \to \langle m\rangle$, erhält man die skeletale Unterkategorie $\mathbf{\Delta} \subset \mathbf{\Phi}$ der augmentierten Simplizes. Weil die Monoidalkategorie $(\mathbf{\Delta}, +, \langle 0\rangle)$ die freie Monoidalkategorie mit dem Monoid $\langle 1\rangle$ ist, beschreiben die monotonen Funktionen endlicher Mengen einen PRO, dessen Algebren die Monoide sind, d.h. man erhält eine Äquivalenz der Kategorie der monoidalen Funktoren und monoidalen natürlichen Transformationen in eine beliebige Monoidalkategorie $\mathbf{C}$ zur Kategorie der Monoide und Monoidhomomorphismen in $\mathbf{C}$, d.h. es gilt

$$\mathrm{Mon}(\mathbf{\Delta}, \mathbf{C}) \cong \mathbf{Mon}(\mathbf{C})$$

Die Kategorie $\mathbf{\Delta}$ ist Ausgangspunkt der algebraischen Topologie und der homologischen Algebra und damit tatsächlich der ganzen postmodernen Mathematik.

Eine weitere wichtige Quelle für Matrizen und lineare Gleichungssysteme ist die Analysis. Vereinfachend kann man sagen, dass Analysis darin besteht, gekrümmte Objekte — man redet von sogenannten glatten Mannigfaltigkeiten — und komplizierte Funktionen, die darauf definiert sind, lokal in "unendlich kleinen" Umgebungen einzelner Punkte — man redet von sogenannten Tangentialräumen in den Punkten — zu studieren. Der entscheidende Vorteil besteht darin, dass in einer solchen "infinitesimal kleinen" Umgebung alles linear wird und sich durch einfacher zu behandelnde flache Objekte und lineare Funktionen approximieren lässt — man redet im Zusammenhang mit diesem natürlichen Prozess auch von Linearisieren. Der Vorteil davon, dass lineare Strukturen natürlicherweise in infinitesimalen Strukturen auftauchen, besteht darin, dass die Probleme der linearen Algebra in vieler Hinsicht einfacher sind und sich viel effizienter lösen lassen, als die ursprünglich von den Analytikern formulierten Probleme. Deshalb werden die meisten Probleme der Mathematik und der Physik zur effektiven Lösung erst einmal linearisiert und landen somit früher oder später in der linearen Algebra. Die Funktional-Matrix und die Hesse-Matrix

$$\partial(f) = (\partial_j f_i), \qquad H(f) = (\partial_i \partial_j f)$$

verallgemeinern die erste f' und die zweite Ableitung f'' bzw. die Steigung und die Krümmung auf höhere Dimensionen. Entscheidend für die Analysis ist dann der Umstand, dass der Linearisierungsprozess funktoriell ist, d.h. dass die Komposition von Funktionen mit der Komposition ihrer Linearisierungen verträglich ist. In der Sprache der Analysis besagt dieser fundamentale Sachverhalt, dass unter der Linearisierung die Verkettung zweier Funktionen f und g zur Verkettung ihrer Funktionalmatrizen wird, wie die Kettenregel

$$\partial(g \circ f) = \partial(g) \cdot \partial(f)$$

besagt. Damit erlaubt es die Linearisierung einer Differentialgleichung, Aussagen über ihr Stabilitätsverhalten zu machen. Nur mit Hilfe von finiten Elementen bzw. Systemen von Differenzengleichungen lassen sich die meisten partiellen

Differentialgleichungen und Systeme von Differentialgleichungen numerisch approximativ lösen[4]. Geometrisch ist die lineare Algebra also die lokale Theorie der Mannigfaltigkeiten. Um mit den flachen Hyperebenen der linearen Algebra allerdings gekrümmte Objekte genügend gut approximieren zu können, wird man in der Regel mit *vielen* linearer Gleichungen rechnen müssen. Deshalb werden solche Objekte als Polytope bezeichnet und basieren die meisten seriösen Anwendungen auf multilinearer Algebra.

Schliesslich ist die lineare Algebra nicht einfach irgend eine weitere Sammlung mathematischer Tricks zum Lösen von Problemen, sondern eine ausgewachsene mathematische Theorie. Der Kosmos der linearen Algebra spielt aus mathematischer Sicht eine prominente Rolle, weil für ihn mit Hilfe der Matrizenrechnung systematische und konstruktiv ziemlich lange befriedigende Konzepte entwickelt werden können, die natürlich miteinander verträglich sind. Fast[5] nur in der linearen Algebra kann das Nebeneinander der algorithmischen, der algebraischen und der geometrischen Seite einer Theorie studiert und bei Bedarf erst noch davon abstrahiert werden. In jüngster Zeit hat zudem die ungewöhnliche Logik der linearen Algebra, die im Gegensatz zur klassischen Logik auf die Resourcen Rücksicht nimmt und beispielsweise das Kopieren nicht erlaubt, zu einem vertieften Verständnis der Rolle von Information in der Informatik und in der Physik geführt. Tieferes Verständnis der erfolgreichsten physikalischen Theorie aller Zeiten — der Quantenmechanik — deren Gleichungen linear sind und de-

[4] Tatsächlich ist der Autor der Meinung, dass einem heutigen Ingenieur-Studenten viel zu viel überholte Analysis und Trigonometrie zugemutet wird. Seiner Ansicht nach würde man im 21. Jahrundert die noch zur Verfügung stehende Zeit effizienter zum Lernen von natürlicher Mathematik, z. B. von mehr linearer Algebra investieren, als Tricks zum Integrieren, Lösen von Extremalaufgaben, Differentialgleichungen, Berechnen von Flächeninhalten oder gar zum Drillen von Dreiecksaufgaben, trigonometrischen Gleichungen usw. zu pauken und mystisch im aussichtslosen numerischen Chaos herumzuwühlen. Wir leben im Zeitalter des Computers und nicht mehr in jenem des Rechenschiebers und der Tabellenwerke. Bei den enorm vielen, in kürzester Zeit durchgeführten arithmetischen Operationen ist man erst recht darauf angewiesen zu verstehen, was eigentlich abläuft und kann nicht einfach gedankenlos die schnell entstanden farbigen Bildchen anstaunen und blind auf einen weiteren Knopf drücken. Paradoxerweise lassen sich die alten Praktiker von der praktischsten aller bekannten Theorien nicht so leicht überzeugen, weil sie sie selten gut genug kennen. Gutes Anwender-Handwerk, das heute meistens als Modellieren bezeichnet wird, würde von ihnen aus dieser Sicht verlangen, die hinter einer beabsichtigten Anwendung liegende mathematische Struktur — also insbesondere das zugehörige lineare Problem — zu erkennen und es direkt, d.h. ohne Umweg über die Analysis, zu modellieren, d.h. abstrakt zu formulieren. Der Rest ist Mathematik. Dass die Wahl einer möglichst natürlichen mathematischen Sprache einen wesentlichen Einfluss darauf hat, wie weit man mit ihr kommt, erkennt man, wenn man die Komplexität der Multiplikation CCLIV · LVII im Zahlensystem der alten Römer, die sich auf ihre Bodenständigkeit auch viel einbildeten, mit jener von $254 \cdot 57$ im Dezimalsystem oder gar mit $11111110_2 \cdot 111001_2 \cdot$ im Binärsystem vergleicht. Ein römischer Praktiker hätte sich vermutlich 254 Eimer verschafft und in jeden 57 Erbsen abgezählt. Dann hätte er alle Eimer zusammengeschüttet und versucht, die Erbsen des ganzen Haufens zu zählen! Nach endlich langer Zeit hätte er das Ergebnis MMMMMMMMMMMMMMCDLXXVIII erhalten, das wir heute im Dezimalsystem kurz durch $14'478$ oder binär durch 11100010001110_2 darstellen und von einer Maschine in wenigen Nanonsekunden berechnen lassen. Andere elementare Beispiele, wo eine bessere Theorie Wunder gewirkt hat, sind die Trigonometrie, die sphärische Trigonometrie oder die Beschreibungen von Drehungen im Raum mit Hilfe von Euler-Winkeln. Diese Theorien werden mit Hilfe der komplexen Zahlen, der Vektorrechnung und den Quaternionen konzeptionell einfacher und damit durchsichtiger und zuverlässiger. Ferner muss man erst noch nicht mehr rechnen wie wild! Warum solche nicht-linearen Theorien einem Anfänger überhaupt noch gelehrt werden, ist also völlig schleierhaft, peinlich und nicht einmal politisch verständlich.

[5] Vergleichbar sind einzig die algebraische Geometrie und die algebraische Topologie, obwohl dort die Effizienz der Algorithmen für den Praktiker (derzeit) zu wünschen übrig lässt.

ren (multi-)linearer Charakter für ihre Verrücktheiten im Zusammenhang mit den Superpositionen, Unbestimmtheitsrelationen und Verschränkungen verantwortlich ist, hat in jüngster Zeit zu einer Fusion der beiden Gebiete in der Quanteninformatik geführt.

Den Interessen und der Vorbildung der Studenten entsprechend werden wir unseren Schwerpunkt auf die algorithmische Seite legen und viele konkrete Probleme lösen. Trotzdem wollen wir ab und zu versuchen, uns aus dem numerischen Morast in die Luft zu erheben, um aus der Vogelperspektive etwas weiter zu sehen und ein tieferes Verständnis für die zugrunde liegenden mathematischen Zusammenhänge und Strukturen zu entwickeln, weil auch viele Anwender hoffentlich früher oder später die rechnerische Knochenarbeit einer Maschine überlassen und über ihr Gebiet lieber nachdenken als stumpfsinnig weiterrechnen werden.

In der Regel werden in einer realistischen Anwendung Gleichungssysteme mit hundertausenden von Unbekannten entstehen. Diese grosse Zahl linearer Gleichungen zeigt schon, dass eine allgemeine Theorie nötig ist, wenn man nicht einfach als Kröte im numerischen Sumpf suhlen will. Historisch war das den Mathematikern lange nicht klar. Die lineare Algebra wurde nämlich später als viele andere — kompliziertere — Theorien zu einer vollständigen Theorie entwickelt. Die Mathematiker wollten die lineare Algebra lange Zeit nicht als eigenständige Theorie anerkennen — sie war ihnen zu einfach! Zum Glück haben Physiker beim Studium physikalischer Theorien, wie Hamilton und Maxwell in der linearen Elektrodynamik oder die Modernen in der linearen Quantenmechanik, nicht aufgehört, immer wieder Vektoren zur Beschreibung ihrer Systeme zu benutzen. Erst als die Mathematiker am Anfang des 20. Jahrhunderts begonnen haben, ihr Augenmerk auf das Studium von Symmetrien zu richten, ist ihnen die zentrale Bedeutung der linearen Algebra klar geworden.

Matrizen haben sich dann später in der Informatik als idealer Datentyp erwiesen: mit Maschinen können Matrizen einfach manipuliert werden. In der Bild- und Signalverarbeitung treten in der Regel sehr grosse Matrizen auf. Ein TV-Signal beispielsweise benötigt etwa 10^7 Pixel pro Sekunde, um alle relevanten Informationen darstellen zu können. Eine einzige Sekunde eines solchen abgetasteten Signals kann als Vektor mit 10^7 Komponenten aufgefasst werden. Ein kurzer Film liefert 10^3 solcher Vektoren. Ein 90-minütiger Film, der mit 24 Bildern pro Sekunde läuft, besteht aus $129'600$ Bildern, von denen jedes $2048 \cdot 872$ Pixel mit 4 Bytes Farbinformation enthält. Ein solcher Film braucht bei der Herstellung also rund $2 \cdot 10^{12}$ [Bytes] d.h. 2 [TB] Speicherplatz. Datenmengen dieser Grössenordnung und erst recht der von den Geheimdiensten vollständig abgehörte weltweite E-Mail-Verkehr, die von der Polizei angelegten Fingerabdruck-Datenbanken oder die Webseiten des Internets müssen dann mit Hilfe der linearen DFT und anderen linearen Datenfiltern, etwa der Wavelet-Transformation, oder von linearen Bewertungsalgorithmen wie dem von Google benutzten PageRank-Algorithmus oder dem HITS-Algorithmus in kürzester Zeit durchsucht, geglättet, gefiltert, komprimiert, sortiert oder sonstwie verarbeitet werden. Wegen der sehr grossen Menge in solchen Anwendungen anfallender Daten ist es wichtig, alle Matrizenoperationen möglichst effizient und stabil durchführen zu können, um Rechenzeit zu sparen und um möglichst wenig Rundungsfehler in Kauf nehmen zu müssen. Um den zu erwarteten Aufwand abschätzen zu können, werden wir regelmässig die Komplexität unserer Verfahren untersuchen und gelegentlich das numerische Verhalten der fundamentalen

Algorithmen unter die Lupe nehmen.
Die DFT kann als Multiplikation mit der speziellen Matrix mit den Elementen

$$F_n = \left(e^{(r-1)(s-1)\frac{2\pi}{n}i}\right) \in \mathbb{C}^{n,n}, \qquad 1 \leq r, s \leq n$$

aufgefasst werden, in der die komplexen n-ten Einheitswurzeln

$$\zeta^k := e^{ik\frac{2\pi}{n}} \in \mathbb{C}, \quad (0 \leq k \leq n-1)$$

die algebraisch die Gleichung $\zeta^n = 1$ lösen und geometrisch auf den Ecken eines regulären n-Ecks liegen, eine prominente Rolle spielen. Diese komplexe Matrix lautet ausgeschrieben

$$F_n = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \zeta & \zeta^2 & \zeta^3 & \cdots & \zeta^{n-1} \\ 1 & \zeta^2 & \zeta^4 & \zeta^6 & \cdots & \zeta^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta^{n-1} & \zeta^{2(n-1)} & \zeta^{3(n-1)} & \cdots & \zeta^{(n-1)\cdot(n-1)} \end{pmatrix} \in \mathbb{C}^{n,n}$$

Ihre Zeilen und die Spalten sind paarweise orthogonal und haben alle die selbe Norm n, was sich in der Matrizengleichung

$$F_n \cdot \overline{F}_n^T = n \cdot E_n = \overline{F}_n^T \cdot F_n$$

manifestiert, auf Grund der die Matrix F_n leicht invertierbar ist. In den Anwendungen taucht die Fourier-Transformation immer dann auf, wenn periodische Phänomene eine Rolle spielen. Dass sich die Fourier-Transformation speziell gut mit periodischen Phänomenen verträgt, hängt damit zusammen, dass einerseits unter F_n ein r-periodisches Signal in ein $\frac{n}{r}$- periodisches Signal übergeht. Mit der Matrix F_n lassen sich andererseits alle zirkulanten Matrizen, die im Zusammenhang mit der zyklischen Faltung auftreten, simultan diagonalisieren. Das hat zur Folge, dass unter Fourier-Transformation das Faltungsprodukt in ein elementweises Produkt übergeht und dass daher die Fourier-Transformation eines beliebigen Vektors auf einfache Art mit der Fourier-Transformation des zyklischen Shift dieses Vektors verträglich ist, weil seine k-te Komponente einfach elementweise mit ζ^k multipliziert wird. Dieser Sachverhalt erlaubt es dann sogar, dass wir zwei (grosse) ganze Zahlen oder zwei Polynome sehr viel schneller multiplizieren können, als wir es in der Schule gelernt haben.

Dass das Matrizenprodukt eines Vektors $\vec{a}$ mit der Matrix F_n speziell effizient berechnet werden kann, hängt mit einer bemerkenswerten Faktorsierungseigenschaft dieser Matrizen zusammen, die von den Symmetrien der regulären n-Ecks und der Struktur des Tensorproduktes herrühren. Ordnen wir nämlich für gerades $n = 2k$ die Komponenten des zu transformierenden Vektors $\vec{a} \in \mathbb{C}^n$ so um, dass wir zuerst seine geraden und dann die ungeraden Komponenten hinschreiben, erhalten wir die Blockzerlegung

$$F_n \cdot \vec{a} = F_n \cdot \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-2} \\ a_{n-1} \end{pmatrix} = \left(\begin{array}{c|c} F_k & \zeta_n F_k \\ \hline F_k & -\zeta_n F_k \end{array}\right) \cdot \left(\begin{array}{c} a_0 \\ \vdots \\ a_{n-2} \\ \hline a_1 \\ \vdots \\ a_{n-1} \end{array}\right)$$

Die spezielle Struktur der Matrizen F_n lässt sich mit Hilfe des Tensorproduktes besonders übersichtlich ausdrücken. Beispielsweise gilt

$$F_{16} = B_{16} \cdot (E_2 \otimes B_8) \cdot (E_4 \otimes B_4) \cdot (E_8 \otimes B_2) \cdot P_{16}$$

Dabei setzt die Permutationsmatrix P_{2m} die m geraden Komponenten vor die m ungeraden (Bitumkehr) und die sogn. Schmetterlingsmatrix B_{2l} ist definiert als Blockmatrix von Diagonalmatrizen

$$B_{2l} = \begin{pmatrix} E_l & W_l \\ E_l & -W_l \end{pmatrix}, \qquad W_l = \operatorname{diag}\Big(1, e^{\frac{\pi i}{l}}, \ldots, e^{k \cdot \frac{\pi i}{l}}, \ldots, e^{(l-1) \cdot \frac{\pi i}{l}}\Big)$$

Dieser rekursive Aufbau der Fourier-Matrix F_n gestattet es, die Rechnung sehr effizient durchzuführen, (FFT) weil dank der Faktorisierung F_{16} mit F_8 und rekursiv F_8 mit F_4 und schliesslich F_4 mit F_2 in sehr enger Beziehung steht. Der FFT-Algorithmus erlaubt es, Matrizenprodukte mit F_n nicht mit der üblichen Komplexität $\mathcal{O}(n^2)$ sondern mit der Komplexität $\mathcal{O}(n \cdot \log(n))$ zu berechnen. Diese enorme Effizienzsteigerung ist der Grund für den Erfolg moderner angewandter Gebiete wie etwa der Signaltechnik und schliesslich Grundlage für mehr als einen Drittel des globalen Sozialproduktes.

Generell stehen Zerlegungseigenschaften gewisser Matrizen sowohl vom theoretischen als auch vom praktischen Standpunkt her im Zentrum des algorithmischen Interesses. Beispielsweise dürfte in der näheren Zukunft die recht gute Parallelisierbarkeit der Matrizenrechnung, die sich in den Blockzerlegungen von Matrizen manifestiert, eine immer wichtigere Rolle spielen.

Weil wir im allgemeinen viele Gleichungen und viele Variablen betrachten müssen, benötigen wir eine kompakte und systematische Bezeichnungsweise für lineare Gleichungssysteme. Ein lineares Gleichungssystem von m Gleichungen in den n Unbekannten $x_1, x_2, \ldots, x_n$ schreiben wir als:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

Dabei werden die $x_1, x_2, \ldots, x_n$ auch als *Unbestimmte*, die a_{ij} als *Koeffizienten* und die b_i als *Konstanten* des Systems bezeichnet. Unter einer *Lösung* dieses Gleichungssystems verstehen wir n Zahlen $x_1, x_2, \ldots, x_n$, die *alle* Gleichungen des Systems erfüllen. Die Gesamtheit aller Lösungen heisst *Lösungsraum* des linearen Gleichungssystems. Die zentralen Fragen in der Theorie der linearen Gleichungssysteme sind die folgenden:

1. Hat ein gegebenes lineares Gleichungssystem *überhaupt eine Lösung*?
2. *Wieviele Lösungen* hat ein lineares Gleichungssystem?
3. Wie kann man alle Lösungen *konstruieren* und *überblicken*?

Bevor wir Antworten auf diese Fragen suchen, wollen wir lineare Gleichungssysteme mit Hilfe von Matrizen übersichtlicher darstellen. Um Schreibarbeit zu sparen, ersetzen wir ein lineares Gleichungssystemystem durch die *erweiterte*

Matrix des Systems. Die erweiterte Matrix von obigem linearen Gleichungssystem besteht aus seinen Koeffizienten und Konstanten, die wir zum rechteckigen Schema

$$(A \mid \vec{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right)$$

zusammenfassen. Beim Aufstellen der erweiterten Matrix müssen die Unbekannten in allen Gleichungen in der selben Reihenfolge auftreten. Gegebenenfalls muss man mit einer 0 andeuten, dass eine Variable nicht vorkommt. Im linken Block A der erweiterten Matrix steht die *Koeffizientenmatrix* und im rechten Block $\vec{b}$ steht der *Konstantenvektor*. In unserem Beispiel lauten sie also

$$A = \left(\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array}\right), \qquad \vec{b} = \left(\begin{array}{c} b_1 \\ b_2 \\ \cdots \\ b_m \end{array}\right)$$

Die doppelte Indizierung der *Elemente* a_{ij} einer Matrix A erlaubt es, diese Komponenten innerhalb des rechteckigen Systems schnell zu lokalisieren — das Element a_{ij} liegt in der i-ten Zeile und in der j-ten Spalte. Der erste Index der Matrix a_{ij} gibt also an, in welcher Zeile der Koeffizient steht (Zeilenindex). Der zweite Index entspricht der Nummer der zugehörigen Spalte (Spaltenindex).

$$\left(\begin{array}{ccccc} 1 & — & \mathrm{j} & \rightarrow & n \\ & & \vdots & & \\ \mid & & & & \\ i & \cdots & a_{ij} & & \\ \downarrow & & & & \\ m & & & & \end{array}\right)$$

Im linearen Gleichungssystem entspricht der Zeilenindex der Gleichungsnummer und der Spaltenindex der Variablennummer. Da in unserer Matrix m Zeilen und n Spalten vorkommen, hat A den Typ $m \times n$. Die erweiterte Matrix $(A \mid \vec{b})$ hingegen hat den Typ $m \times (n+1)$. Die positionelle Schreibweise von linearen Gleichungssystemen wurde bereits 1693 von G.W. Leibniz in einem Brief verwendet.

Man beachte übrigens die aus dem Lateinischen stammenden Singular- und Plural-Formen. Jeder halbwegs gebildete Mensch redet in unserem Zusammenhang von einer Matrix und von vielen Matrizen oder ganz gebildet von Matrizes. Eine Matrize ist ein Begriff aus dem Druckwesen und hat mit unserem Gegenstand nichts zu tun! Ganz entsprechend werden wir es manchmal mit einem Index und notgedrungen öfter mit mehreren Indizes zu tun haben, obwohl wir nach Möglichkeit darauf verzichten, eine Indexschlacht zu veranstalten, sondern zu viele Indizes als Indizien dafür interpretieren, dass eine bessere mathematische Sprache für den betreffenden Sachverhalt gesucht werden sollte.

Wir wollen diese Begriffe an einer jener Textaufgaben demonstrieren, die in der Sekundarschule den Rechnungsunterricht dominiert haben.

Beispiel. In ein Gefäss mit einem Inhalt von V Volumeneinheiten mündet eine Zuleitung mit der u-fachen Leistung der wegführenden Ableitung. Falls die Ab-

leitung t Zeiteinheiten und die Zuleitung v-mal so lange offen sind, fliesst das Gefäss über. Wie gross sind die Leistungen[6] der beiden Leitungen?

Bevor er wie wild zu rechnen beginnt, sollte der Anwender sorgfältig ein vollständiges Modell des Problems ausarbeiten. Dazu wählt er im vorliegenden Fall etwa die Leistung des Abflusses x_1 und die Leistung des Zuflusses x_2 als Unbekannte und bilanziert nun diese beiden Grössen.

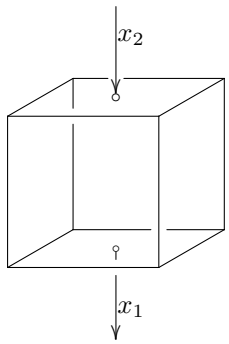


Abbildung 1.2: Das Füllen einer Badewanne.

Er beachtet, dass die Bezeichnungen für die Unbekannten willkürlich sind und welchen Einfluss eine andere Wahl auf die zu erwartende Lösung haben wird. Auch über die Rolle der Wahl der Orientierungen, die in der Figur durch die Pfeilspitzen der beiden Ströme angedeutet wird, ist er sich vollständig im Klaren.

Auf Grund der Aufgabenstellung erkennt er hier ein Gleichgewichtsproblem für die beiden Unbekannten x_1 und x_2 und erhält die beiden Gleichungen $x_2 = ux_1$ und $-tx_1+vtx_2 = V$, die er dann — geordnet — zum linearen Gleichungssystem

$$\left\{ \begin{array}{rcl} ux_1 - & x_2 = & 0 \\ -tx_1 + vtx_2 & = & V \end{array} \right.$$

zusammenfasst und allenfalls seine physikalischen Einheiten kontrolliert. Bisher war — bis auf das Ordnen — keine Mathematik im Spiel! Detailierte Kenntnisse des betreffenden Systems, die Definition der involvierten Grössen und ihrer Einheiten, die Formulierung der relevanten Beziehungen ist Sache des Anwenders — im vorliegenden Fall des Hydraulikers. Dieser Prozess heisst modellieren. Die Aufgabe des Anwenders ist es auch, die notwendigen physikalischen Voraussetzungen an die Parameter zu formulieren. Im vorliegenden Fall wird er etwa u, v, $V > 0$ verlangen. Ferner registriert er, dass er ein *lineares* Problem der Dimension 2 vor sich und damit eine gute Chance hat, es vollständig zu lösen.

In der Mathematik abstrahieren wir nun vom konkreten Problem und formulieren das Modell in Form der zugehörigen erweiterten Matrix

$$\left(\begin{array}{cc|c} u & -1 & 0 \\ -t & v & V \end{array} \right), \qquad \left[\begin{array}{c} Z_{12}(t,u) \\ \\ \end{array} \right]$$

Was diese Symbole bedeuten, spielt schon deshalb keine Rolle, weil verschiedene Anwendungen zum selben System führen können.

[6] Gewisse Physiklehrer reden auch vom Volumenstrom.

Um nun das erhaltene lineare Gleichungssystem zu lösen, wird man es in eine einfachere, aber gleichwertige Form überführen, aus der man die gesuchten Lösungen direkt ablesen kann. Addition des t-fachen der ersten Zeile zum u-fachen der zweiten ist wegen $u > 0$ unbedenklich, da umkehrbar, und hat den Effekt, dass die erste Unbekannte in der zweiten Zeile eliminiert wird. Dieser erste Eliminationsschritt liefert die neue Matrix

$$\left(\begin{array}{cc|c} u & -1 & 0 \\ 0 & (uv-1)t & uV \end{array}\right), \qquad \left[\begin{array}{c} \\ Z_{21}((uv-1)t) \end{array}\right]$$

Man beachte, dass nun dank der neu vorhandenen 0 in der zweiten Zeile dort die erste Unbekannte tatsächlich eliminiert ist. Um auch die zweite Unbekannte in der ersten Zeile noch zu eliminieren, wird nun die zweite Zeile zum $(uv-1)t$-fachen der ersten addiert[7] und man erhält das äquivalente System

$$\left(\begin{array}{cc|c} u(uv-1)t & 0 & uV \\ 0 & (uv-1)t & uV \end{array}\right), \qquad \left(\begin{array}{cc|c} 1 & 0 & \frac{V}{(uv-1)t} \\ 0 & 1 & \frac{uV}{(uv-1)t} \end{array}\right)$$

in dem sich die beiden Unbekannten nicht mehr in die Quere kommen, weil der linke Block dieser Matrix Diagonalgestalt hat. Division der beiden Zeilen durch das jeweilige führende Element liefert die sogn. normierte, reduzierte Stufenform, aus der sich nun die eindeutig bestimmte Lösung des ursprünglichen Systems leicht ablesen lässt. Es ist

$$x_1 = \frac{V}{(uv-1)t}, \qquad x_2 = \frac{uV}{(uv-1)t}.$$

Selbstverständlich muss sie nun durch Einsetzen in das ursprüngliche System kontrolliert werden. Im vorliegenden Beispiel ist die erste Gleichung $x_2 = ux_1$ sofort ersichtlich und die zweite muss nachgerechnet werden.

CAS. Rechnungen wie diese lassen sich mit Hilfe eines **C**omputer-**A**lgebra-**S**ystems leicht durchführen. Bei der Wahl von Maschinen und Programmiersprachen entbrennen unter Anwendern regelmässig heilige Kriege, die der Autor einfach nur belächeln kann, weil er weiss, dass sie alle Turing-vollständig, d.h. gleich mächtig sind. Für ihn sind in erster Linie wissenschaftliche Argumente ausschlaggebend. Wenn er schon Maschinen einsetzt, spielt deshalb für ihn keine Rolle, welches Logo nun darauf klebt und er legt beispielsweise auf möglichst grosse Zuverlässigkeit und Vielseitigkeit sowie auf eine möglichst übersichtliche, intuitive und gut dokumentierte Syntax einer Sprache wert.

Um den fundamentalen Unterschied zwischen symbolischer Mathematik und Numerik zu erkennen, möge der Leser sich zunächst überlegen, was die Auswertung des Ausdrucks

$$(10^{70} + 1) - 10^{70}$$

liefern sollte, bevor er ihn blind mit seinem Taschenrechner auswertet. Fehler solcher Art sind schwer zu entdecken und ob sie sich in einer längeren Rechnung als Zwischenergebnisse ergeben werden, ist theoretisch schwierig vorherzusehen. Deshalb werden in einem CAS-System keine ungefragten numerischen Approximationen verwendet, d.h. ein solches System macht (fast) immer nur exakte Umformungen. Weil Mathematiker in ihren Theorien keine “Lotterie”-Ergebnisse

[7] Wir werden gleich besprechen müssen, was es bedeutet, wenn dieser Faktor verschwindet!

brauchen können und wissen, dass Dezimalzahlen sich meistens anders verhalten als erwartet, machen sie einen grossen Bogen um die unzuverlässige Numerik und überlassen sie grosszügig den Anwendern. Denen können sie dann halt nur empfehlen, mindestens die Ergebnisse ihrer Bemühungen sehr kritisch zu hinterfragen, damit sie nicht ganz wertlos sind.

Wenn das einzige System, das momentan die meisten[8] mathematischen Wünsche befriedigt, erst noch frei zugänglich und quelloffen ist; zusätzlich von hunderten von Freiwilligen auf der ganzen Welt, die in diversen Diskussionforen auch für Fragen und Probleme rund um die Uhr zur Verfügung stehen, gepflegt und weiterentwickelt wird; es eng mit dem Textsatzsystem LaTeX zusammenspielt und erst noch viele der frei zugängliche Packete enthält, mit denen die unter Anwendern verbreiteten, aber auch für Studenten sündhaft teuren Systeme problemlos simmuliert werden können, weil es die Mission

> Creating a viable free open source alternative to Magma, Maple, Mathematica and Matlab.

hat, ist seine Entscheidung gefallen. Er empfiehlt einer Schule, viel Geld für die Lizenzen und den überflüssigen Informatik-Support einzusparen oder anderweitig — beispielsweise als Spende — einzusetzen. Einem Studenten empfiehlt er, seine Zeit während der Einführungswoche nicht mit Software-Installation zu vertrödeln, sondern gleich zur Sache zu kommen und sich nun auf der Web-Adresse umzusehen. Dort wird er sehen, dass Sage entweder sofort individuell als Taschenrechner oder nach einer formlosen Anmeldung — sogar für die sehr empfehlenswerte Zusammenarbeit mit Kommilitionen — in der Cloud benutzt werden kann. Nach Bedarf kann er später die neueste Version des vollständigen Programms immer noch auf seiner eigenen Maschine installieren. Im Web findet er buchstäblich tonnenweise Einführungen, ein empfehlenswertes Tutorial, ein umfangreiches Referenz-Handbuch, mehr oder weniger umfangreiche Lehrbücher — auch für die zugrundeliegende objektorientierte, dynamisch typisierte, weitgehend funktionale und erst noch leicht zu lernende Programmiersprache Python, die heute jeder Praktiker, der sein Salz wert ist, beherrschen sollte, weil bei genauerem Hinsehen auch bei ihm die Entwicklungszeit die Laufzeit eines Programms bei weitem überwiegt.

Für unser Beispiel gibt man etwa folgenden Kode

```
var("V, u, v, t")
b=vector([0,V])
A=matrix([ [u,-1], [-t, v*t] ])
x=A.solve_right(b)
x1=x[0].simplify_rational().factor(); show(x1)
x2=x[1].simplify_rational().factor(); show(x2)
```

[8] Es sei nicht verschwiegen, dass auch CAS-Systeme ihre Tücken haben. Um sie zu erkennen, möge der Leser sich zunächst überlegen, was die Lösungsmenge der linearen Gleichung

$$a \cdot x = a$$

für eine fest gewählte, aber beliebige reelle Zahl a ist und dann einen Kommilitonen bzw. ein CAS-System befragen. Beide vereinfachen den Bruch $\frac{a}{a}$ automatisch zu 1 und berücksichtigen spezielle Werte nur selten. Genau solche Sonderfälle spielen aber in der Mathematik eine wichtige Rolle, wie wir bereits in der linearen Algebra im Zusammenhang mit Parametern erkennen werden. Abhilfe, d.h. eine vollständige Liste aller Sonderfälle ist höchstens von den automatischen Theorembeweisern zu erwarten.

und drückt dann die Taste "Evaluate", um schliesslich den "Lösungsvektor" x mit den von uns berechneten Komponenten x_1 und x_2 abzulesen. Dieser Kode bewirkt:

- Zunächst deklarieren wir die involvierten Parameter.
- Dann definieren wir einen Vektor b mit Hilfe einer Liste[9].
- Dann definieren wir eine Matrix A mit Hilfe einer Liste von Listen.
- Dann erteilen wir dem Programm drei Befehle:
  - Der erste berechnet den Lösungsvektor x.
  - Der zweite und der dritte extrahieren aus dem Lösungsvektor die erste und die zweite Komponente, vereinfachen die entstandenen Brüche und zeigen die Lösungsformeln x_1 und x_2 auf dem Bildschirm an.

Man beachte, dass in Sage mit 0 und nicht wie traditionell in der Matrizenrechnung mit 1 zu zählen begonnen wird!

Das Programm benutzt den von uns verwendeten Algorithmus, der auf der Idee basiert, ein lineares Gleichungssystem solange durch ein einfacheres mit der selben Lösungsmenge zu ersetzen, wie dies möglich ist.

Auch die normierte, reduzierte Stufenform der erweiterten Matrix $B = (A \mid b)$ lässt sich mit Hilfe von Sage einfach berechnen, indem wir die beiden Befehle

```
B=A.augment(b)
C=B.rref().simplify_rational().factor(); show(C)
```

verwenden. Mit dem ersten wird aus A und b die erweiterte Matrix $B = (A \mid b)$ zusammengestellt und mit dem zweiten deren vereinfachte normierte, reduzierte Stufenform berechnet. Wir werden bald sehen, dass die normierte, reduzierte Stufenform einer Matrix noch viel mehr interessante und brauchbare Information liefert, als bloss die Lösung eines linearen Gleichungssystems. Deshalb steht sie und nicht so sehr die Lösung des Gleichungssystems, von Anfang an im Zentrum unseres Interesses. Sie wird unser rechnerisches Arbeitspferd sein und wir werden den Eliminationsalgorithmus im Laufe der Zeit genauer untersuchen.

Selbstverständlich können wir auch die Kontrolle der vermuteten Lösung Sage überlassen indem wir zusätzlich die Befehle

```
ls1=u*x1-x2; show(ls1)
ls2=(-t*x1+v*t*x2).combine().factor(); show(ls2)
```

benutzen. Wir setzen also die vermeintlichen Lösungen x_1 und x_2 in die linken Seiten der beiden Gleichungen des Systems ein und vereinfachen die entstehenden Brüche. ◇

Nun ist es an der Zeit, die gefundenen Lösungen unserer Aufgabe genauer zu betrachten und zu diskutieren. Zunächst beachten wir, dass die beiden Lösungsformeln fast gleich aussehen und insbesondere ihre Nenner, die das Lösbarkeitskriterium beinhalten, übereinstimmen. Dieses symmetrische Verhalten der Lösung und die Einheitenprobe machen das Resultat für den Anwender zusätzlich plausibel.

[9] Wie in jeder für Mathematik optimierten Programmiersprache bilden Listen und nicht etwa Zahlen in Sage den fundamentalen Datentyp.

Bei der Diskussion der Lösung stellt man fest, dass die gefundene Lösung unbrauchbar ist, falls der gemeinsame Nenner der beiden Brüche, die sogn. Determinante

$$\det(A) = (uv - 1)t = 0,$$

verschwindet, weil dann die Division nicht durchgeführt werden kann. Dieser Spezialfall liegt dann vor, wenn $t = 0$ oder $uv = 1$ ist. Die erweiterte Matrix nimmt dann nach dem ersten Eliminationsschritt die spezielle Form

$$\left(\begin{array}{cc|c} u & -1 & 0 \\ 0 & 0 & V \end{array}\right)$$

an, aus deren zweiten Zeile klar wird, dass dieses System keine Lösung haben wird, falls $V > 0$ ist, weil dann ein Widerspruch vorliegt.

Den ersten singulären Fall $t = 0$ müssen wir aus physikalischen Gründen nicht weiter verfolgen. Er widerspricht der Kausalität, d.h. dem physikalischen Umstand, dass gleichzeitig nicht zwei unterschiedliche Verhältnisse herrschen können.

Der zweite singulären Fall $u \cdot v = 1$ entspricht physikalisch der Situation, dass dem Gefäss gleich viel zu- wie abfliesst und es daher gar nie voll werden kann. Offensichtlich ist in diesem Fall das Problem schlecht gestellt. Solche Sonderfälle muss der Informatiker in seinen Programmen abfangen, damit sie nicht bei einer Division durch 0 abstürzen; den Benutzer muss er auf diesen singulären Fall $u \cdot v = 1$ mit Hilfe einer Fehlermeldung besonders hinweisen. Der Ingenieur muss solche Sonderfälle genau kennen, weil dann sein System ein unerwünschtes Verhalten aufweisen wird: es kommt zu einem Kurzschluss oder zu einem Kollaps der beabsichtigten Konstruktion. Leider weisst ihn ein CAS-System wie Sage auf solche Sonderfälle nicht hin und er sollte besser den Kopf bei der Sache haben!

Aus diesen Beobachtungen ergibt sich die verblüffende Möglichkeit, das vorliegende Problem ohne viel Rechnerei durch kluges Raten zu lösen — eine Möglichkeit, die in der modernen Physik, wo die Systeme oft viel zu gross sind, um sie rechnerisch zu lösen, eine zentrale Rolle spielt und die unter Physikern als Feynman-Methode bekannt ist und wie folgt charakterisiert wird:

> You write down the problem. Then you look at it and you think.
> Then you write down the answer.

Auch im Normalfall $u \cdot v \neq 1$ ist in der numerischen Praxis nicht alles gleich rosig. Um eine verlangte Genauigkeit des Resultates garantieren zu können, wird man in der Nähe der gefährlichen Bedingung $u \cdot v = 1$ mit unterschiedlich grosser Präzision rechnen müssen, um diese Genauigkeit zu erreichen, weil dann das Problem schlecht konditioniert ist. Dazu muss man allerdings verstehen, wie genau ein Problem in Abhängigkeit von den Parametern konditioniert ist.

Die numerischen Lösungen linearer Gleichungssysteme $(A \mid \vec{b})$ mit quadratischer Koeffizientenmatrix A können, auch wenn sie existieren, sehr empfindlich von den Daten abhängen und zu vollständig unbrauchbaren Resultaten führen, wenn man diese Abhängigkeit ignoriert. Für die Qualität des numerischen Lösungsverhaltens eines linearen Gleichungssystems ist die *Konditionszahl* $\kappa(A)$ der Koeffizientenmatrix verantwortlich. Darunter versteht man die Zahl, die durch

$$\kappa(A) = |A| \cdot |A^{-1}|$$

definiert ist, wobei A^{-1} die inverse Matrix von A und

$$|A|^2 = \max_{\vec{x}\neq\vec{0}} \frac{|A\cdot\vec{x}|^2}{|\vec{x}|^2} = \lambda_{\max}(A^T\cdot A)$$

das sogn. Normquadrat[10] bezeichnet. Es kann mit Hilfe des sogn. betragsgrössten Eigenwertes der symmetrischen Matrix $A^T\cdot A$ berechnet werden.

Im Fall einer symmetrischen Matrix A ist die Norm $|A| = \lambda_{\max}(A)$ gerade der betragsgrösste Eigenwert und die Konditionszahl wird einfach zum Verhältnis

$$\kappa(A) = |A|\cdot|A^{-1}| = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

des betragsgrössten zum betragskleinsten Eigenwert von A. Zur Untersuchung der Abhängigkeit der numerischen Ergebnisse von den Daten braucht man also die ganze Maschinerie der Matrizenrechnung.

In unserem Beispiel ist für $t = 1$ die Koeffizientenmatrix

$$A = \begin{pmatrix} u & -1 \\ -1 & v \end{pmatrix}$$

symmetrisch. Ihr charakteristisches Polynom

$$\chi_A(\lambda) = \lambda^2 - (u+v)\lambda + (uv-1)$$

hat die beiden Eigenwerte

$$\lambda_1 = \frac{u+v+\sqrt{4+(u-v)^2}}{2}, \qquad \lambda_2 = \frac{u+v-\sqrt{4+(u-v)^2}}{2}$$

als Nullstellen. Damit ist hier die Konditionszahl

$$\kappa(A) = \frac{\lambda_1}{\lambda_2} = \frac{u+v+\sqrt{4+(u-v)^2}}{u+v-\sqrt{4+(u-v)^2}}$$

Sie wächst über alle Grenzen, wenn die singuläre Bedingung $u\cdot v = 1$ erfüllt ist. In der Nähe der Singularität hat diese Funktion ein relativ kompliziertes Verhalten, wie ihre Niveaulinien und ihr Graph in folgender Figur zeigen.

Beim numerischen Lösen eines linearen Gleichungssystems kann man auf Grund von Rundungsfehlern etwa $\log_{10}(\kappa(A))$ Dezimalstellen verlieren.

CAS. Diese Informationen lassen sich mit folgendem Sage-Kode erhalten:

```
var("u, v")
A=matrix([ [u,-1],[-1,v] ]); show(A)
p=characteristic_polynomial(A);show(p)
eig=A.eigenvalues();show(eig)
kondition(u,v)=eig[1]/eig[0];show(kondition)
contour_plot(kondition(u,v),(u,0,4),(v,0,4),labels=False,
```

[10] Für einen Vektor $\vec{x}\in\mathbb{R}^n$ bzw. eine quadratische Matrix $A\in\mathbb{R}^{n,n}$ bezeichnet

$$|\vec{x}| = \max|x_i|, \qquad \text{bzw.} \qquad |A| = \max_{|\vec{x}|=1}|A\cdot\vec{x}|$$

hier die sogn. ∞-Norm.

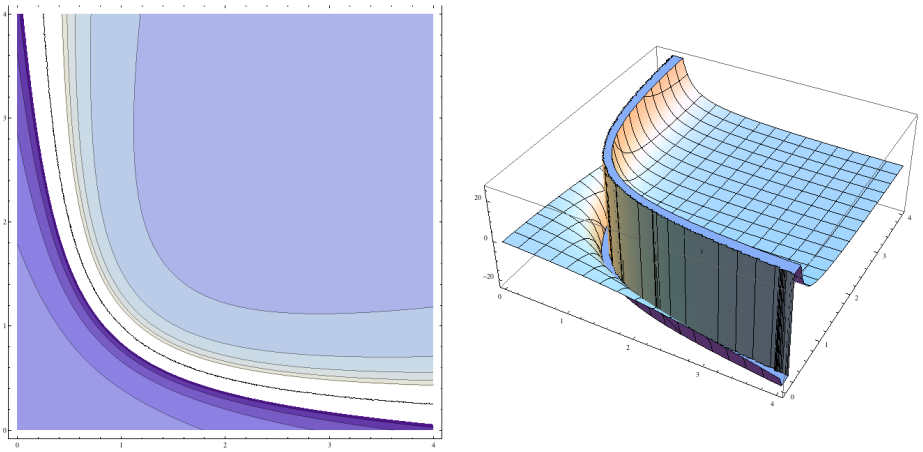


Abbildung 1.3: Die Konditionszahl $\kappa(A)$ der Matrix A.

```
plot_points=900,contours=[-30,-20,-10,-5,-3,3,5,10,20,30],
colorbar=True)
```

Man beachte, dass die Funktion $\kappa(A)$ nicht linear ist. Dieser Umstand weist schon darauf hin, dass die Numerik auch der linearen Algebra alles andere als erfreulich ist und schnell in einen schwer durchschaubaren Dschungel führt. $\diamond$

Das nächste Beispiel soll — genau so wie das simple Beispiel

$$1.1 + 1.1 + 1.1 == 3.3$$

noch einmal den Unterschied zwischen exakter Mathematik und Numerik an Hand der linearen Algebra verdeutlichen.

Beispiel. Wir gehen von der symmetrischen Matrix $A \in \mathbb{R}^{n,n}$ und den konstanten Vektoren $\vec{d}, \vec{b} \in \mathbb{R}^n$ aus, die für $n = 8$ folgendermassen gegeben sind:

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} & \frac{1}{12} \\ \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} & \frac{1}{12} & \frac{1}{13} \\ \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} & \frac{1}{12} & \frac{1}{13} & \frac{1}{14} \\ \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} & \frac{1}{12} & \frac{1}{13} & \frac{1}{14} & \frac{1}{15} \end{pmatrix}, \quad \vec{d} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} \frac{761}{280} \\ \frac{4609}{2520} \\ \frac{3601}{2520} \\ \frac{32891}{27720} \\ \frac{28271}{27720} \\ \frac{323171}{360360} \\ \frac{288851}{360360} \\ \frac{52279}{72072} \end{pmatrix}$$

Weil der Vektor $\vec{b} = A \cdot \vec{d}$ durch Multiplikation mit A erklärt ist, hat das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ die exakte Lösung $\vec{x} = \vec{d}$.

Fassen wir nun diese Daten als Matrizen reeller Zahlen $A_{\mathbb{R}}$, $\vec{d}_{\mathbb{R}}$ und $\vec{b}_{\mathbb{R}}$ auf und lösen dann das entstehende lineare Gleichungssystem $A_{\mathbb{R}} \cdot \vec{x}_{\mathbb{R}} = \vec{b}_{\mathbb{R}}$ numerisch, erhalten wir einen numerischen "Lösungsvektor" $\vec{x}_{\mathbb{R}}$, der sich vom exakten

Lösungsvektor $\vec{x}$ um den "Fehler"-Vektor"

$$\vec{f} = \vec{x} - \vec{x}_{\mathbb{R}}$$

unterscheidet, dessen Norm $|\vec{f}| = 6.82 \cdot 10^{-7}$ noch ganz tragbar erscheint. Die Konditionszahl dieser Matrix A beträgt $\kappa(A) = 3.3 \cdot 10^{10}$, wie man anhand des zugehörigen Kodes erkennt.

Bereits für $n = 16$ beträgt die Konditionszahl der entsprechenden Matrix allerdings $\kappa(A) = 5.1 \cdot 10^{22}$ und die Norm des "Fehler"-Vektors ist dann $|\vec{f}| = 8.34$. In diesem Fall enthält also der numerische "Lösungsvektor" $\vec{x}_{\mathbb{R}}$ Komponenten, von denen keine einzige Dezimalstelle korrekt ist!

Für $n = 32$ beträgt die Konditionszahl $\kappa(A) = 1.3 \cdot 10^{47}$ und die Norm des "Fehler"-Vektors ist nun $|\vec{f}| = 257.66$. Die Konditionszahl $\kappa(A)$ dieser Matrizen wächst mit ihrer Grösse n exponentiell wie $e^{\frac{7n}{2}}$. In diesem Beispiel sind also keine "Messfehler" der Daten involviert. Nur die numerische Approximation und die schwer durchschaubare "Arithmetik" der Dezimalzahlen führen zu Lotterie.

Wir werden numerische Probleme deshalb in Zukunft ausser acht lassen und vorwiegend mit exakten Zahlen rechnen, weil wir schon jetzt erkennen, dass für ein gewisses Verständnis der durch die Numerik in die linearen Algebra eingeschleppten "Fehler" bereits die ganze Theorie erforderlich ist. Abgesehen davon gibt es sehr effiziente Algorithmen für die lineare Algebra über $\mathbb{Q}$, so dass wir dadurch nicht viel an Effizienz einbüssen.

Die Daten einer Anwendung sind allerdings meistens gar nicht exakt gegeben, sondern höchstens approximativ bekannt. Dort handelt es sich meistens um Messwerte oder um Approximationen nichtlinearer Gleichungs- oder Differentialgleichungssysteme $\vec{y}' = f(\vec{y})$, die aus der Analysis stammen und daher mit Hilfe einer Funktional-Matrix $\partial(f)$ formuliert werden müssen, die in der Regel schlecht konditioniert ist. Daher wird sich ein Anwender schon irgend einmal intensiv mit dieser Problematik befassen müssen, wenn er möchte, dass die Ergebnisse seiner Rechnungen Relevanz haben. ◯

Ein Benutzer wird nun den Parametern erlaubte numerische Werte zuweisen. Im vorliegenden Fall misst er etwa $u = v = 2$, $t = 7$ und $V = 21$. Damit hat das lineare Gleichungssystem die numerische Form

$$\begin{cases} 2x_1 - x_2 = 0 \\ -7x_1 + 14x_2 = 21 \end{cases}$$

Die zugehörige erweiterte Matrix lautet dann

$$B = \left(\begin{array}{cc|c} 2 & -1 & 0 \\ -7 & 14 & 21 \end{array}\right), \qquad \left(\begin{array}{cc|c} 2 & -1 & 0 \\ -1 & 2 & 3 \end{array}\right), \qquad \left(\begin{array}{cc|c} -1 & 2 & 3 \\ 2 & -1 & 0 \end{array}\right)$$

Um mit betragsmässig möglichst kleinen ganzen Zahlen weiterrechnen zu können, haben wir zunächst die zweite Zeile durch ihren grössten gemeinsamen Teiler $\mathrm{ggT}(-7, 14, 21) = 7$ geteilt. Man beachte, dass die zugehörige primitive Gleichung $-x_1 + 2x_2 = 3$ die selbe Lösungsmenge wie die ursprüngliche hat. Dann vertauschen wir die beiden Zeilen, wobei sich auch dabei die Lösungsmenge nicht verändert. Schliesslich addieren wir das 2-fache der ersten Zeile zur zweiten, was

wiederum die Lösungmenge nicht verändert, aber in der zweiten Zeile die erste Variable eliminiert.

$$\left(\begin{array}{cc|c} -1 & 2 & 3 \\ 0 & 3 & 6 \end{array}\right), \quad \left(\begin{array}{cc|c} -1 & 2 & 3 \\ 0 & 1 & 2 \end{array}\right), \quad \left(\begin{array}{cc|c} -1 & 0 & -1 \\ 0 & 1 & 2 \end{array}\right), \quad \left(\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \end{array}\right)$$

Um wiederum mit möglichst kleinen Zahlen weiterrechnen zu können, haben wir nun die zweite Zeile durch 3 dividiert. Schliesslich haben wir das (-2)-fache der zweiten Zeile zur ersten addiert, um in der ersten Zeile die zweite Variable zu eliminieren. Schliesslich haben wir im letzten Schritt noch die erste Zeile normiert, indem wir sie mit (-1) multipliziert haben. Aus der damit entstandenen sogn. normierten reduzierten Stufenform lesen wir nun als Lösung unseres Problems

$$x_1 = 1, \qquad x_2 = 2$$

ab. Selbstverständlich lässt sie sich durch Einsetzen in das ursprüngliche Gleichungssystem kontrollieren.

CAS. Die soeben durchgeführten Umformungen der erweiterten Matrix B lassen sich mit Hilfe des folgenden Sage-Kodes realisieren:

```
b=vector([0,21])
A=matrix([ [2,-1],[-7,14] ])
B=A.augment(b); show(B)
B1 = B.with_rescaled_row(1,1/7); show(B1)
B2 = B1.with_swapped_rows(0,1); show(B2)
B3 = B2.with_added_multiple_of_row(1,0,2); show(B3)
B4 = B3.with_rescaled_row(1,1/3); show(B4)
B5 = B4.with_added_multiple_of_row(0,1,-2); show(B5)
B6 = B5.with_rescaled_row(0,-1); show(B6)
```

Für eine Matrix A werden also die Elementaroperationen durch folgende Befehle erzeugt:

I. Vertauschen der i-ten und der j-ten Zeile.

```
A.with_swapped_rows(j-1,i-1)
```

II. Addition des r-fachen der i-ten Zeile zur j-ten Zeile.

```
A.with_added_multiple_of_row(j-1,i-1,r)
```

III. Multiplikation der i-ten Zeile mit dem Faktor r.

```
A.with_rescaled_row(i-1,r)
```

Diese Elementaropertionen lassen sich selbstverständlich auch zusammensetzen. So erhalten wir beispielsweise folgende zusammengesetzte Operation, die wir regelmässig benutzen werden.

IV. Addition des r-fachen der i-ten zum s-fachen der j-ten Zeile.

```
A.with_rescaled_row(j-1,s).with_added_multiple_of_row(j-1,i-1,r)
```

Addieren wir etwa in der Matrix B_1 die erste Zeile zum 2-fachen der zweiten Zeile, erhalten wir eine Matrix, in der die erste Variable eliminiert ist.

$$B_1 = \left(\begin{array}{cc|c} 2 & -1 & 0 \\ -1 & 2 & 3 \end{array}\right), \qquad \left(\begin{array}{cc|c} 2 & -1 & 0 \\ 0 & 3 & 6 \end{array}\right), \qquad \left(\begin{array}{cc|c} 2 & -1 & 0 \\ 0 & 1 & 2 \end{array}\right)$$

Wiederum dividieren wir nun die zweite Zeile durch 3, um mit betragsmässig möglichst kleinen ganzen Zahlen weiterrechnen zu können. Als nächstes addieren wir die zweite Zeile zur ersten, um die zweite Variable in der ersten Zeile zu eliminieren. Schliesslich normieren wir die erste Zeile, indem wir sie durch 2 dividieren.

$$\left(\begin{array}{cc|c} 2 & 0 & 2 \\ 0 & 1 & 2 \end{array}\right), \qquad \left(\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \end{array}\right)$$

Offensichtlich liefert dieser zweite Weg am Schluss die selbe normierte reduzierte Stufenform.

Der Kode des beschriebenen zweiten Weges sieht folgendermassen aus:

```
b=vector([0,21])
A=matrix([ [2,-1],[-7,14] ])
B=A.augment(b); show(B)
B1 = B.with_rescaled_row(1,1/7); show(B1)
B2 = B1.with_rescaled_row(1,2).with_added_multiple_of_row(1,0,1); show(B2)
B3 = B2.with_rescaled_row(1,1/3); show(B3)
B4 = B3.with_added_multiple_of_row(0,1,1); show(B4)
B5 = B4.with_rescaled_row(0,1/2); show(B5)
```

Selbstverständlich liefern beide Wege die selbe Lösung des linearen Gleichungssystems. ◇

Dass $x_1 = 1$ und $x_2 = 2$ tatsächlich das ursprüngliche lineare Gleichungssystem lösen, lasst sich noch auf eine andere Art einsehen, auf die wir nun unser Augenmerk richten wollen. Definieren wir nämlich die Spaltenvektoren bzw. den Konstantevektor der Matrix wie folgt,

$$\vec{a}_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \vec{a}_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

erkennen wir folgende Zerlegung von $\vec{b}$

$$x_1\vec{a}_1 + x_2\vec{a}_2 = \vec{b}$$

als Linearkombination von $\vec{a}_1$ und $\vec{a}_2$, aus der ebenfalls folgt, dass x_1 und x_2 Lösungen des linearen Gleichungssystems sind.

CAS. Die Rechnung lässt sich mit Hilfe des folgenden Sage-Kodes realisieren:

```
a1=A.column(0)
a2=A.column(1)
x1*a1+x2*a2==b
```

Linearkombinationen können in Sage wie im skalaren Fall berechnet werden.

Gelegentlich möchte man umgekehrt aus einer Liste von Vektoren eine Matrix mit diesen Vektoren als Spaltenvektoren machen. Das erreicht man im Beispiel mit dem Befehl

```
D=column_matrix([a1,a2,b])
```

Dass D die ursprüngliche erweiterte Matrix B ist, kontrolliert man mit dem Vergleich `B==D`. Weil Sage Zeilen vor Spalten bevorzugt, erhält man die Matrix mit diesen Zeilen durch den kürzeren Befehl `matrix([a1,a2,b])`. ◇

Wichtig für die lineare Algebra und alle ihre Anwendungen ist der Umstand, dass sich ein lineares System geometrisch auf *zwei* unterschiedliche Arten interpretieren lässt. Diese Dualität zieht sich wie ein roter Faden durch die ganze Theorie. Der Autor muss jedoch zerknirscht zugeben, dass ihn die Rolle der Dualität in der linearen Algebra und ihren Anwendungen in Mathematik und Physik, d.h. der Umstand dass die Kategorie $\mathbf{FinVect}_{\mathbb{R}}$ der endlich dimensionalen Vektorräume selbstdual ist, gelegentlich irritiert oder gar verwirrt.

In der *zeilenweisen* Interpretation des linearen Gleichungssystems fassen wir die Zeilen des Systems als Gleichungen zur Beschreibung von zwei Geraden auf. Das Problem besteht in der zeilenwewisen Interpretation darin, den *Schnittpunkt* dieser beiden Geraden zu bestimmen.

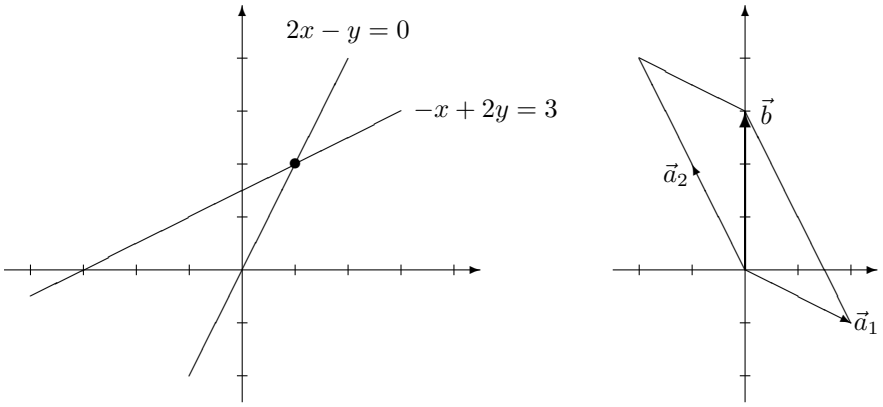


Abbildung 1.4: Zeilenweise und Spaltenweise Interpretation des selben linearen Gleichungssystems.

In der *spaltenweisen* Interpretation des linearen Gleichungssystems fassen wir die Spalten des Systems als Vektoren auf. Das Problem besteht in der spaltenweisen Interpretation darin, eine *Linearkombination*

$$x_1\vec{a}_1 + x_2\vec{a}_2 = \vec{b}$$

des Vektors $\vec{b}$ als Resultierende von $\vec{a}_1$ und $\vec{a}_2$ zu finden. Sie lässt sich mit Hilfe der sogn. Parallelogrammkonstruktion auch graphisch mit Hilfe eines Lineals konstruieren.

Aus den beiden Figuren lesen wir für das vorliegende lineare Gleichungssystem in beiden Interpretationen die eindeutig bestimmte Lösung $x_1 = 1, x_2 = 2$ ab, die wir bereits durch Einsetzen in das ursprüngliche Gleichungssystem algebraisch kontrolliert haben. ◯

Beispiel. In einem Otto-Motor wird n-Octan (C_8H_{18}) unter Sauerstoffzufuhr (O_2) zu Kohlendioxid (CO_2) und Wasser (H_2O) verbrannt, d.h. es läuft eine chemische Reaktion der Art

$$x_1\ C_8H_{18} + x_2\ O_2 \to x_3\ CO_2 + x_4\ H_2O$$

ab. Um die optimale Mischung dieser 4 Moleküle zu finden, müssen die involierten 3 Atomsorten so bilanziert werden, dass auf beiden Seiten des Reaktionspfeils die selbe Anzahl Atome jeder Sorte vorhanden sind. Das führt zu folgenden Gleichungen:

$$\begin{array}{ll} \text{Kohlenstoff } (C): & 8x_1 = x_3 \\ \text{Wasserstoff } (H): & 18x_1 = 2x_4 \\ \text{Sauerstoff } (O): & 2x_2 = 2x_3 + x_4 \end{array}$$

Durch Ordnen dieser (linearen!) Bedingungen erhalten wir das lineare Gleichungssystem

$$\left\{ \begin{array}{rrrrr} 8x_1 & & - & x_3 & = 0 \\ 18x_1 & & & - 2x_4 & = 0 \\ & 2x_2 & - 2x_3 & - x_4 & = 0 \end{array} \right.$$

mit 3 homogenen Gleichungen und 4 Unbestimmten und der (relativ dünn besetzten) erweiterten Matrix

$$\left(\begin{array}{rrrr|r} 8 & 0 & -1 & 0 & 0 \\ 18 & 0 & 0 & -2 & 0 \\ 0 & 2 & -2 & -1 & 0 \end{array} \right)$$

Im vorliegenden Beispiel sieht man eine Lösung „vom Schiff aus“ — die triviale Lösung $x_1 = x_2 = x_3 = x_4 = 0$, die aber für Chemiker ohne Interesse ist. Sie interessieren sich für nichttriviale Lösungen des Systems. Auf Grund des Gesetzes der multiplen Proportion, wonach in allen chemischen Verbindungen gleiche Elemente in einem ganzzahligen Verhältnis stehen, interessieren sie sich für eine nichttriviale Lösung des Systems mit möglichst kleinen natürlichen Zahlen.

Um einen Überblick über sämtliche rationalen Lösungen des Systems zu erhalten, formen wir es in eine äquivalente, aber einfachere Form um. Um mit möglichst kleinen ganzen Zahlen rechnen zu können, kürzen wir zunächst die zweite Zeile durch ihren grössten gemeinsamen Teiler 2 und erhalten die Matrix

$$\left(\begin{array}{rrrr|r} 8 & 0 & -1 & 0 & 0 \\ 9 & 0 & 0 & -1 & 0 \\ 0 & 2 & -2 & -1 & 0 \end{array} \right)$$

Um nun in der zweiten Zeile die erste Unbestimmte x_1 zu eliminieren, addieren wir das 9-fache der ersten Zeile zum (-8)-fachen der zweiten und erhalten

$$\left(\begin{array}{rrrr|r} 8 & 0 & -1 & 0 & 0 \\ 0 & 0 & -9 & 8 & 0 \\ 0 & 2 & -2 & -1 & 0 \end{array} \right)$$

Weil in den ersten beiden Zeilen nun die Unbestimmte x_2 gar nicht vorkommt, d.h. bereits eliminiert ist, vertauschen wir die zweite und die dritte Zeile.

$$\left(\begin{array}{rrrr|r} 8 & 0 & -1 & 0 & 0 \\ 0 & 2 & -2 & -1 & 0 \\ 0 & 0 & -9 & 8 & 0 \end{array} \right)$$

Diese Matrix hat bereits sogn. Stufenform an Hand der wir erkennen, dass wir offenbar die Variable $x_4 = t \in \mathbb{R}$ frei wählen können.

Um mit dieser Wahl für x_4 dann die restlichen 3 Unbestimmten x_1, x_2, x_3 eindeutig ausdrücken zu können, addieren wir das (-2)-fache der dritten Zeile zum 9-fachen der zweiten und die dritte Zeile zum (-9)-fachen der ersten. Damit erhalten wir eine Matrix in sogn. reduzierter Stufenform (rref)

$$\left(\begin{array}{cccc|c} -72 & 0 & 0 & 8 & 0 \\ 0 & 18 & 0 & -25 & 0 \\ 0 & 0 & -9 & 8 & 0 \end{array}\right)$$

Um die gesuchten restlichen 3 Unbestimmten möglichst einfach bestimmen zu können, normieren wir nun die einzelnen Zeilen, indem wir die erste durch (-72) die zweite durch 18 und die dritte durch (-9) dividieren. Schliesslich erhalten wir die sogn. normierte, reduzierte Stufenform

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & -\frac{1}{9} & 0 \\ 0 & 1 & 0 & -\frac{25}{18} & 0 \\ 0 & 0 & 1 & -\frac{8}{9} & 0 \end{array}\right)$$

Offensichtlich ist nun die vierte Variable frei wählbar und nach dieser Wahl sind die anderen drei Variablen eindeutig festgelegt. Damit gilt für die gesuchte allgemeine Lösung offensichtlich

$$x_4 = t, \qquad x_1 = \frac{1}{9}t, \qquad x_2 = \frac{25}{18}t, \qquad x_3 = \frac{8}{9}t, \qquad t \in \mathbb{R}$$

wie man durch Einsetzen leicht kontrolliert. Diese Lösung lässt sich übersichtlich in vektorieller Form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = t \begin{pmatrix} \frac{1}{9} \\ \frac{25}{18} \\ \frac{8}{9} \\ 1 \end{pmatrix}, \qquad t \in \mathbb{R}$$

darstellen. Jede beliebige Wahl des Parameters $t \in \mathbb{R}$ liefert also eine Lösung des linearen Gleichungssysystems. In der Regel wird diese Lösung jedoch aus Brüchen bestehen, die für die beabsichtigte Anwendung unbrauchbar sind. Weil das betrachtete Gleichungssystem homogen ist, sind Vielfache von Lösungen wieder Lösungen und wir dürfen den gefundenen Vektor mit einem beliebigen Skalar multiplizieren. Um alle ganzzahligen Lösungen zu erhalten, multiplizieren wir den gefundenen Vektor mit dem kleinsten gemeinsamen Vielfachen der Nenner der gefundenen Brüche, d.h. mit $\mathrm{kgV}(9, 18, 9) = 18$ und können die ganzzahligen Lösungen des Systems in der Form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \tilde{t} \begin{pmatrix} 2 \\ 25 \\ 16 \\ 18 \end{pmatrix}, \qquad \tilde{t} \in \mathbb{Z}$$

parametrisieren. Jede Wahl von $\tilde{t} \in \mathbb{Z}$ liefert nun eine ganzzahlige Lösung und umgekehrt lassen sich alle ganzzahligen Lösungen so erhalten. Um die kleinste

Lösung mit lauter natürlichen Zahlen zu erhalten, brauchen wir nur $\tilde{t} = 1$ zu wählen und erhalten für die gesuchte Lösung schliesslich

$$x_1 = 2, \qquad x_2 = 25, \qquad x_3 = 16, \qquad x_4 = 18$$

Damit lautet die gesuchte Reaktionsgleichung für die Verbrennung von n-Octan

$$2\ C_8H_{18} + 25\ O_2 \to 16\ CO_2 + 18\ H_2O$$

Sie wäre durch blindes Pröbeln auch zu finden gewesen, weil das vorliegende System spärlich besitzt ist, d.h. relativ viele 0 in der Koeffizientenmatrix hat.

CAS. Die rationale normierte, reduzierte Stufenform dieses Beispiels erhält man in gewohnter Manier:

```
b=vector([0,0,0])
A=matrix([ [8,0,-1,0],[18,0,0,-2],[0,2,-2,-1] ])
B=A.augment(b);show(B)
show(B.rref())
```

Daraus lässt sich dann die angegebene rationale Lösung mühelos ablesen. Der Befehl `A.solve_right(b)` liefert für homogene Gleichungssysteme nur die uninteressante triviale Lösung und ist also nicht sehr lehrreich. Die nichttriviale, ganzzahlige Lösung eines homogenen Systems erhält man mit dem Befehl

```
A.right_kernel()
```

Sage behandelt also die Daten dieses Beispiels als exakte ganze Zahlen und nicht als durch Dezimalbrüche angenäherte reelle Zahlen. ◇

In der zeilenweisen Interpretation dieses Systems haben wir diesmal 3 Hyperebenen in $\mathbb{R}^4$ durch den Ursprung miteinander geschnitten. Die gefundene Lösung beschreibt geometrisch ihre 1-dimensionale Schnittgerade, die selbstverständlich auch durch den Ursprung gehen muss. In der spaltenweisen Interpretation haben wir die homogene Vektorgleichung

$$x_1\vec{a}_1 + x_2\vec{a}_2 + x_3\vec{a}_3 + x_4\vec{a}_4 = \vec{0}$$

der Spaltenvektoren der erweiterten Matrix untersucht und dabei festgestellt, dass sie unendlich viele Lösungen hat. Die für die chemische Reaktion relevante ganzzahlige Lösung besagt etwa, dass die Vektorgleichung

$$2\vec{a}_1 + 25\vec{a}_2 + 16\vec{a}_3 + 18\vec{a}_4 = \vec{0}$$

mit vier nicht trivialen Koeffizienten gilt. Das bedeutet, dass die vier Vektoren $\vec{a}_1, \vec{a}_2, \vec{a}_3, \vec{a}_4$ linear abhängig sind. Mit Hilfe der gefundenen linearen Beziehung zwischen diesen 4 Vektoren lässt sich jeder als rationale Linearkombination der übrigen darstellen. Es gilt

$$\begin{aligned}
\vec{a}_1 &= -\frac{25}{2}\vec{a}_2 - 8\vec{a}_3 - 9\vec{a}_4 \\
\vec{a}_2 &= -\frac{2}{25}\vec{a}_1 - \frac{16}{25}\vec{a}_3 - \frac{18}{25}\vec{a}_4 \\
\vec{a}_3 &= -\frac{1}{8}\vec{a}_1 - \frac{25}{16}\vec{a}_2 - \frac{9}{8}\vec{a}_4 \\
\vec{a}_4 &= -\frac{1}{9}\vec{a}_1 - \frac{25}{18}\vec{a}_2 - \frac{8}{9}\vec{a}_3
\end{aligned}$$

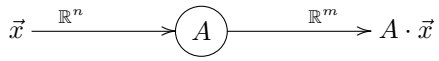
wie man leicht überprüft. Weil in diesem Beispiel bereits 4 Dimensionen involviert sind, lassen sich keine sehr aussagekräftigen ebenen Bilder mehr zeichnen, was aber nicht heisst, dass höherdimensionale Probleme dieser Art nicht doch eine anschauliche Interpretation haben können. ◯

Kapitel 2

Matrizenalgebra

Bisher haben wir Matrizen nur verwendet, um lineare Gleichungssysteme in der Form $(A \mid \vec{b})$ abzukürzen. Matrizen haben aber ein interessantes Doppelleben und erst das macht sie so nützlich. Einserseits können sie, ähnlich wie Listen oder Folgen, als Datentypen und andererseits als verallgemeinerte Zahlen betrachtet werden. Im Gegensatz zu jenen Datentypen sind sie mehrdimensional und für uns eigenständige mathematische Objekte, die wir nun zum Selbstzweck untersuchen wollen. Dieser Standpunkt ist nicht nur für die Kodierung linearer Gleichungssysteme, sondern für die Beschreibung fundamentaler geometrischer Operationen und von Prozessen zwischen Systemen bestens geeignet.

Eine Matrix $A \in \mathbb{R}^{m,n}$ ist aber nicht nur Datenträger. Sie kann nämlich andererseits auch als Operator benutzt werden. Dazu fassen wir in der Gleichung $A \cdot \vec{x} = \vec{b}$ den Vektor $\vec{b}$ als Output eines Prozesses A auf, der auf den Inputdaten, die im Vektor $\vec{x}$ zusammengefasst sind, operiert und den Wert $A \cdot \vec{x}$ produziert.



Viele Aspekte der linearen Algebra werden erst dann intuitiv klar, wenn wir die Dynamik und die Geometrie ins Spiel bringen und die Matrix zur Beschreibung der linearen Abbildung

$$f_A \colon \mathbb{R}^n \to \mathbb{R}^m, \qquad \vec{x} \mapsto A \cdot \vec{x}$$

heranziehen. Um allerdings Matrizen auf Vektoren anwenden zu können, müssen wir zuerst lernen, mit Matrizen zu rechnen.

2.1 Definitionen

Zu jeder neuen Sprache gehört an den Anfang entsprechendes Vokabular. Wir beginnen unsere Untersuchungen mit folgender Definition.

Definition. Für die Indizes $1 \leq i \leq m$ und $1 \leq j \leq n$ heisst ein rechteckiges Schema

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

von Elementen aus einer Zahlenmenge R eine *Matrix*[1]. Sie hat m Zeilen und n Spalten und ist vom Typ $m \times n$. Das Element in der i-ten Zeile und in der j-ten Spalte bezeichnet man mit a_{ij}. Gelegentlich benutzt man auch die Bezeichnung $A = (a_{ij})_{1 \le i \le m, 1 \le j \le n}$ oder etwas weniger pompös $A = (a_{ij})$.

Mit $R^{m,n}$ oder gelegentlich auch mit $\mathrm{Mat}_{m,n}(R)$ bezeichnen wir die Menge aller Matrizen vom Typ $m \times n$, deren Elemente Zahlen aus R sind.

Die betreffenden quadratischen Matrizen, für die $m = n$ ist, werden auch kurz mit $\mathrm{Mat}_m(R)$ bezeichnet.

CAS. In Sage definiert man den Raum aller Matrizen vom Typ 2×3 mit ganzzahligen, rationalen oder mit reellen Koeffizienten, d.h. $\mathbb{Z}^{2,3}$, $\mathbb{Q}^{2,3}$ oder $\mathbb{R}^{2,3}$ mit Hilfe der drei Befehle

```
MSZ=MatrixSpace(ZZ,2,3)
MSQ=MatrixSpace(QQ,2,3)
MSR=MatrixSpace(RR,2,3)
```

Man gibt also zunächst den Ring und dann den Typ der Matrizen an.

Hat man den Raum aller Matrizen eines festen Typs über einem festen Ring einmal erklärt, kann man anschliessend einzelne Matrizen etwas einfacher definieren, indem man bloss eine Liste ihrer Elemente angibt. Der Kode

```
A1=MSZ([1,-3,5,0,-3,2])
```

definiert die selbe ganzzahlige Matrix vom Typ 2×3, wie der Befehl

```
A2=matrix(ZZ,2,3, [[1,-3,5],[0,-3,2]])
```

mit Hilfe einer Liste von Listen.

Analog definieren die beiden Befehle

```
B1=MSQ([1,1/2,1/3,1/4,1/5,1/6])
```

und

```
B2=matrix(QQ,2,3, [[1,1/2,1/3],[1/4,1/5,1/6]])
```

die selben Matrizen mit rationalen Elementen vom Typ 2×3. Um eine zufällige Matrix aus einem solchen Raum zu erzeugen, verwendet man folgenden Befehl:

```
Az=MSZ.random_element()
Az.nrows()
Az.ncols()
```

Die beiden anderen Befehle lieferen ihre Zeilen- bzw. Spaltenzahl. Eine zufällige (2×3)-Matrix mit ganzzahligen Elementen kann alternativ auch mit dem Befehl

```
Az=random_matrix(ZZ, 2, 3)
```

erzeugt werden.

Man beachte sorgfältig, dass die fast gleichlautenden Befehle

```
AQ=MSQ([1,-3,5,0,-3,2])
AQ=matrix(QQ,2,3, [[1,-3,5],[0,-3,2]])
AQz=MSQ.random_element()
```

[1] Etwas formaler würde man eine solche Matrix als Abbildung $[1..m] \times [1..n] \to R$ auffassen. Dabei ist R ein kommutativer Ring und $[1..k] = \{1, 2, \ldots, k\}$ bezeichnet den Prototyp einer endlichen Menge mit k Elementen. Im Gegensatz zu anderen Orten beginnt man also in der Matrizenrechnung aus historischen Gründen die Nummerierung mit 1 statt mit 0.

eine völlig andere Art von Objekten liefern! Diesmal handelt es sich um Matrizen, die zwar den selben Typ haben, aber über dem Körper $\mathbb{Q}$ der rationalen statt über dem Ring $\mathbb{Z}$ der ganzen Zahlen erklärt sind. Gelegentlich sieht man also einem ausgedruckten Sage-Objekt nicht genau an, von welcher Art es ist!

Analog liefern die Befehle

```
AR=MSR([1,-3,5,0,-3,2])
AR=matrix(RR,2,3, [[1,-3,5],[0,-3,2]])
ARz=MSR.random_element()
```

Matrizen vom selben Typ über dem Körper $\mathbb{R}$ der reellen Zahlen.

Genau so liefern die Befehle

```
BQ=MSQ([1,1/2,1/3,1/4,1/5,1/6])
BQ=matrix(QQ,2,3, [[1,1/2,1/3],[1/4,1/5,1/6]])
BQz=MSQ.random_element()
```

Matrizen mit rationalen Koeffizienten, während die fast gleichlautenden Befehle

```
QR=MSR([1,1/2,1/3,1/4,1/5,1/6])
BR=matrix(RR,2,3, [[1,1/2,1/3],[1/4,1/5,1/6]])
BRz=MSR.random_element()
```

reelle Matrizen vom Typ 2×3 liefern.

Will man genau wissen, von welcher Art ein Sage-Objekt ist, benutzt man folgende Befehle:

```
show(parent(A1))
show(parent(B1))
show(parent(AR))
```

Die Vergleiche

```
parent(A1)==parent(B1)
parent(A1)==parent(AR)
parent(B1)==parent(AR)
```

zeigen, dass es sich in der Tat um Objekte unterschiedlicher Art handelt. Mit den beiden Befehlen

```
MSQ(A1)
MSR(A1)
```

lässt sich eine ganzzahlige Matrix in eine rationale bzw. in eine reelle Matrix umwandeln. Analog wandelt man mit Hilfe des Befehls

```
MSR(B1)
```

eine Matrix mit rationalen Elementen in eine mit reellen Elementen um.

Diese Ringwechsel erreicht man auch mit den Befehlen

```
A1.change_ring(QQ)
A1.change_ring(RR)
B1.chage_ring(RR)
```

Durch Überladen muss man bei vielen Befehlen nicht aufpassen, auf welche Art von Matrizen sie wirken. Schliesslich erlauben es die Befehle

```
show(MSZ.base_ring())
show(MSQ.base_ring())
show(MSR.base_ring())
```

den Grundring eines Matrizenraumes abzufragen, ◇

Vorläufig werden die Elemente unserer Matrizen alles ganze, rationale oder reelle Zahlen sein. Man kann aber die Zahlenmengen $\mathbb{Z}$, $\mathbb{Q}$ oder $\mathbb{R}$ der ganzen, rationalen oder reellen Zahlen durch andere Zahlenmengen R ersetzen, in denen man mit den üblichen Rechenregeln Addieren und Multiplizieren kann. Um uns nicht mit Auslöschung, Rundungsfehlern und Überläufen herumschlagen und uns nicht im Sumpf der theoretisch undurchschaubaren Numerik herumtreiben zu müssen, werden wir nach Möglichkeit ganze Zahlen aus $\mathbb{Z}$ und Elemente aus dem exakten Körper der rationalen Zahlen $\mathbb{Q}$ und nicht Näherungen aus $\mathbb{R}$ benutzen[2]. Später werden als Matrizenelemente auch komplexe Zahlen aus $\mathbb{C}$, Polynome mit ganzzahligen Koeffizienten aus $\mathbb{Z}[x]$ bzw. solche mit rationalen Koeffizienten aus $\mathbb{Q}[x]$ oder die Elemente eines Primkörpers $\mathbb{Z}_p$ oder allgemeiner eines endlichen Körpers $\mathbb{K}_q$ mit einer Primzahlpotenz $q = p^n$ in Frage kommen. Ab und zu werden die Elemente einer Matrix gar selber Matrizen sein.

Eine Struktur $(R, +, \cdot, -, 0, 1)$ mit den Möglichkeiten der Schul-Arithmetik bezeichnet man als *Ring*. Ein Ring ist also ein allgemeines Zahlensystem und besteht aus einer Menge R von Elementen zusammen mit zwei 2-stelligen Operationen $+\colon R \times R \to R$ und $\cdot\colon R \times R \to R$, einer 1-stelligen Operation $-\colon R \to R$ und zwei ausgezeichneten Elementen $0, 1 \in R$. Diese Strukturelemente erfüllen die Rechenregeln der Schulmathematik.

$x + (y + z) = (x + y) + z$	(Assoziativgesetz der Addition)
$x \cdot (y \cdot z) = (x \cdot y) \cdot z$	(Assoziativgesetz der Multiplikation)
$x + y = y + x$	(Kommutativgesetz der Addition)
$x + 0 = x$	(Neutralgesetz der Addition)
$x \cdot 1 = x$	(Neutralgesetz der Multiplikation)
$x + (-x) = 0$	(Invesengesetz der Addition)
$x \cdot (y + z) = x \cdot y + x \cdot z$	(linkes Distributivgesetz)
$(y + z) \cdot x = y \cdot x + z \cdot x$	(rechtes Distributivgesetz)

Falls zusätzlich das Kommutativgesetz der Multiplikation $x \cdot y = y \cdot x$ gilt, nennt man den Ring kommutativ. Fehlt wie bei $(\mathbb{N}, +, \cdot, 0, 1)$ das additive Inverse $-x$ und entsprechend das Inversengesetz der Addition, redet man von einem Rig.

CAS. In Sage lassen sich die wichtigsten für die Anwendungen relevanten Ringe mit Hilfe folgender Definitionen erzeugen:

`ZZ`: Ring $\mathbb{Z}$ der ganzen Zahlen.

`QQ`: Körper $\mathbb{Q}$ der rational Zahlen.

`RR`: Körper $\mathbb{R}$ der rellen Zahlen.

`CC`: Körper $\mathbb{C}$ der komplexen Zahlen.

`ZZ["x"]`: Ring $\mathbb{Z}[x]$ der Polynome in der Unbestimmten x mit ganzzahligen Koeffizienten.

`QQ["x"]`: Ring $\mathbb{Q}[x]$ der Polynome in der Unbestimmten x mit rationalen Koeffizienten.

[2] Wer das unbedingt will, muss halt eine Vorlesung über Numerik belegen und wird dann das "number-crunching" erst recht einer Maschine überlassen wollen. In Sage erledigt man das mit dem für die Numerik entwickelte Zusatzpacket `NumPy`, das man zunächst mit dem Befehl `import numpy` importiert, um dann eine Matrix mit dem typischen Befehl `A=numpy.array([[1,2,3],[3,0,1]])` zu definieren.

`Frac(ZZ["x"])`: Quotientenkörper $\mathbb{Z}(x)$ der rationalen Brüche in der Unbestimmten x mit ganzzahligen Koeffizienten.

`ZZ[["x"]]`: Ring $\mathbb{Z}[\![x]\!]$ der Potenzreihen in der Unbestimmten x mit ganzzahligen Koeffizienten.

`Integers(6)`: Ring $\mathbb{Z}_6$ der ganzen Zahlen modulo 6.

`FiniteField(9)`: Endlicher Körper $\mathbb{F}_9$ mit 9 Elementen.

Wie oben kann man damit auch die zugehörigen Matrizenringe definieren und die Matrizen unterschiedlicher Art durch Ringwechsel ineinander umwandeln, falls das sinnvoll ist. ◇

Wir werden nach Möglichkeit unsere Methoden für kommutative Ringe entwickeln, die keine Nullteiler haben. In einem solchen Ring, der als *Integritätsbereich* bezeichnet wird, folgt aus $a \cdot b = 0$, dass $a = 0$ oder $b = 0$ sein muss. In einem (unendlichen) Integritätsbereich kann man also nicht unbedingt durch ein von 0 verschiedenes Element dividieren. Integritätsbereiche brauchen deshalb keine Köper zu sein, in denen zusätzlich jedes von 0 verschiedene Element eine Einheit ist. Wir wollen also nach Möglichkeit das Rechnen mit Brüchen vermeiden und bloss die Teilbarkeitsbeziehung benutzen.

Eine wirklich einfache Teilbarkeitstheorie findet man jedoch erst in einem *Hauptidealbereich*, in dem jedes Ideal von einem einzigen Element erzeugt werden kann. In einem solchen Ring R hat jedes Element eine eindeutige Primfaktorzerlegung in Primelemente und zu je zwei Elementen $x, y \in R$ existiert ihr grösster gemeinsamer Teiler $\mathrm{ggT}(x, y)$ bzw. ihr kleinstes gemeinsames Vielfaches $\mathrm{kgV}(x, y)$ und für teilerfremde x, y kann jedes Element $r \in R$ als Linearkombination $ax+by = r$ dargestellt werden. Als Prototypen von Hauptidealbereicheen denke man an den kommutativen Ring $\mathbb{Z}$ der ganzen Zahlen. Auch jeder Körper K oder allgemeiner der kommutativen Ring $K[x]$ der Polynome mit Koeffizienten in einem Körper und der Ring $K[\![x]\!]$ der formalen Potenzreihen oder der Ring der ganzen Gauss'schen Zahlen $\mathbb{Z}[i]$ sind Hauptidealbereiche. In jedem dieser Integritätsbereiche gibt es eine Normfunktion

$$|-|: R \to \mathbb{N}, \qquad r \mapsto |r|$$

mit folgenden Eigenschaften:

1. $|0| = 0$.
2. Aus $|r| = 0$ folgt, dass $r = 0$ ist.
3. Für beliebige Elemente $0 \neq x, y \in R$ gilt:
   - Entweder ist y ein Teiler von x in R.
   - Oder es gibt Elemente $a, b \in R$ mit $0 < |ax - by| < |y|$.

In der Tat sind die Hauptidealbreiche genau die Integritätsbereiche mit einer solchen Norm. Beispielsweise kann für ein Polynom $p(x) \in K[x]$ über einem Körper K durch $2^{\deg(p(x))}$ eine solche Norm definiert werden, woraus folgt, dass es sich bei $K[x]$ in der Tat um einen Hauptidealbereich handelt.

Falls ein Hauptidealbereich R sogar eine Norm hat, die statt (3) die echt stärkere Eigenschaft

4. Für beliebige Elemente $0 \neq x, y \in R$ gibt es $q, r \in R$ mit der Eigenschaft, dass
$$a = bq + r$$
gilt für $r = 0$ oder $|r| < |y|$.

hat, so kann man in R die Division mit Rest ausführen und redet von einem Euklid'schen Bereich, weil sich dann der grösste gemeinsame Teiler und das kleinste gemeinsame Vielfache in der üblichen Art mit Hilfe des Euklid'schen Algorithmus explizit berechnen lassen. Beispielsweise ist jeder Körper K ein solcher Euklid'scher Bereich, weil man als Norm $|x| = 1$ setzen kann. Auch der Ring $K[x]$ der Polynome mit Koeffizienten in einem Körper ist ein Euklid'scher Bereich, da man als Euklid'sche Norm den Grad $\deg(p(x))$ eines von 0 verschiednen Polynomes verwenden kann. Der Ring $K[\![x]\!]$ der formalen Potenzreihen wird zu einem Euklid'schen Bereich, falls man zu jeder formalen Potenzreihe $0 \neq f(x) \in K[\![x]\!]$ als Norm den Grad des kleinsten, von 0 verschiedenen Koeffizienten benutzt. Der Ring der ganzen Gauss'schen Zahlen $\mathbb{Z}[i]$ wird zu einem Euklidschen Bereich, falls man als Norm $|a + bi| = a^2 + b^2$ verwendet. Schliesslich wird der Ring $\mathbb{Z}$ der ganzen Zahlen zu einem Euklid'schen Bereich, falls man als Norm $|r|$ einer ganzen Zahl $r \in \mathbb{Z}$ ihren Betrag verwendet. Weil im Gegensatz dazu der Integritätsbereich $\mathbb{Z}[x]$ der Polynome mit ganzzahligen Koeffizienten kein Hauptidealbereich ist, da beispielsweise das von 2 und x erzeugte Ideal nicht durch ein einziges Polynom erzeugt werden kann, existiert für ihn auch keine Euklidsche Norm.

Der Abhängigkeit der linearen Algebra vom Grundring wird in Grundvorlesungen oft nicht die nötige Aufmerksamkeit geschenkt, was im Umgang mit linearen Gleichungssystemen zu verwirrenden Ergebnissen führen kann.

Beispiel. In Sage wird eine Matrix A mit ganzzahligen Elementen stillschweigend als Matrix über dem Grundring $\mathbb{Z}$ interpretiert, während die Befehle `solve` und `rref` standardmässig über dem Quotientenkörper $\mathbb{Q}$ durchgeführt werden. Wer so lange wie möglich in einem Euklidschen Bereich wie $\mathbb{Z}$ rechnen will, um etwa die ganzzahligen Lösungen eines linearen Gleichungssystems studieren zu können, braucht statt der normierten, reduzierten Stufenform die sogn. Smith'sche Normalform S, die man in Sage mit dem folgenden Befehl

```
A.smith_form()
A.elementary_divisors()
```

erhält. Die führenden Elemente von S werden als Elementarteiler von A bezeichnet. Weil dabei nicht nur ganzzahlig invertierbare Zeilen- sondern zusätzlich auch solche Spaltenoperationen durchgeführt werden müssen, gilt mit den beiden ganzzahligen unimodularen Transformationsmatrizen U und V dann die Faktorisierung $S = U \cdot A \cdot V$. Lineare ganzzahlige Gleichungssysteme können also im Gegensatz zu beliebigen diophantischen Problemen immer durch Algorithmen gelöst werden. ◯

Die lineare Algebra über den komplexen Zahlen spielt beim Studium von Wechselstromkreisen und in der Quantenmechanik eine fundamentale Rolle. Das Studium der Dynamik eines autonomen linearen dynamischen Systems führt auf das Eigenwertproblem und dann zu Matrizen über einem Polynomring. Den Euklid'sche Bereich $\mathbb{Q}[x]$ schliessen wir also deshalb von Anfang in unsere Untersuchungen ein, weil die Ähnlichkeitstheorie einer quadratischen Matrix A

der Berechnung der Normalform ihrer charakteristischen Matrix $A - xE$ entspricht, deren Elemente Polynome in $\mathbb{Q}[x]$ sind. Die lineare Algebra über endlichen Körpern hat wichtige Anwendungen in der Kodierungstheorie und in der Kryptologie. Oft sind in der Informatik, in der Topologie oder in der Zahlentheorie die Elemente einer Matrix ganze Zahlen aus $\mathbb{Z}$ oder rationale Polynome aus $\mathbb{Q}[x]$. Wie wir gesehen haben, interessieren sich auch viele Anwender — etwa Chemiker oder Soziologen — in erster Linie für ganzzahlige Lösungen.

Die Definition der Matrizen umfasst die in der Schule benutzten Zahlen. Eine reelle 1×1-Matrix ist im Wesentlichen eine Zahl, die wir in der Regel ohne Klammern schreiben. Wir identifizieren also $\mathbb{R}^{1,1}$ mit $\mathbb{R}$. Dieses Vermischen von Datentypen führt kaum zu ernsthaften Problemen, solange man auf den Kontext achtet. Insbesondere werden wir die Rechenoperationen für Matrizen so erklären, dass sie die bereits aus der Schule bekannten Regeln für das Rechnen mit rationalen Zahlen verallgemeinern. Umgekehrt lässt sich dieser Spezialfall dazu verwenden, um bekannte Konzepte von der rationalen Zahlen auf allgemeinere Matrizen zu verallgemeinern.

Die Matrizenrechnung umfasst auch die Vektoralgebra aus der Schule.

Definition. Besteht eine Matrix aus einer einzigen Spalte, heisst sie *Spaltenvektor*. Für einen Spaltenvektor vom Typ $m \times 1$ schreiben wir:

$$\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in \mathbb{R}^{m,1} =: \mathbb{R}^m$$

Die einzelnen Komponenten werden also untereinander geschrieben.

Wir schreiben für einen solchen Spaltenvektor kurz $\vec{a}$, wenn die einzelnen Elemente keine Rolle spielen. Eine Lösung $\vec{x}$ eines linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ ist ein solcher Spaltenvektor. Entsprechend werden die Konstanten zum Konstantenvektor $\vec{b}$ zusammengefasst.

Analog wird eine Matrix vom Typ $1 \times n$, d.h. eine Matrix mit nur einer Zeile

$$(a_1, a_2, \cdots, a_n) \in \mathbb{R}^{1,n} =: (\mathbb{R}^n)^*$$

als *Zeilenvektor*, als n-Tupel oder gelegentlich auch als n-dimensionalen Kovektor bezeichnet. Die Elemente eines n-Tupels schreiben wir also nebeneinander und trennen sie gelegentlich zur besser Übersicht durch Kommas[3]. Ein

[3] Etwas formaler würde man einen solchen Vektor als Abbildung $[1..n] \to R$, d.h. als Wort der Länge n auffassen. Seine Buchstaben stammen aus dem Alphabet R, das wir als kommutativen Ring wählen. Gelegentlich benutzt man auch Folgen, d.h. Vektoren mit abzählbar unendlich vielen Komponenten $(a_0, a_1, \cdots, a_n, \cdots)$. Ein solcher Vektor kann als Abbildung $\mathbb{N} \to R$ aus $R^{\mathbb{N}}$ interpretiert werden und wird in der Analysis oft als Potenzreihe

$$a_0 + a_1 x + a_2 + \cdots + a_n x^n + \cdots = \sum_{j=0}^{\infty} a_k x^k \in R[\![x]\!]$$

bezeichnet. Die Koeffizienten werden dann statt durch Kommas durch die Symbole x^k getrennt. Falls nur endlich viele Koeffizienten von Null verschieden sind, erhält man ein Polynom

$$a_0 + a_1 x + a_2 + \cdots + a_n x^n = \sum_{j=0}^{n} a_k x^k \in R[x]$$

solches n-Tupel schreiben wir auch kurz als Liste mit n Elementen in der Form $\vec{a}^T = (a_1, a_2, \cdots, a_n)$. Physiker sind stolz auf die Unterscheidung zwischen dem Bra $\langle a|$ (Zeilenvektor) und dem zugehörigen Ket $|a\rangle$ (Spaltenvektor).

CAS. In Sage werden Spaltenvektoren mit einem der folgenden Kodes definiert:

```
spalten_vektor=matrix(3,1,[a1,a2,a3])
spalten_vektor=matrix([ [a1],[a2],[a3] ])
```

Im Gegensatz dazu werden Zeilenvektoren folgendermassen definiert:

```
zeilen_vektor=matrix(1,3,[a1,a2,a3])
zeilen_vektor=matrix([ [a1,a2,a3] ])
```

Zwischen den beiden Sorten von Vektoren lässt sich mit Hilfe der Befehle

```
zeilen_vektor=spalten_vektor.transpose()
spalten_vektor=zeilen_vektor.transpose()
```

umrechnen. Man beachte, dass der in Sage eingebaute Befehl

```
vektor=vector([a1,a2,a3])
```

weder einen Spalten-, noch einen Zeilenvektor, sondern ein sogn. Tupel liefert, das sich aber für die meisten Zwecke sowohl als Spalten- als auch als Zeilenvektor auffassen lässt. Entsprechend wird durch den Befehl

```
VS=VectorSpace(QQ,3)
```

oder kürzer durch

```
VS=QQ^3
VS.dimension()
```

der Raum $\mathbb{Q}^3$ aller Vektoren der Dimension 3 mit rationalen Komponenten erklärt. Bei Bedarf kann seine Dimension abgefragt werden. Um einen zufälligen Vektor aus diesem Vektorraum zu erzeugen, verwendet man folgenden Befehl

```
v=VS.random_element()
v.degree()
```

Der zweite Befehl liefert die Dimension des ihn enthaltenden Vektorraumes.

In einem Vektorraum muss der Grundring zwingend ein Körper sein. Um die Menge $\mathbb{Z}^3$ aller Vektoren mit ganzzahligen Koeffizienten (Gitterpunkte im Raum) zu erklären, benötigt man die Befehle

```
FM=ZZ^3
FM=FreeModule(ZZ,3)
```

Objekte dieser Art nennen die Mathematiker freie Moduln. Auf sie lässt sich ein grosser Teil der linearen Algebra verallgemeinern. ◇

Wir wollen also in Zukunft Spalten- und Zeilenvektoren sorgfältig auseinanderhalten. Vektoren treten in zwei unterschiedlichen Sorten auf, obwohl diese Unterscheidung, die ziemlich subtiler Natur ist, aus typographischen Gründen

---

Es wird durch die endliche Koeffizientenliste $(a_0, a_1, \cdots, a_n) \in R^\infty$ kodiert. Die Vektoren dieses Vektorraumes sind also die endlichen Listen von Elementen aus R.

Für die beiden Vektorräume $\mathbb{R}^\infty$ und $\mathbb{R}^\mathbb{N}$ gilt die Dimensionsbeziehung $\dim(\mathbb{R}^\infty) < \dim(\mathbb{R}^\mathbb{N})$. In der Tat ist sogar $(\mathbb{R}^\infty)^* \cong \mathbb{R}^\mathbb{N}$. Für einen Vektor $\vec{w} \in \mathbb{R}^\mathbb{N}$ und einen Vektor $\vec{v} \in \mathbb{R}^\infty$ liefert nämlich das "Skalarprodukt" $\langle \vec{w}, \vec{v} \rangle$ für fests $\vec{w}$ ein lineares Funktional $\mathbb{R}^\infty \to \mathbb{R}$, das für diese Dualitätsbeziehung verantwortlich ist. Sie wird im endlichdimensionalen Fall enger, weil dort sogar $(\mathbb{R}^n)^* \cong \mathbb{R}^n$ und daher auch $(\mathbb{R}^n)^{**} \cong \mathbb{R}^n$ gilt. Diese enge Dualitätsbeziehung spielt im Hintergrund der Matrizenrechnung eine zentrale Rolle.

wider besseren Wissens in vielen Lehrbüchern und aus Ignoranz in kaum einer Programmiersprache gemacht wird. Entsprechend unterscheidet man dann nicht zwischen dem Raum der Spaltenvektoren $\mathbb{R}^{k,1}$ und jenem der Zeilenvektoren $\mathbb{R}^{1,k}$ und bezeichnet beide Räume kurz mit $\mathbb{R}^k$. Der Leser gewöhne sich an, in jeden Fall zu fragen, um welche Sorte von Vektoren es sich nun im konkreten Fall handelt. Die Antwort auf diese Frage ist selbstverständlich vom Kontext abhängig.

Ohne diese sorgfältige Unterscheidung zwischen Spalten- und Zeilenvektoren würde ein zentraler Aspekt der linearen Algebra — die Dualität — auf der Strecke bleiben. Strukturblinde Anwender übersehen damit oft eine fundamentale mathematische Struktur und damit viele interessante Symmetrie-Beziehungen! Dazu gehört etwa die enge Beziehung zwischen Gradient $\mathrm{grad}(f)$ und totalem Differential df, jene zwischen der Spannung U (Potential) und dem Strom I, zwischen parallelen und seriellen Schaltungen, zwischen einem Kurzschluss und einem offenen Schaltkreis, zwischen Widerstand R und Leitfähigkeit G bzw. zwischen den beiden zugehörigen Formen des Ohmschen Gesetzes

$$U = R \cdot I, \qquad I = G \cdot U$$

und allgemeiner zwischen der Kapazität C und der Induktivität L bzw. zwischen den zugehörigen Gleichungen

$$I(t) = C \cdot \frac{dU(t)}{dt}, \qquad U(t) = L \cdot \frac{dI(t)}{dt}$$

etc. in der Elektrizitätslehre oder zwischen den analogen Konzepten in der Hydrodynamik, zwischen Deformation und Spannung in der Elastizitätslehre, zwischen Produktionsmenge und Rohstoffkosten in der Betriebswirtschaftslehre oder zwischen elektrischem und magnetischem Feld in der Elektrodynamik und zwischen Anfangs- und Endzustand in der Quantenmechanik. Wer als Anwender nicht die volle involvierte mathematische Struktur im Auge behält und studiert, darf sich nicht wundern, wenn er sein Kerngeschäft gar nicht richtig versteht und dessen vielfältige Möglichkeiten nur unvollständig nutzen kann. Später werden wir die Struktur linearer Netzwerke beschreiben, aber erst mit Hilfe der multilinearen Algebra erkennen, dass die Elektrodynamik eine Verallgemeinerung davon ist.

Oft ist es nützlich, eine Matrix $A \in \mathbb{R}^{m,n}$ als n-Tupel von Spaltenvektoren

$$A = \begin{pmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{pmatrix}, \qquad \vec{a}_i = \begin{pmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{pmatrix} \in \mathbb{R}^m$$

oder dual als Spaltenvektor von Zeilenvektoren

$$A = \begin{pmatrix} \vec{v}_1 \\ \vdots \\ \vec{v}_m \end{pmatrix}, \qquad \vec{v}_j = \begin{pmatrix} \vec{a}_{j1} & \vec{a}_{j2} & \cdots & \vec{a}_{jn} \end{pmatrix} \in \mathbb{R}^n$$

aufzufassen. Wir werden bald ein Verfahren besprechen, um aus einem Spaltenvektor einen Zeilenvektor mit den selben Elementen zu machen und umgekehrt.

Beispiel. Die Matrix

$$A = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 0 & 4 \end{pmatrix} \in \mathbb{R}^{2,3}$$

ist vom Typ 2×3. Sie hat die beiden Zeilenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 & 3 & 5 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 2 & 0 & 4 \end{pmatrix}$$

und die drei Spaltenvektoren

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad \vec{a}_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \qquad \vec{a}_3 = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$$

und die Elemente $a_{12} = 3$ und $a_{22} = 0$. ◯

CAS. In Sage können einzelne Zeilen und Spalten von Matrizen folgendermassen erhalten werden:

```
A=matrix(2,3,[1,3,5,2,0,4])
v1=A.row(0), v2=A.row(1)
a1=A.column(0), a2=A.column(1), a3=A.column(2)
A[0,1], A[1,1]
```

Die letzte Zeile zeigt, wie man auf einzelne Elemente von Matrizen zugreift. ◇

Konzeptionell entspricht das Vertauschen der Zeilen und Spalten einer Matrix dem Übergang von der Matrizenkategorie $\mathbf{Mat}_{\mathbb{R}}$ zur dualen Kategorie $(\mathbf{Mat}_{\mathbb{R}})^*$. Dabei werden „alle Pfeile umgekehrt“.

Wir wollen zur besseren Übersicht Matrizen mit lateinischen Grossbuchstaben bezeichnen. Für ihre Elemente benutzen wir doppelte Indizes. Spaltenvektoren bezeichnen wir mit Kleinbuchstaben und einem Pfeil darüber. Zahlen schliesslich werden mit Kleinbuchstaben bezeichnet. Im Zusammenhang der Matrizenrechnung bezeichnet man Zahlen auch als *Skalare*, um sie von den Matrizen von grösserem Typ unterscheiden zu können.

Matrizen wurden vor rund hundertfünfzig Jahren vom englischen Mathematiker A. Cayley eingeführt. Im Jahre 1858 veröffentlichte er eine Arbeit mit dem Titel: “A Memoir on the Theory of Matrices”. In grandioser Fehleinschätzung meinte er zu seiner Schöpfung: „Meine Idee wird niemals eine praktische Anwendung finden.“ Heute gehören die Anwendungen der Matrizenalgebra zum täglichen Brot von Physikern, Ingenieuren und Krämern. Während die Physiker im Zusammenhang mit der Quantenmechanik die enorme Bedeutung der Matrizen etwa seit 1930 erkannt haben, werden sie von den Ingenieuren erst etwa seit 1960 wirklich geschätzt, als klar wurde, dass mit ihrer Hilfe sowohl mechanische als auch elektrische Prozesse und ihre Regelungstheorie vorteilhaft beschrieben werden können und sie für die Signaltechnik unverzichtbar sind. Entsprechend haben Matrizen erst seit kurzer Zeit Eingang in die technische Lehrbuchliteratur für Studienanfänger gefunden.

Die Theorie linearer Gleichungssysteme ist viel älter. Die ältesten schriftlichen Spuren führen zu den Chinesen des 2. Jahrhunderts vor unserer Zeitrechnung. Im Kapitel 8 ihrer „neun Bücher über arithmetische Kunst“ verwenden sie bereits den Eliminations-Algorithmus. Im alten Europa wurde vielfach der sogn.

Determinanten-Kalkül zur Lösung linearer Gleichungssystemen benutzt und eine Matrix nicht als eigenständige Struktur, sondern bloss als Vorstufe betrachtet, aus der dann geeignete Determinanten gebildet wurden. Diese Haltung widerspiegelt sich noch in Sylvesters Wortwahl, bezeichnet doch das lateinische Wort "Matrix" den Mutterschoss oder die Gebährmutter. Diese Haltung, den Determinanten gegenüber den Matrizen den Vorzug zu geben, ist aus den zu besprechenden Gründen aus heutiger Sicht alles andere als empfehlenswert und wir werden sie deshalb vermeiden. Wir betrachten Matrizen als eigenständige mathematische Objekte, die vor allem zur Beschreibung linearer Prozesse dienen und verzichten so lange wie möglich auf den Gebrauch von Determinanten.

Wenn schon doppelt indizierte Grössen (a_{ij}) eine zentrale Rolle spielen sollen, wird man sich fragen, ob denn nicht auch „höherdimensionale Gebilde" mit drei und mehr Indizes (a_{ijk}) studiert werden sollten.

Definition. Ein *Tensor* vom Typ $n_1 \times n_2 \times \cdots \times n_d$ ist ein d-dimensionales Gebilde

$$T = (a_{i_1 i_2 \cdots i_d}) \in \mathbb{R}^{n_1, n_2, \cdots, n_d}$$

Tensoren entsprechen den Multilinearformen und spielen in der theoretischen Physik eine zentrale Rolle. Die Ingenieure sind aber noch nicht alle von ihrer Nützlichkeit überzeugt. Deshalb wollen wir uns hier nicht mit der ganzen Tensoralgebra befassen, obwohl viele Fragen der linearen Algebra erst klar werden, wenn man auch die multilineare Algebra heranzieht.

Definition. Zwei Matrizen A und B heissen *gleich*, d.h. $A = B$, falls gilt:

1. A und B sind vom gleichen Typ.
2. Entsprechende Elemente von A und B sind gleich.

Die Gleichheit von Matrizen ist also elementweise durch $A_{ij} = B_{ij}$ definiert.

Man beachte, dass schon die scheinbar harmlose Untersuchung auf Gleichheit von zwei Matrizen allerhand zu tun geben kann. Um die Gleichheit zweier Matrizen vom Typ $m \times n$ zu kontrollieren, sind insgesamt $m \cdot n$ Vergleiche durchzuführen, was auch einen Computer erheblich Rechenzeit kostet, wenn m und n, wie in den typischen Anwendungen, sehr gross sind. Dieses Anwachsen der Komplexität war für die gestrigen — nur mit Bleistift und Rechenschieber ausgerüsteten — Ingenieure der Hauptgrund, warum sie den Matrizen mit Skepsis begegnet sind. Mittlerweile wurde dank der Computer aus der Not eine Tugend.

CAS. Ein Gleichheitstest zeigt, dass in Sage durch folgende Definitionen die selben Objekte erzeugt werden:

```
matrix(2,3,[1,2,3,4,5,6]) == matrix([[1,2,3],[4,5,6]])
```

Im Gegensatz dazu zeigen die Tests

```
vector([1,2,3]) == matrix(1,3,[1,2,3])
vector([1,2,3]) == matrix(3,1,[1,2,3])
matrix(1,3,[1,2,3]) == matrix(3,1,[1,2,3])
```

dass in Sage Vektoren tatsächlich nicht die selben Objekte wie Zeilen- oder Spaltenvektoren sind und dass dort sorgfältig zwischen diesen beiden dualen Arten von Vektoren unterschieden werden kann! ◇

Gewisse spezielle Matrizen spielen eine wichtige Rolle und haben deshalb eigene Namen. Leider kommt man nicht darum herum, sich dieses Vokabular gut einzuprägen; also macht man es besser gleich und gründlich.

Definition. Eine Matrix, deren Elemente alle Null sind, heisst *Nullmatrix*. Entsprechend reden wir vom *Nullvektor*. Die Nullmatrix vom Typ $m \times n$ und der Nullvektor werden mit 0 bzw. mit $\vec{0}$ bezeichnet, falls der Typ aus dem Kontext klar ist. Im Zweifelsfall benutzen wir die präziseren Bezeichnungen $0_{m,n}$, $\vec{0}_m$.

CAS. Selbstverständlich will man eine grosse Nullmatrix nicht elementweise eingeben. Der folgende Kode bietet eine effiziente Alternative:

```
A=matrix(QQ,3,4,0); show(A)
```

Hat man den Raum aller Matrizen vom betreffenden Typ sowieso bereits definiert, kann man in ihm die Nullmatrix leicht auswählen.

```
MS=MatrixSpace(QQ,3,4)
A=MS.zero_matrix()
A.is_zero()
```

Mit dem letzten Befehl kann man überprüfen, ob eine Matrix tatsächlich die Nullmatrix ist.

Für den Nullvektor benutzt man entweder den leicht modifizieren Kode

```
v=vector(QQ,{2:0}); show(v)
```

oder falls man den Vektorraum der betreffenden Dimension bereits definiert hat

```
VS=VectorSpace(QQ,3)
v=VS.zero_vector(); show(v)
v.is_zero()
```

Zunächst erklärt man also den Vektorraum der Dimension 3 mit rationalen Komponenten und wählt dann in ihm den Nullvektor. Mit dem letzten Befehl überprüft man, ob ein Vektor der Nullvektor ist. ◇

Die nächstwichtigen Matrizen haben genau eine 1 und sonst alles 0.

Definition. Für $1 \leq k \leq m$ und $1 \leq l \leq n$ bezeichnen wir mit B_{kl} diejenige $m \times n$-Matrix, die an der Stelle (k, l) eine 1 und sonst alles 0 hat. Für ihre Elemente ist also

$$(B_{kl})_{ij} = \begin{cases} 1 & \text{falls } i = k, j = l \\ 0 & \text{sonst} \end{cases}$$

Diese $m \cdot n$ Matrizen heissen *Standardbasismatrizen*, weil sie eine Basis des Raumes $\mathbb{R}^{m,n}$ aller solcher Matrizen bilden. Mit ihnen können also sämtliche Matrizen von diesem Typ eindeutig als Linearkombinationen zusammengebaut werden.

Beispiel. Falls $n = 1$ ist, bilden die Standardbasisvektoren $\vec{e}_k = B_{k1}$, die durch die Bedingung

$$\left(\vec{e}_k\right)_i = \begin{cases} 1 & \text{falls } i = k \\ 0 & \text{sonst} \end{cases}, \quad 1 \leq k \leq m$$

erklärt sind, eine Basis des m-dimensionalen Vektorraumes $\mathbb{R}^m$. Für jeden Spaltenvektor $\vec{a} \in \mathbb{R}^m$ erhalten wir nämlich die eindeutige Linearkombination

$$\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = a_1\vec{e}_1 + a_2\vec{e}_2 + \cdots + a_m\vec{e}_m = \sum_{k=1}^{m} a_k\vec{e}_k$$

Die Standardbasisvektoren spielen also unter allen Vektoren die Rolle von Atomen, mit denen sämtliche Vektoren auf eindeutig Art linear zusammengebaut werden können. ◯

Beispiel. Im Falle $m = n = 2$ erhalten wir die vier Standard-Basis Matrizen

$$B_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad B_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Mit ihrer Hilfe kann eine beliebige Matrix aus $\mathbb{R}^{2,2}$ linear zusammengebaut werden. In unserem Beispiel erhalten wir die Linearkombination

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11} \cdot B_{11} + a_{12} \cdot B_{12} + a_{21} \cdot B_{21} + a_{22} B_{22}.$$

Zerlegungen von Matrizen in einfache Grundtypen spielen eine fundamentale Rolle in der linearen Algebra. ◯

CAS. In Sage erklärt man im folgenden Kode zunächst den Vektorraum der betreffenden Dimension und berechnet dann die Liste seiner Standardbasisvektoren.

```
VS=VectorSpace(QQ,3)
b=VS.basis()
e1=b[0]; e2=b[1]; e3=b[2]; show(e1,e2,e3)
```

Analog definiert man den Raum der Matrizen von einem gewissen Typ und bestimmt dann die Liste seiner Standardbasismatrizen.

```
MS=MatrixSpace(QQ,2,2)
b=MS.basis()
B11=b[0]; B12=b[1]; B21=b[2]; B22=b[3]; show(B11,B12,B21,B22)
```

Man beachte, wie sich die Indizes der Standardbasismatrizen aus der Binärdarstellung der Nummer der Listenelemente ergeben. ◇

Definition. Eine Matrix A mit n Zeilen und n Spalten heisst *quadratisch*. Die Elemente $a_{11}, a_{22}, \cdots, a_{nn}$ heissen *Diagonalelemente* von A.

CAS. Um in Sage zu testen, ob eine Matrix A quadratisch ist, benutzt man folgenden Befehl

```
A.is_square()
A.diagonal()
```

Mit dem zweiten Befehl erhält man die Liste der Diagonalelemente einer quadratischen Matrix A. ◇

Oberhalb der Diagonalen stehen die Elemente a_{ij} mit $i < j$ und unterhalb diejenigen mit $i > j$.

Definition. Eine quadratische Matrix A heisst *obere Dreiecksmatrix*, falls ihre Elemente unterhalb der Diagonalen alles 0 sind, d.h. es gilt:

$$a_{ij} = 0 \text{ für } i > j$$

Die quadratische Matrix A heisst *untere Dreiecksmatrix*, falls ihre Elemente oberhalb der Diagonalen alles 0 sind, d.h. es gilt:

$$a_{ij} = 0 \text{ für } i < j$$

Die quadratische Matrix A heisst *Diagonalmatrix*, falls sie obere und untere Dreiecksmatrix ist, d.h. falls höchstens die Diagonalelemente von 0 verschieden sind. Für Diagonalmatrizen gilt

$$a_{ij} = 0 \text{ für } i \neq j$$

Wichtig sind Diagonalmatrizen mit lauter 1 in der Diagonalen. Für ihre Elemente gilt also

$$a_{ij} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}$$

Für die *Einheitsmatrix* vom Typ $n \times n$ schreiben wir E_n.

Beispiel. Bei der quadratischen Matrix

$$\begin{pmatrix} 1 & -2 & 3 & 0 \\ 0 & 2 & 4 & 3 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 3 \end{pmatrix} \in \mathbb{R}^{4,4}$$

handelt es sich um eine obere Dreiecksmatrix. Im Gegensatz dazu ist

$$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 4 & 3 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 2 & 0 & 5 & 2 \end{pmatrix} \in \mathbb{R}^{4,4}$$

eine untere Dreiecksmatrix. Schliesslich sind

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \in \mathbb{R}^{4,4}, \qquad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{4,4}$$

zwei Diagonalmatrizen. Bei der rechten Matrix handelt es sich um E_4. ◯

CAS. In Sage erklärt man obige Diagonalmatrix, indem man die Liste der Diagonalelemente angibt, d.h. durch den Kode

```
A=diagonal_matrix([1,2,3,4]); show(A)
```

Die Einheitsmatrix spielt eine grosse Rolle und kann mit folgendem Befehl

```
A=identity_matrix(4); show(A)
A.is_one()
```

erzeugt und mit dem zweiten abgefragt werden. ◇

Definition. Eine quadratische Matrix A heisst *symmetrisch*, falls sie an der Diagonalen spiegelsymmetrisch ist. Für ihre Elemente gilt also $a_{ij} = a_{ji}$.

Beispiel. Die quadratische Matrix

$$\begin{pmatrix} 1 & 2 & -1 & 5 \\ 2 & 2 & 0 & 3 \\ -1 & 0 & 3 & -7 \\ 5 & 3 & -7 & -1 \end{pmatrix} \in \mathbb{R}^{4,4}$$

ist symmetrisch. ◯

Definition. Die quadratische Matrix A heisst *schiefsymmetrisch*, falls für alle Elemente $a_{ij} = -a_{ji}$ gilt.

Beispiel. Die Matrix

$$\begin{pmatrix} 0 & 1 & -3 & -5 \\ -1 & 0 & -4 & 0 \\ 3 & 4 & 0 & 2 \\ 5 & 0 & -2 & 0 \end{pmatrix} \in \mathbb{R}^{4,4}$$

ist schiefsymmetrisch. Man beachte, dass die Diagonalelemente schiefsymmetrischer Matrizen 0 sein müssen. ◯

CAS. In Sage prüft man mit den Befehlen

```
A.is_symmetric()
A.is_skew_symmetric()
```

ob eine quadratische Matrix A symmetrisch bzw. antisymmetrisch ist. ◇

2.2 Grundoperationen

Analog zu den Zahlen gibt es eine „Matrizenarithmetik“, die aus Addition und Multiplikation besteht. Diese Grundoperationen für das Rechnen mit Matrizen wollen wir jetzt beschreiben und ihre Rechengesetze kennenlernen.

2.2.1 Addition

Ausgangspunkt für das Rechnen mit Matrizen ist ihre Addition mit einer naheliegenden Definition.

Definition. Sind A und B zwei Matrizen vom gleichen Typ, so ist ihre *Summe* $A + B$ diejenige Matrix, die durch Addition der einander entsprechenden Elemente entsteht. Es gilt:

$$(A + B)_{ij} = A_{ij} + B_{ij}$$

Man sagt auch, die Addition sei *elementweise* erklärt. Falls die beiden Matrizen nicht den selben Typ haben, ist ihre Addition nicht erklärt.

Beispiel. Folgende Addition ist erklärt:

$$\begin{pmatrix} 3 & -2 & 0 & 1 \\ 2 & 8 & -1 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 0 & -1 & 3 \\ -2 & 1 & 9 & 10 \end{pmatrix} = \begin{pmatrix} 8 & -2 & -1 & 4 \\ 0 & 9 & 8 & 14 \end{pmatrix}$$

und wird elementweise berechnet. Um zwei Matrizen vom Typ $m \times n$ zu addieren, sind $m \cdot n$ skalare Additionen erforderlich. ◯

CAS. In Sage werden die arithmetischen Operationen mit den bereits in der Schule üblichen Symbolen bezeichnet.

```
A=matrix([ [3,-2,0,1],[2,8,-1,4] ])
B=matrix([ [5,0,-1,3],[-2,1,9,10] ])
C=A+B; show(C)
```

Sollten die Typen der beiden Symmanden nicht übereinstimmen, erhält man eine Fehlermeldung. ◇

2.2.2 Multiplikation mit Skalaren

Auch die Definition einer Matrix mit einem Skalar bietet keine Überraschung.

Definition. Es sei A eine Matrix und r ein Skalar. Dann verstehen wir unter dem *Produkt* rA diejenige Matrix, die durch Multiplikation jedes Elementes von A mit r entsteht. Es gilt:

$$(rA)_{ij} = rA_{ij}$$

Man sagt auch, die Multiplikation mit r sei *elementweise* erklärt.

Beispiel. Für die Matrix:

$$A = \begin{pmatrix} 2 & -2 & 4 \\ 1 & 3 & -1 \end{pmatrix}$$

ist

$$2A = \begin{pmatrix} 4 & -4 & 8 \\ 2 & 6 & -2 \end{pmatrix} \quad \text{und} \quad (-1)A = \begin{pmatrix} -2 & 2 & -4 \\ -1 & -3 & 1 \end{pmatrix}$$

Beide Vielfachen wurden elementweise berechnet. Um eine Matrizen vom Typ $m \times n$ mit einem Skalar zu multiplizieren, sind $m \cdot n$ skalare Multiplikationen erforderlich. ◯

CAS. Auch die Multiplikation mit einem Skalar wird in Sage erwartungsgemäss berechnet.

```
A=matrix([ [2,-2,4],[1,3,-1] ])
C=2*A; show(C)
D=(-1)*A; show(D)
```

Selbstverständlich ist hier keine Typenverträglichkeit erforderlich. ◇

Wie üblich nennt man $(-1)A$ die *negative Matrix* von A. Man schreibt dafür kurz $-A$ und erklärt die *Differenz* zweier Matrizen durch $A - B = A + (-B)$.

Beispiel. Zum Beispiel ist folgende Differenz erklärt:

$$\begin{pmatrix} 3 & -2 & 0 & 1 \\ 2 & 8 & -1 & 4 \end{pmatrix} - \begin{pmatrix} 5 & 0 & -1 & 3 \\ -2 & 1 & 9 & 10 \end{pmatrix} = \begin{pmatrix} -2 & -2 & 1 & -2 \\ 4 & 7 & -10 & -6 \end{pmatrix}$$

Sie wurde elementweise berechnet. ◯

CAS. Auch Sage kennt diese Konventionen.

```
A=matrix([ [3,-2,0,1],[2,8,-1,4] ])
B=matrix([ [5,0,-1,3],[-2,1,9,10] ])
show(-A)
C=A-B; show(C)
```

Falls die Typen von Minuend und Subtrahend nicht übereinstimmen, erhält man eine Fehlermeldung. ◇

2.2.3 Matrizenprodukt

Die fundamentale Operation der Matrizenrechnung ist vor allem ein weiteres Produkt. Bisher haben wir nur die Multiplikation einer Matrix mit einem Skalar erklärt; nicht aber die Multiplikation zweier Matrizen. Es ist naheliegend anzunehmen, das Produkt zweier Matrizen sei elementweise zu definieren. Es stellt sich jedoch heraus, dass eine andere Definition eines Produktes von zwei Matrizen viel nützlicher ist. Um zu verstehen, wie die Mathematiker auf diese Definition gekommen sind, halten wir uns noch einmal ein allgemeines lineares Gleichungssystem vor Augen:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

Die zugehörige Koeffizientenmatrix A, der Spaltenvektor $\vec{x}$ der Unbekannten und der Spaltenvektor der Konstanten $\vec{b}$ lauten:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ .. \\ b_m \end{pmatrix}$$

Es liegt also nahe, die Multiplikation $A \cdot \vec{x}$ der Matrix A mit dem Vektor $\vec{x}$ so zu erklären, dass das lineare Gleichungssystem in der Form $A \cdot \vec{x} = \vec{b}$ geschrieben werden kann. Dann wäre nämlich das lineare Gleichungssystem zu einer einzigen Matrizengleichung geworden! Es soll also gelten:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ .. \\ b_m \end{pmatrix}$$

Das stimmt nur dann mit dem gegebenen linearen Gleichungssystem überein, falls wir für das links stehende Produkt definieren:

$$\begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}$$

Offenbar wird also das Element in der i-ten Zeile dadurch gebildet, dass man paarweise die Produkte der Elemente der i-ten Zeile von A mit den Elementen

von $\vec{x}$ bildet und dann die entstandenen Produkte addiert. Symbolisch lautet diese Definition:

Definition. Es sei $A \in \mathbb{R}^{m,n}$ eine $m \times n$-Matrix und $\vec{x} \in \mathbb{R}^n$ ein n-dimensionaler Spaltenvektor. Dann verstehen wir unter dem *Produkt* den m-dimensionalen Spaltenvektor $A \cdot \vec{x}$, dessen Elemente durch die Gleichung

$$(A \cdot \vec{x})_i = \sum_{k=1}^{n} a_{ik}\vec{x}_k$$

erklärt sind.

Beispiel. Die folgenden numerischen Beispiele zeigen dieses Verfahren:

$$\begin{pmatrix} 1 & 3 & -5 \\ 3 & -2 & 4 \\ 2 & 3 & -7 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -4 \\ 1 \end{pmatrix} = \begin{pmatrix} -15 \\ 18 \\ -15 \end{pmatrix}$$

$$\begin{pmatrix} 4 & 1 \\ 5 & 1 \\ 6 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} = 11$$

Das letzte Beispiel kann als Skalarprodukt eines Zeilen- mit einem Spaltenvektor kann als Skalarprodukt interpretiert werden. ◯

CAS. Das Produkt einer Matrix mit einem Vektor berechnet man in Sage wie erwartet. Obige drei Beispiele liefert folgender Kode:

```
A=matrix([[1,3,-5],[3,-2,4],[2,3,-7]])
v=vector([2,-4,1])
w=A*v; show(w)
```

```
A=matrix([[4,1],[5,1],[6,1]])
v=vector([1,3])
w=A*v; show(w)
```

```
A=matrix([[2,3,0]])
v=vector([1,3,5])
w=A*v; show(w)
```

Will man im letzten Beispiel keine Matrix vom Typ 1×1, sondern einen Skalar als Ergebnis, erklärt man besser beide Faktoren als Vektoren, um sie dann zu multiplizieren.

```
a=vector([2,3,0])
v=vector([1,3,5])
w=a*v; show(w)
```

Das letzte Ergebnis erhält man auch mit dem Befehl `w=a.dot_product(v)`. ◇

Speziell nützlich ist es zu wissen, was mit den Standardbasisvektoren passiert, wenn man auf sie eine Matrix anwendet.

Beispiel. Für das bereits benutzte numerische Beispiel erhalten wir etwa

$$\begin{pmatrix} 1 & 3 & -5 \\ 3 & -2 & 4 \\ 2 & 3 & -7 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 3 & -5 \\ 3 & -2 & 4 \\ 2 & 3 & -7 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \\ 3 \end{pmatrix}$$

Offenbar filtert das Matrizenprodukt $A \cdot \vec{e}_j = \vec{a}_j$ mit einem Standardbasisvektor gerade die Spaltenvektoren der Matrix A heraus. Umgekehrt sind die Spalten einer Matrix A gerade die Bilder der Standardbasisvektoren unter dem Prozess $\vec{e}_j \mapsto A \cdot \vec{e}_j$. Das hat den unschätzbaren Vorteil, dass der Prozess A vollständig beschrieben ist, falls wir wissen, wie er auf den (endlich vielen!) Standardbasisvektoren wirkt. ◯

Man beachte, dass das Produkt $A \cdot \vec{x}$ einer Matrix A mit einem Spaltenvektor $\vec{x}$ nur dann sinnvoll ist, wenn eine gewisse Kompatibilität zwischen A und $\vec{x}$ erfüllt ist: Die Matrix A muss gleich viele Spalten haben, wie der Vektor $\vec{x}$ Zeilen hat. Sonst sind Input und Operator inkompatibel.

Will man diese Definition des Produktes von den Spaltenvektoren auf beliebige Matrizen als zweiten Faktor verallgemeinern, ist es nützlich, sich vor Augen zu halten, dass eine Matrix nichts anderes als ein „(Zeilen-) Vektor von (Spalten-) Vektoren“ ist. Die $m \times n$-Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m,n}$$

kann nämlich als Vektor $A = (\vec{a}_1 \ \vec{a}_2 \ \cdots \ \vec{a}_n)$ ihrer Spalten aufgefasst werden. Dabei gilt natürlich für den j-ten Spaltenvektor von A

$$\vec{a}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \in \mathbb{R}^m$$

In dieser Form werden Matrizen in vielen Programmiersprachen (z. B. Sage) auch wirklich behandelt. Der Informatiker redet dann auf Neudeutsch von einem Array. Damit meint er allerdings nur den Datentyp und nicht die dazugehörige Struktur, für die wir uns in der Mathematik in erster Linie interessieren.

Um nun die Definition des Matrizenproduktes von zwei Matrizen $A \in \mathbb{R}^{m,n}$ und $B \in \mathbb{R}^{n,r}$ zu finden, denken wir uns den zweiten Faktor, d.h. die Matrix B in dieser Form als Vektor von Vektoren zerlegt. Es gilt also $B = (\vec{b}_1 \ \vec{b}_2 \ \cdots \ \vec{b}_r)$. Das gesuchte Matrizenprodukt $A \cdot B$ entsteht dadurch, dass man die Matrix A der Reihe nach mit jedem Spaltenvektor von B multipliziert und die entstehenden Spaltenvektoren $A \cdot \vec{b}_1, A \cdot \vec{b}_2, \cdots, A \cdot \vec{b}_r$ zu einer neuen Matrix

$$A \cdot B = (A \cdot \vec{b}_1 \ A \cdot \vec{b}_2 \ \cdots \ A \cdot \vec{b}_r)$$

zusammenfasst.

Folgende fundamentale Definition für die Elemente des Produktes von zwei Matrizen A und B fasst diese Idee für das Matrizenprodukt zusammen.

Definition. Es sei $A \in \mathbb{R}^{m,n}$ eine $m \times n$-Matrix und $B \in \mathbb{R}^{n,r}$ eine Matrix vom Typ $n \times r$. Dann verstehen wir unter ihrem *Produkt* die Matrix $A \cdot B \in \mathbb{R}^{m,r}$ vom Typ $m \times r$, deren Elemente durch die Gleichung

$$(A \cdot B)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}$$

erklärt sind, in der über den zweiten Index des ersten Faktors und über den ersten des zweiten Faktors summiert wird.

Um also das Element in der Zeile i und Spalte j von $A \cdot B$ zu finden, multipliziert man paarweise die Elemente der i-ten Zeile von A und der j-ten Spalte von B und addiert die entstehenden Produkte.

Beispiel. Für die folgende 2×3-Matrix A und die 3×4-Matrix B

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \in \mathbb{R}^{2,3}, \qquad B = \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} \in \mathbb{R}^{3,4}$$

ist das Produkt $A \cdot B$ eine 2×4-Matrix. Um beispielsweise das Element in der Zeile 2 und in der Spalte 3 des Produktes zu berechenen, betrachten wir die zweite Zeile von A und die dritte Spalte von B. Dann multiplizieren wir die einander entsprechenden Elemente und addieren diese Produkte. Folgende schematische Darstellung soll das veranschaulichen:

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} = \begin{pmatrix} \square & \square & \square & \square \\ \square & \square & \boxed{26} & \square \end{pmatrix} \in \mathbb{R}^{2,4}$$

Wir rechnen mit Hilfe der zweiten Zeile von A und der dritten Spalten von B.

$$(2 \cdot 4) + (6 \cdot 3) + (0 \cdot 5) = 26$$

Das Element in der ersten Zeile und vierten Spalte des Produktes $A \cdot B$ wird folgendermassen berechnet:

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} = \begin{pmatrix} \square & \square & \square & \boxed{13} \\ \square & \square & \square & \square \end{pmatrix} \in \mathbb{R}^{2,4}$$

Es gilt also:

$$(1 \cdot 3) + (2 \cdot 1) + (4 \cdot 2) \quad = \quad 13$$

Die restlichen Elemente des Produktes ergeben sich aus den Rechnungen

$$\begin{array}{rcl} (1 \cdot 4) + (2 \cdot 0) + (4 \cdot 2) & = & 12 \\ (1 \cdot 1) - (2 \cdot 1) + (4 \cdot 7) & = & 27 \\ (1 \cdot 4) + (2 \cdot 3) + (4 \cdot 5) & = & 30 \\ (2 \cdot 4) + (6 \cdot 0) + (0 \cdot 2) & = & 8 \\ (2 \cdot 1) - (6 \cdot 1) + (0 \cdot 7) & = & -4 \\ (2 \cdot 3) + (6 \cdot 1) + (0 \cdot 2) & = & 12 \end{array}$$

Zusammenfassend gilt also

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} = \begin{pmatrix} 12 & 27 & 30 & 13 \\ 8 & -4 & 26 & 12 \end{pmatrix} \in \mathbb{R}^{2,4}$$

Matrizenprodukte geben offenbar allerhand zu rechnen. In unserem Beispiel waren dazu $m \cdot n \cdot r = 24$ Multiplikationen und $m \cdot (n-1) \cdot r = 16$ Additionen durchzuführen, da $m = 2$, $n = 3$ und $r = 4$ ist. ◯

Definitionsgemäss kann das Matrizenprodukt $A \cdot B$ nur dann gebildet werden, wenn die Anzahl Spalten des ersten Faktors A mit der Anzahl Zeilen des zweiten Faktors B übereinstimmt; anderenfalls ist das Produkt nicht definiert. Ein einfacher Weg festzustellen, ob ein Matrizenprodukt definiert ist, besteht darin, die Typen der beiden Faktoren nebeneinander zu schreiben:

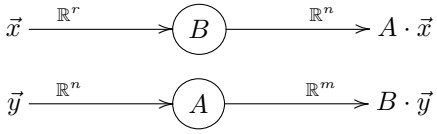
Das Produkt von zwei Matrizen ist genau dann definiert, wenn die inneren beiden Zahlen übereinstimmen. Die beiden äusseren Zahlen geben den Typ des Produktes an.

CAS. Auch die Multiplikation von zwei Matrizen wird in Sage erwartungsgemäss berechnet.

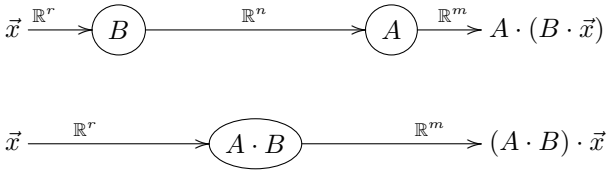
```
A=matrix([ [1,2,4],[2,6,0] ])
B=matrix([ [4,1,4,3],[0,-1,3,1],[2,7,5,2] ])
C=A*B; show(C)
```

Bei einer Typenunverträglichkeit erhält man eine Fehlermeldung. ◇

Interpretieren wir die beiden Matrizen $A \in \mathbb{R}^{m,n}$ und $B \in \mathbb{R}^{n,r}$ als zwei lineare Prozesse



und verketten sie nun zu einem einzigen Prozess, so wird die *Komposition*



durch das Matrizenprodukt $A \cdot B$ beschrieben. Das Matrizenprodukt kann also als Verkettung der beiden Prozesse aufgefasst werden, die durch die beiden Faktoren A und B einzeln beschrieben werden. Es gehört damit zur linearen Abbildung

$$f\colon \mathbb{R}^r \overset{f_B}{\to} \mathbb{R}^n \overset{f_A}{\to} \mathbb{R}^m, \qquad \vec{x} \mapsto f(\vec{x}) = f_A(f_B(\vec{x})) = A \cdot (B \cdot \vec{x}) = (A \cdot B) \cdot \vec{x}$$

Diese Verknüpfung von Prozessen erklärt selbstverständlich auch, warum die beiden Matrizen im erläuterten Sinn kompatibel sein müssen, damit ihr Produkt erklärt ist: die Datentypen müssen an den Schnittstellen übereinstimmen.

Man achte beim Komponieren auf die Reihenfolge, in der die beiden involvierten Teilprozesse B *und dann* A durchgeführt werden, um die Komposition zu erhalten, die zur Matrix $A \cdot B$ gehört! Möchte man die Reihenfolge der Faktoren des Matrizenproduktes $A \cdot B$ sprachlich berücksichtigen, wo wir in unserer Kultur gewohnt sind, von links nach rechts zu lesen, müsste man etwa sagen, dass der Prozess A *nach* dem Prozess B ausgeführt wird.

Beispiel. Um diese Fehlerquelle für immer aus dem Weg zu räumen, betrachten wir ein typisches, konkretes Beispiel. Der erste Prozess werde durch die Matrix

$$B = \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} \in \mathbb{R}^{3,4}$$

beschrieben. Er führt also einen beliebigen Vektor $\vec{x} \in \mathbb{R}^4$ in den Vektor

$$\vec{y} = B \cdot \vec{x} = \begin{pmatrix} 4 & 1 & 4 & 3 \\ 0 & -1 & 3 & 1 \\ 2 & 7 & 5 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4x_1 + x_2 + 4x_3 + 3x_4 \\ -x_2 + 3x_3 + x_4 \\ 2x_1 + 7x_2 + 5x_3 + 2x_4 \end{pmatrix} \in \mathbb{R}^3$$

über. Falls der zweite Prozess durch die Matrix

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \in \mathbb{R}^{2,3}$$

beschrieben wird, geht dabei ein beliebiger Vektor $\vec{y} \in \mathbb{R}^3$ in den Vektor

$$\vec{z} = A \cdot \vec{y} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1 + 2y_2 + 4y_3 \\ 2y_1 + 6y_2 \end{pmatrix} \in \mathbb{R}^2$$

über. Setzt man nun die beiden Prozesse zusammen, so wird der Vektor $\vec{x}$ unter der Komposition insgesamt in den Vektor $\vec{z} = A \cdot (B \cdot \vec{x})$ abgebildet. Er ist durch Substitution der betreffenden Formeln definiert:

$$\begin{pmatrix} (4x_1 + x_2 + 4x_3 + 3x_4) + 2(-x_2 + 3x_3 + x_4) + 4(2x_1 + 7x_2 + 5x_3 + 2x_4) \\ 2(4x_1 + x_2 + 4x_3 + 3x_4) + 6(-x_2 + 3x_3 + x_4) \end{pmatrix}$$

Vereinfachen und Sortieren ergibt daraus den gesuchten Vektor

$$\vec{z} = A \cdot (B \cdot \vec{x}) = \begin{pmatrix} 12x_1 + 27x_2 + 30x_3 + 13x_4 \\ 8x_1 - 4x_2 + 26x_3 + 12x_4 \end{pmatrix} \in \mathbb{R}^2$$

Ein Koeffizientenvergleich zeigt, dass dieser Vektor tatsächlich durch Multiplikation von $\vec{x}$ mit der früher berechnete Produktmatrix

$$A \cdot B = \begin{pmatrix} 12 & 27 & 30 & 13 \\ 8 & -4 & 26 & 12 \end{pmatrix} \in \mathbb{R}^{2,4}$$

beschrieben wird. ◯

Man beachte ferner, dass das Matrizenprodukt als Spezialfall das früher zur Formulierung eines linearen Gleichungssystems erklärte Produkt einer Matrix mit einem Spaltenvektor enthält. Im Spezialfall aus der Schule ist $m = n = r = 1$ und wir können das Produkt von Skalaren, das wegen den wenigen Dimensionen ausnahmsweise von der Reihenfolge unabhängig ist, als Matrizenprodukt interpretieren.

Beispiel. Im Gegensatz zum früher berechneten Skalarprodukt

$$\begin{pmatrix} 2 & 3 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} = 11 \in \mathbb{R}$$

unterscheidet sich das dyadische Produkt

$$\begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} \cdot \begin{pmatrix} 2 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 0 \\ 6 & 9 & 0 \\ 10 & 15 & 0 \end{pmatrix} \in \mathbb{R}^{3,3}$$

in der Reihenfolge der Faktoren. Man beachte, dass das entstehende äussere (dyadische) Produkt der beiden Vektoren in der Regel kein Skalar, sondern eine quadratische Matrix ist. Offensichtlich muss man also beim Multiplizieren von Matrizen, wie beim Komponieren von Prozessen, sorgfältig auf die Reihenfolge der Faktoren achten, weil nicht einmal die Typen der Ergebnisse übereinzustimmen brauchen. ◯

CAS. Um das dyadische Produkt eines Spaltenvektors mit einem Zeilenvektor zu berechnen, muss man die beiden Vektoren mit dem richtigen Typ definieren.

```
v=matrix([[1],[3],[5]])
a=matrix([2,3,0])
w=v*a; show(w)
```

Erklärt man einen der beiden Vektoren einfach kurz als `vector`, funktioniert die Sache im Gegensatz zum Skalarprodukt nicht. ◇

Weil das Matrizenprodukt $C = A \cdot B \in \mathbb{R}^{m,r}$ für zwei Matrizen $A \in \mathbb{R}^{m,n}$ und $B \in \mathbb{R}^{n,r}$ für alle Anwendungen ein zentrale Rolle spielt, stellt sich die Frage, wie man effizient *überprüfen* kann, ob die Berechnung von C korrekt ist, ohne dass man die aufwändige Rechnung wiederholt. Es bietet sich an, dafür einen sogn. probalistischen Algorithmus zu verwenden. Dazu bestimmt man mit Hilfe eines Pseudozufallsgenerators einen Zufallsvektor $\vec{p} \in \mathbb{R}^r$ und berechnet damit die Vektoren $C \cdot \vec{p}$ und $A \cdot (B \cdot \vec{p})$, die selbstverständlich übereinstimmen müssen, falls das Matrizenprodukt $A \cdot B$ richtig berechnet wurde. Falls $C \neq A \cdot B$ sein sollte, wird das allerdings der Algorithmus mit einer Wahrscheinlichkeit

$$\Pr(A \cdot B \cdot \vec{p} = C \cdot \vec{p}) \leq \frac{1}{2}$$

nicht merken. Dieser sogn. Monte-Carlo-Algorithmus wird nun insgesamt k Mal durchgeführt und liefert jedesmal eine korrekte Antwort, wenn das Matrizenprodukt $C = A \cdot B$ stimmt. Falls $C \neq A \cdot B$ sein sollte, wird dieser Algorithmus dies mit einer Wahrscheinlichkeit von mindestens $1 - \left(\frac{1}{2}\right)^k$ zeigen.

Die Versagenswahrscheinlichkeit $\left(\frac{1}{2}\right)^k$ dieses Algorithmus ist beispielsweise bereits für $k = 100$ mit $p = 7.88 \cdot 10^{-31}$ winzig klein. Man müsste ihn rund $2^{100} = 1.26 \cdot 10^{30}$ Mal laufen lassen, bevor er das erste mal versagt. Der Aufwand für dieses Testverfahren beträgt $k \cdot n^2$ und ist damit deutlich geringer als für die Matrizenmultiplikation.

2.2.4 Transposition

Es gibt eine wichtige Matrixoperation, die in der gewöhnlichen Arithmetik kein Pendant hat. Es handelt sich um eine Art innere Symmetrie der Matrizenrechnung, die eine zentrale Rolle spielt und die beiden dualen Seiten der Theorie miteinander in eine enge Beziehung bringt.

Definition. Ist A eine $m \times n$ Matrix, ergibt sich ihre *transponierte Matrix* A^T durch Vertauschen der Zeilen mit den Spalten von A. Die i-te Spalte von A^T ist also die i-te Zeile von A und umgekehrt. Daher ist A^T eine Matrix vom Typ $n \times m$. Für ihre Elemente gilt:

$$(A^T)_{ij} = A_{ji}$$

Die Transposition kann als Abbildung $\_^T \colon \mathbb{R}^{m,n} \to \mathbb{R}^{n,m}$ aufgefasst werden.

Beispiel. Folgende Matrizen entsprechen sich beim Transponieren:

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 6 & 0 \end{pmatrix} \in \mathbb{R}^{2,3}, \qquad A^T = \begin{pmatrix} 1 & 3 \\ 2 & 6 \\ 4 & 0 \end{pmatrix} \in \mathbb{R}^{3,2}$$

Informatiker müssen sich überlegen, wie man die Transposition maschinell effizient durchführt, ohne unnötig Speicherplatz zu vergeuden. ◯

CAS. In Sage erhält man die transponierte Matrix mit dem Befehl

```
A=matrix([ [1,2,4],[2,6,0] ])
A.transpose()
```

Das selbe Ergebnis erhält man auch kürzer mit dem Befehl `A.T`. ◇

Für quadratische Matrizen entspricht das Transponieren einer Spiegelung an der Diagonalen. Für symmetrische Matrizen ist $A = A^T$ und für schiefsymmetrische Matrizen gilt $A = -A^T$. Beim Transponieren gehen Spaltenvektoren in Zeilenvektoren über und umgekehrt. Die Skalare ändern sich beim Transponieren nicht; deshalb hat die Transposition in der Schule keine Rolle gespielt.

Diese Janusköpfigkeit — man spricht von Dualität — der Vektoren, dass sie nämlich in Form von Spalten- oder als Zeilenvektoren in Erscheinung treten, spielt in der linearen Algebra und in ihren Anwendungen eine zentrale Rolle, für deren Verständnis ein waches Auge und etwas Erfahrung erforderlich sein dürfte. Entscheidend ist, dass die beiden Gesichter der Vektorrechnung in einer

engen Beziehung zueinander stehen. Diese Dualität manifestiert sich einerseits im Umstand, dass das dyadische Produkt zweier Standardbasisvektoren

$$\vec{e}_k \cdot \vec{e}_l^{\,T} = B_{kl}, \quad 1 \leq k, l \leq n$$

die Standardbasismatrizen und damit insbesondere die sogn. Auflösung der Identität

$$\sum_{j=1}^{n} \vec{e}_j \cdot \vec{e}_j^{\,T} = \sum_{j=1}^{n} B_{jj} = E_n$$

d.h. eine enge Beziehung zwischen den Standardbasisvektoren und der Einheitsmatrix liefert.

Andererseits kann man diese Dualität mit Hilfe der Transposition als Beziehung zwischen den Spaltenvektoren allein ausdrücken: Einem Paar von zwei Spaltenvektoren $(\vec{a}, \vec{b}) \in \mathbb{R}^n \times \mathbb{R}^n$ kann man nämlich mit Hilfe der Transposition und des Matrizenproduktes auf natürliche Art einen Skalar $\langle \vec{a}, \vec{b} \rangle$ zuordnen.

Definition. Es seien $\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$, $\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$ zwei Spaltenvektoren. Unter ihrem *Skalarprodukt* verstehen wir das Matrizenprodukt

$$\langle \vec{a}, \vec{b} \rangle = \vec{a}^T \cdot \vec{b} = a_1 b_1 + \cdots + a_i b_i + \cdots + a_n b_n = \sum_{i=1}^{n} a_i b_i \in \mathbb{R}$$

Zur Berechnung dieses Skalarproduktes sind n Multiplikationen und $(n-1)$ Additionen durchzuführen. Als wichtigen Spezialfall erwähnen wir die *Norm*

$$N(\vec{a}) = |\vec{a}|^2 = \langle \vec{a}, \vec{a} \rangle = \vec{a}^T \cdot \vec{a} = a_1^2 + \cdots + a_i^2 + \cdots + a_n^2 = \sum_{i=1}^{n} a_i^2 \in \mathbb{R}$$

Sie bezeichnet also das Skalarprodukt eines Vektors $\vec{a}$ mit sich selber.

Man beachte, dass das Skalarprodukt $\langle \vec{a}, \vec{b} \rangle$ zweier Spaltenvektoren ein Skalar und damit der erste Teil seines Namen gerechtfertigt ist. Es handelt sich aber nicht um ein Produkt im üblichen Sinn, weil das Ergebnis des Skalarproduktes *nicht* mehr den selben Datentyp hat wie die Ausgangsdaten. Das ist ein Grund dafür, warum wir das Skalarprodukt mit spitzen Klammern schreiben. Die in Schulbüchern für das Skalarprodukt $\langle \vec{a}, \vec{b} \rangle$ benutze Bezeichnungsweise $\vec{a} \cdot \vec{b}$ vermeiden wir, weil ein solches Matrizenprodukt gar nicht definiert ist und daher diese Notation bei Studenten leicht zu Verwirrung führt. Wie wir gesehen haben, kann man das Skalarprodukt tatsächlich als Matrizenprodukt definieren, braucht dann aber zusätzlich die Transposition.

Das Skalarprodukt zweier Vektoren kann verschwinden, ohne dass einer der beiden Faktoren $\vec{0}$ sein muss.

Beispiel. Das Skalarprodukt der beiden Vektoren

$$\vec{a} = \begin{pmatrix} -1 \\ 3 \\ 2 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$$

verschwindet, obwohl $\vec{a}, \vec{b} \neq \vec{0}$ ist. Solche Vektoren heissen *orthogonal* und wir schreiben in diesem Fall kurz $\vec{a} \perp \vec{b}$. Für die Normen ist $|\vec{a}|^2 = 14$ und $|\vec{b}|^2 = 3$. Nach dem Satz von Pythagoras ist $|\vec{a} + \vec{b}|^2 = 17$. ◯

CAS. In Sage können Skalarprodukt und Norm (Betragsquadrat) wie folgt berechnet werden:

```
a=vector([-1,3,2])
b=vector([-1,-1,1])
show(a*b); show(a*a); show(b*b)
s=a+b; show(s*s)
```

Schulbücher und Sage bezeichnen die Quadratwurzel des Betragsquadrates, d.h. den Betrag als "Norm" und geben ihm damit vor seinem Quadrat den Vorrang. Weil das Betragsquadrat an vielen Orten eine wichtigere Rolle spielt, als der Betrag, den wir auf Grund seiner irrationalen Quadratwurzeln nach Möglichkeit vermeiden werden, ziehen wir die gewählte Sprechweise vor. Sage liefert also mit dem Befehl `a.norm()` den Betrag und nicht die Norm, die stattdessen mit dem Befehl `a.norm()^2` berechnet werden muss. ◇

Ein wichtiges Beispiel paarweise orthogonaler Vektoren ist das n-Bein der Standardbasisvektoren. Für sie gelten also die *Orthogonalitätsrelationen*

$$\langle \vec{e}_i, \vec{e}_j \rangle = \vec{e}_i^T \cdot \vec{e}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Als Folge dieser Relationen lässt sich ein beliebiger Vektor $\vec{a} \in \mathbb{R}^n$ als Linearkombination der Standardbasisvektoren darstellen. Dank der Auflösung der Einheit und der Assoziativität des Matrizenproduktes gilt nämlich

$$\vec{a} = E_n \cdot \vec{a} = \sum_{j=1}^{n} \vec{e}_j \cdot \vec{e}_j^T \cdot \vec{a} = \sum_{j=1}^{n} \vec{e}_j \cdot \langle \vec{e}_j, \vec{a} \rangle \cdot = \sum_{j=1}^{n} \langle \vec{e}_j, \vec{a} \rangle \cdot \vec{e}_j = \sum_{j=1}^{n} a_j \cdot \vec{e}_j$$

wobei der eindeutig bestimmte Koeffizient $a_j \in \mathbb{R}$ durch die *Koeffizientenformel*

$$a_j = \vec{e}_j^T \cdot \vec{a} = \langle \vec{e}_j, \vec{a} \rangle, \quad 1 \leq j \leq n$$

gegeben ist, die Ausgangspunkt für die Fourier-Theorie ist: das Skalarprodukt mit dem j-ten Standardbasisvektoren filtert genau die j-te Komponente heraus.
Die (positive) Norm eines Vekteors $\vec{a}$

$$N \colon \mathbb{R}^n \to \mathbb{R}^{\geq}, \qquad \vec{a} \mapsto N(\vec{a}) = |\vec{a}|^2$$

kann geometrisch als Flächeninhalt des über $\vec{a}$ errichteten Quadrates interpretiert werden.
Aus der Norm eines Vektors $\vec{a}$ ergibt sich durch Wurzelziehen sein *Betrag*

$$|\vec{a}| = \sqrt{\langle \vec{a}, \vec{a} \rangle},$$

der geometrisch als Mass für die Länge von $\vec{a}$ interpretiert werden kann. Um die irrationale Quadratwurzel zu vemeiden, verwendet man allerdings nach Möglichkeit besser statt des Betrags $|\vec{a}|$ sein Quadrat $|\vec{a}|^2 = \langle \vec{a}, \vec{a} \rangle$, d.h. eben die Norm.

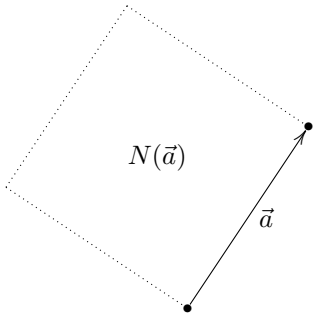


Abbildung 2.1: Norm des Vektors $\vec{a}$ als Flächeninhalt eines Quadrates.

Das Skalarprodukt spielt nicht nur im Zusammenhang mit Filtern eine zentrale Rolle. Wegen des aufgemotzten Satzes von Pythagoras

$$\langle \vec{a}, \vec{b} \rangle = |\vec{a}| \cdot |\vec{b}| \cdot \cos(\gamma)$$

kann das Skalarprodukt

$$\langle \_ , \_ \rangle \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \qquad (\vec{a}, \vec{b}) \mapsto \langle \vec{a}, \vec{b} \rangle$$

zweier Vektoren als Mass für ihre Überlappung interpretiert werden.

Der aufgemotzte Satz von Pythagoras liefert eine Brücke zwischen den rationalen Operationen der linearen Algebra auf der linken Seite und den irrationalen Operationen der Schulgeometrie auf der rechten! Aus dieser für die Euklid'sche Geometrie fundamentalen Beziehung lässt sich nicht nur der Zwischenwinkel $\gamma = \angle(\vec{a}, \vec{b})$ der beiden Vektoren auch in höheren Dimensionen leicht mit Hilfe des Skalarprodukt bestimmen; damit lässt sich die ganze mühsame und mathematisch unbefriedigende Trigonometrie zum Lösen von Massaufgaben umgehen. Der Grund für diese Vereinfachung liegt darin, dass das Skalarprodukt der linken Seite im Gegensatz zu den Wurzel- und Kreisfunktionen auf der rechten Seite eine Reihe einfacher Verträglichkeitseigenschaften hat.

1. Das Skalarprodukt $\langle \vec{a}, \vec{b} \rangle$ ist linear in $\vec{a}$ und $\vec{b}$.
2. Das Skalarprodukt ist symmetrisch: $\langle \vec{a}, \vec{b} \rangle = \langle \vec{b}, \vec{a} \rangle$.
3. Das Skalarprodukt ist regulär: $\langle \vec{a}, \vec{b} \rangle = 0$ für alle $\vec{a} \in \mathbb{R}^n$ impliziert, dass der Vektor $\vec{b} = \vec{0}$ ist.

Die erste Bedingung besagt, dass es sich beim Skalarprodukt um eine Bilinearform auf dem Vektorraum $\mathbb{R}^n$ handelt, die wegen der zweiten Eigenschaft symmetrisch und wegen der dritten regulär ist. Gelegentlich benötigt man sogar die starke Bedingung, dass das Skalarprodukt positiv definit ist.

4. Das reelle Skalarprodukt ist positiv definit: $\langle \vec{a}, \vec{a} \rangle = 0$ impliziert $\vec{a} = \vec{0}$.

Die letzte Bedingung ist für die Skalarprodukte über endlichen Körpern, die vielen Anwendungen eine Rolle spielen, nicht erfüllt und sollte deshalb zurückhaltend benutzt werden.

Dank dieser Verträglichkeitseigenschaft zwischen dem Skalarprodukt und den Grundoperationen eines Vektorraumes darf man mit dem Skalraprodukt weitgehend so rechnen, wie man das aus der Schule von einem Produkt gewohnt ist. Beispielsweise nimmt die erste binomische Formel nun die Form

$$|\vec{a}+\vec{b}|^2 = |\vec{a}|^2 + |\vec{b}|^2 + 2\langle\vec{a},\vec{b}\rangle$$

an. Aus ihr wird ersichtlich, wie sich das Skalarprodukt zweier Vektoren mit Hilfe der Norm ausdrücken lässt. Man erhält so die erste Polarisationsidentität

$$\langle\vec{a},\vec{b}\rangle = \frac{1}{2}\Big(|\vec{a}+\vec{b}|^2 - |\vec{a}|^2 - |\vec{b}|^2\Big) = \frac{1}{2}\big(N(\vec{a}+\vec{b}) - N(\vec{a}) - N(\vec{b})\big)$$

Analog entspricht der zweiten binomische Formel die vektorielle Form

$$|\vec{a}-\vec{b}|^2 = |\vec{a}|^2 + |\vec{b}|^2 - 2\langle\vec{a},\vec{b}\rangle.$$

Daraus erhält man die zweite Polarisationsidentität

$$\langle\vec{a},\vec{b}\rangle = \frac{1}{2}\Big(|\vec{a}|^2 + |\vec{b}|^2 - |\vec{a}-\vec{b}|^2\Big) = \frac{1}{2}\big(N(\vec{a}) + N(\vec{b}) - N(\vec{a}-\vec{b})\big)$$

Man beachte, dass die Polarisationsidentitäten eine unmittelbare geometrische Interpretation des Skalarproduktes als Hinderniss für die Additivität den Norm liefern, die man aus Sicht des Autors in der Schule besser zur geometrischen Definition des Skalarproduktes verwenden würde, statt den dort üblichen Weg mit Hilfe des aufgemotzten Satzes von Pythagoras zu wählen, der eben irrationale Funktionen erfordert.

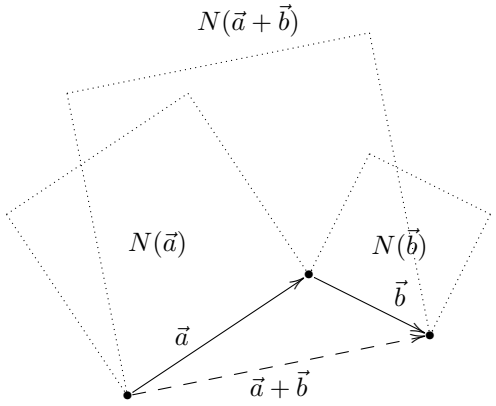


Abbildung 2.2: Skalarprodukt als Hindernis für die Additivität der Norm.

Insbesondere erkennt man an Hand von geometrischen oder numerischen Beispielen leicht, dass die Norm in der Regel nicht additiv sein wird, d.h. dass i.a.

$$N(\vec{a}+\vec{b}) \neq N(\vec{a}) + N(\vec{b})$$

sein wird und damit das Skalarprodukt als Hinderniss in der Regel nicht verschwindet. Ferner vermutet man sofort, dass das nur gerade für orthogonale Vektoren zutrifft, d.h. den Satz von Pythagoras.

Wie wir soeben gesehen haben, erfüllt die Norm keine guten linearen Veträglichkeitseigenschaften! Sie ist nicht nur nicht additiv, sondern auch nicht homogen und erfüllt stattdessen eine *quadratische* Homogenitätsbedingung

$$N(r\vec{a}) = r^2 \cdot N(\vec{a})$$

die sowohl algebraisch als auch geometrisch sofort einleuchtet. Immerhin erlaubt es die Norm, von einem Vektor $\vec{a}$ zu testen, ob er der Nullvektor ist. Es gilt

$$N(\vec{a}) = 0 \text{ genau dann, wenn } \vec{a} = \vec{0}.$$

In dieser Form erfüllt die Norm gelegentlich in Anwendungen nützliche Dienste. Der aus der Norm durch Ziehen der Quadratwurzel abgeleitete Betrag

$$|\vec{a}| := \sqrt{N(\vec{a})} = \sqrt{\langle \vec{a}, \vec{a} \rangle} \geq 0$$

hat folgende Eigenschaften:

1. Für $\vec{a} \in \mathbb{R}^n$ ist $|\vec{a}| = 0$ genau dann, wenn $\vec{a} = \vec{0}$.
2. Für $\vec{a} \in \mathbb{R}^n$ und $r \in \mathbb{R}$ ist $|r\vec{a}| = |r| \cdot |\vec{a}|$.
3. Für $\vec{a}, \vec{b} \in \mathbb{R}^n$ ist $|\vec{a} + \vec{b}| \leq |\vec{a}| + |\vec{b}|$.

Auch der Betrag ist weder additiv noch homogen und kann zum Test herangezogen werden, ob ein Vektor der Nullvektor ist. Die angegebene sogn. Dreiecksungleichung spielt vor allem in der Analysis zum Abschätzen eine wichtige Rolle. Sie ergibt sich durch anwenden der sogn. Schwarzschen Ungleichung

$$\langle \vec{a}, \vec{b} \rangle \leq |\vec{a}| \cdot |\vec{b}|$$

auf die erste binomische Formel

$$|\vec{a} + \vec{b}|^2 = |\vec{a}|^2 + 2\langle \vec{a}, \vec{b} \rangle + |\vec{b}|^2 \leq |\vec{a}|^2 + 2|\vec{a}| \cdot |\vec{b}| + |\vec{b}|^2 = (|\vec{a}| + |\vec{b}|)^2$$

durch Wurzelziehen, das hier umkehrbar ist, weil beide Seiten positiv sind. Um schliesslich die fundamentalere Schwarzsche Ungleichung einzusehen, beachten wir, dass wegen der Positivität der Norm die Ungleichung

$$0 \leq \langle |\vec{b}| \cdot \vec{a} - |\vec{a}| \cdot \vec{b}, |\vec{b}| \cdot \vec{a} - |\vec{a}| \cdot \vec{b} \rangle$$

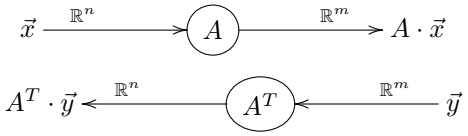
gilt. Daraus wird dank der zweiten binomischen Formel die Ungleichung

$$2 \cdot |\vec{a}| \cdot |\vec{b}| \cdot \langle \vec{a}, \vec{b} \rangle \leq 2 \cdot |\vec{a}|^2 \cdot |\vec{b}|^2$$

und durch Kürzen des positiven Produktes $2 \cdot |\vec{a}| \cdot |\vec{b}|$ die behauptete Ungleichung. Das Gleichheitszeichen tritt in dieser Ungleichung also genau dann auf, wenn $\vec{a}$ und $\vec{b}$ Vielfache sind. Auch die Schwarzsche Ungleichung braucht man in der Analysis immer wieder und sie spielt auch in den Anwendungen — etwa in der Wahrscheinlichkeitstheorie eine wichtige Rolle. In der Quantenmechanik gibt sie etwa Anlass zur fundamentalen Heisenbergschen Unschärferelation.

Die Transposition liefert zu jeder Matrix $A \in \mathbb{R}^{m,n}$ ihre eindeutige Transponierte $A^T \in \mathbb{R}^{n,m}$ und ordnet deshalb jedem linearen Prozess $A\colon \mathbb{R}^n \to \mathbb{R}^m$ einen

eindeutig bestimmten adjungierter Prozess $A^T\colon \mathbb{R}^m \to \mathbb{R}^n$ zu, der in die andere Richtung geht! Dieses adjungierte Paar von Prozessen, das wir schematisch in der Form



ausdrücken, ist eines der Charakteristika der linearen Algebra. Diese Theorie unterscheidet sich deshalb wesentlich von der Theorie **Set** der Mengen und den Funktionen dazwischen. Matrizenrechnung mit ihrer etwas unerwarteten Definition des Matrizenproduktes und der Möglichkeit zur Transposition hat eher den Charakter der symmetrischen Theorie **Rel** der Mengen mit den Relationen[4]. Die Möglichkeit, jedem linearen Prozess seinen transponierten Prozess zuzuordnen, spielt in den Anwendungen der linearen Algebra deshalb eine zentrale und gelegentlich überraschende Rolle, weil in linearen Prozessen dank der Transposition kaum zwischen Input und Output unterschieden werden kann.

Das Skalarprodukt (in unterschiedlichen Dimensionen!) verknüpft die beiden adjungierten Prozesse $A \in \mathbb{R}^{m,n}$ und $A^T \in \mathbb{R}^{n,m}$, weil für alle Vektoren $\vec{x} \in \mathbb{R}^n$ und $\vec{y} \in \mathbb{R}^m$ die fundamentale *Adjunktion*

$$\langle \vec{y}, A\cdot\vec{x}\rangle_{\mathbb{R}^m} = \langle A^T\cdot\vec{y}, \vec{x}\rangle_{\mathbb{R}^n}$$

gilt, die eine Verträglichkeit des Matrizenproduktes mit der Transposition ausdrückt. Das Skalarprodukt klammert also die beiden Seiten der Dualität eng zusammen. Wegen seiner Symmetrie gilt auch die symmetrische Adjunktion

$$\langle \vec{x}, A^T\cdot\vec{y}\rangle_{\mathbb{R}^n} = \langle A\cdot\vec{x}, \vec{y}\rangle_{\mathbb{R}^m}$$

Wir erwarten daher, dass Matrizen, die mit ihrer Transponierten übereinstimmen, d.h. symmetrisch sind, speziell interessante Eigenschaften haben. Auch antisymmetrische Matrizen spielen eine zentrale Rolle, weil sie kontinuierlichen Systemen entsprechen, die Norm-erhaltend oder in der Sprache der Physik Energie-erhaltend sind. Schliesslich beschreiben die orthogonalen Matrizen, für die definitionsgemäss $A\cdot A^T = E$ gilt, die Bewegungen des Euklidschen Raumes.

Diese drei wichtigen Familien von Matrizen erfüllen die allgemeine Bedingung

$$A^T\cdot A = A\cdot A^T$$

die eine enge Beziehung zwischen A und ihrer Transponierten A^T ausdrückt.

Eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ mit dieser Eigenschaft wird als *normal* bezeichnet und kann mit Hilfe der Norm dadurch charakterisiert werden, dass für alle Vektoren $\vec{x} \in \mathbb{R}^n$ die Bedingung

$$|A\cdot\vec{x}|^2 = |A^T\cdot\vec{x}|^2$$

gilt. Der Raum der normalen Matrizen

$$N_n(\mathbb{R}) = \{A \in \mathbb{R}^{n,n} \mid A^T\cdot A = A\cdot A^T\}$$

[4]Eine Relation $f\colon X \rightsquigarrow Y$ kann als Teilmenge von $X\times Y$ beschrieben und als mehrwertige Funktion aufgefasst werden.

enthält also die Teilräume der symmetrischen Matrizen

$$S_n(\mathbb{R}) = \{A \in \mathbb{R}^{n,n} \mid A^T = A\},$$

der antisymmetrischen Matrizen

$$A_n(\mathbb{R}) = \{A \in \mathbb{R}^{n,n} \mid A^T = -A\}$$

und der orthogonalen Matrizen

$$O_n(\mathbb{R}) = \{A \in \mathbb{R}^{n,n} \mid A^T \cdot A = E = A \cdot A^T\}.$$

Für normale Matrizen existiert eine Spektraltheorie, bestehend aus einer Spektralzerlegung und einem Spektralmass.

Viele Anwendungen liefern lineare Gleichungssysteme mit normalen Matrizen. Eine der wichtigsten Quellen für solche Gleichungen ist die sogn. Ausgleichsrechnung, wo es darum geht, für ein gegebenes lineares Gleichungssystem

$$A \cdot \vec{x} = \vec{b}, \qquad A \in \mathbb{R}^{m,n}, \vec{b} \in \mathbb{R}^m$$

das keine Lösung hat, die "beste Ersatzlösung" $\vec{x}^*$ zu bestimmen. Um dabei den "Fehler" möglichst klein zu halten[5], müssen wir den *Residuenvektor*

$$\vec{r}(\vec{x}) = \vec{b} - A \cdot \vec{x} \in \mathbb{R}^m$$

möglichst kurz machen. Wir wählen also $\vec{x}^* \in \mathbb{R}^n$ so, dass der zugehörige Residuenvektor

$$\vec{r}(\vec{x}^*) = \vec{b} - A \cdot \vec{x}^* \in \mathbb{R}^m$$

minimale Norm $N(\vec{x}^*)$ hat. Diese Minimalitätsbedingung für den Residuenvektor ist dank des Satzes von Pythagoras geometrisch gleichwertig damit, dass wir die orthogonale Projektion $\vec{b}^\circ$ von $\vec{b}$ auf den linearen Raum $\mathrm{Im}(A) \subseteq \mathbb{R}^m$, der von den Spaltenvektoren der Matrix A aufgespannt wird[6], bestimmen und dann das lineare Ersatzsystem

$$A \cdot \vec{x}^* = \vec{b}^\circ$$

lösen.

Um nun $\vec{x}^*$, die Orthogonalprojektion $\vec{b}^\circ$ und den minimalen Residuenvektor $r(\vec{x}^*)$ matriziell bestimmen zu können, erinnern wir uns, dass es darum geht, jenen Vektor $A \cdot \vec{x}^*$ aus dem Bild $\mathrm{Im}(A)$ zu bestimmen, für den der Residuenvektor $\vec{r}(\vec{x}^*)$ orthogonal auf $\mathrm{Im}(A)$ steht. Das heisst also, dass $\vec{r}(\vec{x}^*)$ zu allen Vektoren der Form $A \cdot \vec{x}$ orthogonal sein soll. Aus der Orthogonalitätsbedingung

$$\langle A \cdot \vec{x}, \vec{r}(\vec{x}^*)\rangle = 0, \qquad \forall \vec{x} \in \mathbb{R}^n$$

wird dank der Adjunktion die äquivalente Bedingung, dass für alle $\vec{x} \in \mathbb{R}^n$

$$\langle \vec{x}, A^T \cdot \vec{r}(\vec{x}^*)\rangle = 0, \qquad \forall \vec{x} \in \mathbb{R}^n$$

[5] Deshalb redet man manchmal auch von der "Methode der kleinsten Fehlerquadrate".

[6] sogn. Bild von A. Es besteht aus den Linearkombinationen der Spaltenvektoren von A.

gelten soll. Weil dank der Regularität des Skalarproduktes nur der Nullvektor zu *allen* Vektoren von $\mathbb{R}^n$ orthogonal ist, muss also

$$A^T \cdot \vec{r}(\vec{x}^*) = \vec{0}$$

sein. Einsetzen von $\vec{r}(\vec{x}^*) = \vec{b} - A \cdot \vec{x}^*$ liefert für $\vec{x}^*$ die gesuchte Bedingung

$$A^T \cdot \vec{r}(\vec{x}^*) = A^T \cdot (\vec{b} - A \cdot \vec{x}^*) = 0$$

die umgeformt die Form

$$(A^T \cdot A) \cdot \vec{x}^* = A^T \cdot \vec{b}$$

eines linearen Gleichungssystems — der sogn. *Normalengleichungen* — annimmt. Es kann mit Hilfe der Inversen explizit durch

$$\vec{x}^* = (A^T \cdot A)^{-1} \cdot A^T \cdot \vec{b}$$

gelöst werden.

Man kann sich leicht überzeugen, dass die Koeffizientenmatrix $A^T \cdot A$ dieses linearen Gleichungssystems symmetrisch und damit in der Tat normal ist. Man kann zeigen, dass eine solche sogn. Gram-Matrix der Form $A^T \cdot A$, deren Einträge die Standardprodukte der Spaltenvektoren von A sind, immer positv definit ist, d.h. dass sie immer dann invertierbar ist, falls die Spaltenvektoren von A linear unabhängig sind.

Wie in der Einleitung angedeutet, könnte man den optimalen Vektor $\vec{x}^*$ selbstverständlich auch mit Hilfe von Differentialrechnung finden, indem man die Norm $\langle \vec{r}(x), \vec{r}(x) \rangle = |\vec{r}(x)|^2$ des Residuenvektors nach allen Komponenten von $\vec{x}$ ableitet und dann 0 setzt. In der linearen Algebra macht aber der Satz von Pythagoras nicht nur die Differentialrechnung, sondern auch den fälligen Nachweis für die Existenz und die Eindeutigkeit des Minimums obsolet.

Aus dieser Formel ergibt sich, dass die Orthogonalprojektion $\vec{b}^\circ$ von $\vec{b}$ auf das Bild $\mathrm{Im}(A)$ der Matrix A wegen der gefundenen Beziehung

$$\vec{b}^\circ = A \cdot \vec{x}^* = A \cdot (A^T \cdot A)^{-1} \cdot A^T \cdot \vec{b}$$

durch Multiplikation mit der Projektionsmatrix

$$\mathrm{Proj}_{\mathrm{Im}(A)} = A \cdot (A^T \cdot A)^{-1} \cdot A^T$$

erhalten werden kann. Für den Residuenvektor minimaler Länge gilt

$$\vec{r}(\vec{x}^*) = \vec{b} - A\vec{x}^* = \vec{b} - A \cdot (A^T \cdot A)^{-1} \cdot A^T \cdot \vec{b}$$

Er kann also als Produkt von $\vec{b}$ mit der Projektionsmatrix

$$\mathrm{Proj}_{\mathrm{Im}(A)^\perp} = E - \mathrm{Proj}_{\mathrm{Im}(A)} = E - A \cdot (A^T \cdot A)^{-1} \cdot A^T$$

auf den Orthogonalraum $\mathrm{Im}(A)^\perp$ von $\mathrm{Im}(A)$ erhalten werden.

Die soeben bestimmten Formeln liefern auch eine Beschreibung für die Spiegelung von $\vec{b}$ am Teilraum $\mathrm{Im}(A)$. Ein Blick auf die geometrische Situation zeigt nämlich, dass für diese Spiegelung

$$\begin{aligned} S_{\mathrm{Im}(A)}(\vec{b}) &= \vec{b}^\circ - \vec{r}(\vec{x}^*) = \vec{b}^\circ - (\vec{b} - A \cdot \vec{x}^*) = \vec{b}^\circ - \vec{b} + A \cdot \vec{x}^* = 2\vec{b}^\circ - \vec{b} \\ &= 2\mathrm{Proj}_{\mathrm{Im}(A)} \cdot \vec{b} - \vec{b} = (2\mathrm{Proj}_{\mathrm{Im}(A)} - E) \cdot \vec{b} \end{aligned}$$

gilt. Die Spiegelung am Teilraum $\mathrm{Im}(A)$ kann also durch Multiplikation mit der Matrix

$$S_{\mathrm{Im}(A)} = 2\mathrm{Proj}_{\mathrm{Im}(A)} - E = 2A \cdot (A^T \cdot A)^{-1} \cdot A^T - E$$

beschrieben werden.

Beispiel. Im allereinfachsten Fall besteht die Matrix A aus einem einzigen Spaltenvektor $\vec{a} \neq \vec{0}$. Dann wird die Matrix $A^T \cdot A$ zur Norm $\langle \vec{a}, \vec{a} \rangle$ und die Projektionsmatrix vereinfacht sich zum dyadischen Produkt des normierten Vektors

$$\mathrm{Proj}_{\vec{a}} = \frac{1}{\langle \vec{a}, \vec{a} \rangle} \vec{a} \cdot \vec{a}^T$$

Die Orthogonalprojektion eines beliebigen Vektors $\vec{b}$ auf die vom Vektor $\vec{a}$ aufgespannte Gerade durch den Ursprung wird also durch die *Projektionsformel*

$$\mathrm{Proj}_{\vec{a}}(\vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{\langle \vec{a}, \vec{a} \rangle} \vec{a}$$

beschrieben.

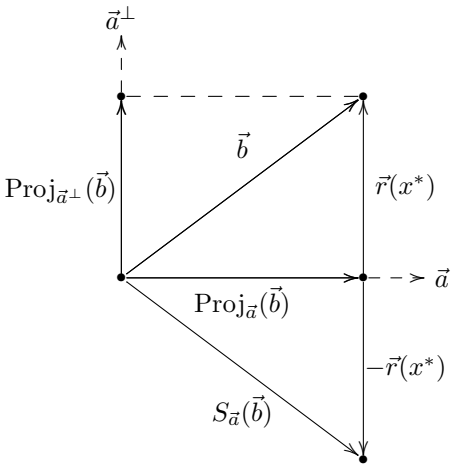


Abbildung 2.3: Orthogonalprojektion von $\vec{b}$ auf den Vektor $\vec{a}$.

Die Orthogonalprojektion eines beliebigen Vektors $\vec{b}$ auf den Orthogonalraum $\vec{a}^{\perp}$ wird durch die Matrix

$$\mathrm{Proj}_{\vec{a}^{\perp}} = E - \frac{1}{\langle \vec{a}, \vec{a} \rangle} \vec{a} \cdot \vec{a}^T$$

beschrieben.

Das zugehörige Ausgleichsproblem tritt dann auf, wenn in einer Proportionalität der Form (man denke an das Hookesche oder Ohmsche Gesetz)

$$a \cdot x = b$$

die Grössen a und b mehrfach gemessen werden und mit Hilfe dieser Messungen ein optimaler Wert für x bestimmt werden soll. Fasst man die Messwerte von a

zum Spaltenvektor $\vec{a}$ und die Messwerte von b zum Spaltenvektor $\vec{b}$ zusammen, so wird der optimale Wert durch den Quotienten

$$x^* = \frac{\langle \vec{a}, \vec{b} \rangle}{\langle \vec{a}, \vec{a} \rangle}$$

gegeben. Er ist nur dann das arithmetische Mittel der Komponenten von $\vec{b}$, wenn die Komponenten von $\vec{a}$ alle 1 sind.

Der zugehörige Residuenvektor hat dann die Form

$$\vec{r}(x^*) = \vec{b} - \frac{\langle \vec{a}, \vec{b} \rangle}{\langle \vec{a}, \vec{a} \rangle} \vec{a}$$

und ist orthogonal zu $\vec{b}$, wie man leicht nachrechnet. Seine Norm ist

$$|\vec{r}(x^*)|^2 = \langle \vec{b}, \vec{b} \rangle - 2\frac{\langle \vec{a}, \vec{b} \rangle^2}{\langle \vec{a}, \vec{a} \rangle} + \frac{\langle \vec{a}, \vec{b} \rangle^2}{\langle \vec{a}, \vec{a} \rangle} = \langle \vec{b}, \vec{b} \rangle - \frac{\langle \vec{a}, \vec{b} \rangle^2}{\langle \vec{a}, \vec{a} \rangle} \geq 0$$

Daraus ergibt sich durch Umstellen die Ungleichung

$$\langle \vec{a}, \vec{b} \rangle^2 \leq \langle \vec{a}, \vec{a} \rangle \cdot \langle \vec{b}, \vec{b} \rangle$$

die sich auch aus der Schwarzschen Ungleichung

$$\langle \vec{a}, \vec{b} \rangle \leq |\vec{a}| \cdot |\vec{b}|$$

durch quadrieren ergibt. ◯

Beispiel. Häufig wird man auf eine Hyperebene H von $\mathbb{R}^n$ projizieren oder an ihr spiegeln wollen. Dann wird also das Bild $\mathrm{Im}(A) = H$ von $n-1$ linear unabhängigen Vektoren $\vec{a}_1, \ldots, \vec{a}_{n-1} \in \mathbb{R}^n$ aufgespannt, die dann die Spalten von A bilden.

Eine solche Hyperebene H kann auch mit Hilfe eines Normalenvektors $\vec{n}$ beschrieben, der also die Bedingungen

$$\langle \vec{a}_i, \vec{n} \rangle = 0, \qquad 1 \leq i \leq n-1$$

erfüllt. In diesem Fall wird der Orthogonalraum $\mathrm{Im}(A)^\perp = H^\perp$ durch den einzigen Vektor $\vec{n}$ aufgespannt. Die Orthogonalprojektion auf die Hyperebene $H = \vec{n}^\perp$ in Richtung $\vec{n}$ wird also nach dem letzten Beispiel durch die Matrix

$$\mathrm{Proj}_{\vec{n}^\perp} = E - \frac{1}{\langle \vec{n}, \vec{n} \rangle} \vec{n} \cdot \vec{n}^T$$

beschrieben. Die Spiegelung an der Hyperebene $H = \vec{n}^\perp$ wird dann entsprechend durch die Spiegelungsformel

$$S_{\vec{n}^\perp} = 2\mathrm{Proj}_{\vec{n}^\perp} - E = E - \frac{2}{\langle \vec{n}, \vec{n} \rangle} \vec{n} \cdot \vec{n}^T$$

beschrieben. ◯

Beispiel. Das konkrete lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ für

$$A = \begin{pmatrix} 3 & 4 \\ 3 & -1 \\ 2 & 1 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} 0 \\ 10 \\ 7 \end{pmatrix},$$

ist unlösbar. Addieren wir nämlich in der zugehörigen erweiterten Matrix

$$\left(\begin{array}{cc|c} 3 & 4 & 0 \\ 3 & -1 & 10 \\ 2 & 1 & 7 \end{array}\right)$$

das (-1)-fache der ersten Zeile zur zweiten und das (-2)-fache der erste Zeile zum 3-fachen der dritten, erhalten wir die Matrix

$$\left(\begin{array}{cc|c} 3 & 4 & 0 \\ 0 & -5 & 10 \\ 0 & -5 & 21 \end{array}\right), \qquad \left(\begin{array}{cc|c} 3 & 4 & 0 \\ 0 & -5 & 10 \\ 0 & 0 & 11 \end{array}\right)$$

aus der wir diesen Sachverhalt bereits ablesen können. Ganz klar wird die Sache sicher dann, wenn wir die Stufenform herstellen, indem wir noch das (-1)fache der zweiten Zeile zur dritten dazuzählen und dann beachten, dass wir keine reelle Zahl y finden werden, die mit 0 multipliziert 11 ergibt! Mit der linearen Gleichung $0 \cdot y = 11$ sind auf ein unüberwindbares Hindernis für die Lösbarkeit des urprünglichen linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ gestossen.
In diesem Beispiel gilt

$$A^T \cdot A = \begin{pmatrix} 22 & 11 \\ 11 & 18 \end{pmatrix}, \qquad A^T \cdot \vec{b} = \begin{pmatrix} 44 \\ -3 \end{pmatrix}$$

und wie wir später sehen werden

$$(A^T \cdot A)^{-1} = \frac{1}{275} \begin{pmatrix} 18 & -11 \\ -11 & 22 \end{pmatrix}$$

Das System der Normalengleichungen mit der erweiterten Matrix

$$\left(\begin{array}{cc|c} 22 & 11 & 44 \\ 11 & 18 & -3 \end{array}\right) \qquad \text{bzw.} \qquad \left(\begin{array}{cc|c} 2 & 1 & 4 \\ 11 & 18 & -3 \end{array}\right)$$

hat in diesem Fall die eindeutige Lösung

$$\vec{x}^* = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

Die Orthogonalprojektion auf die Ebene $\mathrm{Im}(A)$ durch den Ursprung, die von den beiden Spaltenvektoren von A aufgespannt wird, lässt sich durch die Matrix

$$\mathrm{Proj}_{\mathrm{Im}(A)} = \frac{1}{11} \begin{pmatrix} 10 & -1 & 3 \\ -1 & 10 & 3 \\ 3 & 3 & 2 \end{pmatrix}$$

beschreiben. Man beachte, dass diese Matrix zwei spezielle Eigenschaften hat. Sie ist symmetrisch und idempotent, d.h. sie stimmt mit ihrem Quadrat überein.

Matrizen mit diesen beiden Eigenschaften heissen Projektoren und beschreiben genau die Orthogonalprojektionen auf einen Teilraum W von $\mathbb{R}^n$.

Mit diesem Projektor lässt sich also die Orthogonalprojektion des Vektors $\vec{b}$ auf die von $\vec{a}_1$ und $\vec{a}_2$ aufgespannte Ebene $\mathrm{Im}(A)$ durch den Ursprung mit Hilfe von

$$\vec{b}^{\circ} = \mathrm{Proj}_{\mathrm{Im}(A)} \cdot \vec{b} = \begin{pmatrix} 1 \\ 11 \\ 4 \end{pmatrix}$$

leicht bestimmen. Dieser Vektor ist tatsächlich im Bild von A, da er als Linearkombination der beiden Spaltenvektoren $\vec{a}_1$ und $\vec{a}_2$ von A dargestellt werden kann, da nämlich gilt:

$$\vec{b}^{\circ} = 3\vec{a}_1 - 2\vec{a}_2.$$

Das lineare Ersatzsystem

$$A \cdot \vec{x}^{\,*} = \vec{b}^{\circ}$$

hat auch tatsächlich die oben gefundene eindeutige Lösung $\vec{x}^*$, wie man leicht nachrechnet.

Die Projektion auf den Orthogonalraum wird durch die Matrix

$$\mathrm{Proj}_{\mathrm{Im}(A)^{\perp}} = E - \mathrm{Proj}_{\mathrm{Im}(A)} = \frac{1}{11} \begin{pmatrix} 1 & 1 & -3 \\ 1 & 1 & -3 \\ -3 & -3 & 9 \end{pmatrix}$$

beschrieben, die also auch symmetrisch und idempotent ist.

Der Residuenvektor minimaler Länge

$$\vec{r}(\vec{x}^{\,*}) = \vec{b} - \vec{b}^{\circ} = \mathrm{Proj}_{\mathrm{Im}(A)^{\perp}} \cdot \vec{b} = \begin{pmatrix} -1 \\ -1 \\ 3 \end{pmatrix}$$

ist in der Tat orthogonal zum Bild von A, weil sein Skalarprodukt mit den beiden Spaltenvektoren $\vec{a}_1$ und $\vec{a}_2$ von A verschwindet. Seine Norm $N(\vec{r}(\vec{x}^{\,*})) = 11$ bzw. sein Betrag $|\vec{r}(\vec{x}^{\,*})| = \sqrt{11}$ liefert ein Mass für den gemachten Fehler.

Schliesslich wird die Spiegelung an der von den beiden Vektoren $\vec{a}_1$ und $\vec{a}_2$ aufgespannten Ebene durch die Matrix

$$S_{\mathrm{Im}(A)} = 2\mathrm{Proj}_{\mathrm{Im}(A)} - E = \frac{1}{11} \begin{pmatrix} 9 & -2 & 6 \\ -2 & 9 & 6 \\ 6 & 6 & -7 \end{pmatrix}$$

beschrieben. Sie hat neben der Symmetrie eine weitere wichtige Eigenschaft. Sie ist nämlich involutiv, d.h. ihr Quadrat ist die Einheitsmatrix. Matrizen mit diesen beiden Eigenschaften beschreiben Spiegelungen an einem Teilraum $W \subseteq \mathbb{R}^n$. Als Konsequenz davon bilden ihre Spaltenvektoren ein orthonormiertes (negativ orientiertes) Dreibein.

All das lässt sich auch erhalten, wenn man beachtet, dass ein Vektor in Richtung des Residuenvektors minimaler Länge

$$\vec{n} = \vec{r}(\vec{x}^{\,*}) = \begin{pmatrix} -1 \\ -1 \\ 3 \end{pmatrix}, \qquad \langle \vec{n}, \vec{n} \rangle = 11$$

als Normalenvektor der Ebene $\mathrm{Im}(A)$ gewählt werden kann. Die Projektion auf diese Ebene kann deshalb auch durch die Matrix

$$\begin{aligned}\mathrm{Proj}_{\vec{n}^\perp} &= E - \frac{1}{\langle \vec{n}, \vec{n}\rangle}\vec{n}\cdot\vec{n}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{11}\begin{pmatrix} 1 & 1 & -3 \\ 1 & 1 & -3 \\ -3 & -3 & 9 \end{pmatrix} \\ &= \frac{1}{11}\begin{pmatrix} 10 & -1 & 3 \\ -1 & 10 & 3 \\ 3 & 3 & 2 \end{pmatrix}\end{aligned}$$

erhalten werden, die selbstverständlich mit der früher erhaltenen übereinstimmt. Die Spiegelung an dieser Ebene ergibt sich auch aus der Spiegelungsformel

$$S_{\vec{n}^\perp} = 2\mathrm{Proj}_{\vec{n}^\perp} - E = \frac{1}{11}\begin{pmatrix} 9 & -2 & 6 \\ -2 & 9 & 6 \\ 6 & 6 & -7 \end{pmatrix}$$

und ist uns auch bereits bekannt. Je nach beabsichtigter Anwendung wird man eine Hyperebene besser durch einen Normalenvektor beschreiben bzw. dual mit Hilfe eines Systems linear unabhängiger Vektoren aufspannen. ◯

Genau so, wie sich lineare Gleichungssysteme auf zwei Arten interpretieren lassen, kann man das Matrizenprodukt auf zwei duale Arten verstehen, die beide eine zentrale Rolle spielen. Zunächst betrachten wir nochmals den Spezialfall des Matrizenproduktes $A\cdot\vec{x}$ einer Matrix mit einem Vektor am numerischen Beispiel

$$A\cdot\vec{x} = \begin{pmatrix} 1 & 3 & -5 \\ 3 & -2 & 4 \end{pmatrix}\cdot\begin{pmatrix} 2 \\ -4 \\ 1 \end{pmatrix} = \begin{pmatrix} -15 \\ 18 \end{pmatrix}$$

In der *zeilenweisen* Interpretation des Matrizenproduktes fassen wir die Zeilen des linken Faktors A als Zeilenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 & 3 & -5 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 3 & -2 & 4 \end{pmatrix}$$

auf. Dann kann das Matrizenprodukt als *Skalarenprodukte der Zeilen* von A mit dem Vektor $\vec{x}$ aufgefasst werden, da $\langle\vec{v}_1,\vec{x}\rangle = -15$ und $\langle\vec{v}_2,\vec{x}\rangle = 18$ gilt.

In der *spaltenweisen* Interpretation des Matrizenproduktes fassen wir die Spalten der Matrix A als Vektoren auf.

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad \vec{a}_2 = \begin{pmatrix} 3 \\ -2 \end{pmatrix}, \qquad \vec{a}_3 = \begin{pmatrix} -5 \\ 4 \end{pmatrix}$$

Dann besteht das Matrizenprodukt aus einer *Linearkombination der Spalten* von A mit den Komponenten des Vektors $\vec{x}$ als Gewichte, weil gilt:

$$2\vec{a}_1 - 4\vec{a}_2 + 1\vec{a}_3 = 2\begin{pmatrix} 1 \\ 3 \end{pmatrix} - 4\begin{pmatrix} 3 \\ -2 \end{pmatrix} + 1\begin{pmatrix} -5 \\ 4 \end{pmatrix} = \begin{pmatrix} -15 \\ 18 \end{pmatrix}$$

Diese duale Auffassung lässt sich nun auf ein beliebiges Matrizenprodukt verallgemeinern. Das Matrizenprodukt $A\cdot B$ einer $m\times r$ Matrix A mit einer $r\times n$-Matrix B besteht in der zeilenweisen Interpretation aus $m\cdot n$ vielen Skalarprodukten der m Zeilen von A mit den n Spalten von B. Im Gegensatz dazu

entsteht es bei der spaltenweisen Interpretation aus n Linearkombination der r Spaltenvektoren von A mit den Gewichtsfaktoren aus den Spaltenvektoren von B. Dem Leser wird empfohlen, sich diese duale Sichtweise des Matrizenproduktes nochmals an einem Beispiel zu veranschaulichen und es sich gut einzuprägen.

Offenbar handelt es sich auch bei *Linearkombination* eines Systems von Vektoren $\vec{a}_1, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ um ein zentrales Konzept der linearen Algebra. Darunter verstehen wir einen Vektor $\vec{v} \in \mathbb{R}^n$, der sich als Vielfachsumme der gegebenen Vektoren in der Form

$$\sum_{j=1}^{k} x_j \vec{a}_j$$

darstellen lässt. Intuitiv wird dadurch der Vektor $\vec{v}$ aus den Vektoren des Systems $\{\vec{a}_j\}_{j \in J}$ zusammengebaut.

Es kann nun sein, dass einer der Vektoren des gegebenen Systems $\vec{a}_1, \ldots, \vec{a}_k$ bereits aus den übrigen zusammengebaut werden kann und daher als Grundbaustein für den Zusammenbau weiterer Vektoren unnötig ist. Die folgende Definition beschreibt diese Sitation in möglichst symmetrischer Form:

Definition. Ein System von Vektoren $\vec{a}_1, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ heisst *linear unabhängig*, falls die Koeffizienten x_j einer verschwindenden Linearkombination

$$\sum_{j=1}^{k} x_j \vec{a}_j = \vec{0}$$

alle trivial, d.h. $x_j = 0$ sein müssen. Sonst heisst es linear abhängig.

Diese Bedingung ist genau dann erfüllt, wenn keiner der Vektoren des Systems als Linearkombination der übrigen darstellbar ist. Die Vektoren des Systems sind also genau dann linear abhängig, wenn einer von ihnen als Linearkombination der übrigen dargestellt werden kann. Lineare Abhängigkeit lässt sich äquivalent mit Hilfe des homogenen linearen Gleichungssystems $A \cdot \vec{x} = \vec{0}$ formulieren, dessen Koeffizientenmatrix A die Spaltenvektoren $\vec{a}_j$ hat und besagt dann, dass dieses Gleichungssystem nichttriviale Lösungen hat, die zur Formulierung der linearen Abhängigkeiten zwischen den Vektoren des Systems benutzt werden können.

Beispiel. Die Untersuchung der linearen Abhängikeit des Systems der Vektoren aus $\mathbb{R}^4$, die durch die vier Spaltenvektoren

$$\vec{a}_1 = \begin{pmatrix} 1 \\ a \\ -2 \\ 1 \end{pmatrix}, \quad \vec{a}_2 = \begin{pmatrix} 1 \\ 1 \\ -4 \\ 2 \end{pmatrix}, \quad \vec{a}_3 = \begin{pmatrix} 0 \\ 0 \\ 2 \\ -1 \end{pmatrix}, \quad \vec{a}_4 = \begin{pmatrix} -2 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

gegeben sind, läuft auf die Untersuchung der Vektorgleichung

$$x_1\vec{a} + x_2\vec{a}_2 + x_3\vec{a}_3 + x_4\vec{a}_4 = \vec{0}$$

hinaus, die sich matriziell als homogenes, lineares Gleichungssystem der Form $A \cdot \vec{x} = \vec{0}$ präsentiert, dessen Koeffizientenmatrix A als Spalten die vier gegebenen

Vektoren hat und daher

$$A = \begin{pmatrix} 1 & 1 & 0 & -2 \\ a & 1 & 0 & 0 \\ -2 & -4 & 2 & 1 \\ 1 & 2 & -1 & 1 \end{pmatrix}$$

lautet. Die vier Vektoren in einem Raum der Dimension 4 werden für einen typischen Parameter $a \in \mathbb{R}$ linear unabhängig sein und damit den ganzen 4-dimensionalen Raum aufspannen. Nur für gewisse (wenige!) Wahlen von a werden diese vier Vektoren linear abhängig sein und dann einen Raum niedrigerer Dimension auspannen. Es geht darum, diese a's zu bestimmen.

Um die linearen Abhängigkeiten zwischen diesen Vektoren zu finden, untersuchen wir das homogene lineare Gleichungssystem $A \cdot \vec{x} = \vec{0}$ mit der Koeffizientenmatrix A und bestimmen seine Lösungsmenge.

Addition des $(-a)$-fachen der ersten zur zweiten Zeile, des 2-fachen der ersten zur dritten Zeile und des (-1)-fachen der ersten zur vierten Zeile liefert

$$\begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1-a & 0 & 2a \\ 0 & -2 & 2 & -3 \\ 0 & 1 & -1 & 3 \end{pmatrix}$$

Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die zweite mit der vierten Zeile und erhalten die Matrix

$$\begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & -2 & 2 & -3 \\ 0 & 1-a & 0 & 2a \end{pmatrix}$$

Addition des 2-fachen der zweiten Zeile zur dritten und des $(a-1)$-fachen der zweiten Zeile zur vierten liefert die Matrix

$$\begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 1-a & 5a-3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 1-a & 5a-3 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

deren dritte und vierte Zeile wir noch vertauscht haben. Aus dieser Matrix entnehmen wir, dass $a = 1$ sein muss, damit das Gleichungssystem nichttriviale Lösungen haben kann. Die vier gegebenen Vektoren sind also genau dann linear abhängig, falls $a = 1$ ist.

Um alle möglichen linearen Abhängigkeiten zwischen den vier Vektoren zu bestimmen, suchen wir nun die Lösungen des homogenen linearen Systems und lösen dazu das Gleichungssystem für $a = 1$ zu Ende. Die gefundene Matrix hat dann die Gestalt

$$\begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

Addieren des (-3)-fachen der dritten Zeile zum 2-fachen der vierten liefert die Stufenmatrix

$$\begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

mit Hilfe der man bestätigt, dass die Matrix A für $a = 1$ in der Tat nur 3 unabhängige Zeilen hat. Die vier Vektoren spannen dann also nur einen 3-dimensionalen Unterraum auf. Um mit möglichst einfach Zahlen weiterrechnen zu können, haben wir noch die dritte Zeile durch 2 dividiert.

Addition des (-3)-fachen der dritten Zeile zur zweiten und des 2-fachen der dritten Zeile zur ersten liefert die Matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Addition des (-1)-fachen der zweiten Zeile zur ersten liefert die reduzierte Stufenmatrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

aus der wir nun die allgemeine Lösung ablesen können. Wählen wir für den freien Parameter $x_3 = t$, erhalten wir nämlich $x_1 = -t$, $x_2 = t$ und $x_4 = 0$ bzw. in vektorieller Form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = t \begin{pmatrix} -1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \qquad t \in \mathbb{R}$$

Offenbar gibt es also im Sonderfall $a = 1$ eine 1-dimensionale Lösungsschar für unser Gleichungssystem und im Normallfall $a \neq 1$ hat es nur die triviale Lösung.

Die zugehörige nichttriviale Darstellung des Nullvektors als Linearkombination ist bis auf Vielfache eindeutig bestimmt und lautet

$$-\vec{a}_1 + \vec{a}_2 + \vec{a}_3 = \vec{0}$$

Diese nicht triviale Vektorgleichung drückt also eine lineare Beziehung (bis auf Vielfache die einzig mögliche!) zwischen den vier Vektoren aus.

Offenbar sind für $a = 1$ schon die ersten drei Vektoren $\vec{a}_1$, $\vec{a}_2$ und $\vec{a}_3$ linear abhänging, weil der vierte Vektor in der gefundenen Beziehung gar nicht vorkommt. Mit Hilfe der gefundenen linearen Beziehungen zwischen den Vektoren lässt sich durch Auflösen jeder der ersten drei Vektoren als Linearkombination der beiden anderen ausdrücken,

$$\begin{aligned} \vec{a}_1 &= \vec{a}_2 + \vec{a}_3 \\ \vec{a}_2 &= \vec{a}_1 - \vec{a}_3 \\ \vec{a}_3 &= \vec{a}_1 - \vec{a}_2 \end{aligned}$$

wie man an Hand der gegebenen Vektoren numerisch überprüfen kann.

Es ist aber unmöglich, den Vektor $\vec{a}_4$ als Linearkombination der restlichen Vektoren darzustellen.

Falls man nur die lineare Abhängigkeit von $\vec{a}_1$ von $\vec{a}_2$, $\vec{a}_3$ und $\vec{a}_4$ untersuchen will, macht man den Ansatz

$$\vec{a}_1 = t_2\vec{a}_2 + t_3\vec{a}_3 + t_4\vec{a}_4$$

und untersucht das entstehende lineare Gleichungssystem auf Lösbarkeit. ◯

Die Grundoperationen der Matrizenrechnung d.h. Linearkombinationen, Skalar- und Matrizenprodukte sind uns aus dem täglichen Leben vertraut. Dort spielen nämlich oft indizierte Mengenangaben in Form von Listen eine Rolle, mit denen dann gerechnet wird.

Beispiel. Laut Kochbuch braucht man für einen Butterzopf bzw. für ein Brot die folgenden Zutaten:

$$\begin{array}{l}\text{Mehl [gr]}\\ \text{Salz [gr]}\\ \text{Butter [gr]}\\ \text{Hefe [gr]}\\ \text{Milch [dl]}\end{array} \quad : \quad \vec{z}_{\text{Zopf}} = \begin{pmatrix} 500 \\ 12 \\ 60 \\ 20 \\ 3 \end{pmatrix} \in \mathbb{R}^5, \qquad \vec{z}_{\text{Brot}} = \begin{pmatrix} 700 \\ 16 \\ 0 \\ 20 \\ 0 \end{pmatrix} \in \mathbb{R}^5.$$

Will der Bäcker 12 Zöpfe und 23 Brote backen, muss er von diesen Zutaten insgesamt die Mengen aus folgender Linearkombination in $\mathbb{R}^5$

$$12\vec{z}_{\text{Zopf}} + 23\vec{z}_{\text{Brot}} = 12 \begin{pmatrix} 500 \\ 12 \\ 60 \\ 7 \\ 3 \end{pmatrix} + 23 \begin{pmatrix} 700 \\ 16 \\ 0 \\ 20 \\ 0 \end{pmatrix} = \begin{pmatrix} 22'100 \\ 512 \\ 720 \\ 700 \\ 36 \end{pmatrix} \begin{array}{l}\text{Mehl [gr]}\\ \text{Salz [gr]}\\ \text{Butter [gr]}\\ \text{Hefe [gr]}\\ \text{Milch [dl]}\end{array}$$

bereitstellen.

Der Bäcker bezieht die erforderlichen Zutaten vom Krämer, der sie in seinem Laden feilbietet. Bezeichnen wir mit p_i den Einheitspreis der i-ten Zutat, so verlangt der momentan:

$$\begin{array}{l}\text{Mehl [Rp.]/[gr]}\\ \text{Salz [Rp.]/[gr]}\\ \text{Butter [Rp.]/[gr]}\\ \text{Hefe [Rp.]/[gr]}\\ \text{Milch [Rp.]/[dl]}\end{array} \quad : \quad \vec{p} = \begin{pmatrix} 0.29 \\ 1.84 \\ 1.02 \\ 8.57 \\ 18.5 \end{pmatrix} \in \mathbb{R}^5$$

Die Materialkosten eines Zopfes berechnen sich als Skalarprodukt

$$k_{\text{Zopf}} = \langle \vec{z}_{\text{Zopf}}, \vec{p} \rangle = 455.18 \text{ [Rp.]}.$$

Nehmen wir allgemeiner an, in einer Bäckerei enthalte jedes der insgesamt m erhältlichen Gebäcke eine gewisse Menge aus dem Vorrat an r Zutaten. Mit dem Vektor $\vec{z}_j \in \mathbb{R}^r$ kodieren wir die für das j-te Gebäck erforderlichen Mengenangaben. Seine Komponente z_{ij} bezeichnet also die benötigte Menge der i-ten Zutat

im j-ten Gebäck. Diese Daten werden zweckmässig zur Zutatenmatrix $Z \in \mathbb{R}^{r,m}$ zusammengefasst.

$$\text{Zutat} \begin{pmatrix} & & \text{Gebäck} & & \\ & & \text{j} & \to & m \\ & & \vdots & & \\ i & \cdots & z_{ij} & & \\ \downarrow & & & & \\ r & & & & \end{pmatrix} = Z \in \mathbb{R}^{r,m}$$

In ihren j-ten Spalte steht also der Vektor $\vec{z}_j$. Das Matrixelement Z_{ij} gibt die benötigte Menge der i-ten Zutat im j-ten Gebäck an.

Der Bäcker wird die gewünschten Produktionsmengen der Gebäcke zum Produktionsvektor $\vec{x} \in \mathbb{R}^m$ zusammenfassen, dessen j-te Komponenten x_j angibt, wieviel vom j-ten Gebäck hergestellt werden soll. Um herauszufinden, wieviel er von der jeweiligen Zutat bereitstellen muss, bildet er das Produkt $\vec{b} = Z \cdot \vec{x}$. Seine i-te Komponente b_i gibt an, wieviel er von der i-ten Zutat benötigt und die Gesamtkosten der Produktion belaufen sich auf $\langle Z \cdot \vec{x}, \vec{p} \rangle$.

Ein Versicherer, der die Bäckerei gewissermassen von aussen betrachtet, würde in dualer Sichtweise den Wert der einzelnen Gebäcke durch die Skalarprodukte $\vec{w} = Z^T \cdot \vec{p}$ berechnen. In der j-ten Zeile der transponierten Matrix $Z^T \in \mathbb{R}^{m,r}$ stehen nämlich die erforderlichen Zutaten des j-ten Gebäcks und daher gibt die j-te Komponente dieses Vektors den Wert des j-ten Gebäcks an. Aus seiner Sicht wird der Gesamtwert der Produktion durch das Skalarprodukt $\langle \vec{x}, Z^T \cdot \vec{p} \rangle$ berechnet. Die Adjunktion

$$\langle Z \cdot \vec{x}, \vec{p} \rangle = \langle \vec{x}, Z^T \cdot \vec{p} \rangle$$

besagt, dass Gesamtkosten und Gesamtwert übereinstimmen bzw. dass Geld ausgeben und verdienen dual sind, weil dazwischen eine Symmetrie besteht, die sich in einem Erhaltungssatz manifestiert.

Auch allgemeinere Matrizenprodukte kommen im täglichen Leben vor. Die Preisüberwacherin wird die Einheitspreise der Backzutaten der letzten n Jahre in einer Matrix $P \in \mathbb{R}^{r,n}$ speichern, deren k-te Spalte $\vec{p}_k$ die Einheitspreise im k-ten Jahr angibt. Das Element p_{ik} gibt die Kosten der i-ten Zutat im k-ten Jahr an.

$$\text{Zutat} \begin{pmatrix} & & \text{Jahr} & & \\ & & \text{j} & \to & n \\ & & \vdots & & \\ i & \cdots & p_{ik} & & \\ \downarrow & & & & \\ r & & & & \end{pmatrix} = P \in \mathbb{R}^{r,n}$$

Wenn sie die Preise für Backwaren in den letzten Jahren vergleichen will, bildet sie die Matrix $B = Z^T \cdot P \in \mathbb{R}^{m,n}$, deren Element b_{jk} den Preis des j-ten

Gebäcks im k-ten Jahr angibt.

$$\text{Gebäck} \begin{pmatrix} & & \text{Jahr} & & \\ & & \text{k} & \to & n \\ & & \vdots & & \\ j & \cdots & b_{jk} & & \\ \downarrow & & & & \\ m & & & & \end{pmatrix} = B \in \mathbb{R}^{m,n}$$

Die Operationen der Matrizenrechnung sind also für die Anwendungen in der Krämerei bestens geeignet. ◯

Was für den Ökonomen die Kostenbilanz ist, spielt für den Physiker die Rolle der Arbeit. Beide sind in erster Linie am Optimieren interessiert! Skalarprodukte und Linearkombinationen spielen bereits in der klassischen Physik (versteckt) eine zentrale Rolle.

Beispiel. Tatsächlich wird in der klassischen Mechanik das Skalarprodukt verwendet, um die Arbeit W zu definieren, die von einer Kraft $\vec{F}$ verrichtet wird, wenn sie ein Objekt in Richtung des Vektors $\vec{s}$ verschiebt. Es gilt

$$W = \langle \vec{F}, \vec{s} \rangle$$

Insbesondere ist als W eine vorzeichenbehaftete reelle Zahl. ◯

Beispiel. Allgemein lassen sich die klassischen physikalischen Systeme mathematisch derart beschreiben, dass man ihre skalare Lagrange[7]-Funktion

$$\mathcal{L}\big(\vec{x}(t), \vec{x}\,'(t)\big)$$

angibt, die von der momentanten Lagen $\vec{x}(t) \in \mathbb{R}^n$ der Teilchen und ihrer momentanten Geschwindigkeiten $\vec{x}\,'(t) \in \mathbb{R}^n$, d.h. vom momentanen Zustand $(\vec{x}(t), \vec{x}\,'(t)) \in \mathbb{R}^{2n}$ des Systems abhängt. Will man wissen, auf welchem Weg sich nun die Teilchen bewegen, die sich zur Zeit t_A an der Stelle $A = \vec{x}(t_A)$ befinden mögen, und die sich zur Zeit t_B an der Stelle $B = \vec{x}(t_B)$ befinden sollen, so betrachtet man *alle* möglichen Kurven γ, die in dieser Zeit von A nach B führen und berechnet für jede solche Kurve die sogn. *Wirkung*

$$S(\gamma) = \int_{t_A}^{t_B} \mathcal{L}\big(\vec{x}(t), \vec{x}\,'(t)\big)\, dt$$

die das Produkt einer Energie mit einer Zeit als Einheit hat. Unter allen Wegen γ vom Anfangszustand A zum Endzustand B wählen die Teilchen nun denjenigen $\tilde{\gamma}$, für den die Wirkung stationär ist. Dieses *Optimalitätsprinzip der kritischen Wirkung* läuft in vielen Fällen darauf hinaus, dass die Wirkung minimal wird.

Mit Hilfe von etwas Variationsrechnung, einer Verallgemeinerung der Analysis einer Variablen, findet man, dass jede glatte Lösung des Optimierungsproblems für alle Zeiten $t \in [t_A, t_B]$ die Differentialgleichung

$$\frac{d}{dt}\Big(\mathcal{L}_{\vec{x}'}\big(\vec{x}(t), \vec{x}\,'(t)\big)\Big) = \mathcal{L}_{\vec{x}}\big(\vec{x}(t), \vec{x}\,'(t)\big)$$

[7] Der Leser entschuldige das Name-dropping. Es ist historisch begründet.

von Euler-Lagrange erfüllt. Sie liefert ein System zweiter Ordnung

$$\mathcal{L}_{\vec{x}'\vec{x}'}\frac{d^2\vec{x}}{dt^2}+\mathcal{L}_{\vec{x}'\vec{x}}\frac{d\vec{x}}{dt}=\mathcal{L}_{\vec{x}}$$

Definiert man mit Legendre den Impuls durch

$$\vec{p}(t)=\mathcal{L}_{\vec{x}'}\big(\vec{x}(t),\vec{x}\,'(t)\big)$$

und damit dann die Energie mit Hilfe des Skalarproduktes durch

$$H(t,\vec{x},\vec{x}\,')=\langle\vec{p}(t),\vec{x}\,'(t)\rangle-\mathcal{L}\big(\vec{x}(t),\vec{x}\,'(t)\big)$$

so gilt auf Grund der Euler-Lagrange-Gleichung

$$\frac{d}{dt}\Big(\vec{p}(t)\Big)=\mathcal{L}_{\vec{x}}\big(\vec{x}(t),\vec{x}\,'(t)\big),\qquad \vec{p}\,'=\mathcal{L}_{\vec{x}}=-H_{\vec{x}}$$

und die Euler-Lagrange-Gleichung nimmt die äquivalente Form

$$\left\{\begin{array}{rcl}\frac{d\vec{x}}{dt} &=& H_{\vec{p}}\\ \frac{d\vec{p}}{dt} &=& -H_{\vec{x}}\end{array}\right.$$

an. Damit und wegen der Definition des Impuls $\vec{p}=\mathcal{L}_{\vec{x}'}$ gilt für die zeitliche Ableitung der Energiefunktion

$$\begin{array}{rcl}\frac{dH(t)}{dt} &=& \langle\vec{p}\,',\vec{x}\,'\rangle+\langle\vec{p},\vec{x}\,''\rangle-\langle\mathcal{L}_x,\vec{x}\,'\rangle-\langle\mathcal{L}_{\vec{x}'},\vec{x}\,''\rangle\\ &=& \langle(\vec{p}-\mathcal{L}_{\vec{x}\,'}),\vec{x}\,''\rangle+\langle(\vec{p}\,'-\mathcal{L}_x),\vec{x}\,'\rangle=0\end{array}$$

Daher gilt für jedes solche System der Energiesatz

$$\frac{dH(t)}{dt}=0,\quad t\in[t_A,t_B]$$

Er besagt, dass die Energie entlang des optimalen Weges $\tilde{\gamma}$ erhalten bleibt. Eine genau Untersuchung zeigt, dass dieser Erhaltungssatz mit der Tatsache zusammenhängt, dass die Lagrange-Funktion $\mathcal{L}$ nicht explizit von der Zeit t abhängt und deshalb unter Zeitverschiebung $t\mapsto t+c$ invariant bleibt. Diese Symmetrieeigenschaft des Systems ist also eng mit dem Energieerhaltungssatz verknüpft. Allgemeiner ist mit jeder Symmetrie-Invarianten der Lagrange-Funktion eine Erhaltungsgrösse verknüpft. Ist sie unter Raumverschiebungen invariant, führt dies zum Impulserhaltungssatz und ist sie unter Raumdrehungen invariant, führt dies zum Drehimpulserhaltungssatz. Diese drei fundamentalen Erhaltungssätze der Mechanik sind also Konsequenzen von drei geometrischen Symmetrien.

Befinden sich beispielsweise in der Mechanik die Teilchen der Massen $m_j>0$ in einem Potential $U(\vec{x})$, so gilt definitionsgemäss

$$\mathcal{L}=\frac{m}{2}\langle\vec{x}\,',\vec{x}\,'\rangle-U(\vec{x}),\qquad\text{bzw.}\qquad H=\sum_{j=1}^{n}\frac{p_j^2}{2m_j}+U(\vec{x})$$

wobei der erste, mit Hilfe des Skalarproduktes definierte, Summand die sogn. kinetische und der zweite Summand die sogn. potentielle Energie angibt. Die

Euler-Lagrange-Differentialgleichung wird dann zum System von Differentialgleichungen

$$m_j \vec{x}_j'' = -U_{\vec{x}_j}, \qquad 1 \le j \le n$$

das als Formel $m_j \vec{a}_j = \vec{F}_j$ jedem Schüler als Newtonsches Gesetz bekannt ist.

Die Lagrange-Funktion für ein Teilchen der Masse m, das sich reibungsfrei auf einer Fläche mit der positiv definiten Metrik $g_{ij}(x_1, x_2) = g_{ij}(x)$ bewegt, die von der Wahl des Koordinatensystems und vom Ort auf der Fläche abhängt und mit der das Intervall (Linienelement) die Form

$$(d\vec{s})^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} g_{ij}(x)\ dx_i\, dx_j$$

hat, gegeben durch die kinetische Energie

$$\mathcal{L}\Big(x_1, x_2, \frac{dx_1}{dt}, \frac{dx_2}{dt}\Big) = \frac{1}{2} m \vec{v}^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} g_{ij}(x) \frac{dx_i}{dt} \frac{dx_j}{dt}$$

weil ausser der senkrechten Kraft, die das Teilchen auf die Fläche zwingt, keine weiteren Kräfte wirken. Die Berechnung der Lagrange-Gleichung liefert in diesem Fall ein System gekoppelter Differentialgleichungen zweiter Ordnung

$$\sum_{j=1}^{2} g_{lj}(x) \frac{d^2 x_j}{dt^2} + \frac{1}{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \Big(\frac{\partial g_{lk}(x)}{\partial x_j} + \frac{\partial g_{lj}(x)}{\partial x_k} - \frac{\partial g_{jk}(x)}{\partial x_l} \Big) \frac{dx_j}{dt} \frac{dx_k}{dt} = 0, l = 1, 2$$

Die Masse kürzt sich heraus und die Dynamik des Teilchens, das heisst der Ort auf der Fläche, wo sich das Teilchen zur Zeit t befindet, hängt vom Anfangszustand $(x_1(0), x_2(0), \dot{x}_1(0), \dot{x}_2(0))$ ab. Es stellt sich heraus, dass das die selbe Differentialgleichung ist, die entsteht, wenn man sich für ein scheinbar ganz anderes Problem interessiert. Eine Geodäte, d.h. ein Weg extremalen Abstandes zwischen zwei beliebigen Punkten auf der Fläche wird nämlich durch die genau gleiche Differentialgleichung beschrieben. Weil die Teilchenbewegung und die Geodäten die gleichen Differentialgleichungen erfüllen, bietet sich die Möglichkeit einer Analogie und wir können annehmen, das Teilchen folge einem Gummiband, das zwischen zwei Punkten auf der Fläche gestreckt wird. Das physikalische Problem ist zu reiner Geometrie geworden! Dank dieser Einsicht von Einstein wird Physik in den letzten hundert Jahren immer mehr zu einem Teilgebiet der Geometrie und damit globaler und eleganter (d.h. weniger analytisch). ◯

In der Elektrodynamik wird ein geladenes Teilchen der Masse $m > 0$ und der Ladung e in einem elektromagentischen Feld durch die Lagrange-Funktion

$$\mathcal{L}(\vec{x}, \vec{x}\,') = \frac{m}{2} \langle \vec{x}\,', \vec{x}\,' \rangle - e\varphi + \frac{e}{c} \langle \vec{x}\,', \vec{A} \rangle$$

beschrieben. Dabei bezeichnet $\varphi(\vec{x})$ das elektrische Potential und $\vec{A}(\vec{x})$ das magnetische Vektorpotential.

Auch in der Wellenlehre und in der Quantenmechanik spielt das Skalarprodukt eine zentrale Rolle. Verdreht man zwei parallele ideale linear-polarisierende Filter mit den Polarisationsrichtungen $\vec{a}, \vec{b} \neq \vec{0}$ um den Winkel γ zueinander, beträgt der Anteil der insgesamt durchgelassenen Strahlung

$$\cos^2(\gamma) = \frac{\langle \vec{a}, \vec{b} \rangle \cdot \langle \vec{a}, \vec{b} \rangle}{\langle \vec{a}, \vec{a} \rangle \cdot \langle \vec{b}, \vec{b} \rangle}$$

Stehen die beiden Polarisationsrichtungen aufeinander senkrecht ($\vec{a} \perp \vec{b}$), wird also kein Licht durchgelassen. Sind die beiden Polariationsrichtungen parallel ($\vec{a} \parallel \vec{b}$), so wird alles Licht durchgelassen. Selbstverständlich muss man dazu zeigen, dass die rechte Seite dieser Gleichung im Einheitsintervall liegt, d.h. dass für beliebige Vektoren die fundamentale Schwarzsche Ungleichung

$$\langle \vec{a}, \vec{b} \rangle \cdot \langle \vec{a}, \vec{b} \rangle \leq \langle \vec{a}, \vec{a} \rangle \cdot \langle \vec{b}, \vec{b} \rangle, \quad \text{bzw.} \quad (\langle \vec{a}, \vec{b} \rangle)^2 \leq |\vec{a}|^2 \cdot |\vec{b}|^2$$

gilt.

Diese Gleichung erfordert eine völlig neue Interpretation, wenn man den diskreten Charakter der elektromagnetische Energie in Betracht zieht. Dazu reduziert man die Intensität des Lichtes so, dass jeweils nur ein einzelnes — durch den ersten Polarisator polarisiertes — Photon am zweiten, um γ verdrehten, Polarisator ankommt. Weil Photonen nicht halbiert werden können, wird das Photon den zweiten Polarisator durchqueren oder nicht. Weil nichts ersichtlich ist, womit der zweite Polarisator die ankommenden Photonen in zwei Teilmengen aufspalten könnte, bleibt nichts anderes übrig als zu sagen, dass eine gewisse *Wahrscheinlichkeit* bestehe, dass das polarisierte Photon den zweiten Filter passiert. Dank der fundamentalen Ungleichung lässt sich der Ausdruck auf der rechten Seite in obiger Gleichung tatsächlich als Wahrscheinlichkeit interpretieren.

Diese Interpretation spielt in der korpuskularen Auffassung der Strahlung eine fundamentale Rolle. Die Gleichung gibt die Wahrscheinlichkeit an, mit der ein einzelnes Photon zwei verdrehte Polarisatoren durchläuft. Die Hypothese, dass jedes polarisierte Photon beim Durchgang durch den zweiten Polarisator in zwei Teilchen aufgespalten wird, von denen eines den Polarisator durchquert und das andere absorbiert wird, ist nicht haltbar, weil sonst nach dem Polarisator im Strahl gleich viele Teilchen wie im einfallenden Strahl vorhanden sein sollten. Um die beobachtete kleinere Intensität des durchgelassenen Strahl zu erklären, müsste eines der durchgelaufenen Teilchen eine geringere Energie als die ankommenen Photonen haben. Wegen Einsteins Beziehung $E = h\nu$ müsste dann aber die Frequenz der durchgelaufenen Teilchen kleiner als jene der ankommenen Teilchen sein. Weil aber beim Durchgang durch den Polarisator keine Farbwechsel beobachtet wird, kann diese Hypothese nicht zutreffen und die Energie muss beim Durchgang durch den Polarisator die selbe geblieben sein. Daher enthält der durchgelaufene Strahl *weniger* Photonen als der einfallende Strahl, was sich experimentell mit einzelnen Photonen bestätigen lässt. Daher muss $\cos^2(\gamma)$ als *Anteil* der Photonen interpretiert werden, die den Polarisator durchqueren! Für ein einzelnes polarisiertes Photon gibt $\cos^2(\gamma)$ die Wahrscheinlichkeit an, dass es durchgelassen wird.

Der Anteil Photonen, der absorbiert wird, beträgt $1 - \cos^2(\gamma) = \sin^2(\gamma)$. Die Konsequenzen dieser Beobachtung sind bemerkenswert. Obwohl alle Photonen nach dem ersten Polarisationsfilter im selben Polarisationszustand sind, werden

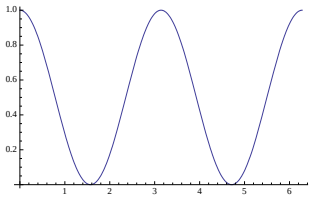


Abbildung 2.4: Durchlasswahrscheinlichkeit bei einem linearen Polarisator.

bei verdrehten Filtern einige vom zweiten Filter gestoppt und andere durchgelassen und der klassische deterministische Glaube, dass gleiche Bedingungen immer die selben Resultate verursachen, ist schlicht unhaltbar. In der Natur wird offensichtlich mit Hilfe des Skalarproduktes gewürfelt! Es gehört zu den grossen Überraschungen der modernen Physik, dass wir nur in ganz speziellen Situationen zum Vorneherein wissen können, was Teilchen tun werden und dass physikalische Theorien ein probalistisches Element enthalten müssen. Dieser Indeterminismus des Mikrokosmos zeigt, dass sich die Natur letztlich nur mit Wahrscheinlichkeiten beschreiben lässt und dass sich diese Wahrscheinlichkeiten nicht mehr weiter reduzieren lassen. Fast noch überraschender als die Notwendigkeit, in unserem physikalischen Weltbild deterministische durch stochastische Gesetze ersetzen zu müssen, ist allerdings der weitere Umstand, dass die notwendige Wahrscheinlichkeitstheorie der Quantentheorie bei der Beschreibung unbeobachteter Objekte nicht auf der klassischen Booleschen Logik der Theorie der Mengen basieren kann, sondern auf einer Logik aufgebaut ist, die stattdessen zur Theorie der Vektorräume, d.h. zur linearen Algebra gehört, wie man erkennt, wenn man einen dritten solchen Polarisationsfilter mit der Polarisationsrichtung $\vec{c} \neq \vec{0}$ zwischen die beiden ersten stellt und sorgfältig beschreibt, was man dann feststellt.

Ein Polarisationsfilter kann als analoger Computer (Qubit) interpertiert werden, der sich nicht durch ein klassisches Bit diskretisieren lässt.

Erst beim Nachdenken über das Verhalten des Mikrokosmos sind die Physiker auf die zentrale Bedeutung des Matrizenproduktes zum Komponieren von Prozessen gestossen.

Beispiel. Ursprünglich ging es darum, die im 19. Jahrundert beobachteten Atomspektren zu verstehen. Gemäss Bohr und Einstein hängen die gemessenen Kreisfrequenzen mit den möglichen Energien des Systems durch die Beziehung

$$\omega_{ij} = \frac{E_i - E_j}{\hbar}$$

zusammen, wenn es vom Energiezustand E_i in den Energiezustand E_j übergeht. Das Ziel bestand darin, die möglichen Energieniveaus $E_1, E_2, \ldots$ zu bestimmen. Bereits früher hatte Balmer gezeigt, dass Spektrallinien im Wasserstoff-Spektrum in Serien auftreten und und 1890 hatte Rydberg gesehen, dass sich deren Kreisfrequenzen durch die bemerkenswerte Differenzenformel

$$\omega_{ij} = \frac{2\pi c}{\lambda_{ij}} = 2\pi c R\Big(\frac{1}{i^2} - \frac{1}{j^2}\Big), \qquad (i, j = 1, 2 \ldots)$$

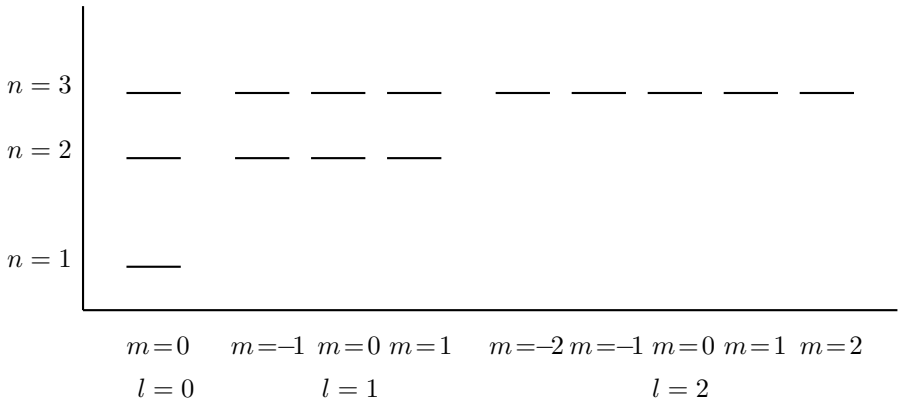


Abbildung 2.5: Termschema von Wasserstoff.

ausdrücken lassen. Dabei bezeichnet c die Lichtgeschwindigkeit und R die Rydberg-Konstante. Daher lassen sich zwei Frequenzen ω_{ij} und ω_{jk} nach dem fundamentalen Kompositionsprinzip

$$\omega_{ik} = \omega_{ij} + \omega_{jk}$$

kombinieren. In der älteren Quantentheorie gelang es Bohr 1913, das Atomspektrum von Wasserstoff, das traditionell in Form eines Termschemas beschrieben wird, theoretisch herzuleiten.

Die Zustände des Wasserstoffatoms $\vec{v}_{n,l,m}$ können in der damaligen Sprechweise durch drei gazzahlige Indizes (n, m, l) (Quantenzahlen) beschrieben werden, die den Einschränkungen

$$\begin{cases} n & = & 1, 2, \dots \\ l & = & 0, 1, \dots, n-1 \\ m & = & -l, -(l-1), \dots, (l-1), l \end{cases}$$

genügen. Die möglichen Energienieveaus sind nach Bohr durch die Formel

$$E_{n,m,l} = \frac{R}{n^2}$$

gegeben.

In späterer Formulierung liefern die horizontalen Linien im Termschema eine graphische Illustration der Basisvektoren $\vec{v}_{n,l,m}$, die als simmultane Eigenvektoren der drei linearen Operatoren

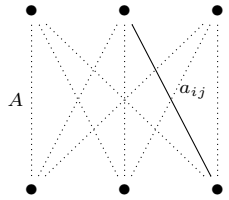
E : die Energie
L : der Betrag des Bahn-Drehmomentes
L_z : die z-Komponente des Bahn-Drehmomentes

entstehen. Wir können also sagen, dass wenn das Wasserstoffatom im Zustand $\vec{v}_{nlm}$ ist, dass dann diese drei physikalischen Grössen einen festen Wert haben. In der Tat gilt für die Eigenwerte von $\vec{v}_{nlm}$:

$$\begin{aligned} E \cdot \vec{v}_{nlm} &= \tfrac{R}{n^2} \vec{v}_{nlm} \\ L \cdot \vec{v}_{nlm} &= \sqrt{l(l+1)} \vec{v}_{nlm} \\ L_z \cdot \vec{v}_{nlm} &= m \vec{v}_{nlm} \end{aligned}$$

Diese Eigenzustände des Energieoperators entsprechen also den „stationären Bahnen“ in der älteren Sprache, die aber bereits bei der Beschreibung des Spektrums des Helium-Atoms versagte.

Als dann 1925 Heisenberg versuchte, allgemeinere Spektren zu beschreiben, ging er von einem schematischen Termschema der Art



aus und ordnete einem Übergang vom Zustand i (Energieniveau) in den Zustand j die *komplexe Amplitude* $a_{ij} \in \mathbb{C}$ zu.

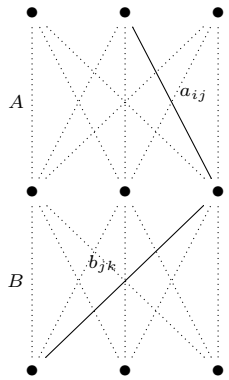
Im einfachsten Fall ging er von der unendlichen Matrix

$$q(t) = \left(q_{ij} e^{\mathfrak{i}\omega_{ij} t}\right), \qquad (i, j = 1, 2 \ldots)$$

der Kreisfrequenzen ω_{ij} und der komplexen Amplituden q_{ij} aus. Dabei bezeichnet $\mathfrak{i}$ die später zu besprechende imaginäre Einheit mit der Eigenschaft $\mathfrak{i}^2 = -1$. Bei seinen Überlegungen benutzte Heisenberg also nur Grössen, die direkt zu den physikalischen Messungen aus der Spektroskopie von Atomen und Molekülen in Beziehung stehen und benutzte insbesondere weder den Begriff der Bahnkurve noch jenen der Geschwindigkeit der Teilchen. Die Ritz'sche Kombinationsregel

$$\omega_{ik} = \omega_{ij} + \omega_{jk}$$

führte ihn zum Kombinieren von Übergängen nach dem Schema



Es lässt sich in symbolischer Form durch die Gleichung

$$(AB)_{ik} = \sum_j a_{ij} b_{jk}$$

ausdrücken. Dabei handelt es sich ganz offensichtlich um die beschriebene Matrizenmultiplikation, die Heisenberg aber vorher allem Anschein nach nicht kannte.

Sein Doktorvater Born erinnerte sich an eine Vorlesung über Matrizenrechnung, erkannte den Zusammenhang und und vermutete für den Impuls, den er analog durch die Matrix

$$p(t) = m\dot{q}(t) = \left(\mathrm{i}m\omega_{ij}q_{ij}e^{\mathrm{i}\omega_{ij}t}\right)$$

beschrieb, die kanonische Vertauschungsrelation

$$q(t) \cdot p(t) - p(t) \cdot q(t) = \mathrm{i}\hbar E$$

Diese Auffassung, Prozesse dadurch zu komponieren, dass man über alle Pfade via Zwischenzustände summiert und das Betragsquadrat $|a_{ij}|^2$ der komplexen Amplitude als *Wahrscheinlichkeit* für den Übergang vom Zustand i in den Zustand j zu interpretieren, war der Ausgangspunkt der Quantenmechanik und das Ende der deterministischen Weltsicht, die für die klassische Physik zentral war. Weil wegen der Dreiecksungleichung

$$|\vec{a}+\vec{b}| \leq |\vec{a}| + |\vec{b}|, \qquad |\vec{a}+\vec{b}|^2 \leq (|\vec{a}| + |\vec{b}|)^2 = |\vec{a}|^2 + |\vec{b}|^2 - 2|\vec{a}| \cdot |\vec{b}|$$

eine Summe von komplexen Ampltituden einen Betrag haben kann, der kleiner als die Summe der Summanden ist, führt die Quantenmechanik zu destruktiver Interferenz: erlaubt man einem Prozess auf mehr Arten stattzufinden, kann sich die Chance, dass er abläuft, dabei verkleinern!

Die Idee der Summe über alle Pfade führte später im kontinuierlichen Grenzfall zu den Pfadintegralen, die heute die Quantenfeldtheorien dominieren. Leider ist diese Theorie mathematisch weder einfach noch haltbar. Dem Leser sei dazu das elementar geschriebene Büchlein von Feynman mit dem Titel “QED. Die seltsame Theorie des Lichts und der Materie” zur Lektüre empfohlen. Es beruht auf einer populären Vorlesung, die man im Web findet und zu den Perlen dieses sonst riesigen Misthaufens gehört. ◯

2.3 Eigenschaften der Matrixoperationen

Viele der von der Arithmetik d.h. vom „Rechnen mit Zahlen“ her bekannte Regeln gelten auch für das Rechnen mit Matrizen.

Satz. Alle Matrizen seien so gewählt, dass die Operationen definiert sind. Dann gelten folgende Rechenregeln:

- Regeln der Addition:

  1. $A + (B + C) = (A + B) + C$ (Assoziativgesetz der Addition)
  2. $A + B = B + A$ (Kommutativgesetz der Addition)
  3. $A + 0 = A = 0 + A$ (Neutralelement der Addition)

  Aus der letzten Regel folgt natürlich $0 - A = -A$. Eine Struktur, die diese drei Regeln erfüllt, heisst *kommutatives Monoid*.

- Regel des Negativen:

  4. $A - A = 0$

Man beachte, dass sich die Nullmatrix analog zur Zahl 0 verhält. Eine Struktur, die diese vier Regeln erfüllt, heisst *kommutative Gruppe*.

- Regeln der Multiplikation mit Skalaren:

  5. $r(sA) = (rs)A$
  6. $1A = A$

- Verträglichkeitsregeln zwischen Addition und Multiplikation mit Skalaren:

  7. $r(A + B) = rA + rB$
  8. $(r + s)A = rA + sA$

  Eine Struktur, die diese acht Regeln erfüllt, heisst *Vektorraum*.

- Regeln der Multiplikation:

  9. $A \cdot (B \cdot C) = (A \cdot B) \cdot C$ (Assoziativgesetz der Multiplikation)
  10. $A \cdot E = A = E \cdot A$ (Neutralelement der Multiplikation)

  Man beachte, dass sich die Einheitsmatrix E analog zur Zahl 1 verhält. Eine Struktur, die die letzten beiden Regeln erfüllt, heisst *Monoid*.

- Verträglichkeitsregeln zwischen Addition und Multiplikation:

  11. $A \cdot (B + C) = (A \cdot B) + (A \cdot C)$ (linkes Distributivgesetz)
  12. $(B + C) \cdot A = (B \cdot A) + (C \cdot A)$ (rechtes Distributivgesetz)

  Eine Struktur, die diese letzten vier und die ersten vier Regeln erfüllt, heisst *Ring*. In jedem Ring gilt $A \cdot 0 = 0 = 0 \cdot A$.

- Verträglichkeitsregel zwischen Multiplikation mit Skalaren und Multiplikation:

  13. $r(A \cdot B) = (rA) \cdot B = A \cdot (rB)$

  Eine Struktur, die diese dreizehn Regeln erfüllt, heisst *Algebra*.

- Verträglichkeitsregel zwischen Transponieren und den Grundoperationen:

  14. $0^T = 0$
  15. $E^T = E$
  16. $((A)^T)^T = A$
  17. $(A + B)^T = A^T + B^T$
  18. $(rA)^T = rA^T$
  19. $(A \cdot B)^T = B^T \cdot A^T$

  In der letzten Regel achte man auf die korrekte Reihenfolge!

Der Beweis dieser Rechenregeln ist in den meisten Fällen einfach. Ausnahmen bilden die Regeln $9, 11, 12$ und 19, die das Matrizenprodukt betreffen, das ja nicht elementweise definiert ist. Da aber auch diese beiden Beweise nicht sonderlich lehrreich sind, schenken wir sie uns! Hat man ein geometrisches Bild der Matrizenoperationen oder interpretiert Matrizen als lineare Prozesse, sind diese Regeln ohnehin ziemlich trivial. Diese Regeln der Matrizenrechnung werden wir im Laufe dieser Lehrveranstaltung tausendfach benutzen; es lohnt sich also, sich diese genau zu merken und an einzelnen konkreten Zahlenbeispielen einzuüben. Es empfiehlt sich sogar dringend, diese Regeln auswendig zu lernen und nicht einfach darüber hinwegzugehen mit der Begründung: „Das habe ich schon immer so gemacht“. Insbesondere erfinde man keine weiteren Regeln für das Rechnen mit Matrizen! Aus diesen Regeln folgen allerdings weitere Regeln — etwa die Verträglichkeitsregeln für das Rechnen mit dem Skalarprodukt.

Obwohl viele Regeln der Arithmetik auch für das Rechnen mit Matrizen gelten, sind sie nämlich nicht alle erfüllt. Der aufmerksame Leser wird sich gefragt haben, ob denn die Matrizenmultiplikation nicht kommutativ ist. In der Tat gilt selbst in dem Fall, wo $A \cdot B$ und $B \cdot A$ beide definiert sind, im allgemeinen das Kommutativgesetz nicht. Dies geht aus folgendem Gegenbeispiel hervor.

Beispiel. Man betrachte die beiden Matrizen

$$A = \begin{pmatrix} -1 & 0 \\ 2 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix} \in \mathbb{R}^{2,2}$$

Durch Multiplikation erhält man die Matrizen

$$A \cdot B = \begin{pmatrix} -1 & -2 \\ 11 & 4 \end{pmatrix} \quad \text{und} \quad B \cdot A = \begin{pmatrix} 3 & 6 \\ -3 & 0 \end{pmatrix}$$

Ihre Differenz verschwindet also nicht, sondern es gilt

$$[A, B] = A \cdot B - B \cdot A = \begin{pmatrix} -4 & -8 \\ 14 & 4 \end{pmatrix}$$

Hier ist also $A \cdot B \neq B \cdot A$. Die Matrizenalgebra ist also nicht kommutativ! Wir müssen beim Bilden eines Matrizenproduktes jeweils angeben, ob wir die Matrix B mit der Matrix A von links oder von rechts multiplizieren.

Der *Kommutator* $[A, B] = A \cdot B - B \cdot A$ der beiden Matrizen A und B enthält die Information darüber, ob die beiden Matrizen kommutieren oder nicht und spielt eine zentrale Rolle bei der Untersuchung kontinuierlicher Symmetrien mit Hilfe von Linearisierungen. Allgemein ist also

$$A \cdot B = B \cdot A + [A, B]$$

Unter einer *Lie-Algebra* versteht man einen Vektorraum zusammen mit einer Operation, die dem Kommutator von Matrizen nachgebildet ist und folgende Verträglichkeitseigenschaften hat:

1. $[A, B] = -[B, A]$ (Schiefsymmetrie)
2. $[A, B + C] = [A, B] + [A, C]$ (Additivität)

3. $[A, rB] = r[A, B]$ (Homogenität)

4. $[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0$ (Jacobi Identität)

Tatsächlich überprüft man durch Nachrechnen leicht, dass der Kommutator alle diese Eigenschaften hat. Selbstverständlich gilt wegen der Schiefsymmetrie $[A, A] = 0$. Die reellen $n \times n$-Matrizen zusammen mit dem Kommutator bilden also eine Lie-Algebra, die unter dem Namen $\mathfrak{gl}(n, \mathbb{R})$ bekannt ist.

Der Kommutator ist auch mit den restlichen Operationen verträglich. Die Verträglichkeit mit der Transposition drückt sich in der Beziehung

$$[A, B]^T = [B^T, A^T]$$

aus. Die Verträglichkeit mit dem Matrizenprodukt zeigt sich in der Beziehung

$$[A \cdot B, C] = [A, C] \cdot B + A \cdot [B, C]$$

Alle diese Beziehungen lassen sich leicht beweisen, indem man die involvierten Kommutatoren ausschreibt. Dank diesern Regeln ist es in Zukunft kaum je notwendig, den Inhalt eines Kommutators auszuschreiben. Man beachte die Analogie der letzten Beziehung zur Produktregel zum Ableiten eines Produktes.

$$\frac{d(a \cdot b)}{d(c)} = \frac{d(a)}{d(c)} \cdot b + a \cdot \frac{d(b)}{d(c)}$$

In der Tat übernehmen Kommutatoren die Rolle einer Ableitung, wenn Symmetrien im Spiel sind, weil sie insbesondere die Information kodieren, wie sich solche Symmetrien zu einer weiteren Symmetrie zusammensetzen lassen.

Beispiel. Im Zusammenhang mit Kugeldrehungen erzeugen die drei Matrizen

$$Q(\vec{e}_1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad Q(\vec{e}_2) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad Q(\vec{e}_3) = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

die Drehungen um die drei Koordinatenachsen. Ihre Elemente lassen sich knapp durch

$$Q(\vec{e}_i)_{jk} = -\varepsilon_{ijk}$$

beschreiben. Dabei liefert für jede Variation $(k_1, k_2, \ldots, k_n) \in \mathbb{N}^n, 1 \leq k_j \leq n$ das sogn. *Permutationssymbol*

$$\varepsilon_{k_1 k_2 \ldots k_n}$$

den Wert $\varepsilon_{1\ldots n} = 1$ und ändert das Vorzeichen bei der Vertauschung zweier Indizes. Es ist also

$$\varepsilon_{k_1 k_2 \ldots k_n} = \begin{cases} 1 & \text{falls } (k_1 k_2 \ldots k_n) \text{ eine gerade Permutation ist.} \\ -1 & \text{falls } (k_1 k_2 \ldots k_n) \text{ eine ungerade Permutation ist.} \\ 0 & \text{falls in } (k_1 k_2 \ldots k_n) \text{ mindestens zwei Indizes gleich sind.} \end{cases}$$

Beispiel. Im Fall $n = 3$ gilt also für das Permutationssymbol $\varepsilon_{ijk} = 1$, falls die Permutation (i, j, k) von $(1, 2, 3)$ gerade ist. Falls sie ungerade ist, ist $\varepsilon_{ijk} = -1$.

Sind mindestens zwei der Indizes i, j, k gleich, so ist $\varepsilon_{ijk} = 0$. Das liefert für die 3! Permutationen von $\{1, 2, 3\}$ folgende Werte[8]:

$$\varepsilon_{123} = 1, \quad \varepsilon_{132} = -1, \quad \varepsilon_{213} = -1, \quad \varepsilon_{231} = 1, \quad \varepsilon_{312} = 1, \quad \varepsilon_{321} = -1$$

Bei einer Permutation (i, j, k) von $(1, 2, 3)$ bezeichnet man ε_{ijk} auch als Vorzeichen der Permutation. ◯

Diese drei Generatoren der Drehgruppe ergeben sich, indem man mit Hilfe der Drehmatrizen

$$D_1(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix}$$

$$D_2(\varphi) = \begin{pmatrix} \cos(\varphi) & 0 & \sin(\varphi) \\ 0 & 1 & 0 \\ -\sin(\varphi) & 0 & \cos(\varphi) \end{pmatrix}$$

$$D_3(\varphi) = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) & 0 \\ \sin(\varphi & \cos(\varphi) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

"ein ganz kleines Bisschen" um die jeweilige Achse dreht, d.h. diese Matrizen nach φ ableitet und dann den Winkel $\varphi = 0$ setzt. Die drei so erhaltenen Erzeugermatrizen $Q(\vec{e}_j$ erfüllen die zyklischen Vertauschungsrelationen

$$[Q(\vec{e}_1), Q(\vec{e}_2)] = Q(\vec{e}_3), \qquad [Q(\vec{e}_2), Q(\vec{e}_3)] = Q(\vec{e}_1), \qquad [Q(\vec{e}_3), Q(\vec{e}_1)] = Q(\vec{e}_2)$$

oder knapper

$$[Q(\vec{e}_i), Q(\vec{e}_j)] = \varepsilon_{ijk} \cdot Q(\vec{e}_k)$$

Diese sogn. fundamentalen Vertauschungsrelationen spielen in der Quantenmechanik eine zentrale Rolle, weil die drei Komponenten des Drehimpulses die Vertauschungsrelationen

$$[L_1, L_2] = L_3, \qquad [L_2, L_3] = L_1, \qquad [L_3, L_1] = L_2, \qquad [L_i, L_j] = \varepsilon_{ijk} L_k$$

erfüllen. Entsprechende Vertauschungsrelationen werden auch von den Pauli-Matrizen

$$S_1 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad S_2 = \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad S_3 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

erfüllt, die den Elektronenspin beschreiben, weil für einen normierten Vektor $\vec{n} \in \mathbb{R}^3$ die Matrix

$$S_{\vec{n}} := n_1 S_1 + n_2 S_2 + n_3 S_3 = \frac{1}{2} \begin{pmatrix} n_3 & n_1 - i n_2 \\ n_1 + i n_2 & -n_3 \end{pmatrix}$$

den Spin in Richtung $\vec{n}$ beschreibt.

Allgemein wird eine Drehung um die Achse $\vec{a}$ durch die fundamentale, antisymmetrische Matrix

$$Q(\vec{a}) := a_1 Q(\vec{e}_1) + a_2 Q(\vec{e}_2) + a_3 Q(\vec{e}_3) = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix}$$

[8] Merkregel: 123123; wie beim Walzertanzen. Man erhält 1, wenn man von links nach rechts abliesst und -1, wenn man von rechts nach links abliesst.

erzeugt. Die Lie-Algebra $\mathfrak{so}(3,\mathbb{R})$, und nicht das von vielen Anwendern leider immer noch mit Vorliebe verwendete Kreuzprodukt

$$\vec{a}\times\vec{x} := Q(\vec{a})\cdot\vec{x} = \begin{pmatrix} a_2x_3 - a_3x_2 \\ a_3x_1 - a_1x_3 \\ a_1x_2 - a_2x_1 \end{pmatrix}$$

beschreiben die Kugeldrehgruppe auf natürliche Weise. Um allerdings die Beziehung zwischen dieser Lie-Algebra und dem Vektorprodukt genau zu erkennen, d.h. warum eine Drehung um die normierte Drehachse $\vec{a}$ im positiven Umlaufsinn um den Drehwinkel φ durch die orthogonale Matrix

$$D_{\varphi,\vec{a}} = e^{\varphi Q(\vec{a})} = E + \sin(\varphi)Q(\vec{a}) + (1-\cos(\varphi)Q^2(\vec{a}), \quad |\vec{a}| = 1$$

die mit Hilfe des Vektorproduktes geschrieben die Form

$$D_{\varphi,\vec{a}}\cdot\vec{x} = \cos(\varphi)\vec{x} + (1-\cos(\varphi))\langle\vec{x},\vec{a}\rangle\vec{a} + \sin(\varphi)(\vec{a}\times\vec{x}), \quad |\vec{a}| = 1$$

annimmt, beschrieben wird, müssen wir zunächst die matrizielle Exponentialfunktion e^A studieren. Bei $\mathfrak{so}(3,\mathbb{R})$ handelt sich um die einfachste interessante Lie-Algebra, die nicht nur eine enge Beziehung zum Vektorprodukt, zu den Raumdrehungen und den Quaternionen, sondern auch zum Drehmoment, zur Lorentz-Kraft und zum Elektronenspin hat. ◯

Beispiel. Die Linearkombinationen der drei Matrizen

$$M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad N = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathbb{R}^{2,2}$$

bilden die Matrizen aus dem 3-dimensionalen Vektorraum $\mathfrak{sl}(2,\mathbb{R})$

$$rM + sN + tP = \begin{pmatrix} t & r \\ s & -t \end{pmatrix} \in \mathbb{R}^{2,2}$$

Diese (2×2)-Matrizen werden dadurch charakterisiert, dass ihre Spur, d.h die Summe der Diagonalelemente verschwindet. Dieser Vektorraum wird zu einer Lie-Algebra, in der für die Basisvektoren die Kommutator-Beziehungen

$$[P,M] = 2M, \qquad [P,N] = -2N, \qquad [M,N] = P$$

gelten, wie man leicht nachrechnet. ◯

Kommutatoren spielen in der Quantenmechanik eine grosse Rolle, weil sie beschreiben, wie stark zwei Prozesse nicht vertauschbar sind. Tatsächlich war historisch die Tatsache, dass gewisse Messprozesse nicht miteinander kommutieren der Ausgangspunkt für die Quantenmechanik. Dass sich die Messungen des Ortes Q und des Impuls P eines Teilchens nicht vertauschen lassen, drückt sich in der kanonischen Vertauschungsrelation

$$[Q,P] = i\hbar E$$

aus, die Heisenberg mit Hilfe (unendlicher) Matrizen gelöst hat. Dabei bezeichnet die reduzierte Planck-Konstante $\hbar = \frac{h}{2\pi}$ die kleinst mögliche Wirkung, die also quantisiert ist. Das Wirkungsquantum

$$h = 6.260'075'5\cdot 10^{-34}\ [\mathrm{Js}]$$

ist eine Naturkonstante, die im Vergleich zu den Wirkungen, die im täglichen Leben auftreten und die bei einer typischen Energie von

$$1\ [\mathrm{J}] = 1\ \left[\mathrm{kg}\cdot\frac{\mathrm{m}^2}{\mathrm{s}^2}\right] = 1\ [\mathrm{Ws}] = 1.2418\cdot 10^{18}\ [\mathrm{eV}]$$

und der Dauer 1 [s] von der Grössenordnung von 1 [Js] sind, phantastisch klein.

Allgemein wird die Symmetriegruppe G eines Quantensystems weitgehend durch ihre zugehörige Lie-Algebra $\mathfrak{g}$ beschrieben. In vielen Fällen legt die Symmetriegruppe ein solches System sogar vollständig fest. ◯

2.4 Inverse Matrix

Auch was die Teilbarkeit angeht, gelten für Matrizen andere Gesetze als für das Rechnen mit Skalaren. Zum Beispiel ist die Kürzungsregel nicht erfüllt und die Matrizenalgebra hat Nullteiler. Die den folgenden beiden Gesetzen entsprechenden Eigenschaften sind also für Matrizen i.a. nicht erfüllt:

- Ist $a\cdot b = a\cdot c$ und $a\neq 0$, so ist $b = c$. (Kürzungsregel)
- Ist $a\cdot d = 0$, so ist mindestens ein Faktor 0. (Nullteilerfreiheit)

Beispiel. Um diese beiden Gesetze zu widerlegen, betrachte man folgende Matrizen:

$$A = \begin{pmatrix} 3 & 6 \\ 1 & 2 \end{pmatrix},\ B = \begin{pmatrix} -3 & 5 \\ 6 & 2 \end{pmatrix},\ C = \begin{pmatrix} 1 & 3 \\ 4 & 3 \end{pmatrix},\ D = \begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix}$$

Es ist

$$A\cdot B = A\cdot C = \begin{pmatrix} 27 & 27 \\ 9 & 9 \end{pmatrix}, \qquad A\cdot D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Obwohl $A\neq 0$ ist, darf man A *nicht* kürzen! Die Kürzungsregel gilt also nicht für alle Matrizen. Ferner kann das Produkt zweier von der Nullmatrix verschiedener Matrizen sehr wohl die Nullmatrix liefern. Obwohl nämlich $A\neq 0$ und $D\neq 0$ sind, ist $A\cdot D = 0$. Man sagt, A und D seien *Nullteiler*. ◯

Immerhin sind die beiden Phänomene miteinander eng verwandt. Es gilt nämlich folgendes Resultat.

Satz. In einem Ring sind die Kürzungsregel und die Nullteilerfreiheit äquivalent.

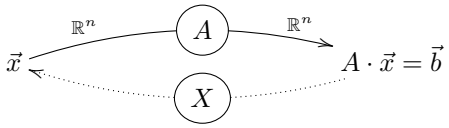
Beweis. Für die eine Richtung nehmen wir an, dass die Kürzungsregel erfüllt ist. Falls nun $a\cdot d = 0$ ist und $a\neq 0$ gilt, so ist $a\cdot d = 0 = a\cdot 0$. Wegen der vorausgesetzten Kürzungsregel dürfen wir daraus auf $d = 0$ schliessen. Daher ist der Ring Nullteilerfrei.

Für die andere Richtung nehmen wir nun an, dass der Ring keine Nullteiler hat. Falls $a\neq 0$ ist mit $a\cdot b = a\cdot c$ so folgt daraus, dass $a\cdot b - a\cdot c = a\cdot(b-c) = 0$ gilt. Daraus folgt aber mit der Nullteilerfreiheit, dass $b - c = 0$ d.h. dass $b = c$ gilt. Damit ist die Kürzungsregel gezeigt. □

In der Praxis ist es meistens einfacher, einen Ring auf Nullteiler zu untersuchen. Daraus erkennt man dann, ob die Kürzungsregel gilt oder nicht.

Diese Gegenbeispiele zeigen schon, dass die Division von Matrizen nur mit Einschränkungen möglich ist. Insbesondere wird man mit Matrizen nur schon wegen der fehlenden Kommutativität nicht gedankenlos Bruchrechnen können. Wir beginnen mit einer präzisen Definition eines Begriffes, der dem Stammbruch aus der Arithmetik entspricht. Weil vermutlich in der Schule das arithmetische Pendant nie sorgfältig besprochen wurde, ist der Leser gut beraten, sich bei dieser Gelegenheit mit dem Bruchrechnen sorgfältig auseinanderzusetzen.

Um die passende Definition einer Inversen zu finden, fassen wir die Matrix A wieder als linearen Prozess auf. Dabei wird der Input $\vec{x} \in \mathbb{R}^n$ auf den Output $\vec{b} = A \cdot \vec{x} \in \mathbb{R}^n$ abgebildet. Beim Umkehren geht es um die Frage, ob der Input aus dem Output rekonstruiert werden kann. Falls dies für jeden Output $\vec{b}$ der Fall ist, nennen wir den Prozess A umkehrbar und der Prozess X, der in die umgekehrte Richtung läuft und jedem Output $\vec{b}$ seinen eindeutig bestimmten Input zuordnet, heisst dann der zu A *inverse* Prozess.



Invertierbare Prozesse erhalten also die Information.

Der Prototyp eines invertierbaren Prozesses ist die Identität.

Er ist seine eigene Inverse und wird mit Hilfe der Einheitsmatrix E beschrieben.

Die lineare Abbildung

$$f_A \colon \mathbb{R}^n \to \mathbb{R}^n, \qquad \vec{x} \mapsto A \cdot \vec{x}$$

heisst also invertierbar, falls eine lineare Abbildung

$$f_X \colon \mathbb{R}^n \to \mathbb{R}^n, \qquad \vec{b} \mapsto X \cdot \vec{b}$$

existiert, so dass die beiden Kompositionen

$$f_X \circ f_A = \mathrm{Id}_{\mathbb{R}^n}, \qquad f_A \circ f_X = \mathrm{Id}_{\mathbb{R}^n}$$

die Identität liefern. Die linke Bedingung verlangt, dass die Abbildung f_A injektiv ist. Wegen der rechten Bedingung muss sie auch surjektiv sein, falls sie umkehrbar sein soll. Umkehrbare Prozesse werden oft auch als Isomorphismen und in der alten Literatur gelegentlich als Koordinatentransformationen bezeichnet.

In die Matrizensprache übersetzt erhalten wir folgenden Begriff.

Definition. Es sei A eine quadratische Matrix. Gibt es eine Matrix X mit der Eigenschaft $A \cdot X = X \cdot A = E$, so heisst A *invertierbar*. Die Matrix X wird als *Inverse* von A bezeichnet.

Man beachte, dass diese Beziehung symmetrisch ist. Falls also X Inverse von A ist, so ist umgekehrt A auch Inverse von X. Wir werden später zeigen, dass

man nur eine der beiden Gleichungen $A \cdot X = E$ bzw. $X \cdot A = E$ überprüfen muss um zu garantieren, dass A und X inverse Matrizen sind. Das ist wieder eine Besonderheit der Matrizenrechnung: eine Inverse von einer Seite ist automatisch Inverse von der anderen, bzw. Surjektionen sind automatisch Injektionen und umgekehrt!

Beispiel. Die Matrix $X = \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix}$ ist eine Inverse von $A = \begin{pmatrix} 2 & -5 \\ -1 & 3 \end{pmatrix}$, wie folgende Rechnungen zeigen.

$$A \cdot X = \begin{pmatrix} 2 & -5 \\ -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{E}$$

$$X \cdot A = \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 2 & -5 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{E}$$

Hat man also einmal einen geeigneten Kandidaten für die Inverse, so ist die Kontrolle einfach und wird dem Leser dringend empfohlen.

Er wird sich selbstverständlich gefragt haben, wie man die angegebene Matrix X finden kann. Wie wir noch besprechen werden, ist die Antwort auf diese Frage gar nicht so dringend, wie es scheint. Selten braucht man die Inverse wirklich!

Um aber die Neugier des Lesers zu befriedigen, machen wir für X den Ansatz als beliebige (2×2)-Matrix:

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Dann gilt für das Produkt:

$$A \cdot X = \begin{pmatrix} 2 & -5 \\ -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 2a - 5c & 2b - 5d \\ -a + 3c & -b + 3d \end{pmatrix}$$

Damit dieses Produkt die Einheitsmatrix liefert, müssen die beiden linearen Gleichungssysteme

$$\begin{cases} 2a - 5c = 1 \\ -a + 3c = 0 \end{cases} \qquad \begin{cases} 2b - 5d = 0 \\ -b + 3d = 1 \end{cases} \qquad \left(\begin{array}{cc|cc} 2 & -5 & 1 & 0 \\ -1 & 3 & 0 & 1 \end{array}\right)$$

erfüllt sein. Diese beiden Systeme löst man am besten simultan indem man die beiden Konstantenvektoren zu einem einzigen Block zusammenfasst. Addition der ersten Zeile zum 2-fachen der zweiten liefert

$$\left(\begin{array}{cc|cc} 2 & -5 & 1 & 0 \\ 0 & 1 & 1 & 2 \end{array}\right)$$

Addition des 5-fachen der zweiten Zeile zur ersten ergibt nach dem Kürzen der ersten Zeile durch ihren grössten gemeinsamen Teiler 2 eine Blockmatrix

$$\left(\begin{array}{cc|cc} 2 & 0 & 6 & 10 \\ 0 & 1 & 1 & 2 \end{array}\right), \qquad \left(\begin{array}{cc|cc} 1 & 0 & 3 & 5 \\ 0 & 1 & 1 & 2 \end{array}\right)$$

in deren linkem Block die Einheitsmatrix entstanden ist. Aus ihr lesen wir die gesuchten Lösungen ab. Die beiden Gleichungen des linken Systems haben die eindeutig bestimmte Lösung $a = 3$ und $c = 1$. Entsprechend hat das rechte

System die eindeutige Lösung $b = 5$ und $d = 2$. Man beachte, dass diese Werte im rechten Block gerade die Elemente der gesuchten Inversen X liefern, die hier eindeutig bestimmt ist. ◯

Beispiel. Wie von der gewöhnlichen Arithmetik aus der Schule her gewöhnt, kann die Nullmatrix nicht invertierbar sein, da keine Matrix X existieren kann mit der Eigenschaft $X \cdot 0 = E$. ◯

Ebenfalls wie in der Arithmetik im Ring $\mathbb{Z}$ der ganzen Zahlen, aber nicht wie beim Rechnen im Körper $\mathbb{R}$ der reellen Zahlen gibt es neben der 0 weitere Matrizen, die nicht invertierbar sind und es geht zunächst darum, sich einen Überblick über die invertierbaren Matrizen zu verschaffen.

Beispiel. Beispielsweise ist die quadratische Matrix $A = \begin{pmatrix} 3 & 6 \\ 1 & 2 \end{pmatrix}$ nicht invertierbar.

Um das einzusehen, machen wir für X den Ansatz als beliebige 2×2-Matrix:

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Dann gilt für das Produkt:

$$A \cdot X = \begin{pmatrix} 3 & 6 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 3a + 6c & 3b + 6d \\ a + 2c & b + 2d \end{pmatrix}$$

Damit dieses Produkt die Einheitsmatrix liefert, müssen die beiden linearen Gleichungssysteme

$$\begin{cases} 3a + 6c = 1 \\ a + 2c = 0 \end{cases} \qquad \begin{cases} 3b + 6d = 0 \\ b + 2d = 1 \end{cases} \qquad \left(\begin{array}{cc|cc} 3 & 6 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array}\right)$$

erfüllt sein. Ein Blick auf die beiden Gleichungen des linken Systems zeigt, dass sich diese beiden Gleichungen widersprechen. Entsprechendes gilt für das rechte System. Daher kann für diese Matrix A, die wir bereits als Nullteiler erkannt haben, keine Inverse existieren.

Man beachte bei dieser Gelegenheit, dass die entstandenen Gleichungssysteme beide die Matrix A als Koeffizientenmatrix haben und sich nur in den Konstantenvektoren unterscheiden. ◯

Obwohl also nicht jede von 0 verschiedene quadratische Matrix invertierbar ist, kann eine invertierbare Matrix immerhin nicht mehr als eine Inverse haben. Der folgende Satz zeigt, dass jede invertierbare Matrix genau eine Inverse hat.

Satz. Sind X und X' inverse Matrizen der quadratischen Matrix A, so ist $X = X'$.

Beweis. Da X eine Inverse von A ist, gilt $X \cdot A = E$. Multiplizieren wir beide Seiten dieser Gleichung von rechts mit X', so folgt $(X \cdot A) \cdot X' = E \cdot X' = X'$. Andererseits ist $A \cdot X' = E$. Multiplizieren wir die Gleichung von links mit X, so ergibt sich $X \cdot (A \cdot X') = X \cdot E = X$. Wegen des Assoziativgesetzes ist also $X = X'$. □

Als Folge dieses Satzes können wir jetzt von *der* Inversen einer invertierbaren Matrix A reden. Wir bezeichnen diese Inverse mit A^{-1}. Es gilt als:

$$A \cdot A^{-1} = A^{-1} \cdot A = E$$

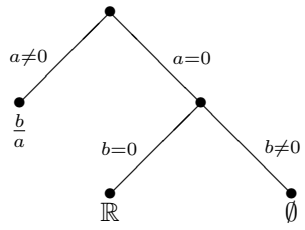
Die Inverse einer Matrix A entspricht dem Kehrwert einer Zahl in der Arithmetik. Sie ist eindeutig bestimmt, falls sie existiert.

Weil man einer quadratischen Matrix in der Regel nicht ansieht, ob sie invertierbar ist und wie ihre Inverse aussieht, stellen sich zwei Fragen.

1. Gibt es Kriterien für die Invertierbarkeit einer quadratischen Matrix?
2. Gibt es einen Algorithmus für die Berechnung der inversen Matrix, falls sie existiert?

Beide Fragen lassen sich mit Hilfe des Eliminationsverfahrens allgemein beantworten. Aus der Definition der Inversen zeigt sich ferner, dass ein enger Zusammenhang zwischen dem Berechnen inverser Matrizen und dem Lösen gewisser linearer Gleichungssysteme besteht, der für das effiziente Berechnen von Inversen von Bedeutung ist.

Tatsächlich lässt sich die inverse Matrix zur Lösung der meisten linearen Gleichungssystemen benutzen, die in der Schule unter der Rubrik Dreisatzrechnung aufgetaucht sind. Die Kurzschreibweise $A \cdot \vec{x} = \vec{b}$ für ein lineares Gleichungssystem erinnert einen an die simpelste skalare Gleichung der Form $a \cdot x = b$, deren Lösung sich für $a \neq 0$ durch Multiplikation mit $\frac{1}{a}$ ergibt. In diesem Normalfall ist nämlich $x = \frac{b}{a}$. Falls allerdings $a = 0$ ist, ist eine weitere Fallunterscheidung notwendig. Falls zusätzlich $b = 0$ ist, hat die Gleichung alle Zahlen aus $\mathbb{R}$ d.h. unendlich viele Lösungen. Falls aber $b \neq 0$ so ist die Lösungsmenge der Gleichung die leere Menge $\emptyset$. Dieses für lineare Gleichungen typische Lösungsverhalten erkennt man in folgendem Baum.



Der Leser sei jetzt schon gewarnt, dass die beiden Sonderfälle für die Anwendungen nicht etwa unnütz sind, sondern eine grosse Rolle spielen werden. Beispielsweise sind viele statische Systeme statisch unbestimmt oder gar dynamisch unbestimmt, was der Tatsache entspricht, dass man in einem der beiden Sonderfälle ist.

Einen entsprechenden Lösungsweg kann man auch für gewisse Matrizengleichungen $A \cdot \vec{x} = \vec{b}$ einschlagen. Für quadratische Koeffizientenmatrizen kann man den folgenden Normalfall nämlich analog behandeln.

Satz. Falls die quadratische Matrix A invertierbar ist, so hat das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ eine eindeutige Lösung $\vec{x} = A^{-1} \cdot \vec{b}$.

Beweis. Multiplikation der Gleichung $A \cdot \vec{x} = \vec{b}$ von links mit der Matrix A^{-1} liefert $\vec{x} = A^{-1} \cdot \vec{b}$. Die Lösung ist eindeutig, denn falls $\vec{y}$ eine weitere Lösung wäre, so wäre also $A \cdot \vec{y} = \vec{b}$. Wie oben ist aber $\vec{y} = A^{-1} \cdot \vec{b}$ und daher $\vec{x} = \vec{y}$. □

Dieses Verfahren zur Lösung von linearen Gleichungssystemen mit Hilfe der inversen Matrix ist in vielen Taschenrechnern implementiert. Es ist aus folgenden Gründen jedoch *nicht* zu empfehlen:

1. Das Verfahren funktioniert nur für gewisse Gleichungssysteme.

2. Die Berechnung der inversen Matrix erfordert etwa dreimal so viel Rechenaufwand als die später zu besprechende direkte Lösungsmethode mit Hilfe des Eliminationsverfahrens

Muss man also ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ für eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ lösen, so ist der direkte Weg über das Eliminationsverfahren zu wählen. Der Umweg über die Berechnung der Inversen A^{-1} führt auf n lineare Gleichungsysteme, was sicher mehr Arbeit gibt! Der Aufwand ist zugegebenermassen nicht n-mal, sondern nur 3-mal so gross, wie beim direkten Vorgehen, weil die n Gleichungssystem alle die selbe Koeffizientenmatrix A haben und deshalb simultan gelöst werden können.

Der empfehlenswerte Algorithmus zum Lösen linearer Gleichungssystemen ist das Eliminitionsverfahren, das einen Aufwand von $\frac{n^3}{3}$ Schritten benötigt und es kann bei Bedarf auch zum Berechnen der Inversen mit einem Aufwand von n^3 Schritten benutzt werden. Man überlegt es sich also besser dreimal, ob man zur Lösung eines Problems wirklich die Inverse benötigt. In der Regel genügt es zu wissen, ob sie existiert.

Auf ein nützliches Kriterium zur Invertierbarkeit einer quadratischen Matrix wollen wir hier schon hinweisen.

Satz. Die quadratische Matrix A hat genau dann eine Inverse, wenn das homogene Gleichungssystem $A \cdot \vec{x} = \vec{0}$ nur die triviale Nulllösung $\vec{x} = \vec{0}$ hat.

Beweis. Falls A invertierbar ist, so hat das Gleichungssystem $A \cdot \vec{x} = \vec{0}$ nach dem letzten Satz die eindeutig bestimmte Lösung $\vec{x} = A^{-1} \cdot \vec{0} = \vec{0}$.

Hat das lineare Gleichungssystem $A \cdot \vec{x}$ eine nicht triviale Lösung $\vec{x} \neq \vec{0}$, so kann die Matrix A nicht invertierbar sein, weil ja sonst die Lösung eindeutig bestimmt wäre, wir aber zwei verschiedene Lösungen gefunden haben. □

Wir wollen den Zusammenhang zwischen dem Berechnen der Inversen und dem Lösen linearer Gleichungssysteme noch an einem etwas grösseren Beispiel demonstrieren.

Beispiel. Falls etwa die Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix}$$

invertierbar ist, muss eine gewisse Matrix X existieren mit der Eigenschaft, dass $A \cdot X = E_3$ gilt. Für X machen wir nun einen Ansatz mit vorläufig unbekannten

Elementen x_{ij}. Dann muss die Matrizengleichung

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix} \cdot \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

gelten. Durch Ausmultiplizieren und Vergleichen entsprechender Elemente geht diese Matrizengleichung in ein lineares Gleichungssystem über, das aus den folgenden 9 Gleichungen besteht, die wir etwas unorthodox angeordnet haben. Rechts daneben findet man dieses System in knapper Form als Blockmatrix dargestellt. In ihrem linken Block steht die Ausgangsmatrix A und in ihrem rechten Block findet man die Einheitsmatrix E_3.

$$\left\{\begin{array}{rcrcrcl} x_{11} & + & 2x_{21} & + & 3x_{31} & = & 1 \\ 2x_{11} & + & 5x_{21} & + & 3x_{31} & = & 0 \\ x_{11} & & & + & 8x_{31} & = & 0 \end{array}\right.$$

$$\left\{\begin{array}{rcrcrcl} x_{12} & + & 2x_{22} & + & 3x_{32} & = & 0 \\ 2x_{12} & + & 5x_{22} & + & 3x_{32} & = & 1 \\ x_{12} & & & + & 8x_{32} & = & 0 \end{array}\right. \qquad \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 5 & 3 & 0 & 1 & 0 \\ 1 & 0 & 8 & 0 & 0 & 1 \end{array}\right)$$

$$\left\{\begin{array}{rcrcrcl} x_{13} & + & 2x_{23} & + & 3x_{33} & = & 0 \\ 2x_{13} & + & 5x_{23} & + & 3x_{33} & = & 0 \\ x_{13} & & & + & 8x_{33} & = & 1 \end{array}\right.$$

Die gesuchte Inverse ergibt sich also aus den Lösungen dieses linearen Gleichungssystems. Man beachte, dass diese 9 Gleichungen nur schwach gekoppelt sind, da sie sich in Teilsysteme von je 3 Gleichungen gruppieren lassen, deren Koeffizientenmatrix jeweils die gegebene Matrix A ist. In der ersten Gruppe sind nur die Unbekannten der ersten Spalte, in der zweiten Gruppe nur jene der zweiten Spalte und in der dritten Gruppe nur diejenigen der dritten Spalte miteinander verknüpft. Die drei Teilsysteme unterscheiden sich also nur in ihrem Konstantenvektor, der jeweils einer der Standardbasisvektoren ist, was zur angegebenen Blockstruktur $(A \mid E)$ führt.

Auch diese Gleichungssysteme lassen sich nun mit dem Eliminationsverfahren simultan lösen. Addition des (-2)-fachen der ersten Zeile zur zweiten Zeile und des (-1)-fachen der ersten Zeile zur dritten Zeile liefert:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 1 & 0 \\ 0 & -2 & 5 & -1 & 0 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ Z_{13}(-1) \end{array}\right]$$

Addition der 2-fachen der zweiten Zeile zur dritten liefert:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 1 & 0 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(2) \end{array}\right]$$

Addition des (-3)-fachen der dritten Zeile zur zweiten und des 3-fachen der dritten Zeile zur ersten ergibt:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 0 & -14 & 6 & 3 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} Z_{31}(3) \\ Z_{32}(-3) \\ \end{array}\right]$$

Addition des (-2)-fachen der zweiten Zeile zur ersten liefert schliesslich im linken Block reduzierte Stufenform.

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -40 & 16 & 9 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(-2) \\ \\ \\ \end{array}\right]$$

Zum Normieren wird nun noch die dritte Zeile mit (-1) multipliziert und wir erhalten die Matrix

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -40 & 16 & 9 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & 1 & 5 & -2 & -1 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ \\ S_3(-1) \end{array}\right]$$

Daraus können wir die gesuchten Lösungen sofort ablesen. Für das erste Teilsystem ergibt sich die eindeutig bestimmte Lösung $x_{11} = -40$, $x_{21} = 13$, $x_{31} = 5$. Für das zweite Teilsystem lautet die Lösung $x_{12} = 16$, $x_{22} = -5$, $x_{32} = -2$. Für das dritte Teilsystem schliesslich findet man die Lösung $x_{13} = 9$, $x_{23} = -3$, $x_{33} = -1$, wie der Leser durch Einsetzen kontrollieren mag. Die gesuchte Inverse muss daher die Matrix

$$A^{-1} = \left(\begin{array}{ccc} -40 & 16 & 9 \\ 13 & -5 & -3 \\ 5 & -2 & -1 \end{array}\right)$$

sein, wie sich durch Kontrolle der Gleichung $A^{-1} \cdot A = E_3 = A \cdot A^{-1}$ direkt bestätigen lässt. Es handelt sich um den rechten Block der erhaltenen Blockmatrix $(E|A^{-1})$, falls im linken Block die Einheitsmatrix E entstanden ist. ◯

Für den Fall der 2×2-Matrizen gibt der folgende nützliche Satz über das Invertieren erschöpfend Auskunft:

Satz. Die Matrix

$$A = \left(\begin{array}{cc} a & b \\ c & d \end{array}\right) \in \mathbb{R}^{2,2}$$

ist genau dann invertierbar, wenn die *Determinante* $\det(A) = ad - bc \neq 0$ ist. In diesem Fall gilt:

$$A^{-1} = \frac{1}{ad - bc} \left(\begin{array}{cc} d & -b \\ -c & a \end{array}\right) \in \mathbb{R}^{2,2}$$

Insbesondere ist eine beliebige solche Matrix mit grosser Wahrscheinlichkeit invertierbar.

Beweis. Wir gehen von der Blockmatrix

$$(A|E_2) = \left(\begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array}\right)$$

aus. Addition des $(-c)$-fachen der ersten Zeile zum a-fachen der zweiten liefert im linken Block die Stufenform

$$\left(\begin{array}{cc|cc} a & b & 1 & 0 \\ 0 & \det(A) & -c & a \end{array}\right)$$

Addition des b-fachen der zweiten Zeile zum $(-\det(A))$-fachen der ersten liefert wegen $-\det(A) - bc = -ad$ im linken Block eine Diagonalmatrix

$$\left(\begin{array}{cc|cc} -a\det(A) & 0 & -ad & ab \\ 0 & \det(A) & -c & a \end{array}\right), \qquad \left(\begin{array}{cc|cc} \det(A) & 0 & d & -b \\ 0 & \det(A) & -c & a \end{array}\right)$$

deren erste Zeile durch $(-a)$ gekürzt haben. Division der beiden Zeilen durch $\det(A)$ liefert die Blockmatrix

$$(E_2|A^{-1}) = \left(\begin{array}{cc|cc} 1 & 0 & \frac{d}{\det(A)} & -\frac{b}{\det(A)} \\ 0 & 1 & -\frac{c}{\det(A)} & \frac{a}{\det(A)} \end{array}\right)$$

in deren linkem Block die Einheitsmatrix und im rechten Block der angegebene Kandidat für die Inverse steht. Weil wir hier salopp über die notwendigen Fallunterscheidung $a \neq 0$ hinweggegangen sind und nur den generischen Fall betrachtet haben, sollte der Leser nun das Ergebnis mindestens kontrollieren und durch Nachrechnen überprüfen, dass $A^{-1} \cdot A = A \cdot A^{-1} = E_2$ gilt. Als Student tut man gut daran, sich diesen einfachen Spezialfall zu merken. □

Die Tatsache, dass man die meisten linearen Prozesse $f_A\colon \mathbb{R}^n \to \mathbb{R}^n$ invertieren kann, kommt etwas unerwartet und ist ein weiteres Charakteristikum der Matrizenrechnung. Immerhin ist uns die analoge Situation vom skalaren Fall der Dreisatzrechnung $a \cdot x = b$ geläufig.

Wir wollen uns nun überlegen, wie sich die Inverse mit den Grundoperationen der Matrizenrechnung verträgt.

Satz. Für Matrizen aus $\mathbb{R}^{n,n}$ gilt:

1. Die Einheitsmatrix E_n ist invertierbar und es gilt $E_n^{-1} = E_n$.
2. Falls A invertierbar ist, so ist auch die Matrix A^{-1} ist invertierbar und es gilt $(A^{-1})^{-1} = A$.
3. Falls A invertierbar ist, so ist für jeden Skalar $c \neq 0$ die Matrix cA invertierbar und es gilt $(cA)^{-1} = \frac{1}{c}A^{-1}$.
4. Mit A ist auch A^T invertierbar. Es gilt $(A^T)^{-1} = (A^{-1})^T$.
5. Für zwei invertierbare Matrizen A, B ist das Produkt $A \cdot B$ invertierbar und es ist $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$.

Beweis. Die Beweise ergeben sich durch Einsetzen der Definitionen.

1. Wegen $E_n \cdot E_n = E_n$ gilt $E_n^{-1} = E_n$.
2. Wegen $A \cdot A^{-1} = A^{-1} \cdot A = E$ ist A^{-1} definitionsgemäss invertierbar und es gilt $(A^{-1})^{-1} = A$.
3. Ist $c \neq 0$, so gilt $(cA)\cdot(\frac{1}{c}A^{-1}) = \frac{1}{c}(cA)\cdot A^{-1} = 1E = E$. Analog erhält man $(\frac{1}{c}A^{-1}) \cdot (cA) = E$. Daher ist cA invertierbar und es gilt die behauptete Beziehung $(cA)^{-1} = \frac{1}{c}A^{-1}$.

4. Es genügt zu zeigen, dass $A^T \cdot (A^{-1})^T = (A^{-1})^T \cdot A^T = E$ gilt. Es ist aber $A^T \cdot (A^{-1})^T = (A^{-1} \cdot A)^T = E^T = E$. Analog zeigt man die andere Gleichung.

5. Wir zeigen, dass $(A \cdot B) \cdot (B^{-1} \cdot A^{-1}) = (B^{-1} \cdot A^{-1}) \cdot (A \cdot B) = E$ gilt. Dann ist die letzte Behauptung gezeigt. Auf Grund des Assoziativgesetzes ist aber $(A \cdot B) \cdot (B^{-1} \cdot A^{-1}) = A \cdot (B \cdot B^{-1}) \cdot A^{-1} = AEA^{-1} = A \cdot A^{-1} = E$, was die eine Hälfte zeigt. Die andere Hälfte zeigt man analog.

   Das letzte Ergebnis kann rekursiv auf mehrere Faktoren übertragen werden.

Simple Beweise wie diesen, die sich unmittelbar durch Einsetzen der Definitionen ergeben, sollte jeder Student in Zukunft selbstständig führen können. Geeignete numerische Beispiele zur Illustration der Konzepte und als Gegenbeispiele kann er auch selbstständig wählen. □

Weil die Menge der invertierbare Matrizen vom selben Typ nach der letzten Behauptung unter Multiplikation und nach der zweiten Behauptung unter Bilden von Inversen abgeschlossen ist, spielen invertierbare Matrizen eine grosse Rolle als Symmetrien des affinen Raumes $\mathbb{R}^n$ und wir geben dieser Menge einen eigenen Namen.

Definition. Mit $\mathrm{Gl}_n(\mathbb{R})$ bezeichnen wir die Menge aller invertierbaren Matrizen vom Typ $n \times n$. Es ist $\mathrm{Gl}_n(\mathbb{R}) \subseteq \mathbb{R}^{n,n}$.

Aus dem ersten und letzten Teil des soeben bewiesenen Satz folgt, dass $\mathrm{Gl}_n(\mathbb{R})$ eine Gruppe ist. Es handelt sich sogar um eine Lie-Gruppe. Ihr Tangentialraum an E ist die früher angetroffene Lie-Algebra $\mathfrak{gl}(n, \mathbb{R})$ der Dimension n^2.

In der Geometrie werden Abbildungen $f\colon \mathbb{R}^n \to \mathbb{R}^n$, die durch invertierbare Matrizen $A \in \mathrm{Gl}_n(\mathbb{R})$ beschrieben werden, als urprungserhaltende Affinitäten bezeichnet. Es handelt sich um geradentreue, parallelentreue und teilverhältnistreue Automorphismen linearer Räume. Bezeichnet man allgemein eine bijektive und geradentreue Bijektion als Affinitiät, so zeigt man mit Hilfe des Strahlensatzes leicht, dass für $n \geq 2$ jede Affinität von der Art

$$f\colon \mathbb{R}^n \to \mathbb{R}^n, \qquad \vec{x} \mapsto A \cdot \vec{x} + \vec{v}$$

darstellbar ist und daher parallelentreu und teilverhältnistreu ist.

Wir weisen speziell auf die vielleicht unerwartete Reihenfolge der Faktoren in der letzten Eigenschaft hin. Das Produkt invertierbarer Matrizen ist invertierbar und die Inverse des Produktes ist das Produkt der Inversen in umgekehrter Reihenfolge. Das folgende Beispiel soll dies illustrieren.

Beispiel. Für die beiden Matrizen $A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$ und ihr Produkt

$$A \cdot B = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 8 \\ 7 & 11 \end{pmatrix}$$

lassen sich die Inversen mit obigem Satz im Kopf berechnen. Es ist
$A^{-1} = \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix}$, $B^{-1} = \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix}$ und $(A \cdot B)^{-1} = \begin{pmatrix} -11 & 8 \\ 7 & -5 \end{pmatrix}$.

Andererseits gilt in der Tat:

$$B^{-1} \cdot A^{-1} = \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} -11 & 8 \\ 7 & -5 \end{pmatrix}$$

Man achte auf die korrekte Reihenfolge der beiden Faktoren. Das Produkt

$$A^{-1} \cdot B^{-1} = \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix} = \begin{pmatrix} -13 & 8 \\ 5 & -3 \end{pmatrix}$$

hätte ein völlig falsches Ergebnis geliefert. Das wissen wir aus dem Alltag. Während wir am Morgen zuerst die Unterhosen und dann die Hosen anziehen, müssen wir uns am Abend in umgekehrter Reihenfolge ausziehen.

Es ist übrigens auch nicht wahr, dass invertierbare Matrizen unter Addition abgeschlossen sind. Als Gegenbeispiel betrachte man die Matrix

$$A + B = \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix}$$

die nicht invertierbar ist, weil ihre Determinante verschwindet. ◯

Von einer wichtigen Klasse von Matrizen lassen sich die Inversen besonders leicht bestimmen, weil dann die Dualität zu Hilfe kommt.

Definition. Eine quadratische Matrix A heisst *orthogonal*, falls sie die Gleichung $A \cdot A^T = E = A^T \cdot A$ erfüllt. Mit $\mathrm{O}_n(\mathbb{R})$ bezeichnen wir die Menge aller orthogonalen Matrizen vom Typ $n \times n$.

Jede orthogonale Matrix A ist invertierbar und es ist $A^{-1} = A^T$. Daher gilt also $\mathrm{O}_n(\mathbb{R}) \subseteq \mathrm{Gl}_n(\mathbb{R})$. Tatsächlich handelt es sich dabei um eine weitere wichtige Lie-Gruppe, mit der die Symmetrien des Euklidschen Raumes $(\mathbb{R}^n, \langle -, - \rangle)$ beschrieben werden. Ihr Tangentialraum an E ist die Lie-Algebra $\mathfrak{so}(n, \mathbb{R})$ der quadratischen antisymmetrischen Matrixen $A \in \mathbb{R}^{n,n}$ der Dimension $\binom{n}{2} = \frac{n(n-1)}{2}$.

Beispiel. Nicht jede invertierbare Matrix ist orthogonal. Als Gegenbeispiel benutzen wir die invertierbare Matrix A aus dem letzten Beispiel. Es gilt:

$$A^T \cdot A = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 5 \\ 5 & 13 \end{pmatrix} \neq E$$

Daher ist A nicht orthogonal. ◯

Beispiel. Die Matrix[9]

$$A = \begin{pmatrix} \frac{3}{7} & \frac{2}{7} & \frac{6}{7} \\ -\frac{6}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{6}{7} & -\frac{3}{7} \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 3 & 2 & 6 \\ -6 & 3 & 2 \\ 2 & 6 & -3 \end{pmatrix}$$

[9] Die Rationalität von A beruht auf der bemerkenswerten zahlentheoretischen Identität $2^2 + 3^2 + 6^2 = 7^2$, nach der sich die Summe dreier Quadrate manchmal wieder als Quadrat schreiben lässt. Die rationalen Punkte der Einheitssphäre $S^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}$ lassen sich durch die stereographische Projektion

$$x = \frac{2u}{u^2 + v^2 + 1}, \qquad x = \frac{2v}{u^2 + v^2 + 1}, \qquad x = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}$$

parametrisieren. Deshalb lassen sich alle sogn. primitiven pythagoräischen Quadrupel $(a, b, c, d) \in \mathbb{N}^4$, die den Bedingungen $a^2 + b^2 + c^2 = d^2$ und $\mathrm{ggT}(a, b, c) = 1$ genügen, durch $a = m^2 + n^2 - p^2 - q^2$, $b = 2(m \cdot q + n \cdot p)$, $c = 2(n \cdot q - m \cdot p)$, $d = m^2 + n^2 + p^2 + q^2$ prametrisieren. Weitere solche Quadrupel sind $(2, 2, 1, 3), (4, 4, 7, 8), (2, 6, 9, 11)$ und $(1, 4, 8, 9)$.

hingegen ist orthogonal, denn

$$A^T \cdot A = \begin{pmatrix} \frac{3}{7} & -\frac{6}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{3}{7} & \frac{6}{7} \\ \frac{6}{7} & \frac{2}{7} & -\frac{3}{7} \end{pmatrix} \cdot \begin{pmatrix} \frac{3}{7} & \frac{2}{7} & \frac{6}{7} \\ -\frac{6}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{6}{7} & -\frac{3}{7} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = A \cdot A^T$$

Man beachte, dass die Orthogonalität dieser Matrix besagt, dass ihre Spaltenvektoren ein orthonormiertes 3-Bein bilden. ◯

Allgemein wird eine orthogonale Matrix $A \in \mathrm{O}_n(\mathbb{R})$ das orthonormierte n-Bein der Standardbasisvektoren $\vec{e}_1, \ldots, \vec{e}_n$ in das orthonormierte n-Bein ihrer Spaltenvektoren $\vec{a}_1, \ldots, \vec{a}_n$ überführen, für das also definitionsgemäss

$$\langle \vec{a}_i, \vec{a}_j \rangle = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases}$$

gilt. Umgekehrt lässt sich jedes solche orthonormierte n-Bein zu einer orthogonalen Matrix zusammengefassen.

Fasst man allgemeiner ein orthonormiertes k-Bein $\vec{a}_1, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ spaltenweise zur Matrix A zusammen, erhält man eine Matrix $A \in \mathbb{R}^{n,k}$ mit der Eigenschaft $A^T \cdot A = E_k$. Diese Matrizen bilden die sogn. *Stiefel-Mannigfaltigkeit*

$$V_{n,k} = \{A \in \mathbb{R}^{n,k} \mid A^T \cdot A = E_k\} \qquad (k \leq n)$$

der k-Beine in $\mathbb{R}^n$. Ihre Dimension ist $\binom{n}{2} - \binom{n-k}{2} = \frac{(2n-k-1)k}{2}$. Insbesondere ist also $V_{n,1} = S^{n-1}$ die $(n-1)$-Sphäre und $V_{n,n-1} = \mathrm{SO}_n(\mathbb{R})$ ist die spezielle orthogonale Gruppe, weil jedes orthonormierte $(n-1)$-Bein aus $\mathbb{R}^n$ auf genau zwei Arten zu einem orthonormierten n-Bein erweitert werden kann, von denen genau eines positiv orientiert ist. Weil dieses durch Drehung aus dem orthonormierten n-Bein der Standardbasisvektoren erhalten werden kann, beschreibt die Lie-Gruppe $\mathrm{SO}_n(\mathbb{R})$ genau die Drehungen von $\mathbb{R}^n$ bzw. der Sphäre S^{n-1}. Schliesslich gilt $V_{n,n} = \mathrm{O}_n(\mathbb{R})$. Durch Weglassen von n-Beinen erhalten wir eine Folge von surjektiven Abbildungen

$$\mathrm{O}_n(\mathbb{R}) = V_{n,n} \to V_{n,n-1} \to \cdots \to V_{n,2} \to V_{n,1} = S^{n-1}$$

zwischen den Stiefel-Mannigfaltigkeiten. Die Faser der Projektion $V_{n,k+1} \to V_{n,k}$ ist die Sphäre S^{n-k-1}.

Ein orthonormiertes k-Bein von Vektoren $\vec{a}_1, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ ist immer linear unabhängig. Berechnet man nämlich das Skalarprodukt der Linearkombination

$$\sum_{j=1}^{k} x_j \vec{a}_j = \vec{0}$$

mit dem Vektor $\vec{a}_i$, filtern die Regeln für das Rechnen mit Skalarprodukten dank der Orthonormalität der Vektoren des Systems den Koeffizienten x_i heraus und liefern die Beziehung $x_i = 0$, was die Behauptung zeigt.

Umgekehrt lässt sich jedes System linear unabhängiger Vektoren $\vec{a}_1, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ so *orthonormalisieren*, dass ein orthonormiertes System von Vektoren

$\vec{u}_1, \ldots, \vec{u}_k$ aus $\mathbb{R}^n$ entsteht, die aber immer noch den selben Raum aufspannen. Diese Vektoren definiert man rekursiv wie folgt:

$$\vec{u}_1 := \frac{1}{|\vec{a}_1|}\vec{a}_1$$

Um $\vec{u}_{j+1}$ rekursiv zu berechnen, projizieren wir zunächst den gegebenen Vektor $\vec{a}_{j+1}$ orthogonal auf den Unterraum U_j, der von den bereits konstruiereten Vekoren $\vec{u}_1, \ldots, \vec{u}_j$ aufgespannt wird und der mit dem von den gegebenen Vaktoren $\vec{a}_1, \ldots, \vec{a}_j$ aufgespannten Unterraum übereinstimmt. Dann wählen wir für $\vec{w}_{j+1}$ die Differenz von $\vec{a}_{j+1}$ und dieser Orthogonalprojektion und erhalten durch Normieren den Vektor $\vec{u}_{j+1}$. Diese Konstruktion macht $\vec{u}_{j+1}$ orthogonal zum Raum U_j und kann formelmässig wie folgt beschrieben werden:

$$\vec{w}_{j+1} := \vec{a}_{j+1} - \sum_{l=1}^{j} \langle \vec{a}_{j+1}, \vec{u}_l \rangle \vec{u}_l, \qquad u_{j+1} := \frac{1}{|\vec{w}_{j+1}|}\vec{w}_{j+1}$$

Beispiel. Die drei Vektoren in $\mathbb{R}^4$

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{a}_2 = \begin{pmatrix} -1 \\ 4 \\ 4 \\ -1 \end{pmatrix}, \qquad \vec{a}_3 = \begin{pmatrix} 4 \\ -2 \\ 2 \\ 0 \end{pmatrix}$$

sind zwar linear unabhängig, aber nicht orthonormiert. Der Orthonormierungsprozess liefert der Reihe nach die Vektoren

$$\vec{u}_1 = \frac{1}{|\vec{a}_1|}\vec{a}_1 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

$$\vec{w}_2 = \vec{a}_2 - \langle \vec{a}_2, \vec{u}_1 \rangle \vec{u}_1 = \begin{pmatrix} -1 \\ 4 \\ 4 \\ -1 \end{pmatrix} - 3\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{5}{2} \\ \frac{5}{2} \\ \frac{5}{2} \\ -\frac{5}{2} \end{pmatrix}, \qquad \vec{u}_2 = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$$

$$\begin{aligned} \vec{w}_3 &= \vec{a}_3 - \langle \vec{a}_3, \vec{u}_1 \rangle \vec{u}_1 - \langle \vec{a}_3, \vec{u}_2 \rangle \vec{u}_2 = \begin{pmatrix} 4 \\ -2 \\ 2 \\ 0 \end{pmatrix} - 2\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} - (-2)\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ -2 \\ 2 \\ -2 \end{pmatrix}, \qquad \vec{u}_3 = \frac{1}{|\vec{w}_3|}\vec{w}_3 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \end{aligned}$$

Es ist leicht nachzurechnen, dass die drei Vektoren $\vec{u}_1, \vec{u}_2, \vec{u}_3$ orthonormiert sind. Ferner spannt $\vec{u}_1$ nach Konstruktion die selbe Gerade U_1 auf wie $\vec{a}_1$. Analog spannen die beiden Vektoren $\vec{u}_1, \vec{u}_2$ die selbe Ebene U_2 auf wie $\vec{a}_1, \vec{a}_2$ und die drei Vektoren $\vec{u}_1, \vec{u}_2, \vec{u}_3$ spannen den selben Raum U_3 auf wie $\vec{a}_1, \vec{a}_2, \vec{a}_3$. Die zugehörige Matrix

$$A = (\vec{u}_1, \vec{u}_2, \vec{u}_3) = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \in V_{4,3}$$

hat tatsächlich die Eigenschaft $A^T \cdot A = E_3$. ◯

Orthogonale Matrizen werden in der älteren Literatur vielfach als Koordinatentransformationen bezeichnet und können geometrisch aus Drehungen und Spiegelungen zusammengesetzt werden und hängen deshalb eng mit den Kongruenzabbildungen zusammen, die in der Schulgeometrie im Zentrum standen.

Beispiel. Aus der folgenden Untersuchung orthogonaler Matrizen folgt, dass die orthogonale Matrix A des letzten Beispiels eine Faktorisierung der Form

$$A = \begin{pmatrix} \frac{3}{7} & \frac{2}{7} & \frac{6}{7} \\ -\frac{6}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{6}{7} & -\frac{3}{7} \end{pmatrix} = \begin{pmatrix} \frac{16}{21} & \frac{13}{21} & \frac{4}{21} \\ -\frac{11}{21} & \frac{16}{21} & -\frac{8}{21} \\ -\frac{8}{21} & \frac{4}{21} & \frac{19}{21} \end{pmatrix} \cdot \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix} = B \cdot C$$

besitzt. Dabei sind die beiden Faktoren B und C wiederum orthogonal, wie man leicht bestätigt. Sie kommutieren dank der Beziehung $B \cdot C = A = C \cdot B$ bzw. $[B, C] = 0$. Die beiden Wege im folgenden kommutativen Diagramm liefern also das selbe Ergebnis.

$$\begin{array}{ccc} \mathbb{R}^3 & \xrightarrow{C} & \mathbb{R}^3 \\ {\scriptstyle B}\downarrow & & \downarrow{\scriptstyle B} \\ \mathbb{R}^3 & \xrightarrow[C]{} & \mathbb{R}^3 \end{array}$$

Weil sie etwas einfacher zu verstehen sind, untersuchen wir nun die beiden orthgonalen Faktoren B und C und verwenden dann die entwickelten Methoden zur Untersuchung und zur Berechnung der angegebenen Faktorisierung von A.

Der zweite orthogonale Faktor

$$C = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & 2 \\ 2 & 2 & -1 \end{pmatrix}$$

wird sich als Spiegelung an der Ebene mit der Gleichung $x + y - 2z = 0$ entpuppen. Weil Ebenenspiegelungen die Orientierung umkehren, ist also das orthonormierte Dreibein der Spaltenvektoren von C (von links nach rechts!) negativ orientiert. (Linksdreibein) Weil eine Ebenenspiegelung idempotent ist, gilt

$C^2 = E$ und damit ist $C = C^{-1} = C^T$, d.h. C muss insbesondere symmetrisch sein.

Um die angegebene Gleichung der Spiegelebene zu finden, benötigen wir entweder ihren Normalenvektor

$$\vec{a} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

oder zwei typische Vektoren

$$\vec{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

mit denen die Spiegelebene aufgespannt werden kann und die daher orthogonal zu $\vec{a}$ sind. Wegen $C \cdot \vec{v}_j = \vec{v}_j$ bleiben die beiden Vektoren $\vec{v}_j$ unter der Spiegelung C tatsächlich fix und wegen $C \cdot \vec{a} = -\vec{a}$ wird der Vektor $\vec{a}$ unter C in den entgegengesetzten Vektor abgebildet.

Um nun diese beiden Vektoren aus der gegebenen Matrix C zu bestimmen, lösen wir das Eigenwertproblem, d.h. wir suchen die nichttrvialen Lösungen $\vec{v} \neq \vec{0}$ des linearen, homogenen Gleichungssystems

$$C \cdot \vec{v} = \lambda \vec{v}, \qquad (C - \lambda E) \cdot \vec{v} = \vec{0}$$

Um ganzzahlig rechnen zu können, multiplizieren wir diese Gleichung mit 3 und suchen also nun die nichttrivialen Lösungen des homogenen Gleichungssystem $(3C - 3\lambda E) \cdot \vec{v} = \vec{0}$ mit der Koeffizientenmatrix[10]:

$$(3C - 3\lambda E) = \begin{pmatrix} 2-3\lambda & -1 & 2 \\ -1 & 2-3\lambda & 2 \\ 2 & 2 & -1-3\lambda \end{pmatrix}$$

Um keine lästigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die ersten beiden Zeilen

$$\begin{pmatrix} -1 & 2-3\lambda & 2 \\ 2-3\lambda & -1 & 2 \\ 2 & 2 & -1-3\lambda \end{pmatrix}$$

und addieren nun das $(2-3\lambda)$-fache der ersten Zeile zur zweiten und das 2-fachen der ersten zur dritten und erhalten die Matrix

$$\begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 3-12\lambda+9\lambda^2 & 6-6\lambda \\ 0 & 6-6\lambda & 3-3\lambda \end{pmatrix}, \qquad \begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 1-4\lambda+3\lambda^2 & 2-2\lambda \\ 0 & 2-2\lambda & 1-\lambda \end{pmatrix}$$

deren zweite und dritte Zeile wir durch 3 dividiert haben um mit betragsmässig kleineren ganzen Zahlen weiterrechnen zu können. Um weiterhin keine Fallunterscheidungen durchführen zu müssen, vertauschen wir nun die letzten beiden

[10] Um Schreibarbeit zu sparen, unterdrücken wir in Zukunft bei homogenen Systemen die 0-en auf der rechten Seite und arbeiten nur mit der Koeffizientenmatrix weiter.

Zeilen und addieren dann das 3λ-fache der zweiten Zeile zum 2-fachen der dritten, was unbedenklich ist und die folgende Matrix liefert:

$$\begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 2-2\lambda & 1-\lambda \\ 0 & 1-4\lambda+3\lambda^2 & 2-2\lambda \end{pmatrix}, \qquad \begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 2-2\lambda & 1-\lambda \\ 0 & 2-2\lambda & 4-\lambda-3\lambda^2 \end{pmatrix}$$

Addieren wir nun das (-1)-fache der zweiten zur dritten Zeile und kürzen die dritte Zeile durch 3, erhalten wir eine Matrix $S(\lambda)$ in Dreiecksform.

$$\begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 2-2\lambda & 1-\lambda \\ 0 & 0 & 3-3\lambda^2 \end{pmatrix}, \qquad S(\lambda) = \begin{pmatrix} -1 & 2-3\lambda & 2 \\ 0 & 2-2\lambda & 1-\lambda \\ 0 & 0 & 1-\lambda^2 \end{pmatrix}$$

Aus ihr entnehmen wir, dass das Gleichungssystem nur dann nicht triviale Lösungen haben kann, wenn λ eine der beiden Gleichungen $2(1-\lambda)=0$ oder $1-\lambda^2=0$ erfüllt, d.h. eine der Nullstellen des kubischen Polynoms

$$f(\lambda) = 2(1-\lambda)\cdot(1-\lambda^2) = 2(1-\lambda-\lambda^2+\lambda^3)$$

also einer beiden Werte $\lambda_1 = 1$ oder $\lambda_2 = -1$ ist. En passant beachten wir noch, dass die Matrix C Nullstelle dieses Polynoms ist, d.h. es gilt

$$f(C) = 0$$

Weil es sich bei C um eine Ebenenspiegelung handelt, die immer idempotent ist, gilt sogar $C^2 = E$ und daher ist C sogar Nullstelle des sogn. Minimalpolynoms

$$\mu_C(\lambda) = 1-\lambda^2$$

das wir aus obiger Stufenform auch ablesen können.

Die beiden für λ gefundenen Fälle untersuchen wir nun separat.

1. Fall: $\lambda = 1$. Dann lautet die gefundene Dreiecksmatrix

$$S(1) = \begin{pmatrix} -1 & -1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

und wir erkennen aus der ersten Gleichung die angegebene Koordinatengleichung der Spiegelebene, aus der sich die beiden ebenfalls bereits angegebenen Vektoren $\vec{v}_1$ und $\vec{v}_2$ leicht erhalten lassen.

2. Fall: $\lambda = -1$. Dann lautet die gefundene Dreiecksmatrix

$$S(-1) = \begin{pmatrix} -1 & 5 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} -1 & 5 & 2 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} -2 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

wobei wir zunächst die zweite Zeile durch 2 gekürzt und dann das (-5)-fache der zweiten Zeile zum 2-fachen der ersten addiert haben. Daraus entnehmen wir, dass in diesem Fall als Lösungen der Eigenwertgleichung nur die Vielfachen des bereits angegebenen Vektor $\vec{a}$ in Frage kommen.

Selbstverständlich lässt sich aus der Kenntnis dieses Normalenvektors bzw. der Koordinatengleichung der Spiegelebene die Matrix C der Spiegelung berechnen. Diese Ebenenspiegelung kann nämlich durch die lineare Zuordnungsvorschrift

$$\vec{x} \mapsto \vec{x} - 2\frac{\langle \vec{x}, \vec{a} \rangle}{|\vec{a}|^2}\vec{a}$$

berechnet werden. Man erkennt leicht, dass dadurch der Vektor $\vec{a}$ in den entgegengesetzten Vektor $-\vec{a}$ übergeführt wird und jeder Vektor $\vec{x} \perp \vec{a}$ fix bleibt. Diese lineare Abbildung kann durch die Matrix

$$S_{\vec{a}} = E - \frac{2}{|\vec{a}|^2}\vec{a} \cdot \vec{a}^T$$

beschrieben werden. Für den gefundenen Eigenvektor

$$\vec{a} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}, \qquad |\vec{a}|^2 = 6$$

ist tatsächlich

$$\begin{aligned} S_{\vec{a}} &= E_3 - \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ -2 & -2 & 4 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix} = C. \end{aligned}$$

die Matrix, von der wir ausgegangen sind.

Der erste orthogonale Faktor on obiger Faktorisierung

$$B = \begin{pmatrix} \frac{16}{21} & \frac{13}{21} & \frac{4}{21} \\ -\frac{11}{21} & \frac{16}{21} & -\frac{8}{21} \\ -\frac{8}{21} & \frac{4}{21} & \frac{19}{21} \end{pmatrix} = \frac{1}{21}\begin{pmatrix} 16 & 13 & 4 \\ -11 & 16 & -8 \\ -8 & 4 & 19 \end{pmatrix}$$

beschreibt im Gegensatz dazu eine Drehung um die punktweise fixe Drehachse in Richtung des Vektors

$$\vec{a} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

der senkrecht auf der Spiegelebene steht. In der Tat lässt B diesen Vektor $\vec{a}$ fix. Daher ist die Spiegelebene der Spiegelung C auch Drehebene der Drehung B. Weil Drehungen die Orientierung erhalten, ist also das orthonormierte Dreibein der Spaltenvektoren von B (von links nach rechts!) positiv orientiert. (Rechtsdreibein)

Um den Vektor $\vec{a}$ aus der gegebenen Matrix B zu bestimmen, lösen wir diesmal das Eigenwertproblem

$$B \cdot \vec{v} = \lambda\vec{v}, \qquad (B - \lambda E) \cdot \vec{v} = \vec{0}$$

Um auch hier wiederum ganzzahlig rechnen zu können, multiplizieren wir diese Gleichung mit 21 und suchen also nun die nichttrivialen Lösungen des homogenen Gleichungssystem $(21B - 21\lambda E) \cdot \vec{v} = \vec{0}$ mit der Koeffizientenmatrix

$$(21B - 21\lambda E) = \begin{pmatrix} 16 - 21\lambda & 13 & 4 \\ -11 & 16 - 21\lambda & -8 \\ -8 & 4 & 19 - 21\lambda \end{pmatrix}$$

Um keine lästigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die ersten beiden Zeilen

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 16 - 21\lambda & 13 & 4 \\ -8 & 4 & 19 - 21\lambda \end{pmatrix}$$

Dann addieren dann das $(16 - 21\lambda)$-fache der ersten Zeile zum 11-fach der zweiten Zeile und das (-8)-fache der ersten Zeile zum 11-fachen der dritten und erhalten die Matrix

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & 399 - 672\lambda + 441\lambda^2 & -84 + 168\lambda \\ 0 & -84 + 168\lambda & 273 - 231\lambda \end{pmatrix}$$

Um mit betragsmässig kleineren ganzen Zahlen weiterrechnen zu können, dividieren wir ihre zweite und dritte Zeile durch 21.

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & 19 - 32\lambda + 21\lambda^2 & -4 + 8\lambda \\ 0 & -4 + 8\lambda & 13 - 11\lambda \end{pmatrix}$$

Um weiterhin keine Fallunterscheidungen durchführen zu müssen, haben wir dann die die letzten beiden Zeilen vertauscht.

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & -4 + 8\lambda & 13 - 11\lambda \\ 0 & 19 - 32\lambda + 21\lambda^2 & -4 + 8\lambda \end{pmatrix}$$

Nun addieren wir das (-21λ)-fache der zweiten zum 8-fachen der dritten Zeile und erhalten die Matrix

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & -4 + 8\lambda & 13 - 11\lambda \\ 0 & 152 - 172\lambda & -32 - 209\lambda + 231\lambda^2 \end{pmatrix}$$

Addition des 43-fachen der zweiten Zeile zum 2-fachen der dritten liefert die Matrix:

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & -4 + 8\lambda & 13 - 11\lambda \\ 0 & 132 & 495 - 891\lambda + 462\lambda^2 \end{pmatrix}$$

Um mit betragsmässig möglichst kleinen ganzen Zahlen weiterrechnen zu können, dividieren wir deren dritte Zeile durch 33.

$$\begin{pmatrix} -11 & 16 - 21\lambda & -8 \\ 0 & -4 + 8\lambda & 13 - 11\lambda \\ 0 & 4 & 15 - 27\lambda + 14\lambda^2 \end{pmatrix}$$

Um keine Fallunterscheidungen machen zu müssen, vertauschen wir nun die letzten beiden Zeilen.

$$\begin{pmatrix} -11 & 16-21\lambda & -8 \\ 0 & 4 & 15-27\lambda+14\lambda^2 \\ 0 & -4+8\lambda & 13-11\lambda \end{pmatrix}$$

Addition des $(1-2\lambda)$-fachen der zweiten Zeile zur dritten Zeile liefert die Dreiecksmatrix

$$\begin{pmatrix} -11 & 16-21\lambda & -8 \\ 0 & 4 & 15-27\lambda+14\lambda^2 \\ 0 & 0 & 28-68\lambda+68\lambda^2-28\lambda^3 \end{pmatrix}$$

deren dritte Zeile wir schliesslich noch durch 4 dividiert haben.

$$S(\lambda) = \begin{pmatrix} -11 & 16-21\lambda & -8 \\ 0 & 4 & 15-27\lambda+14\lambda^2 \\ 0 & 0 & 7-17\lambda+17\lambda^2-7\lambda^3 \end{pmatrix}$$

Aus ihr entnehmen wir, dass das Gleichungssystem nur dann nicht triviale Lösungen haben kann, wenn das Polynom

$$f(\lambda) = 7-17\lambda+17\lambda^2-7\lambda^3 = (1-\lambda)\cdot(7-10\lambda+7\lambda^2)$$

verschwindet. Das ist im Reellen nur für $\lambda = 1$ der Fall. Wiederum stellen wir wir fest, dass unsere Matrix B Nullstelle dieses Polynoms ist, d.h. es gilt

$$f(B) = 0.$$

Daher müssen nichttriviale Vektoren $\vec{v}$ existieren, für die $B\cdot\vec{v} = \vec{v}$ gilt. Um die Menge all dieser Vektoren zu finden, untersuchen wir nun diesen Spezialfall separat und lösen das lineare Gleichungssystem fertig. Für $\lambda = 1$ lautet die gefundene Dreiecksmatrix

$$S(1) = \begin{pmatrix} -11 & -5 & -8 \\ 0 & 4 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -11 & -5 & -8 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

wobei wir die zweite Zeile durch 2 gekürzt haben. Addition des 5-fachen der zweiten Zeile zum 2-fachen der ersten liefert schliesslich eine Matrix in reduzierter Stufenform

$$\begin{pmatrix} -22 & 0 & -11 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

deren erste Zeile wir noch durch (-11) gekürzt haben. Daraus entnehmen wir, dass in diesem Fall als Lösungen der Eigenwertgleichung nur die Vielfachen des angegebenen Vektors

$$\vec{a} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

in Frage kommen, so dass also die durch diesen Vektor aufgespannte Gerade durch den Ursprung unter B fix bleiben muss. Weil es sich dabei um eine Fixpunktgerade handelt, muss B eine eine Drehung beschreiben, deren Achse in

Richtung von $\vec{a}$ zeigt. Daraus bestimmt man nun leicht wie oben die Orthogonalebene durch den Ursprung mit der Koordinatengleichung $x+y-2z=0$. Sie wird beispielsweise durch die beiden Vektoren

$$\vec{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

aufgespannt. Sie werden unter B in die Vektoren

$$B \cdot \vec{v}_1 = \frac{1}{7} \begin{pmatrix} -1 \\ 9 \\ 4 \end{pmatrix} = A \cdot \vec{v}_1, \quad \text{bzw.} \quad B \cdot \vec{v}_2 = \frac{1}{7} \begin{pmatrix} 12 \\ -10 \\ 1 \end{pmatrix} = A \cdot \vec{v}_2$$

übergeführt. Der Drehwinkel ist ihr Zwischenwinkel $\varphi = \angle(\vec{v}_1, B \cdot \vec{v}_1)$ bzw. $\varphi = \angle(\vec{v}_2, B \cdot \vec{v}_2)$, für den wir

$$\cos(\varphi) = \frac{\langle \vec{v}_j, A \cdot \vec{v}_j \rangle}{|\vec{v}_j| \cdot |A \cdot \vec{v}_j|} = \frac{5}{7}, \qquad \varphi = 0.775 = 44.41^\circ$$

erhalten.

Dank der angegebenen Faktorisierung $A = B \cdot C = C \cdot B$ mit $[B, C] = 0$ erkennen wir also nun, dass die ursprüngliche orthogonale Matrix

$$A = \begin{pmatrix} \frac{3}{7} & \frac{2}{7} & \frac{6}{7} \\ -\frac{6}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{6}{7} & -\frac{3}{7} \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 3 & 2 & 6 \\ -6 & 3 & 2 \\ 2 & 6 & -3 \end{pmatrix}$$

als eine Drehspiegelung, bestehend aus der Drehung B um die Achse $\vec{a}$ mit dem Drehwinkel φ, gefolgt von der Ebenenspiegelung C an der Normalenebene zu $\vec{a}$ durch den Urspung. Weil Drehspiegelungen die Orientierung umkehren, ist also das orthonormierte Dreibein der Spaltenvektoren von A (von links nach rechts!) negativ orientiert. (Linksdreibein) In der Tat bildet auch A den Vektor $\vec{a}$ in den negativen Vektor $A \cdot \vec{a} = -\vec{a}$ ab und die beiden zu $\vec{a}$ orthgonalen Vektoren $\vec{v}_1$ und $\vec{v}_2$ werden unter A in Vektoren $A \cdot \vec{v}_1$ und $A \cdot \vec{v}_2$ um den Winkel φ gedreht, die orthogonal zu $\vec{a}$ sind.

Falls man die Achse $\vec{a}$, die Spiegelebene, den Drehwinkel φ und die Faktorisierung $A = B \cdot C$ direkt aus der Matrix A bestimmen will, muss man das Eigenwertproblem

$$A \cdot \vec{v} = \lambda \vec{v}, \qquad (A - \lambda E) \cdot \vec{v} = \vec{0}$$

lösen. Um auch hier wieder ganzzahlig rechnen zu können, multiplizieren wir diese Gleichung mit 7 und suchen also nun die nichttrivialen Lösungen des homogenen Gleichungssystem $(7A - 7\lambda E) \cdot \vec{v} = \vec{0}$ mit der Koeffizientenmatrix

$$(7A - 7\lambda E) = \begin{pmatrix} 3-7\lambda & 2 & 6 \\ -6 & 3-7\lambda & 2 \\ 2 & 6 & -3-7\lambda \end{pmatrix}$$

Um keine lästigen Fallunterscheidungen durchführen zu müssen, vertauschen wir auch hier die beiden ersten Zeilen.

$$\begin{pmatrix} -6 & 3-7\lambda & 2 \\ 3-7\lambda & 2 & 6 \\ 2 & 6 & -3-7\lambda \end{pmatrix}$$

Nun addieren wir das $(3-7\lambda)$-fache der ersten Zeile zum 6-fachen zweiten und die erste Zeile zum 3-fachen dritten und erhalten die Matrix

$$\begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 21-42\lambda+49\lambda^2 & 42-14\lambda \\ 0 & 21-7\lambda & -7-21\lambda \end{pmatrix}, \quad \begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 3-6\lambda+7\lambda^2 & 6-2\lambda \\ 0 & 3-\lambda & -1-3\lambda \end{pmatrix}$$

deren zweite und dritte Zeile wir durch 7 dividiert haben um mit betragsmässig kleineren ganzen Zahlen weiterrechnen zu können. Um weiterhin keine Fallunterscheidungen durchführen zu müssen, vertauschen wir nun die letzten beiden Zeilen und erhalten die Matrix

$$\begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 3-\lambda & -1-3\lambda \\ 0 & 3-6\lambda+7\lambda^2 & 6-2\lambda \end{pmatrix}$$

Addition des (7λ)-fachen der zweiten Zeile zu dritten ist unbedenklich und liefert folgende Matrix:

$$\begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 3-\lambda & -1-3\lambda \\ 0 & 3+15\lambda & 6-9\lambda-21\lambda^2 \end{pmatrix}, \quad \begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 3-\lambda & -1-3\lambda \\ 0 & 1+5\lambda & 2-3\lambda-7\lambda^2 \end{pmatrix}$$

deren dritte Zeile wir noch durch 3 gekürzt haben, um mit betragsmässig möglichst kleinen ganzen Zahlen weiterrechnen zu können. Addition des 5-fachen der zweiten Zeile zur dritten ist unbedenklich und liefert die Matrix

$$\begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 3-\lambda & -1-3\lambda \\ 0 & 16 & -3-18\lambda-7\lambda^2 \end{pmatrix}, \quad \begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 16 & -3-18\lambda-7\lambda^2 \\ 0 & 3-\lambda & -1-3\lambda \end{pmatrix}$$

deren letzte beiden Zeilen wir vertauscht haben, um keine Fallunterscheidungen machen zu müssen. Addieren wir nun das $(3-\lambda)$-fache der zweiten zum (-16)-fachen der dritten Zeile erhalten wir eine Matrix in Dreiecksform.

$$S(\lambda) = \begin{pmatrix} -6 & 3-7\lambda & 2 \\ 0 & 16 & -3-18\lambda-7\lambda^2 \\ 0 & 0 & 7-3\lambda-3\lambda^2+7\lambda^3 \end{pmatrix}$$

Aus ihr entnehmen wir, dass das Gleichungssystem nur dann nicht triviale Lösungen haben kann, wenn das Polynom

$$f(\lambda) = 7-3\lambda-3\lambda^2+7\lambda^3 = (1+\lambda)\cdot(7-10\lambda+7\lambda^2)$$

verschwindet. Das ist im Reellen nur für $\lambda = -1$ der Fall. Daher müssen nichttriviale Vektoren $\vec{v}$ existieren, für die $A\cdot\vec{v} = -\vec{v}$ gilt. Um die Menge all dieser Vektoren zu finden, untersuchen wir nun diesen Spezialfall separat und lösen das lineare Gleichungssystem fertig. Für $\lambda = -1$ lautet die gefundene Dreiecksmatrix

$$S(-1) = \begin{pmatrix} -6 & 10 & 2 \\ 0 & 16 & 8 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} -3 & 5 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

wobei wir die erste Zeile durch 3 und die zweite durch 8 gekürzt haben. Addition des (-5)-fachen der zweiten Zeile zum 2-fachen der ersten liefert schliesslich eine Matrix in reduzierter Stufenform

$$\begin{pmatrix} -6 & 0 & -3 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

deren erste Zeile wir noch durch (-3) gekürzt haben. Daraus entnehmen wir, dass in diesem Fall als Lösungen der Eigenwertgleichung nur die Vielfachen des angegebenen Vektors

$$\vec{a} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

in Frage kommen, so dass also die durch diesen Vektor aufgespannte Gerade durch den Ursprung unter A fix bleiben muss. Weil es sich dabei allerdings nur um eine Fixgerade und nicht um eine Fixpunktgerade handelt, muss A eine Ebenenspiegelung oder eine Drehspiegelung beschreiben, deren Normelenvektor der Spiegelebene in Richtung von $\vec{a}$ zeigt. Daraus bestimmt man nun leicht wie oben die orthogonale Spiegelebene durch den Ursprung mit der Koordinatengleichung $x + y - 2z = 0$ und bestätigt mit Hilfe eines der aufsspannenden Vektoren $\vec{v}_1$ oder $\vec{v}_2$, dass es sich tatsächlich um eine Drehspiegelung handeln muss und bestimmt den Drehwinkel $\varphi = \angle(\vec{v}_1, A \cdot \vec{v}_1)$.

Die an diesem Beispiel beobachteten Sachverhalte gelten allgemein. Tatsächlich lässt sich mit der besprochenen Methode jede orthogonale Matrix $A \in O_3(\mathbb{R})$ mit Hilfe der Lösung des zugehörigen Eigenwertproblems in ein Produkt $A = B \cdot C$ einer Drehung B und ev. einer Ebenenspiegelung C zerlegen. Weil die Drehachse dabei sogar orthogonal zur Spiegelebene gewählt werden kann, gilt $[B, C] = 0$ und bei A handelt es sich, je nach vorhandenen reellen Eigenwerten um eine Ebenenspiegelung ($\lambda = \pm 1$), eine Drehung ($\lambda = 1$) oder um eine Drehspiegelung ($\lambda = -1$).

Eine analoge Klassifizierung der orthogonalen Matrizen findet man auch in höheren Dimensionen, benutzt aber dazu besser die komplexen Zahlen. ◯

Das folgende Resultat besagt, dass orthogonale Matrizen unter Multiplikation und unter Bildung von Inversen abgeschlossen sind.

Satz. Für Matrizen aus $\mathbb{R}^{n,n}$ gilt:

1. Die Einheitsmatrix E_n ist orthogonal.
2. Zu orthogonalen Matrizen A, B ist das Produkt $A \cdot B$ orthogonal.
3. Für eine orthogonale Matrix A ist die inverse Matrix A^{-1} orthogonal.

Man drückt diesen Satz so aus, dass man sagt, $\mathrm{O}_n(\mathbb{R})$ sei eine Untergruppe der Gruppe der invertierbaren Matrizen $\mathrm{Gl}_n(\mathbb{R})$.

Beweis. Die erste Aussage folgt aus den Bedingung $E_n \cdot E_n$ und $E_n^T = E_n$.

Für den Beweis der zweite Aussage müssen wir überlegen, ob die Gleichung $(A \cdot B)^T \cdot (A \cdot B) = E = (A \cdot B) \cdot (A \cdot B)^T$ gilt. Aus den Rechenregeln für Matrizen

folgt aber $(A \cdot B)^T \cdot (A \cdot B) = (B^T \cdot A^T) \cdot (A \cdot B) = B^T \cdot E \cdot B = B^T \cdot B = E$. Die andere Gleichung beweist man analog.

Um die letzte Aussage einzusehen, überlegt man sich, dass die transponierte Matrix einer orthogonalen Matrix erneut orthogonal ist. Dies folgt aus der Gleichung $(A^T)^T \cdot A^T = E = A^T \cdot (A^T)^T$. Für orthogonale Matrizen stimmen aber Transponierte und Inverse überein. □

Orthogonale Abbildungen spielen in der Geometrie eine wichtige Rolle, weil mit ihnen die Kongruenzabbildungen, d.h. die möglichen Bewegungen eines Euklid'schen Raumes $(\mathbb{R}^n, \langle -, - \rangle)$ beschrieben werden. Die Adjunktion nimmt nämlich für eine orthogonale Matrix A die Form

$$\langle A^T \cdot A \cdot \vec{x}, \vec{y} \rangle = \langle A \cdot \vec{x}, A \cdot \vec{y} \rangle = \langle \vec{x}, \vec{y} \rangle$$

an, aus der hervorgeht, dass orthogonale Abbildungen das Skalarprodukt (und damit alle damit ausdrückbaren metrischen Eigenschaften wie Längen, Winkel, Volumina) erhalten. Der Speziallfall $\vec{x} = \vec{y}$ liefert für alle Vektoren die Längenerhaltung

$$|A \cdot \vec{x}|^2 = \langle A \cdot \vec{x}, A \cdot \vec{x} \rangle = \langle \vec{x}, \vec{x} \rangle = |\vec{x}|^2$$

Aus ihr folgt also, dass sich unter orthogonalen Abbildungen die Längen von Vektoren nicht ändern.

Umgekehrt muss eine längentreue lineare Abbildung automatisch orthogonal sein. Für zwei beliebige Vektoren $\vec{x}$ und $\vec{y}$ gilt nämlich dank der ersten Binomischen Formel des Skalarproduktes

$$|\vec{x} + \vec{y}|^2 = |\vec{x}|^2 + |\vec{y}|^2 + 2\langle \vec{x}, \vec{y} \rangle$$

und wegen der Linearität entsprechend

$$|A \cdot (\vec{x} + \vec{y})|^2 = |A \cdot \vec{x} + A \cdot \vec{y}|^2 = |A \cdot \vec{x}|^2 + |A \cdot \vec{y}|^2 + 2\langle A \cdot \vec{x}, A \cdot \vec{y} \rangle$$

Aus der Längentreue von A folgt also, dass unter A das Skalarprodukt erhalten bleibt und wegen der Adjuktion die Bedingung

$$\langle \vec{x}, \vec{y} \rangle = \langle A \cdot \vec{x}, A \cdot \vec{y} \rangle = \langle A^T \cdot A \cdot \vec{x}, \vec{y} \rangle.$$

Dank der Nichtdegeneriertheit des Skalarproduktes folgt also, dass A orthogonal sein muss.

Falls zu einer orthogonaler Matrix A ein Eigenvektor $\vec{v}$ existiert mit $A \cdot \vec{v} = \lambda \vec{v}$, so gilt dank der Längentreue

$$|\vec{v}|^2 = |A \cdot \vec{v}|^2 = |\lambda \vec{v}|^2 = \lambda^2 \cdot |\vec{v}|^2$$

d.h. $\lambda^2 = 1$. Reelle Eigenwerte orthogonaler Matrizen müssen also ± 1 sein.

Das Skalarprodukt und die Orthogonalität haben neben der geometrischen auch algebraische und numerische Bedeutung, weil mit ihnen lineare Probleme speziell effizient und stabil gelöst werden können.

Aus diesen Sätzen folgt, dass die Menge der orthogonalen Matrizen eine Untergruppe $\mathrm{O}_n(\mathbb{R}) \subset \mathrm{Gl}_n(\mathbb{R})$ ist. Es handelt sich um eine weitere (kompakte) Lie-Gruppe. Ihr Tangentialraum an E ist ebenfalls die Lie-Algebra $\mathfrak{so}(n, \mathbb{R})$ der

antisymmetrischen Matrizen und ihre Dimension ist $\frac{n(n-1)}{2}$. Ein Blick auf diese Dimensionen zeigt die interessante Tatsache

n	1	2	3	4	5	$\cdots$
$\dim(\mathrm{O}_n(\mathbb{R}))$	0	1	3	6	10	$\cdots$

dass sich die Dimension 3 unseres Anschauungsraumes dadurch auszeichnet, dass nur gerade dort Raumdimension und Dimension der orthogonalen Gruppe (Kongruenzabbildungen) übereinstimmen. Dieser Artefakt hat einige Überraschungen und einige Irrtümer zur Folge.

2.5 Potenzen

Viele Anwendungen der Matrizenrechnung im Umkreis dynamischer Systeme führen auf Potenzen einer quadratischen Matrix A. Ihre Potenzen definieren wir, genau wie bei den Skalaren, rekursiv.

Definition. Für eine quadratische Matrix A und eine natürliche Zahl $k \in \mathbb{N}$ definieren wir die Potenzen A^k:

$$\left\{ \begin{array}{lcl} A^0 & = & E \\ A^{k+1} & = & A \cdot A^k \end{array} \right.$$

Wie bei den Skalaren, ist es vernünftig, die Inverse A^{-1} als Potenz mit negativem Exponenten zu interpretieren, falls sie existiert. Damit sind dann für invertierbare Matrizen durch $A^{-k} = (A^{-1})^k$ Potenzen für beliebige ganzzahlige Exponenten erklärt.

Beispiel. Die quadratische Matrix $A = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}$ hat die Potenzen

$$\begin{array}{rcl} A^2 & = & \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} \\ A^3 & = & \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 3 & 2 \end{pmatrix} \\ A^4 & = & \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 5 & 6 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} 11 & 10 \\ 5 & \mathbf{6} \end{pmatrix} \end{array}$$

Interpretieren wir die Matrix $A \in \mathbb{N}^{n,n}$, wie jede quadratische Matrix mit natürlichen Zahlen als Elemente, als Adjazenzmatrix des zugehörigen gerichteten Graphen

indem wir vom Knoten i zum Knoten j genau a_{ji} Kanten zeichnen, so beschreiben die Elemente der Potenzen A^k die Anzahl Wege der Länge genau k in diesem Graphen. In obigem Graphen gibt es also genau **6** geschlossene Wege der Länge 4 vom Knoten 2 zu sich selber zurück. Diese Schleifen werden gegeben durch

$$\begin{array}{ll} 2 \longrightarrow 1 \longrightarrow 1 \longrightarrow 1 \longrightarrow 2, & 2 \longrightarrow 1 \longrightarrow 2 \longrightarrow 1 \longrightarrow 2, \\ 2 \longrightarrow 1 \longrightarrow 2 \dashrightarrow 1 \longrightarrow 2, & 2 \dashrightarrow 1 \longrightarrow 1 \longrightarrow 1 \longrightarrow 2, \\ 2 \dashrightarrow 1 \longrightarrow 2 \longrightarrow 1 \longrightarrow 2, & 2 \dashrightarrow 1 \longrightarrow 2 \dashrightarrow 1 \longrightarrow 2. \end{array}$$

Diese Interpretation durch Pfade hätte eine weitere Motivation für das Matrizenprodukt geliefert und hat seinerzeit am Anfang der Quantenmechanik, die damals noch Matrizenmechanik hiess, eine zentrale Rolle gespielt. Die Matrizenelemente dienen dort dazu, die Übergangswahrscheinlichkeiten zwischen Quantenzuständen zu beschreiben.

Auch die Matrizenaddition lässt sich mit Hilfe der Anzahl von Pfaden interpretieren. Interessiert man sich nämlich im Beispiel für die geschlossenen Pfade des Knotens 2, deren Länge höchstens 4 beträgt, so kommen zu den bereits bestimmten der genauen Länge 4 noch jene der Länge 1, 2 und 3 dazu. Sie werden durch die Elemente der Summe

$$S(4) = A + A^2 + A^3 + A^4 = \sum_{k=1}^{4} A^k = \begin{pmatrix} 20 & 20 \\ 10 & \mathbf{10} \end{pmatrix}$$

gezählt. Wegen $S(4)_{22} = \mathbf{10}$ sollte der Knoten 2 des Graph neben den bereits aufgelisteten 6 Schleifen der exakten Länge 4 noch vier kürzere Schleifen haben. Man findet sie in der folgenden Tabelle aufgelistet:

Länge	1	2	3
	$\emptyset$	$2 \longrightarrow 1 \longrightarrow 2$	$2 \longrightarrow 1 \longrightarrow 1 \longrightarrow 2$
		$2 \dashrightarrow 1 \longrightarrow 2$	$2 \dashrightarrow 1 \longrightarrow 1 \longrightarrow 2$

Ebenso leicht findet man die $S(4)_{21} = 10$ Wege der Länge höchstens 4 vom Knoten 1 zum Knoten 2 dieses Graphen.

Die vorliegende Matrix A ist invertierbar und ihre Inverse hat die Potenzen

$$A^{-1} = \frac{1}{2}\begin{pmatrix} 0 & 2 \\ 1 & -1 \end{pmatrix}, \quad A^{-2} = \frac{1}{4}\begin{pmatrix} 2 & -2 \\ -1 & 3 \end{pmatrix}, \quad A^{-3} = \frac{1}{8}\begin{pmatrix} -2 & 6 \\ 3 & -5 \end{pmatrix}.$$

Man beachte, dass wie beim Rechnen mit Zahlen, Invertieren und Potenzieren vertauschbar sind. Selbstverständlich spielt die Reihenfolge der beiden Faktoren beim Potenzieren wegen des Assoziativgesetzes keine Rolle. ◯

Für das Rechnen mit Matrizenpotenzen gelten die üblichen Potenzgesetze.

Satz. Für eine quadratische Matrix A erfüllt die Matrizenpotenz folgende Funktionalgleichungen.

- Potenzgesetze: Für $r, s \in \mathbb{Z}$ gilt:

  20. $A^r \cdot A^s = A^{r+s}$
  21. $(A^r)^s = A^{r \cdot s}$
  22. $(A^r)^{-1} = \left(A^{-1}\right)^r$

Diese Gesetze lassen sich leicht mit Hilfe der rekursiven Definition und der Definition der Inversen begründen. Es gilt aber im allgemeinen

$$(A \cdot B)^n \neq A^n \cdot B^n$$

weil ja Matrizen in der Regel nicht kommutieren.

Zur Berechnung der Potenz A^k werden auf die naive Art $(k-1)$ Matrixmultiplikationen benötigt, weil durch wiederholte Multiplikation mit A sukzessive die Potenzen $A^2, A^3, \ldots, A^k$ berechnet werden müssen. Obwohl der Aufwand also linear mit dem Exponenten wächst, kann er schnell unerträglich gross werden, weil die Berechnung jedes einzelnen Produktes bei grossen Matrizen einen erheblichen Aufwand erfordert.

Wegen der grossen praktischen Bedeutung der Matrizenpotenzen beschreiben wir jetzt ein Verfahren zur schnellen numerischen Berechnung der Potenz A^k. Dieses Verfahren und Varianten davon spielen an diversen Orten — etwa in der Kryptographie — eine Rolle.

Wir gehen von der Binärentwicklung der natürlichen Zahl k aus. Es gilt

$$k = \sum_{j=0}^{n} b_j 2^j, \qquad b_j \in \mathbb{Z}_2, b_n = 1$$

Die Koeffizienten b_j sind also entweder 0 oder 1 und $n+1$ ist die Anzahl der Binärstellen (Binärlänge) der natürlichen Zahl k. Damit erhalten wir für die gesuchte Potenz:

$$\begin{aligned} A^k &= A^{(\sum_{j=0}^{n} b_j 2^j)} = \prod_{j=0}^{n} (A^{2^j})^{b_j} = (A^{2^0})^{b_0} \cdot (A^{2^1})^{b_1} \cdot (A^{2^2})^{b_2} \cdots (A^{2^k})^{b_n} \\ &= \prod_{0 \le j \le n, b_j = 1} A^{2^j} \end{aligned}$$

Diese Formel liefert eine einfache Idee zur Berechnung von Potenzen.

1. Man berechne die sukzessiven Quadrate A^{2^j} für $0 \le j \le n$.
2. Bestimme A^k als Produkt derjenigen A^{2^j}, für die $b_j = 1$ ist.

Aus den Potenzgesetzen folgt die Verdoppelungsformel $A^{2k} = \left(A^k\right)^2 = A^k \cdot A^k$ und daraus die Identität

$$A^{2^{j+1}} = (A^{2^j})^2$$

mit deren Hilfe $A^{2^{j+1}}$ aus A^{2^j} mittels einer Quadrierung berechnet werden kann. Wir erläutern dieses Verfahren an einem Beispiel, das zeigt, dass der Algorithmus viel effizienter ist, als die naive Multiplikationsmethode.

Beispiel. Für die oben bereits benutzte Matrix $A = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}$ suchen wir die Potenz A^{41}. Die Binärentwicklung des Exponenten lautet

$$41 = 32 + 8 + 1 = 2^5 + 2^3 + 2^0 = 101001_2.$$

Nun berechnen wir sukzessive Quadrate

$$A^2 = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix}$$

$$A^4 = (A^2)^2 = A^{2^2} = \begin{pmatrix} 11 & 10 \\ 5 & 6 \end{pmatrix}$$

$$A^8 = \left(A^{2^2}\right)^2 = A^{2^3} = \begin{pmatrix} 171 & 170 \\ 85 & 86 \end{pmatrix}$$

$$A^{16} = \left(A^{2^3}\right)^2 = A^{2^4} = \begin{pmatrix} 43'691 & 43'690 \\ 21'845 & 21'846 \end{pmatrix}$$

$$A^{32} = \left(A^{2^4}\right)^2 = A^{2^5} = \begin{pmatrix} 2'863'311'531 & 2'863'311'530 \\ 1'431'655'765 & 1'431'655'766 \end{pmatrix}$$

Für die gesuchte Matrix gilt $A^{41} = A \cdot A^{2^3} \cdot A^{2^5}$. Multiplikation liefert die gesuchte Potenz

$$A^{41} = \begin{pmatrix} 1'466'015'503'701 & 1'466'015'503'702 \\ 733'007'751'851 & 733'007'751'850 \end{pmatrix}$$

Wir haben also insgesamt 5 Quadrierungen und zwei Multiplikationen statt der naiven 40 Multiplikationen benötigt. Selbstverständlich werden die involvierten Zahlen immer grösser. Deshalb verwendet man in den technischen Anwendungen oft Skalare aus einem endliche Körper. ◯

Im schlechtestens Fall besteht die Binärdarstellung des Exponenten k aus lauter 1 und dann ist $k = 2^n - 1$ bzw. $n = \log_2(k+1)$. In diesem Fall sind mit diesem Algorithmus insgesamt $n-1$ Quadrierungen und $n-1$ Multiplikationen d.h insgesamt $2n-2 = 2\log_2(k+1)-2$ Multiplikationen statt der $k-1$ erforderlich.

Für praktische Zwecke ist das beschriebene Verdoppelungsverfahren bequem, weil es leicht zu programmieren ist, wie das folgende Fragment zeigt:

```
function matrixpotenz(A:Matrix, k:integer): matrix;
var P: Matrix; m: integer;
begin
      if k = 0 then return E; end;
      P:=A;
      for m:=bitlength(k)-2 to 0 by -1 do
          P := P * P;
          if bittest(k, m) then
             P := P * A;
          end;
      end;
      return P;
end.
```

Die Schleife wird wird der Reihe nach für $m = n-1, n-2, \cdots 1, 0$ d.h. insgesamt `bitlength`$(k) = n+1$ mal durchlaufen. Die Funktion `bittest(k,m)` testet, ob in der Binärdarstellung des Exponenten k die Ziffer $b_j = 1$ ist. Daher sind in jedem Schritt höchstens 2 Multiplikationen nötig, so dass also dieser Algorithmus tatsächlich nur $O\big(\log_2(k)\big)$ Multiplikationen benötigt.

Um den zu Grunde liegenden Algorithmus etwas genauer zu verstehen und um uns von seiner Korrektheit zu überzeugen, definieren wir zur vektoriellen Binärdarstellung des Exponenten

$$k = \big(b_n b_{n-1} \ldots b_1 b_0\big)_2 = \sum_{j=0}^{n} b_j 2^j, \quad b_j \in \mathbb{Z}_2, b_n = 1$$

die approximierende Folge

$$k_m = \big(b_n b_{n-1} \dots b_{m+1} b_m\big)_2 = \sum_{j=m}^{n} b_j 2^j, \qquad 0 \le m \le n$$

Es ist $k_n = 1$ und $k_0 = k$ sowie

$$k_m = \begin{cases} 2k_{m+1} + 1 & \text{falls } b_m = 1 \\ 2k_{m+1} & \text{falls } b_m = 0 \end{cases}$$

Daher gilt mit Hilfe der Verdoppelungsformeln

$$A^{k_m} = \begin{cases} A^{2k_{m+1}+1} = \left(A^{k_{m+1}}\right)^2 \cdot A & \text{falls } b_m = 1 \\ A^{2k_{m+1}} = \left(A^{k_{m+1}}\right)^2 & \text{falls } b_m = 0 \end{cases}$$

Damit kann man also aus $A^{k_{m+1}}$ die Matrix A^{k_m} berechnen und man erhält durch absteigende Iteration über $m = n-1, \dots, 0$ in n Schritten aus $A^{k_n} = A$ die gesuchte Matrix $A^{k_0} = A^k$.

Beispiel. Um A^{100} zu berechnen, gehen wir von der Binärdarstellung

$$100 = 64 + 32 + 4 = (1100100)_2$$

aus. Hier ist $n = 6$ und wir erhalten die approximierende Folge

$k_6 = 1$	$k_3 = (1100)_2 = 2k_4 = 12$
$k_5 = (11)_2 = 2k_6 + 1 = 3$	$k_2 = (11001)_2 = 2k_3 + 1 = 25$
$k_4 = (110)_2 = 2k_5 = 6$	$k_1 = (110010)_2 = 2k_2 = 50$
	$k_0 = (1100100)_2 = 2k_1 = 100$

Mit Hilfe der Verdoppelungsformeln berechnen wir die folgenden Matrizen:

m	b_m	A^{k_m}
6	$b_6 = 1$	$A^{k_6} = A$
5	$b_5 = 1$	$A^{k_5} = A^{2k_6+1} = A^{2k_6} \cdot A = \left(A^{k_6}\right)^2 \cdot A = A^2 \cdot A = A^3$
4	$b_4 = 0$	$A^{k_4} = A^{2k_5} = \left(A^{k_5}\right)^2 = \left(A^3\right)^2 = A^6$
3	$b_3 = 0$	$A^{k_3} = A^{2k_4} = \left(A^{k_4}\right)^2 = \left(A^6\right)^2 = A^{12}$
2	$b_2 = 1$	$A^{k_2} = A^{2k_3+1} = A^{2k_3} \cdot A = \left(A^{k_3}\right)^2 \cdot A = \left(A^{12}\right)^2 \cdot A = A^{25}$
1	$b_1 = 0$	$A^{k_1} = A^{2k_2} = \left(A^{k_2}\right)^2 = \left(A^{25}\right)^2 = A^{50}$
0	$b_0 = 0$	$A^{k_0} = A^{2k_1} = \left(A^{k_1}\right)^2 = \left(A^{50}\right)^2 = A^{100}$

Tatsächlich lässt sich also A^{100} durch 6 Quadrierungen und 2 Multiplikationen d.h. insgesamt durch 8 Matrizenprodukte berechnen, wie es die Faktorisierung $A^{100} = A^{2^6} \cdot A^{2^5} \cdot A^{2^2} = A^{64} \cdot A^{32} \cdot A^4$ erwarten lässt. ◯

Für konkrete Fälle gibt es sogar noch schnellere Exponentiationsalgorithmen, als das beschriebene Verdoppelungsverfahren. Als Beispiel betrachte man die Berechnung von A^{15}. Mit dem Verdoppelungsverfahren sind wegen der Binärdarstellung $15 = (1111)_2$ und der zugehörigen Faktorisierung

$$A^{15} = A \cdot A^2 \cdot A^{2^2} \cdot A^{2^3}$$

3 Quadrierungen und 3 Multiplikationen d.h. insgesamt 6 Multiplikationen erforderlich. Nun lässt sich die Matrix $X = A^3 = A^2 \cdot A$ mit Hilfe von 1 Quadrierung und 1 Multiplikation berechnen. Daraus gewinnt man die gesuchte Potenz $A^{15} = X^5 = X \cdot X^{2^2}$ mit Hilfe von 2 Quadrierungen und 1 Multiplikation. Insgesamt sind also in diesem Fall nur 5 Multiplikationen erforderlich. Das Problem, eine Potenz A^k auf möglichst effiziente Weise zu berechnen, ist erstaunlich kompliziert. Wir verweisen den Leser auf die Literatur [**?**].

Das beschriebene Verdoppelungsverfahren lässt zur numerischen Berechnung von Matrizenpotenzen kaum Wünsche offen. Vom konzeptionellen Standpunkt liefert es, wie die meisten Verfahren der Numerik, jedoch keine Einsichten. Wir wollen uns deshalb um Verfahren kümmern, mit denen sich die Matrizenpotenzen explizit durch elementare Formeln beschreiben lassen. Entscheidend dabei wird sein, dass sich hohe Potenzen einer quadratischen Matrix $A \in \mathbb{R}^{n,n}$ mit Hilfe von Matrizenpotenzen von niedrigerem Grad linear kombinieren lassen. Dass dies möglich sein muss, folgt aus der Tatsache, dass die Menge dieser Matrizen einen Vektorraum der Dimension n^2 bilden, in dem also die $n^2 + 1$ Elemente

$$E, A, A^2, \ldots, A^{n^2}$$

linear abhängig sind und daher eine normierte polynomiale Gleichung der Form

$$a_0 E + a_1 A + a_2 A^2 + \cdots + a_{n^2-1} A^{n^2-1} + A^{n^2} = 0$$

existieren muss. Erstaunlicherweise genügen aber dazu bereits Polynome von viel kleinerem Grad — nämlich von höchstens Grad n. Wir werden sehen, dass das das normierte Polynom kleinsten Grades mit dieser Eigenschaft — das sogn. Minimalpolynom von A — interessante Information über die Matrix A enthält.

Beispiel. In unserem numerischen Beispiel

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

sollte es also eine Linearkombination der Form $A^2 = a_0 E + a_1 A$ geben. Es ist

$$\begin{aligned} A^2 &= a_0 E + a_1 A = a_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a_0 + a_1 & 2a_1 \\ a_1 & a_0 \end{pmatrix} \\ &\stackrel{!}{=} \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} \end{aligned}$$

Um die beiden Koeffizienten a_0 und a_1 zu bestimmen, vergleichen wir in diesem Ansatz die Koeffizienten und erhalten das zugehörige lineare Gleichungssystem

$$\left\{ \begin{array}{rrcl} a_0 & & = & 2 \\ & a_1 & = & 1 \\ a_0 + & a_1 & = & 3 \\ & 2a_1 & = & 2 \end{array} \right.$$

Es hat erstaunlicherweise eine Lösung, die in unserem Beispiel sogar noch eindeutig bestimmt ist! Sie lautet $a_0 = 2$, $a_1 = 1$. Daher ergibt sich für die gesuchte Linearkombination die charakteristische Gleichung

$$A^2 = 2E + A, \qquad -2E - A + A^2 = 0$$

Für die Matrix A spielt also die *charakteristische Gleichung*

$$\chi_A(\lambda) = \det(\lambda E - A) = -2 - \lambda + \lambda^2 = 0 \qquad \text{bzw.} \qquad \lambda^2 = 2 + \lambda$$

eine zentrale Rolle. Das zugehörige *charakteristische Polynom*

$$\chi_A(\lambda) = \det(\lambda E - A) = -2 - \lambda + \lambda^2$$

besitzt die Matrix A als Nullstelle, weil wir soeben gesehen haben, dass die charakteristische Gleichung

$$\chi_A(A) = -2E - A + A^2 = 0 \qquad \text{bzw.} \qquad A^2 = 2E + A.$$

gilt. Es handelt sich sogar um die einfachste Gleichung dieser Art und dank der Normierung ist sie eindeutig bestimmt. ◯

Nach einem fundamentale Resultat von Cayley-Hamilton ist jede quadratische Matrix $A \in \mathbb{R}^{n,n}$ Nullstelle eines gewissen normierte Polynoms vom Grad n. Es wird als *charakteristische Polynom*

$$\chi_A(\lambda) = \det(\lambda E - A) \in \mathbb{R}[\lambda]$$

von A bezeichnet.

Satz. Jede quadratische Matrix $A \in \mathbb{R}^{n,n}$ erfüllt ihr normiertes, charakteristisches Polynom $\chi_A(\lambda)$ vom Grad n. Es gilt also $\chi_A(A) = 0$.

Wir werden diesen überraschenden Satz hier nicht beweisen, da sein Beweis etwas heikel ist. Weil es in den Lehrbüchern weit verbreitet ist, werden wir das charakteristische Polynom von A gelegentlich benutzen. In der Tat interessiert man sich aber gar nicht für das charakteristische Polynom von A, sondern präziser für das eindeutig bestimmte normierte Polynom *kleinsten Grades*, das A als Nullstelle hat, d.h.für das Minimalpolynom von A.

Definition. Das *Minimalpolynom* der Matrix A ist das normierte[11] Polynom $\mu_A(\lambda)$ kleinsten Grades, für das $\mu_A(A) = 0$ gilt.

Um zu sehen, dass das Minimalpolynom tatsächlich im allgemeinen echt kleineren Grad als das charakteristische Polynom haben kann, betrachten wir folgendes enfaches Beispiel.

Beispiel. Die Diagonalmatrix

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

hat das charakteristische Polynom

$$\chi_A(\lambda) = \det(\lambda - A) = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2.$$

[11] Sein Leitkoeffizient ist 1 und daher ist insbesonder $\mu \neq 0$.

Es ist normiert, vom Grad 2 und es gilt $\chi_A(A) = 0$, wie man mit Hilfe der charakteristischen Gleichung $A^2 = 4A - 4E$ leicht bestätigt.

Es ist aber ebenfalls leicht zu sehen, dass das Polynom $\mu_A(\lambda) = \lambda - 2$ auch normiert ist und die Eigenschaft $\mu_A(A) = 0$ hat, da tatsächlich $A - 4E = 0$ gilt. Sein Grad ist jedoch $\deg(\mu) = 1 < 2$. Weil es kein solches Polynom von noch kleineren Grades geben kann, muss

$$\mu_A(\lambda) = \lambda - 1$$

das Minimalpolynom von A sein und es stimmt nicht mit dem charakteristischen Polynom überein.

Der Ansatz als Linearkombination kleinerer Potenzen nach obigem Muster

$$\begin{aligned} A^2 &= a_0E + a_1A = a_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} a_0 + 2a_1 & 0 \\ 0 & a_0 + 2a_1 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} \end{aligned}$$

liefert für die beiden Koeffizienten a_0 und a_1 die lineare Gleichung $a_0 + 2a_1 = 4$. Ihre Lösungsmenge

$$\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

ist nicht eindeutig bestimmt, sondern 1-dimensional. Insbesondere liefert die Wahl $r = 4$ den Werten $a_0 = -4$ und $a_1 = 4$, was den Koeffizienten des charakteristischen Polynoms entspricht. Für $r = 2$ hingegen erhalten wir die Werte $a_0 = 0$ und $a_1 = 2$, was der Linearkombination

$$A^2 = 2A$$

entspricht. Aus ihr ergibt sich durch Kürzen der invertierbaren Matrix A die einfachere Gleichung ersten Grades.

$$A = 2E, \qquad \text{bzw.} \quad A - 2E = 0$$

Sie entspricht dem gefundenen Minimalpolynom $\mu(\lambda) = \lambda - 2$. ◯

Beispiel. Die Matrix

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

hat das kubische Polynom $\chi_A(\lambda) = \lambda^3 - 3\lambda^2 + 3\lambda - 1 = (\lambda - 1)^3$ als charakteristisches Polynom. Weil die Matrix A aber auch Nullstelle des normierten Polynoms $\mu_A(\lambda) = (\lambda - 1)^2 = \lambda^2 - 2\lambda + 1$ vom Grad 2 ist und nicht Nullstelle eines normierten Polynoms vom Grad 1 sein kann, da es sich bei A nicht um das Vielfache der Einheitsmatrix handelt, muss sie das Minimalpolynom $\mu_A(\lambda)$ haben. ◯

Beispiel. Die vier Matrizen

$$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

haben alle das charakteristische Polynom $\chi_A(\lambda) = (\lambda - 2)^4$. Ihre Minimalpolynome sind $(\lambda - 2)$, $(\lambda - 2)^2$, $(\lambda - 2)^3$, $(\lambda - 2)^4$ und damit verschieden. ◯

Auf Grund dieser Beispiele vermutet man zu recht, dass das Minimalpolynom $\mu_A(\lambda) \in \mathbb{R}[\lambda]$ der Matrix A ein Teiler des charakteristischen Polynoms $\chi(\lambda)$ sein muss. Dividieren wir nämlich das charakteristische Polynom durch das Minimalpolynom, erhalten wir die Beziehung

$$\chi = q \cdot \mu + r, \qquad \deg(r) < \deg(\mu)$$

Setzen wir die Matrix A in diese Beziehung ein, erhalten wir nach dem Satz von Cayley-Hamilton

$$0 = \chi(A) = q(A) \cdot \mu(A) + r(A), \qquad \deg(r) < \deg(\mu)$$

Wegen $\mu(A) = 0$ muss $r(A) = 0$ sein, was nur dann der Definition des Minimalpolynoms nicht widerspricht, wenn $r = 0$ und daher $\chi = q \cdot \mu$ ist.

Die Nullstellen des Minimalpolynoms, die man aus seiner Faktorisierung

$$\mu_A(\lambda) = \prod_{i=1}^{k} (\lambda - \lambda_i)^{m_i}$$

ablesen kann, spielen eine zentrale Rolle im Zusammenhang mit der Matrix A, wie wir sehen werden. Man nennt eine solche Nullstelle einen Eigenwert λ_i von A und bezeichnet m_i als seine Vielfachheit. Wir werden später ein Verfahren zur Berechnung des Minimalpolynoms beschreiben. Das Minimalpolynom enthält sehr viel Information über die Matrix A. Es charakterisiert (und nicht das charakteristische Polynom!) das Verhalten der Matrizenpotenzen und der dynamischen Systeme und ist deshalb für die Anwendungen von zentraler Bedeutung.

Mit Hilfe einer polynomialen Gleichung von A vom Grad k lassen sich alle Potenzen von A rekursiv als Linearkombination der Matrizenpotenzen A^k von kleinerem Grad $0 \leq j \leq k - 1$ ausdrücken.

Beispiel. In unserem numerischen Beispiel

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

lassen sich dank der bestimmten charakteristischen Gleichung

$$A^2 = 2E + A$$

alle Potenzen von A als Linearkombination von E und A ausdrücken. Es ist

$$A^k = x_k E + y_k A, \quad k \geq 0$$

mit den rekursiv berechneten konkreten Werten

$$\begin{array}{rcrcr} A^0 & = & 1E & + & 0A \\ A^1 & = & 0E & + & 1A \\ A^2 & = & 2E & + & 1A \\ A^3 & = & 2E & + & 3A \\ A^4 & = & 6E & + & 5A \\ A^5 & = & 10E & + & 11A \\ \multicolumn{5}{c}{\ldots\ldots\ldots\ldots\ldots\ldots} \end{array}$$

Offenbar bilden die Koeffizienten x_k und y_k gewisse Zahlenfolgen

k	0	1	2	3	4	5	6	7	8	9	$\cdots$
x_k	1	0	2	2	6	10	22	42	86	170	$\cdots$
y_k	0	1	1	3	5	11	21	43	85	171	$\cdots$

Um für diese beiden Zahlenfolgen eine rekursive Beschreibung zu finden, die man in diesem einfach Beispiel leicht erraten kann, benutzen wir die Rekursion der Potenzen und erhalten

$$\begin{aligned} A^{k+1} &= A \cdot A^k = A \cdot (x_k E + y_k A) = x_k A + y_k A^2 = x_k A + y_k (A + 2E) \\ &= 2y_k E + (x_k + y_k) A \end{aligned}$$

Daraus und aus der Linearkombination $A^{k+1} = x_{k+1}E + y_{k+1}A$ lesen wir durch Koeffizientenvergleich das System von gekoppelten linearen Differenzengleichungen

$$\left\{ \begin{array}{lcll} x_{k+1} & = & 2y_k & x_0 = 1, y_0 = 0 \\ y_{k+1} & = & x_k + y_k & \end{array} \right.$$

ab. Dieses System lässt sich matriziell formulieren. Definiert man den Zustand $\vec{y}(k)$ und die Matrix B durch

$$\vec{y}(k) = \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}$$

so lässt sich dieses System in der Form

$$\vec{y}(k+1) = B \cdot \vec{y}(k)$$

beschreiben. Man beachte, dass die gefundene Matrix B eine einfache Struktur hat und sich sofort aus dem charakteristischen Polynom

$$\chi_A(\lambda) = -2 - \lambda + \lambda^2$$

von A ablesen lässt. Ihre letzte Spalte besteht aus den negativen Koeffizienten des charakteristischen Polynoms $\chi_A(\lambda)$ von A und ihre restlichen Spalten bestehen aus den Standardbasisvektoren mit Ausnahme von $\vec{e}_1$. Wir werden solche Matrizen als transponierte Begleitermatrizen des charakteristischen Polynoms bezeichnen, d.h. es gilt jeweils

$$B = B(\chi_A(\lambda))^T$$

Durch Elimination je einer Variablen erhalten wir entkoppelte lineare Rekursionsgleichungen

$$x_{k+2} = x_{k+1} + 2x_k, \qquad x_0 = 1, x_1 = 0$$

und

$$y_{k+2} = y_{k+1} + 2y_k, \qquad y_0 = 0, y_1 = 1$$

in denen jede Folge einzeln vorkommt dafür aber die Rekursion zweiter Ordnung ist. Man beachte, dass die Struktur der beiden Rekursionsgleichungen übereinstimmt und sie sich sofort aus dem charakteristischen Polynom der Matrix A ablesen lassen. Die beiden Folgen unterscheiden sich also einzig in ihren Anfangsbedingungen, für die man die Einheitsvektoren benutzen kann.

Mit Hilfe dieser Rekursionsgleichungen lassen sich die Koeffizienten berechnen. In unserem Beispiel ist

$$x_{41} = 733'007'751'850, \qquad y_{41} = 733'007'751'851$$

Damit erhalten wir für die Potenz $A^{41} = x_{41}E + y_{41}A$ die bereits mit dem Verdoppelungsverfahren berechnete Matrix

$$A^{41} = \begin{pmatrix} 1'466'015'503'701 & 1'466'015'503'702 \\ 733'007'751'851 & 733'007'751'850 \end{pmatrix}.$$

Mit den gleich zu besprechenden Methoden werden wir sehen, dass sich diese beiden Folgen sogar explizit durch die Formeln

$$x_k = \frac{1}{3}\cdot 2^k + \frac{2}{3}\cdot(-1)^k, \qquad y_k = \frac{1}{3}\cdot 2^k - \frac{1}{3}\cdot(-1)^k$$

beschreiben lassen, wie man durch Einsetzen in die Rekursionsgleichungen und der Anfangsbedingungen leicht verifiziert. Wir werden gleich besprechen, wie man diese Formeln systematisch produzieren kann, beachten aber schon jetzt, dass es sich um Linearkombinationen der Potenzen der beiden Zahlen $\lambda_1 = 2$ und $\lambda_2 = -1$ handelt, die Lösungen der charakteristischen Gleichung

$$\chi_A(\lambda) = -2 - \lambda + \lambda^2 = 0 = (\lambda - 2)\cdot(\lambda + 1)$$

sind.

Damit erhalten wir für gesuchte Matrizenpotenz die explizite Beschreibung

$$\begin{aligned} A^k &= x_k E + y_k A = x_k \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + y_k \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} x_k + y_k & 2y_k \\ y_k & x_k \end{pmatrix} \\ &= \frac{1}{3}\begin{pmatrix} 2\cdot 2^k + (-1)^k & 2\cdot 2^k - 2\cdot(-1)^k \\ 2^k - (-1)^k & 2^k + 2\cdot(-1)^k \end{pmatrix} \end{aligned}$$

Der Leser möge diese Beziehung zunächst an einigen der bereits berechneten Potenzen kontrollieren und insbesondere durch Einsetzen von $k = 41$ den früher gefundenen Wert

$$A^{41} = \begin{pmatrix} 1'466'015'503'701 & 1'466'015'503'702 \\ 733'007'751'851 & 733'007'751'850 \end{pmatrix}$$

bestätigen. Weil die Matrix A invertierbar ist, liefert die Formel für $k = -1$ auch die Inverse

$$A^{-1} = -\frac{1}{2}\begin{pmatrix} 0 & -2 \\ -1 & 1 \end{pmatrix}$$

Mit Hilfe der expliziten Formel sind also zum Berechnen der Potenzen von A nur noch elementare Formeln auszuwerten und das Langzeitverhalten der Matrizenpotenzen ist sehr gut erkennbar.

Zur Kontrolle soll der Leser nun die gefundene explizite Formel für A^k durch Einsetzen in die rekursive Definition der Matrizenpotenz $A^{k+1} = A \cdot A^k$ verifizieren. Kontrolliert er dann noch die Anfangsbedingung $A^0 = E$, so gilt die Formel allgemein, d.h. für beliebige Potenzen $k \in \mathbb{N}$. ◯

Ist man also bereit, von den ganzen Zahlen auf reelle und später auf komplexe Zahlen auszuweichen, die als Lösungen polynomialer Gleichungen auftreten können, kann man Matrizenpotenzen sogar mit Hilfe expliziter Formeln berechnen. Wir werden nun ein etwas anderes Vorgehen zur expliziten Berechnung von A^k beschreiben, das Methoden erfordert, die unabhängig vom Problem, Matrizenpotenzen zu berechnen, eine zentrale Rolle spielen.

Beispiel. Um in unserem Beispiel A eine explizite Beschreibung der Potenz A^k zu finden, benötigen wir die skalaren Lösungen der charakteristischen Gleichung

$$\chi_A(\lambda) = -2 - \lambda + \lambda^2 = (\lambda - 2) \cdot (\lambda + 1) = 0.$$

Sie hat in unserem Fall die beiden *Eigenwerte* $\lambda_1 = 2$ und $\lambda_2 = -1$ mit den beiden zugehörigen *Eigenvektoren*

$$\vec{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Die Eigenvektoren erfüllen die fundamentale *Eigenwertgleichung*

$$A \cdot \vec{v}_j = \lambda_j \vec{v}_j \qquad \text{bzw.} \qquad \big(A - \lambda_j E\big) \cdot \vec{v}_j = \vec{0}, \qquad j = 1, 2$$

wie man durch Nachrechnen leicht bestätigt. Es ist nämlich

$$A \cdot \vec{v}_1 = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \mathbf{2} \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Daraus folgt, dass $\mathbf{2}$ tatsächlich ein Eigenwert von A zum Eigenvektor $\vec{v}_1$ ist. Entsprechend zeigt die Rechnung

$$A \cdot \vec{v}_2 = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -\mathbf{1} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

dass auch $-\mathbf{1}$ ein Eigenwert von A zum Eigenvektor $\vec{v}_2$ ist.

Die Eigenwertgleichung erlaubt es, die Eigenwerte und die zugehörigen Eigenvektoren von A durch Lösen eines linearen Gleichungssystems zu bestimmen.

Die Koeffizientenmatrix des homogenen Gleichungssystem der Eigenwertgleichung $(A - \lambda E) \cdot \vec{v} = \vec{0}$ lautet in unserem Beispiel

$$A - \lambda E = \begin{pmatrix} 1 - \lambda & 2 \\ 1 & 0 - \lambda \end{pmatrix}, \qquad \left[\begin{array}{c} T_{12} \\ \\ \end{array}\right]$$

Vertauschen der Zeilen ergibt die Matrix

$$\begin{pmatrix} 1 & -\lambda \\ 1 - \lambda & 2 \end{pmatrix}, \qquad \left[\begin{array}{c} \\ Z_{12}(\lambda - 1) \end{array}\right]$$

Addition des $(\lambda - 1)$-fachen der ersten Zeile zur zweiten Zeile erfordert keine Fallunterscheidung und liefert die Stufenform

$$S(\lambda) = \begin{pmatrix} 1 & -\lambda \\ 0 & \chi(\lambda) \end{pmatrix}$$

Dabei haben wir das *charakteristische Polynom* durch

$$\chi(\lambda) = -2 - \lambda(\lambda - 1) = -2 - \lambda + \lambda^2 = (\lambda - 2) \cdot (\lambda + 1)$$

definiert. Seine Nullstellen sind die Lösungen der charakteristischen Gleichung. Damit die Eigenwertgleichung $(A - \lambda E) \cdot \vec{v} = \vec{0}$ bzw. das nach äquivalenten Umformungen erhaltene gleichwertige System $S(\lambda) \cdot \vec{v} = \vec{0}$ nichttriviale Lösungen hat, muss λ eine Nullstelle dieses charakteristischen Polynoms sein. Dafür kommen also nur die beiden Eigenwerte $\lambda_1 = 2$ und $\lambda_2 = -1$ in Frage. Setzt man diese Eigenwerte in der gefundenen Stufenform ein, erhält man die zugehörigen Lösungen.

1. Fall: $\lambda_1 = 2$. Dann nimmt die Stufenform $S(\lambda)$ folgende Gestalt an:

$$S(2) = \begin{pmatrix} 1 & -2 \\ 0 & 0 \end{pmatrix}$$

Als mögliche nichttriviale Lösung dieses Systems kann der erste Eigenvektor

$$\vec{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

benutzt werden. Wir haben oben bereits kontrolliert, dass es sich um einen Eigenvektor von A zum Eigenwert $\lambda_1 = 2$ handelt.

2. Fall: $\lambda_2 = -1$. Dann nimmt die Stufenform $S(\lambda)$ folgende Gestalt an:

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

Als mögliche nichttriviale Lösung dieses Systems kann man den zweiten Eigenvektor

$$\vec{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

benutzen, von dem wir auch bereits wissen, dass es sich um einen Eigenvektor von A zum Eigenwert $\lambda_2 = -1$ handelt.

Mit Hilfe der gefundenen Eigenvektoren als Spalten lässt sich nun die invertierbare Matrix

$$X = (\vec{v}_1, \vec{v}_2) = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}, \qquad X^{-1} = -\frac{1}{3} \begin{pmatrix} -1 & -1 \\ -1 & 2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$$

bilden. Die Eigenwertgleichung

$$A \cdot \vec{x} = \lambda \cdot \vec{x} = \vec{x} \cdot \lambda$$

nimmt dann die matrizielle Form[12] einer Faktorisierung

$$A \cdot X = X \cdot \Lambda$$

an, wobei

$$\Lambda = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$$

[12] Auf der rechten Seite Reihenfolge der Faktoren beachten!

die Diagonalmatrix bezeichnet, die auf der Diagonalen die gefundenen Eigenwerte in der gewählten Reihenfolge hat. Tatsächlich liefert die Koordinatentransformation in die Eigenbasis die Diagonalmatrix

$$\Lambda = X^{-1} \cdot A \cdot X = \frac{1}{3} \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$$

Durch Umstellen ergibt sich daraus die Rücktransformation

$$A = X \cdot \Lambda \cdot X^{-1}$$

Aus dieser sogn. *Diagonalisierung* von A findet man nun leicht die gesuchte Matrizenpotenz als Teleskopprodukt

$$A^k = X \cdot \Lambda^k \cdot X^{-1}$$

Weil Potenzen von Diagonalmatrizen sehr leicht elementweise zu berechnen sind, lässt sich die Matrizenpotenz mit Hilfe des Eigensystems durch die Gleichung

$$\begin{aligned} A^k &= \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2^k & 0 \\ 0 & (-1)^k \end{pmatrix} \cdot \frac{1}{3} \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 2 \cdot 2^k + (-1)^k & 2 \cdot 2^k - 2 \cdot (-1)^k \\ 2^k - (-1)^k & 2^k + 2 \cdot (-1)^k \end{pmatrix} \end{aligned}$$

explizite beschreiben. Sie erfüllt tatsächlich die Rekursion $A^{k+1} = A \cdot A^k$ und $A^0 = E$ der Matrizenpotenz.

Die Diagonalisierung von A liefert für das Matrizenexponential die Matrix

$$e^{At} = \frac{1}{3} \begin{pmatrix} 2 \cdot e^{2t} + e^{-t} & 2 \cdot e^{2t} - 2 \cdot e^{-t} \\ e^{2t} - e^{-t} & e^{2t} + 2 \cdot e^{-t} \end{pmatrix}$$

Sie erfüllt tatsächlich die Differentialgleichung $\left(e^{At}\right)' = A \cdot e^{At}$ und $E^{A0} = E$.

Man beachte rückblickend noch einmal, welch zentrale Rolle die Eigenwerte von A in dieser Formel für A^k spielen: sämtliche Einträge der gefundenen Matrix A^k sind Linearkombinationen

$$c_1 \cdot 2^k + c_2 \cdot (-1)^k, \qquad c_1, c_2 \in \mathbb{R}$$

der Potenzen der Eigenwerte $\lambda_1 = 2$ und $\lambda_2 = (-1)$ von A und die Eigenvektoren legen die Koeffizienten c_1 und c_2 fest. Analog sind die Einträge des Matrizenenxonentials e^{At} die zugehörigen Linearkombinationen

$$c_1 \cdot e^{2t} + c_2 \cdot e^{-t}, \qquad c_1, c_2 \in \mathbb{R}$$

von Exponentialfunktionen.

Man vergleiche die hier vorgestellten Wege zum Lösen linearer Differenzen- bzw. Differentialgleichungen mit konstanten Koeffizienten. Aus mathematisch-geometrischen Gründen ist es vorteilhafter mit einem gekoppelten System erster Ordnung, d.h. mit Matrizen zu arbeiten. Anwender ziehen es hingegen oft vor, das System zunächst zu entkoppeln und müssen dann mit höheren Ordnungen rechnen. Ist das System einmal entkoppelt, bzw. kennt man das charakteristische

Polynom, kann man dank der bekannten Struktur der Lösung einen der beschriebenen Ansätze als Linearkombination von Exponentialfunktionen machen und muss nun mit Hilfe der Anfangswerte die fehlenden Koeffizienten berechnen. Anwender sagen dann oft, sie machten einen Exponentialansatz der Form

$$\lambda^k, \qquad e^{\lambda t}$$

Hat man die strukturellen Zusammenhänge verstanden, erkennt man, dass es sich dabei um äquivalente Vorgehensweisen handelt.

Ein Vorteil der expliziten Beschreibung der Matrizenpotenz A^k gegenüber der blossen numerischen Berechnung besteht darin, dass wir beispielsweise nun leicht untersuchen können, wie sich die Matrizenpotenz A^k für sehr grosse k verhalten wird. In unserem Beispiel gilt für die Beträge der beiden Eigenwerte $|\lambda_1| = 2 > 1$ und $|\lambda_2| = 1$. Für sehr grosses k wird also die geometrische Folge $\lambda_1^k = 2^k$ sehr schnell anwachsen und die zweite geometrische Folge $\lambda_2^k = (-1)^k$ wird um 0 oszillieren. Insgesamt erhalten wir für grosse k die asymptotische Näherung

$$A^k \sim \frac{1}{3} \begin{pmatrix} 2 \cdot 2^k & 2 \cdot 2^k \\ 2^k & 2^k \end{pmatrix}$$

Im numerischen Beispiel ist

$$A^{41} \sim \begin{pmatrix} 1.466015504 \cdot 10^{12} & 1.466015504 \cdot 10^{12} \\ 7.330077519 \cdot 10^{11} & 7.330077519 \cdot 10^{11} \end{pmatrix}$$

eine extrem gute Näherung an den seinerzeit berechneten exakten Wert. Wir erkennen daran auch das Stabilitätsverhalten der Potenzen. Offenbar stabilisieren sich die Potenzen A^k für grosses k genau dann, wenn für die Beträge sämtlicher Eigenwerte

$$|\lambda_j| < 1$$

gilt. An Hand der Lage der Eigenwerte lässt sich also das asymptotische und das Stabilitätsverhalten eines Systems voraussagen. ◯

CAS. Selbstverständlich lassen sich auch hier wieder sämtliche Rechnungen mit Hilfe eines Sage-Kodes maschinell durchführen. Um beispielsweise die ersten 6 Potenzen einer gegebenen Matrix A numerisch zu bestimmen, kann man den folgenden Befehl verwenden:

```
A=matrix([ [1,2], [1, 0] ]); show(A)
A.powers(6)
```

Um das charakteristische Polynom einer Matrix A zu berechnen, definiert man zunächst die Matrix in der üblichen Weise und erteilt dann den Befehl

```
A=matrix([ [1,2], [1, 0] ]); show(A)
f=A.characteristic_polynomial("lambda")
```

Der optionale Parameter in Apostroph bezeichnet die gewünschte Unbestimmte.

Matrizen werden in Polynome eingesetzt, wie erwartet. Um beispielsweise den Satz von Cayley-Hamilton zu überprüfen, verwendet man einfach den zusätzlichen Befehl

```
f(A)
```

Soll das charakteristische Polynom gleich faktorisiert ausgegeben werden, kann man den kurzen Befehl

```
A=matrix([ [1,2], [1, 0] ]); show(A)
f=A.fcp("lambda");show(f)
```

benutzen, dessen Akronym für "factorized characteristic polynomial" steht.

Das Minimalpolynom einer Matrix berechnet man entsprechend:

```
A=matrix([ [1,2], [1, 0] ]); show(A)
mp=A.minimal_polynomial('lambda')
fmp=mp.factor()
```

Um eine Liste der Eigenwerte (das sogn. Spektrum) einer Matrix A auszugeben, verwendet man folgenden Befehl

```
A=matrix([ [1,2], [1, 0] ]); show(A)
A.eigenvalues()
```

Benötigt man zusätzlich zugehörige Eigenvektoren, verwendet man den Befehl

```
A=matrix([ [1,2], [1, 0] ]); show(A)
A.eigenvectors_right()
```

Er liefert je ein Tripel (λ, V_λ, n), bestehend aus einem Eigenwert λ, einer Liste V_λ von Basisvektoren des zugehörenden Eigenraumes

$$V_\lambda = \{\vec{x} \mid A \cdot \vec{x} = \lambda\vec{x}\}$$

und der algebraischen Vilefachheit n von λ.

Die Diagonalisierung einer Matrix A erhält man mit dem Befehl

```
A=matrix([ [1,2], [1, 0] ]); show(A)
A.eigenmatrix_right()
```

Falls die Matrix A diagonalisierbar ist, liefert er ein Paar (Λ, X) bestehend aus einer Diagonalmatrix Λ mit den Eigenwerten von A als Diagonalelemente und einer Transformationsmatrix X mit zugehörigen Eigenvektoren als Spalten. Falls die Matrix A nicht diagonalisierbar ist, enthält die Matrix X Nullspalten und ist deshalb nicht invertierbar, wie der selbe Befehl an Hand einer anderen, nicht diagonalisierbaren, Matrix B zeigt:

```
B=matrix([ [1,1], [0, 1] ]); show(B)
B.eigenmatrix_right()
```

Die Matrizenpotenz A^n einer diagonalisierbaren Matrix A kann damit folgendermassen berechnet werden:

```
var("n")
A=matrix([[1,2],[1,0]])
eigensystem=A.eigenmatrix_right()
D=eigensystem[0]; X=eigensystem[1];
l=D.diagonal()
h(x,n)=x**n
Dpot=diagonal_matrix([h(x,n) for x in l])
pot=X*Dpot*X.inverse();show(pot)
```

Numerische Werte für festes n erhält man dann durch Substitution.

```
pot.subs(n=1)
```

Das Matrizenexponential e^{At} kann mit Hilfe des folgenden Befehls berechnet werden:

```
var("t")
A=matrix([ [1,2], [1, 0] ]); show(A)
prop=(t*A).exp(); show(prop.expand())
```

Numerische Werte für festes t erhält man dann wieder durch Substitution. ◇

Wie bei der Inversen ist es auch beim Eigensystem nützlich, den zweidimensionalen Fall symbolisch vollständig zu kennen.

Beispiel. Wir berechnen das Eigensystem einer beliebigen (2×2)-Matrix.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Dazu gehen wir von der erweiterten Matrix

$$A - \lambda E_2 = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}$$

aus und müssen dafür sorgen, dass das zugehörige homogene System eine nichttriviale Lösung hat. Um keine unnötigen Fälle untersuchen zu müssen, vertauschen wir die beiden Zeilen.

$$\begin{pmatrix} c & d-\lambda \\ a-\lambda & b \end{pmatrix} \qquad \begin{bmatrix} T_{12} \\ T_{21} \end{bmatrix}$$

Addition des $(a-\lambda)$-fachen der ersten Zeile zum $(-c)$-fachen zweiten Zeile liefert

$$\begin{pmatrix} c & d-\lambda \\ 0 & \lambda^2-(a+d)\lambda+(ad-bc) \end{pmatrix} \qquad \begin{bmatrix} \\ ZS_{12}(a-\lambda,-c) \end{bmatrix}$$

Daraus lesen wir das Minimalpolynom

$$\mu_A(\lambda) = \lambda^2 - (a+d)\lambda + (ad-bc) = \lambda^2 - \operatorname{tr}(A)\lambda + \det(A)$$

ab. Seine Koeffizienten lassen sich leicht direkt aus der Matrix ablesen: beim konstanten Koeffizienten handelt es sich um die Determinante $\det(A) = ad - bc$ der Matrix A. Den Koeffizienten der zweithöchsten Potenz von λ bildet man als Summe der Diagonalelemente. Er heisst *Spur* der Matrix A, die als $\operatorname{tr}(A) = a+d$ bezeichnet wird.

Man beachte, dass der Ausnahmefall $c = 0$ das selbe Minimalpolynom liefert. Dann ist nämlich $\mu_A(\lambda) = (a-\lambda)(b-\lambda) = \lambda^2 - \operatorname{tr}(A) + a \cdot d$.

Das Minimalpolynom hat die beiden Nullstellen

$$\lambda_{1/2} = \frac{\operatorname{tr}(A) \pm \sqrt{\operatorname{tr}^2(A) - 4\det(A)}}{2}$$

die verschieden sind, falls die Diskriminante

$$\operatorname{tr}^2(A) - 4\det(A) = (a-d)^2 + 4bc \neq 0$$

ist. Für die zugehörigen Eigenvektoren gilt

$$\vec{v}_1 = \begin{pmatrix} \lambda_1 - d \\ c \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} \lambda_2 - d \\ c \end{pmatrix}$$

wie man leicht bestätigt. Damit erhalten wir die invertierbare Transformationsmatrix

$$X = (\vec{v}_1, \vec{v}_2) = \begin{pmatrix} \lambda_1 - d & \lambda_2 - d \\ c & c \end{pmatrix}, \quad X^{-1} = \frac{1}{c \cdot (\lambda_1 - \lambda_2)} \begin{pmatrix} c & -c \\ d - \lambda_2 & \lambda_1 - d \end{pmatrix}$$

mit ihrer Inversen und die Diagonalmatrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Aus der Diagonalisierung

$$A = X \cdot \Lambda \cdot X^{-1}$$

erhalten wir die Potenz

$$A^k = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} (\lambda_1 - d)(\lambda_1^k - \lambda_2^k) & (d - \lambda_1)(\lambda_1^k - \lambda_2^k) \\ (\lambda_2 - d)(\lambda_1^k - \lambda_2^k) & (d - \lambda_2)\lambda_1^k + (\lambda_1 - d)\lambda_2^k \end{pmatrix}$$

und das Matrizenexponential

$$e^{At} = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} (\lambda_1 - d)(e^{\lambda_1 t} - e^{\lambda_2 t}) & (d - \lambda_1)(e^{\lambda_1 t} - e^{\lambda_2 t}) \\ (\lambda_2 - d)(e^{\lambda_1 t} - e^{\lambda_2 t}) & (d - \lambda_2)e^{\lambda_1 t} + (\lambda_1 - d)e^{\lambda_2 t} \end{pmatrix}$$

wobei für die Differenz der beiden Eigenwerte gilt:

$$\lambda_1 - \lambda_2 = \sqrt{\mathrm{tr}^2(A) - 4\det(A)}.$$

Der Autor kann es einem Studenten nicht verübeln, wenn er diese Formeln nicht auswändig kennt, obwohl er seinerzeit während seines Grundstudiums froh gewesen ist, wenn er sie jeweils schnell zur Hand hatte, um damit neue Ideen zu illustrieren oder auszuprobieren. ◯

Mit expliziten Formeln für Matrizenpotenzen kann das Langzeitverhalten und mit dem Eigenwertsystem auch das Stabilitätsverhalten linearer Prozesse prognostiziert werden. Lineare Prozesse werden durch autonome, lineare Differenzengleichungssysteme beschrieben.

Definition. Unter einem *autonomen, homogenen linearen Differenzengleichungssystem* verstehen wir ein System von n linearen Differenzengleichungen der Art

$$\begin{cases} y_1(k+1) = a_{11}y_1(k) + a_{12}y_2(k) + \cdots + a_{1n}y_n(k) \\ y_2(k+1) = a_{21}y_1(k) + a_{22}y_2(k) + \cdots + a_{2n}y_n(k) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n(k+1) = a_{n1}y_1(k) + a_{n2}y_2(k) + \cdots + a_{nn}y_n(k) \end{cases}$$

Zum besseren Überblick fassen wir die *Zustandsvariablen* $y_j(k)$, d.h. die n gesuchten reellen Folgen $y_j(k)$ zum *Zustandsvektor*

$$\vec{y}(k) = \begin{pmatrix} y_1(k) \\ y_2(k) \\ \cdots \\ y_n(k) \end{pmatrix} \in \mathbb{R}^n, k \in \mathbb{N}$$

der Vektorfolge $k \mapsto \vec{y}(k)$ zusammen.

Zur vollständigen Beschreibung eines diskreten dynamischen Systems benötigen wir noch die Anfangsbedingungen $y_1(0) = a_1, y_2(0) = a_2, \cdots, y_n(0) = a_n$, die wir zum *Anfangszustand*

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$$

zusammenfassen. Definieren wir nun die quadratische *Systemmatrix*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

so lässt sich dieses diskrete *Anfangswertproblem* in der kompakten Form

$$\vec{y}(k+1) = A \cdot \vec{y}(k), \quad \vec{y}(0) = \vec{a}$$

schreiben.

Beim Studium diskreter dynamischer Prozesse geht es darum, einen von k abhängigen Zustand $\vec{y}(k)$ aus dem Anfangszustand $\vec{a}$ so zu bestimmen, dass dieses Differenzengleichungssystem erfüllt ist. In vielen Anwendungen wird k als (diskrete) Zeit interpretiert. Ausgehend vom Anfangszustand $\vec{a}$ erzeugt die Rekursionsgleichung also die Vektorfolge

$$\vec{a}, A\vec{a}, A^2\vec{a}, A^3\vec{a}, A^4\vec{a}, \ldots$$

so dass für den Zustand $\vec{y}(k)$ des Systems nach k solchen diskreten Zeitschritten die Beziehung

$$\vec{y}(k) = A^k \cdot \vec{a}, \qquad k \in \mathbb{N}$$

gilt. Die Zustände zeitdiskreter System sind also nur zu bestimmten Zeitpunkten bestimmt und messbar – etwa so wie die einzelnen Bilder eines Filmes.

Autonome Systeme homogener linearer Differenzengleichungen erster Ordnung verallgemeinern also die geometrischen Folgen auf höhere Dimensionen und zur Beschreibung des Zustandes $\vec{y}(k)$ nach k Zeitschritten benötigen wir die Matrizenpotenz A^k. Ein solcher diskreter dynamischer Prozess kann auch als dynamisches System

$$\Phi\colon \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n, \qquad (\vec{a}, k) \mapsto \vec{y}(k) = A^k \cdot \vec{a}$$

aufgefasst werden, in dem $\mathbb{R}^n$ den Zustandsraum und $\mathbb{N}$ den Zeitraum bezeichnet. Seine Zeitentwicklung geschieht mit Hilfe der Matrizenpotenz A^k, die sich sich mit Hilfe der Diagonalisierung

$$A = X \cdot \Lambda \cdot X^{-1}, \qquad A^k = X \cdot \Lambda^k \cdot X^{-1}$$

von A explizit berechnen lässt. Falls A ausnahmsweise nicht diagonalisierbar sein sollte, benutzt man statt der Digonalmatrix Λ die sogn. Jordan-Matrix J für die entprechend $A = X \cdot J \cdot X^{-1}$ bzw. $A^k = X \cdot J^k \cdot X^{-1}$ gilt und die Potenz J^k ebenfalls einfach berechnet werden kann.

Im Gegensatz dazu wird ein zeitkontinuierlicher dynamischen Prozess durch das *Anfangswertproblem*

$$\vec{y}\,'(t) = A \cdot \vec{y}(t), \qquad \vec{y}(0) = \vec{a}$$

beschrieben. Diesmal wird also die Zeit kontinuierlich interpretiert d.h. ein solches System ist für einen beliebigen Zeitpunkt $t \in \mathbb{R}_{\geq 0}$ definiert und es geht darum, den von der Zeit t abhängigen Zustand $\vec{y}(t)$ aus dem Anfangszustand $\vec{a}$ zu bestimmen. Die Lösung eines solchen autonomen Systems homogener Differentialgleichungen erster Ordnung lässt sich bekanntlich in der Form

$$\vec{y}(t) = e^{At} \cdot \vec{a}, \qquad t \geq 0$$

beschreiben, wobei man zur Berechnung des Propagators e^{At} (Matrizenexponential) wegen seiner Potenzreihendarstellung bzw. auf Grund des Eulerschen Grenzwertes

$$e^{tA} = E + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \frac{t^4}{4!}A^4 + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k = \lim_{n\to\infty} \big(1 + \frac{At}{n}\big)^n$$

ebenfalls die Matrizenpotenzen von A benötigt. Ein solcher kontinuierlicher dynamischer Prozess kann auch als dynamisches System

$$\Phi\colon \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n, \qquad (\vec{a}, t) \mapsto \vec{y}(t) = e^{At} \cdot \vec{a}$$

aufgefasst werden, in dem $\mathbb{R}^n$ den Zustandsraum und $\mathbb{R}_{\geq 0}$ den Zeitraum bezeichnet. Seine Zeitentwicklung geschieht mit Hilfe des Propagators e^{At}, der sich mit Hilfe der Diagonalisierung in der Form

$$e^{At} = X \cdot e^{\Lambda t} \cdot X^{-1}$$

explizit berechnen lässt.

Beispiel. Das konkrete, diskrete dreidimensionale System

$$\left\{ \begin{array}{lcrl} x(k+1) & = & 4x(k) + y(k) + 2z(k) & x(0) = 1 \\ y(k+1) & = & 2y(k) - 4z(k) & y(0) = 0 \\ z(k+1) & = & y(k) + 6z(k) & z(0) = -1 \end{array} \right.$$

wird durch folgende Systemmatrix A, den Zustandvektor $\vec{y}(k)$ und den Anfangszustand $\vec{a}$ beschrieben.

$$A = \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad \vec{y}(k) = \begin{pmatrix} x(k) \\ y(k) \\ z(k) \end{pmatrix} \in \mathbb{R}^3, \quad \vec{a} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \in \mathbb{R}^3$$

Die Komponenten des Anfangszustandes $\vec{a}$ beschreiben die drei Anfangsbedingungen $x(0) = 1$, $y(0) = 0$ und $z(0) = -1$.
Die vektorielle Rekursionsgleichung erster Ordnung

$$\vec{y}(k+1) = A \cdot \vec{y}(k), \quad \vec{y}(0) = \vec{a} \qquad k \in \mathbb{N}$$

des diskreten Anfangswertproblems liefert rekursiv die Vektorfolge

$$\vec{y}(0) = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \vec{y}(1) = \begin{pmatrix} 2 \\ 4 \\ -6 \end{pmatrix}, \quad \vec{y}(2) = \begin{pmatrix} 0 \\ 32 \\ -32 \end{pmatrix},$$
$$\vec{y}(3) = \begin{pmatrix} -32 \\ 192 \\ -160 \end{pmatrix}, \quad \vec{y}(4) = \begin{pmatrix} -256 \\ 1024 \\ -768 \end{pmatrix}, \quad \vec{y}(5) = \begin{pmatrix} -1536 \\ 5120 \\ -3584 \end{pmatrix}, \quad \cdots$$

Um in diesem Beispiel das Eigenwertproblem $(A - \lambda E) \cdot \vec{x} = \vec{0}$ zu lösen, müssen wir untersuchen, wann die Matrix

$$A - \lambda E = \begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 2-\lambda & -4 \\ 0 & 1 & 6-\lambda \end{pmatrix}$$

nicht invertierbar ist und führen sie dazu — möglichst ohne Fallunterscheidungen — in Stufenform über. Um keine Fallunterscheidungen durchführen zu müssen, vertauschen wir die zweite und die dritte Zeile und erhalten die Matrix

$$\begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 1 & 6-\lambda \\ 0 & 2-\lambda & -4 \end{pmatrix}$$

Nun Addieren wir das $(\lambda - 4)$-fache der zweiten Zeile zur dritten und erhalten die Matrix

$$S(\lambda) = \begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 1 & 6-\lambda \\ 0 & 0 & p(\lambda) \end{pmatrix}$$

in Stufenform. Das dabei auftretende quadratische Polynom ist

$$p(\lambda) = (6-\lambda)(\lambda-2) - 4 = -\lambda^2 + 8\lambda - 16 = -(\lambda-4)^2$$

In diesem Fall erhalten wir für das normierte, kubische charakteristische Polynom

$$\chi_A(\lambda) = (\lambda-4)^3 = \lambda^3 - 12\lambda^2 + 48\lambda - 64$$

Die Matrix A erfüllt tatsächlich die charakteristische Gleichung

$$A^3 = 64E - 48A + 12A^2$$

dritter Ordnung, wie man leicht bestätigt und ihr Spektrum $\sigma_A = \{4\}$ besteht aus einem einzigen Eigenwert. Man beachte, dass für ihr Minimalpolynom allerdings

$$\mu_A(\lambda) = \lambda^2 - 8\lambda + 16 = (\lambda-4)^2$$

gilt, weil A sogar die quadratische Gleichung

$$A^2 = 16E + 8A$$

erfüllt. Um zugehörige Eigenvektoren zu bestimmen, untersuchen wir die Matrix

$$S(4) = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

deren Rang offensichtlich 1 ist. Die Lösungsmenge ist also 2-dimensional und hat die vektorielle Darstellung

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$$

Offenbar wird der Eigenraum

$$H_1(4) = \ker(A - 4E)$$

des einzigen Eigenwertes $\lambda_1 = 4$ von A durch die beiden linear unabhängigen Vektoren

$$\vec{v}_{1,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{1,2} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$$

der Basis $\mathcal{B}_1(4)$ aufgespannt.

CAS. Eine Basis des Kerns einer quadratischen Matrix lässt sich sich in Sage mit folgendem Befehl

```
A=matrix([ [0,1,2], [0,1,2], [0,0,0] ]); show(A)
A.right_kernel()
```

leicht bestimmen. ◇

Man beachte, dass diese beiden Vektoren $\vec{v}_{1,1}$ und $\vec{v}_{1,2}$ bloss eine 2-dimensionale Ebene $H_1(4)$ durch den Ursprung aufspannen, die als Eigenraum zum Eigenwert $\lambda = 4$ bezeichnet wird. Weil aber mit diesen beiden Vektoren keine invertierbare Transformationsmatrix $X \in \mathbb{R}^{3,3}$ gebildet werden kann und keine weitere Eigenwerte zur Verfügung stehen, ist die Matrix A nicht diagonalisierbar.

Man beachte auch, dass dank der Homogenität der Eigenwertgleichung

$$(A - \lambda E) \cdot \vec{x} = \vec{0}$$

insbesondere auch die Summe der beiden Eigenvektoren, d.h. der Vektor

$$\vec{w}_{1,2} = \vec{v}_{1,1} + \vec{v}_{1,2} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \in H_1(4)$$

ein Eigenvektor von A zum Eigenwert $\lambda = 4$ ist, der von den beiden gewählten Basisvektoren linear abhängig ist und in der Folge eine Rolle spielen wird.

Weil die Matrix A nicht diagonalisierbar ist, können wir im Moment unser System von Differenzengleichungen nicht wie gewohnt lösen und müssen uns etwas

Neues einfallen lassen. Mit der an Hand dieses Beispiel detaliert zu besprechenden Verallgemeineung der Eigenwert-Methode werden wir für die vorliegende Matrix A eine Faktorisierung

$$A \cdot X = X \cdot J \quad \text{(Jordan-Zerlegung)}$$

finden, wobei diesmal die sogn. Jordanschen Normalform

$$J = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \Lambda + N$$

nur „fast“ eine Diagonalmatrix mit dem einzigen vorhandenen Eigenwert $\lambda = 4$ auf der Diagonalen ist. Dass diesmal nicht einfach die Diagonalmatrix Λ als Normalform verwendet werden kann, liegt daran, dass mit den vorhandenen Eigenvektoren keine invertierbare Transformationsmatrix gebildet werden kann.

Weil die Matrix A nicht diagonalisierbar ist, muss die Diagonalmatrix Λ also noch additiv durch die nilpotente Matrix N, die also „fast“ 0 ist, korrigiert werden. Man beachte, dass die beiden Matrizen Λ und N kommutieren, d.h. dass

$$\Lambda \cdot N = N \cdot \Lambda \qquad \text{bzw.} \qquad [\Lambda, N] = 0$$

gilt[13]. Ferner hat auch die nilpotente Matrix N die Eigenschaft, dass ihre Potenzen einfach zu berechnen sind, weil nämlich $N^2 = 0$ gilt, was ihre Bezeichnungsweise rechtfertigt.

Definition. Eine quadratische Matrix N heisst *nilpotent vom Nilpotenzgrad* $k \geq 1$, falls $N^k = 0$ und $N^{k-1} \neq 0$ gilt.

Eine solche Matrix hat das Minimalpolnom $\mu_N(\lambda) = \lambda^k$ und daher den einzigen Eigenwert $\lambda = 0$.

Als invertierbare Transformationsmatrix kann man diesmal etwa die invertierbare Matrix

$$X = (\vec{w}_{1,2}, \vec{w}_{1,1}, \vec{w}_{2,1}) = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad X^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & -1 \end{pmatrix}$$

verwenden. Für die angegeben Normalform J gilt dann nämlich tatsächlich die behauptete Faktorisierung

$$A \cdot X = X \cdot J.$$

Sie wird genügen, um auch in diesem Fall die Matrizenpotenz A^k bzw. das Matrizenexponential e^{At} von A explizit berechnen und damit Systeme von lineare DIfferenzen- bzw. Differenzialgleichugen explizit lösen zu können.

CAS. Diese sogn. Jordan-Zerlegung (J, X) einer quadratischen Matrix $A \in \mathbb{R}^{n,n}$ kann als Verfeinerung der Diagonalisierung betrachtet werden. Man findet sie in Sage mit dem Befehl

[13] Diagonalmatrizen mit lauter gleicher Elementen auf der Diagonalen kommutieren mit allen Matrizen vom selben Typ! Sie bilden das sogn. Zentrum der Matrizenalgebra.

```
A=matrix([ [4,1,2], [0,2,-4], [0,1,6] ]); show(A)
jordan=A.jordan_form(transformation=True)
J=jordan[0]; show(J)
X=jordan[1]; show(X)
```

Der optionale Parameter bewirkt, dass neben der Jordan-Matrix J auch eine Transformationsmatrix X ausgegebenen wird. ◇

Bevor wir einen Algorithmus besprechen, mit dem man man eine solche Jordan-Zerlegung findet, wollen an dem Beispiel beleuchten, welche speziellen Eigenschaften die Spaltenvektoren der Transformationsmatrix X haben, um damit im Lauf der Zeit zu erkennen, wie man sie systematisch produzieren kann. Für die Transformationsmatrix[14]

$$X = (\vec{w}_{1,2}, \vec{w}_{1,1}, \vec{w}_{2,1})$$

gilt dank der erwünschten Faktorisierung $A \cdot X = X \cdot J$ die Matrixgleichung

$$A \cdot (\vec{w}_{1,2}, \vec{w}_{1,1}, \vec{w}_{2,1}) = (\vec{w}_{1,2}, \vec{w}_{1,1}, \vec{w}_{2,1}) \cdot J$$

die ausmultipliziert den drei Vektorgleichungen

$$A \cdot \vec{w}_{1,2} = 4\vec{w}_{1,2}, \qquad A \cdot \vec{w}_{1,1} = 4\vec{w}_{1,1} + \vec{w}_{1,2}, \qquad A \cdot \vec{w}_{2,1} = 4\vec{w}_{2,1}$$

entspricht. Sie lassen sich nachrechnen, wenn man für die drei sogn. *Hauptvektoren* (verallgemeinerte Eigenvektoren) von A die Spaltenvektoren von X wählt:

$$\vec{w}_{1,2} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \vec{v}_{1,1} + \vec{v}_{1,2}, \qquad \vec{w}_{1,1} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{w}_{2,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \vec{v}_{1,1}$$

Wegen der ersten und der dritten dieser Gleichungen, d.h. wegen

$$A \cdot \vec{w}_{1,2} = 4\vec{w}_{1,2}, \qquad A \cdot \vec{w}_{2,1} = 4\vec{w}_{2,1}$$

sind $\vec{w}_{1,2}$ und $\vec{w}_{2,1}$ Eigenvektoren (Hauptvektoren der Stufe 1) von A zum Eigenwert $\lambda = 4$. Es sind also Lösungen der homogenen Gleichungssysteme

$$(A - 4E) \cdot \vec{w}_{1,2} = \vec{0}, \qquad (A - 4E) \cdot \vec{w}_{2,1} = \vec{0}$$

Diese beiden Vektoren sind linear unabhängig und bilden daher auch eine Basis des Eigenraums $H_1(4)$, d.h. diese beiden Vektoren spannen auch die Ebene $H_1(4)$ auf.

Der dritte Vektor zeigt ein neues Phänomen. Wegen

$$A \cdot \vec{w}_{1,1} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \neq 4\vec{w}_{1,1} = \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix}, \qquad \text{bzw.} \qquad (A - 4E) \cdot \vec{w}_{1,1} \neq \vec{0}$$

[14] Im Laufe der Zeit wird die verwendete — scheinbar eigenartige — Bezeichnungsweise ihrer Spaltenvektoren mit Hilfe von Doppelindizes klar: $\vec{v}_{s,j}(\lambda)$ bezeichnet den j-ten Basisvektor des Basis $\mathcal{B}_s(\lambda)$ des Hauptraumes $H_s(\lambda)$ der Stufe s zum Eigenwert λ und $\vec{w}_{k,s}(\lambda)$ bezeichnet den Vektor in der k-ten Hauptvektorkette auf der Stufe s zum Eigenwert λ. Die Vektoren $\vec{w}_{k,s}$ — und nicht etwa die Vektoren $\vec{v}_{s,j}$ — tauchen in der Regel in den Spalten von X auf! Es geht gerade darum, aus den berechneten Basisvektoren $\vec{v}_{s,j}$ *genügend viele geeignete* Basisvektoren $\vec{w}_{k,s}$ zu produzieren um damit X zu erklären.

ist $\vec{w}_{1,1}$ jedoch *kein* Eigenvektor von A — er liegt, geometrisch gesprochen, nicht in der Ebene $H_1(4)$. Ein Blick auf die mittlere der drei Vektorgleichungen lässt vermuten, dass stattdessen folgende Gleichung gilt:

$$A \cdot \vec{w}_{1,1} - 4\vec{w}_{1,1} = (A - 4E_3) \cdot \vec{w}_{1,1} = \vec{w}_{1,2}$$

Diese Vermutung lässt sich durch eine simple Rechnung bestätigen. Es ist:

$$\begin{aligned} A \cdot \vec{w}_{1,1} - 4\vec{w}_{1,1} = (A - 4E_3) \cdot \vec{w}_{1,1} &= \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \\ &= \vec{w}_{1,2} \neq \vec{0}. \end{aligned}$$

Aus dem Vektor $\vec{w}_{1,1}$ entsteht also unter der linearen Abbildung $A - 4E$ *nicht*, wie für einen Eigenvektor, der Nullvektor $\vec{0}$, sondern dank der dritten Gleichung

$$A \cdot \vec{w}_{1,1} = 4\vec{w}_{1,1} + \vec{w}_{1,2}, \quad \text{bzw.} \quad (A - 4E) \cdot \vec{w}_{1,1} = \vec{w}_{1,2}$$

immerhin ein Eigenvektor von A! Der Vektor $\vec{w}_{1,1}$ ist gewissermassen ein tempierter Eigenvektor: der Zeitzünder ist auf eine Zeiteinheit eingestellt.

Vektorfolgen der Form $\vec{w}_{1,1}$, $\vec{w}_{1,2}$ mit der gefundenen Eigenschaft, dass

$$(A - \lambda E) \cdot \vec{w}_{1,1} = \vec{w}_{1,2}, \qquad (A - \lambda E) \cdot \vec{w}_{1,2} = \vec{0}$$

gilt, bezeichnet man als Folgen verallgemeinerter Eigenvektoren der Länge 2. Dieser Begriff ist für beliebige Längen sinnvoll.

Definition. Unter einer *Folge verallgemeinerter Eigenvektoren der Länge l* der quadratischen Matrix A zum Eigenwert λ verstehen wir eine Folge von Vektoren

$$\vec{w}_1, \vec{w}_2, \vec{w}_3, \ldots, \vec{w}_{l-1}, \vec{w}_l, \vec{0}$$

mit der Eigenschaft, dass die Bedingungen

$$(A - \lambda E) \cdot \vec{w}_k = \vec{w}_{k+1}, \quad 1 \leq k \leq (l-1), \qquad (A - \lambda E) \cdot \vec{w}_l = \vec{0}$$

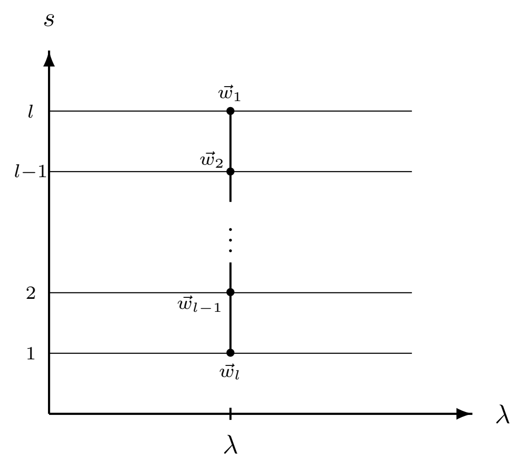
gelten.

Wegen der Bedingung am Ende der Kette für $k = l$, d.h. wegen

$$(A - \lambda E) \cdot \vec{w}_l = \vec{0}$$

ist der Vektor $\vec{w}_l$ am Ende der Kette ein Eigenvektor von A zum Eigenwert λ. Eine solche Folge von Vektoren stellen wir schematisch durch ein Hassediagramm mit l übereinander total angeordneten Knoten dar, die durch die

Vektoren der Kette markiert sind.



In der Jordan-Form entsprechen solche Ketten den Blöcken vom Typ $l \times l$. Weil die Jordan-Matrix einer beliebigen Matrix A mehrere Blöcke zum selben Eigenwert haben kann, werden wir auch mehrere Vektorketten benötigen, die wir dann halt im Hasse-Diagramm nebeneinander zeichnen.

Die Vektoren einer solchen Kette erfüllen also definitionsgemäss folgendes System linearer Gleichungen:

$$\begin{array}{llll}
(A-\lambda E)\cdot\vec{w}_1 &= \vec{w}_2 & \text{bzw.} & A\cdot\vec{w}_1 = \lambda\vec{w}_1+\vec{w}_2 \\
(A-\lambda E)\cdot\vec{w}_2 &= \vec{w}_3 & \text{bzw.} & A\cdot\vec{w}_2 = \lambda\vec{w}_2+\vec{w}_3 \\
\dots\dots\dots & & & \\
(A-\lambda E)\cdot\vec{w}_{l-2} &= \vec{w}_{l-1} & \text{bzw.} & A\cdot\vec{w}_{l-2} = \lambda\vec{w}_{l-2}+\vec{w}_{l-1} \\
(A-\lambda E)\cdot\vec{w}_{l-1} &= \vec{w}_l & \text{bzw.} & A\cdot\vec{w}_{l-1} = \lambda\vec{w}_{l-1}+\vec{w}_l \\
(A-\lambda E)\cdot\vec{w}_l &= \vec{0} & \text{bzw.} & A\cdot\vec{w}_l = \lambda\vec{w}_l
\end{array}$$

Man beachte, dass wir die markierenden Vektoren der Kette *von oben nach unten* und nicht mit ihrer Stufe, von unten nach oben, nummeriert haben, weil wir sie in dieser Reihenfolge rekursiv konstruieren werden. Die Spalten der gesuchten Transformationsmatrix X bestehen aus solchen Vektoren verallgemeinerter Eigenvektoren. In den Spalten von X werden sie dann allerdings in umgekehrter Reihenfolge, d.h. von unten nach oben d.h. wie üblich von links nach rechts angeordnet in der dualen Form

$$\vec{w}_l, \vec{w}_{l-1}, \cdots, \vec{w}_2, \vec{w}_1$$

auftauchen. Der Vektor $\vec{w}_1$ am Anfang der Kette ist im Hasse-Diagramm der oberste und der Eigenvektor $\vec{w}_l$ ist der unterste Vektor der betreffenden Kette. Der Vektor $\vec{w}_k$ einer solchen Kette kann als "tempierter Eigenvektor" aufgefasst werden, dessen Zeitzünder auf $l-k$ Zeiteinheiten eingestellt ist. Weil der Vektor $\vec{w}_1$ am Anfang dieser Kette die ganze Kette eindeutig festlegt, geht es also darum, zu einer gegebenen Matrix A und zu jedem ihrer Eigenwerte $\lambda \in \sigma_A$ solche maximalen Elemente von möglichst langen Ketten verallgemeinerter Eigenvektoren zu konstruieren, mit denen man dann die gesuchte Transformationsmatrix X bilden kann.

Damit die so gebildete Matrix X am Schluss auch wirklich invertierbar ist, müssen wir zusätzlich dafür sorgen, dass die Elemente der Ketten linear unabhängig sind. Um das garantieren zu können, müssen die verallgemeinerten Eigenvektoren der Ketten noch eine zusätzliche Bedingung erfüllen. Um sie zu formulieren, beachten wir, dass der Vektor $\vec{w}_k$ die homogene, lineare Gleichung

$$(A - \lambda E)^{l-k+1} \cdot \vec{w}_k = \vec{0}, \qquad 1 \leq k \leq l$$

erfüllt, die als "tempierte Eigenwertgleichung" betrachtet werden kann. Um die gewünschte lineare Unabhängigkeit der verallgemeinerten Eigenvektoren einer Kette zu gewährleisten, müssen wir dafür sorgen, dass der Vektor $\vec{w}_k$ genau auf der Stufe k und nicht etwa schon auf einer niedrigeren Stufe liegt. Deshalb darf er das lineare, homogene Gleichungssystem

$$(A - \lambda E)^{l-k} \cdot \vec{w}_k = \vec{0}$$

nicht erfüllen. Das führt uns zu folgendem fundamentalen Konzept, in dessen Definition wir die Abkürzung $s = l - k + 1$ benutzen.

Definition. Für eine Matrix A vom Typ $n \times n$ versteht man unter einem *Hauptvektor der Stufe* s zum Eigenwert λ einen Vektor $\vec{x} \neq \vec{0}$ mit der Eigenschaft, dass

$$(A - \lambda E)^s \cdot \vec{x} = \vec{0}, \qquad (A - \lambda E)^{s-1} \cdot \vec{x} \neq \vec{0}$$

gilt. Die Lösungsmenge des auftretenden linearen homogenen Gleichungssystems

$$(A - \lambda E)^s \cdot \vec{x} = \vec{0}, \quad s \geq 0$$

d.h. der Kern

$$\mathrm{H}_s(\lambda) := \mathrm{Ker}(A - \lambda E)^s, \quad s \geq 0$$

wird entsprechend als *Hauptraum der Stufe* s *des Eigenwertes* λ bezeichnet.

Man beachte, dass für einen Hauptvektor $\vec{x}$ der Stufe s zum Eigenwert λ allgemeiner die Gleichung

$$(A - \lambda E)^t \cdot \vec{x} = \vec{0}, \quad t \geq s$$

gilt. Einen Hauptvektor könnte man also auch als Vektor $\vec{x} \neq \vec{0}$ definierien, der die homogene, lineare Gleichung

$$(A - \lambda E)^t \cdot \vec{x} = \vec{0}$$

für irgend eine natürlich Zahl $t \in \mathbb{N}$ erfüllt und seine Stufe s ist dann die kleinste natürliche Zahl, für die dies der Fall ist. Man beachte, dass diese Gleichung nur für einen Eigenwert λ erfüllt sein kann. Falls sie nämlich für einen Vektor $\vec{x} \neq 0$ erfüllt ist, kann die Koeffizientenmatrix $(A - \lambda E)^t$ nicht invertierbar sein. Daher ist dann auch die Matrix $A - \lambda E$ nicht invertierbar, was den Skalar λ zwangsläufig zu einem Eigenvektor von A macht. Eine Matrix A kann also verallgemeinerte Eigenvektoren aber keine "verallgemeinerten Eigenwerte" haben!

Die Haupträume $H_s(\lambda)$ der Dimension $h_s(\lambda)$ spielen für die Jordan-Zerlegung und für viele Anwendungen im Zusammenhang mit linearer Darstellungstheorie eine fundamentale Rolle. Die Dimension des Hauptraumes $H_s(\lambda)$ bezeichnen wir mit $h_s(\lambda)$. Es ist also

$$h_s(\lambda) := \dim(\mathrm{Ker}(A - \lambda E)^s) = n - \mathrm{Rang}(\mathrm{Ker}(A - \lambda E)^s), \quad s \geq 0$$

Zum Raum $H_s(\lambda)$ gehört also ein System von h_s linear unabhängigen Hauptvektoren der Stufe s,

$$\vec{v}_{s,1}(\lambda), \vec{v}_{s,2}(\lambda), \vec{v}_{s,3}(\lambda), \ldots, \vec{v}_{s,h_s(\lambda)}(\lambda)$$

die den Raum $H_s(\lambda)$ aufspannen, d.h. die eine Basis $\mathcal{B}_s(\lambda)$ von $H_s(\lambda)$ bilden. Selbstverständlich liefert der zur Lösung des homogenen, linearen Gleichungssystems

$$(A - \lambda E)^s \cdot \vec{x} = \vec{0}$$

geeignete Eliminationsalgorithmus effektiv eine solche Basis. Wie sich aus diesen Basen dann die für die Jordanzerlegung (J, X) benötigten Hauptvektorketten konstruieren lassen, werden bald sehen.

Vorher halten wir fest, dass wir mit diesem Begriff das Gewünschte erreicht haben. Das besagen folgende Resultate, die mit Hilfe der Definition einfach zu beweisen sind.

Satz. Es sei

$$S : \vec{w}_1, \vec{w}_2, \ldots, \vec{w}_{l-1}, \vec{w}_l$$

eine Folge von Hauptvektoren zum Eigenwert λ und der Vektor $\vec{w}_s$ habe die Stufe s. Dann sind diese Vektoren linear unabhängig und bilden eine Folge verallgemeinerter Eigenvektoren.

Solche Folgen von Hauptvektoren werden wir in Zukunft kurz als *Hauptvektorketten* bezeichnen.

Satz. Die Vektoren von zwei Hauptvektorketten

$$S_1 : \vec{w}_{1,1}\vec{w}_{1,2}, \ldots, \vec{w}_{1,l-1}, \vec{w}_{1,l}$$

und

$$S_2 : \vec{w}_{2,1}\vec{w}_{2,2}, \ldots, \vec{w}_{2,m-1}, \vec{w}_{2,m}$$

zum selben Eigenwert λ sind linear unabhängig, falls die beiden Eigenvektoren $\vec{w}_{1,l}$ und $\vec{w}_{2,m}$ linear unabhängig sind.

Es ist leicht einzusehen, dass Eigenvektoren, die zu verschiedenen Eigenwerten gehören linear unabhängig sind. Dieser Sachverhalt verallgemeinert sich sofort auf Hauptvektorketten.

Satz. Die Vektoren von zwei Hauptvektorketten

$$S_1 : \vec{w}_{1,1}\vec{w}_{1,2}, \ldots, \vec{w}_{1,l-1}, \vec{w}_{1,l}$$

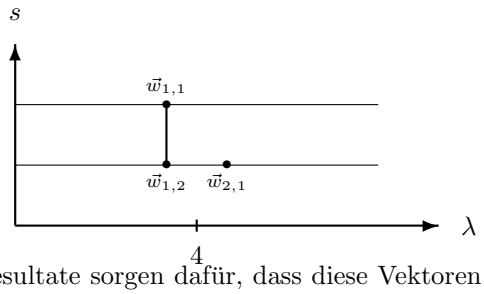
und

$$S_2 : \vec{w}_{2,1}\vec{w}_{2,2}, \ldots, \vec{w}_{2,m-1}, \vec{w}_{2,m}$$

zu verschiedenen Eigenwerten λ_1 und λ_2 sind linear unabhängig.

In unserem laufenden Beispiel liefert der Vektor $\vec{w}_{1,1}$ einen Hauptvektor der Stufe 2 und $\vec{w}_{1,2}$ liefert einen Hauptvektor der Stufe 1, der zur ersten Kette gehört. Der dritte Spaltenvektor $\vec{w}_{2,1}$ gehört nicht zu dieser Kette und bildet eine zweite Kette. Insgesamt bilden also die drei Spaltenvektoren der angegebenen Transformationsmatrix X eine Hauptvektorkette $\vec{w}_{1,1}, \vec{w}_{1,2}$ der Länge 2 und eine

Hauptvektorkette $\vec{w}_{2,1}$ der Länge 1, die sich durch folgendes Hasse-Diagramm veranschaulichen lassen.



Die behaupteten Resultate sorgen dafür, dass diese Vektoren in der Tat linear unabhängig und damit die mit ihnen gebildete Matrix X invertierbar ist.

Die Haupträume $H_s(\lambda)$ sind nicht unabhängig voneinander, sondern bilden ineinander geschachtelte Ketten. Zwischen ihnen gilt auf Grund ihrer Definition nämlich die fundmentale Ordnungs-Beziehung

$$H_s(\lambda) \subseteq H_{s+1}(\lambda),$$

denn für einen Vektor $\vec{x}$ aus $H_s(\lambda)$ gilt ja definitionsgemäss

$$(A - \lambda E)^s \cdot \vec{x} = \vec{0}$$

und daher erst recht

$$(A - \lambda E)^{s+1} \cdot \vec{x} = \vec{0}.$$

Daher bilden die Haupträume von λ eine aufsteigende Kette ineinandergeschachtelter Vektorräumen (Flagge). Eine solche Kette kann aber nicht beliebig lang werden. Sie muss nämlich aus Dimensionsgründen stationär werden und damit haben wir mit ihr keine Grenzwertprobleme! Es ist also:

$$\{\vec{0}\} = \mathrm{H}_0(\lambda) \subset \mathrm{H}_1(\lambda) \subset \mathrm{H}_2(\lambda) \subset \cdots \subset \mathrm{H}_s(\lambda) \subset \cdots \subset \mathrm{H}_{\tilde{s}}(\lambda) \cong \mathrm{H}_{\tilde{s}+1}(\lambda) \cong \cdots$$

Mit $\tilde{s}(\lambda)$ bezeichnen wir also die kleinste natürliche Zahl, für die diese Kette das erste Mal stationär bleibt. Diese Zahl $\tilde{s}(\lambda) \geq 1$ stimmt mit der Vielfachheit von λ im Minimalpolynom von A überein und wird die Grösse des grössten Blocks in der Jordanmatrix J zum Eigenwert λ angeben.
Weil $H_0(\lambda) = \{\vec{0}\}$ gilt, ist insbesondere $h_0(\lambda) = 0$. Der Raum $\mathrm{H}_1(\lambda)$ wird als *Eigenraum von* λ und seine Dimension $h_1(\lambda)$ wird als *geometrische Vielfachheit* von λ bezeichnet. Definitionsgemäss ist genau dann $h_1(\lambda) \geq 1$, wenn das lineare, homogene Gleichungssystem

$$(A - \lambda E) \cdot \vec{x} = \vec{0}$$

eine nichttriviale Lösung hat, d.h. wenn λ ein Eigenwert von A ist.
Die erwähnte Kettenbedingung der Haupträume hat als Konsequenz, dass die Folge $h_s(\lambda)$ ihrer Dimensionen strikt monoton wächst und ab einem gewissen Wert $\tilde{s}(\lambda)$ stationär wird. Wir erhalten also eine strikt monoton wachsend Folge

$$0 = h_0(\lambda) < h_1(\lambda) < h_2(\lambda) < \cdots < h_s(\lambda) < \cdots < h_{\tilde{s}(\lambda)} = h_{\tilde{s}+1(\lambda)} = \cdots$$

d.h. $\tilde{s}(\lambda)$ ist auch die kleinste natürliche Zahl, ab der sich die Folge $h_s(\lambda)$ nicht mehr ändert.

Rekursiv lässt sich nun folgendes Resultat von Jordan[15] zeigen.

Satz. Zu jeder (komplexen) quadratischen Matrix A vom Typ $n \times n$ gibt es eine Jordanmatrix J und eine invertierbare Matrix X mit folgenden Eigenschaften:

$$A \cdot X = X \cdot J, \qquad J = \Lambda + N$$

Die Matrix Λ ist diagonal und die Matrix N ist nilpotent. Ferner gilt die Vertauschungsrelation $[\Lambda, N] = 0$.

Die Jordanmatrix J hat Blockdiagonalform

$$J = \begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_m) \end{pmatrix} = \mathrm{diag}(J(\lambda_1), \ldots, J(\lambda_m)) = J(\lambda_1) \oplus \cdots \oplus J(\lambda_m)$$

und die Matrizen $J(\lambda_j)$ (sogn. Jordanblock zum Eigenwert λ_j) haben auf der Diagonalen den selben Eigenwert und auf der oberen Nebendiagonalen lauter 1.

Die Jordanblöcke bestehen also aus sogn. Kästchen der Form

$$J(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & 0 & \cdots & 0 \\ 0 & \lambda_j & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & \lambda_j & 1 \\ 0 & & & & \lambda_j \end{pmatrix} = \mathrm{diag}(\lambda_j) + J(0) \in \mathbb{R}^{n,n}$$

in deren oberen Nebendiagonalen lauter 1 sind. Für $n = 1$ erhalten wir als Spezialfall die bisher benutzten Diagonalmatrizen. Als weiteren wichtigen Spezialfall erwähnen wir den nilpotenten Jordan-Block

$$J(0) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & 0 & 1 \\ 0 & & & & 0 \end{pmatrix} = (\vec{0}, \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_{n-1}) \in \mathbb{R}^{n,n}$$

mit der Eigenschaft, dass die Standardbasisvektoren

$$\vec{e}_n, \vec{e}_{n-1}, \ldots, \vec{e}_2, \vec{e}_1$$

eine Hauptvektorkette zum Eigenwert $\lambda = 0$ bilden, d.h. dass

$$J(0) \cdot \vec{e}_1 = \vec{0}, \qquad J(0) \cdot \vec{e_k} = \vec{e}_{k-1}, \qquad k = 2, 3, \ldots, n$$

gilt. Damit lassen sich ihre Potenzen einfach berechnen. Es ist nämlich

$$J^m(0) = (\underbrace{\vec{0}, \ldots, \vec{0}}_{m}, \vec{e}_1, \ldots, \vec{e}_{n-m}), \qquad m = 1, \ldots, n-1$$

[15] Marie Ennemond Camille Jordan 1838 - 1922.

Insbesondere ist die Matrix $J(0) \in \mathbb{R}^{n,n}$ nilpotent vom Nilpotenzgrand n.

Die Spalten der invertierbaren Transformationsmatrix X bestehen aus Hauptvektorketten von A.

Die Jordanmatrix J ist im wesentlichen, d.h. bis auf Vertauschung der Blöcke, eindeutig bestimmt. Die Anzahl Blöcke zum Eigenwert λ beträgt $h_1(\lambda)$ und die Anzahl Blöcke der Grösse s zum Eigenwert λ beträgt[16]

$$a_s(\lambda) := 2h_s(\lambda) - h_{s-1}(\lambda) - h_{s+1}(\lambda), \quad s \geq 1, \quad a_0(\lambda) = 0$$

Daher ist

$$\sum_{s=0}^{\tilde{s}(\lambda)} a_s(\lambda) = h_1(\lambda), \qquad \sum_{s=0}^{\tilde{s}(\lambda)} s \cdot a_s(\lambda) = \text{alg-mult}(\lambda).$$

Die Anzahl Blöcke mit dem Diagonalelemt λ ist die geometrische Vielfachheit $h_1(\lambda)$. Der grösste Block mit dem Diagonalelement λ ist vom Typ $\tilde{s}(\lambda) \times \tilde{s}(\lambda)$ und die Summe der Grössen der Blöcke mit dem Diagonalelement λ ist die algebraische Vielfachheit alg-mult(λ) von λ im charakteristischen Polynom $\chi(\lambda)$.

Die Jordansche Normalform einer Matrix A lässt sich wie folgt berechnen:

1. Berechne $h_s(\lambda) := \dim(\text{Ker}(A - \lambda E)^s) = n - \text{Rang}(\text{Ker}(A - \lambda E)^s)$ für $s = 0, 1, \ldots$ und höre auf, wenn sich $h_s(\lambda)$ stabilsiert. Der Index $\tilde{s}(\lambda)$ ist der kleinste Wert, für den $h_s(\lambda) = h_{s+1}(\lambda)$ gilt.

2. Berechne die Folge $a_s(\lambda)$ mit Hilfe der Rekursion

   $$a_s(\lambda) = 2h_s(\lambda) - h_{s-1}(\lambda) - h_{s+1}(\lambda), \quad a_0(\lambda) = 0$$

   In der Jordanschen Normalform von A hat es $a_s(\lambda)$ Blöcke der Form $J(\lambda) \in \mathbb{R}^{s,s}$ für $s = 1, \ldots, \tilde{s}(\lambda)$.

In unserem laufenden Beispiel hat die Matrix A das charakteristische Polynom

$$\chi_A(\lambda) = -(\lambda - 4)^3$$

und daher den einzigen[17] Eigenwert $\lambda = 4$. Seine algebraische Vielfachheit alg-mult(λ) = 3 liefert a priori eine obere Schranke für die Länge der längsten Kette dieses Eigenwerts. Weil das Minimalpolynom von A die Faktorisierung

$$\mu_A(\lambda) = (\lambda - 4)^2$$

hat, wird die Kette sogar ab $\tilde{s}(4) = 2$ stationär. Die zugehörige längste Kette hat also einen grössten Block vom Typ 2×2 zur Folge. Ein Rückblick auf das zugehörige Hasse-Diagramm zeigt, dass in diesem Fall tatsächlich je ein 2-er und

[16] Die Anzahl Blöcke der Grösse mindestens s zum Eigenwert λ beträgt

$$b_s(\lambda) = h_s(\lambda) - h_{s-1}(\lambda).$$

Daher beträgt die Anzahl Blöcke der genauen Grösse s tatsächlich

$$a_s(\lambda) = b_s(\lambda) - b_{s+1}(\lambda) = \Big(h_s(\lambda) - h_{s-1}(\lambda)\Big) - \Big(h_{s+1}(\lambda) - h_s(\lambda)\Big) = 2h_s(\lambda) - h_{s-1}(\lambda) - h_{s+1}(\lambda).$$

[17] Für weitere Eigenwerte $\lambda \in \sigma_A$ führt man das beschriebenen Prozedere sinngemäss durch.

ein 1-er Block vorhanden sein müssen und daher die Jordansche Normform die angegebene Gestalt

$$J(A) = \left(\begin{array}{cc|c} 4 & 1 & 0 \\ 0 & 4 & 0 \\ \hline 0 & 0 & 4 \end{array}\right)$$

haben muss.

CAS. Jordan-Blöcke wie in diesem Beispiel können in Sage mit Hilfe des Befehls

```
JB=jordan_block(4,2); show(JB)
```

leicht erzeugt werden. Sein erstes Argument erfordert die Angabe des Diagonalelementes und das zweite seinen Typ.

Damit lassen sich dann Jordanmatrizen leicht zusammenbauen. In unserem Beispiel leistet das der Kode

```
JB1=jordan_block(4,2)
JB2=jordan_block(4,1))
J=block_diagonal_matrix(JB1,JB2); show(J)
```

in dem zuerst die beiden Jordanblöcke erzeugt und dann damit die erforderliche Blockdiagonalmatrix zusammengebaut wird. ◇

Um diese Normalform J, zugehörende Hauptvektorketten und damit eine Transformationsmatrix X aus der Matrix A zu berechnen, bestimmen wir in unserem laufenden Beispiel zunächst den Eigenraum $H_1(4)$ und lösen dazu das homogene lineare Gleichungssystem

$$(A - 4E) \cdot \vec{x} = \vec{0}, \qquad (A - 4E) = \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix}$$

Als Basis $\mathcal{B}_1(4)$ von $H_1(4)$ können wir die beiden oben bereits angegebenen Vektoren

$$\vec{v}_{1,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{1,2} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$$

benutzen. Wir stellten fest, dass der Eigenraum $H_1(4)$ durch diese beiden linear unabhängigen Vektoren aufgespannt werden kann und daher $h_1(4) = 2$ gilt. Insbesondere muss die Anzahl Blöcke zum Eigenwert $\lambda = 4$ auch 2 betragen.

Als nächstes bestimmen wir eine Basis $\mathcal{B}_2(4)$ des Hauptraumes $H_2(4)$ der nächsten Stufe und und lösen dazu das homogene lineare Gleichungssystem

$$(A - 4E)^2 \cdot \vec{x} = \vec{0}, \qquad (A - 4E)^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Wir stellen fest, dass dieses homogene Gleichungssystem als Lösungsraum $H_2(4)$ den ganzen Raum $\mathbb{R}^3$ haben muss und dieser beispielsweise durch die drei linear unabhängigen Standardbasisvektoren

$$\vec{v}_{2,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{2,2} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_{2,3} := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

der Basis $\mathcal{B}_2(4)$ aufgespannt werden kann. Insbesondere ist also $h_2(4) = 3$. Ab $\tilde{s}(4) = 3$ muss die Folge der Haupträume aus Dimensionsgründen konstant werden. Insgesamt erhalten wir also die Werte folgender Tabelle:

s	0	1	2	3	$\cdots$
$h_s(\lambda)$	0	2	3	3	$\ldots$
$a_s(\lambda)$	0	1	1	0	$\cdots$

Aus dieser Tabelle entnehmen wir sofort, dass die Jordansche Normalform $J(A)$ von A also $a_1 = 1$ Block der Grösse 1 und $a_2 = 1$ Block der Grösse 2 haben und daher tatsächlich die behauptete Form

$$J(A) = \left(\begin{array}{cc|c} 4 & 1 & 0 \\ 0 & 4 & 0 \\ \hline 0 & 0 & 4 \end{array}\right).$$

haben muss.

Es bleibt also mit Hilfe der gefundenen Basen der Haupträume noch eine zugehörige Transformationsmatrix X zu bestimmen. Zunächst wählen wir einen Hauptvektor aus dem grössten Hauptraum $H_2(4)$, der zu allen Vektoren des Unterraums $H_1(4)$ linear unabhängig ist, d.h. der nicht bereits im Unterraum $H_1(4)$ liegt. Dazu kommt der ersten Basisvektor $\vec{v}_{2,1}$ von $H_2(4)$ nicht in Frage. Erst der zweite erfüllt diese Bedingung und wir wählen daher

$$\vec{w}_{1,1} := \vec{v}_{2,2} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Selbstverständlich hätten wir diese Wahl auf viele andere Arten treffen können. Beispielsweise erfüllt auch $\vec{v}_{2,3}$ diese Bedingung. Dank dieser Wahlmöglichkeiten wird die Transformationsmatrix X sicher nicht eindeutig bestimmt sein.

Nun setzen wir die angefangene Hauptvektorkette nach unten fort und berechnen mit Hilfe des soeben gewählten Vektors $\vec{w}_{1,1}$ aus $H_2(4)$ den Vektor

$$\vec{w}_{1,2} = (A - 4E) \cdot \vec{w}_{1,1} = \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \in H_1(4)$$

Dabei handelt es sich um einen Eigenvektor aus dem Eigenraum $H_1(4)$. Dieser Vektor muss also eine eindeutige Linearkombination der beiden Basisvektoren von $H_1(4)$ sein und es gilt tatsächlich

$$\vec{w}_{1,2} = \vec{v}_{1,1} + \vec{v}_{1,2}.$$

Weil wir mit unserer Wahl dafür gesorgt haben, dass $\vec{w}_{1,1} \in H_2(4)$ *nicht* im Unterraum $H_1(4)$ liegt, sind die beiden Vektoren $\vec{w}_{1,1}$ und $\vec{w}_{1,2}$ linear unabhängig und bilden eine Hauptvektorkette der Länge 2, die in der Matrix J zum Block der Grösse 2 Anlass gibt.

Weil wir diese Kette nicht mehr weiter nicht-trivial nach unten verlängern können, fahren wir mit der nächstkürzeren Kette fort. Dazu gehen wir zum Hauptraum $H_1(4)$ und suchen nun eine angepasste Basis für $H_1(4)$, die also den

von oben erzwungenen Vektor $\vec{w}_{1,2}$ enthält. Wir wählen also einen Vektor $\vec{w}_{2,1}$ in $H_1(4)$, der mit dem bereits errechneten Vektor $\vec{w}_{1,2} \in H_1(4)$ der vorherigen Kette ein linear unabhängiges System bildet. Ein Blick auf die Basis von $H_1(4)$ zeigt, dass wir dazu den Eigenvektor $\vec{v}_{1,1}$ wählen können. Wir wählen also

$$\vec{w}_{2,1} = \vec{v}_{1,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Weil es sich dabei bereits um einen Eigenvektor von A handelt, ist auch die Konstruktion dieser Hauptvektorkette zu Ende. Weil wir nun aber insgesamt drei linear unabhängige Vektoren

$$\vec{w}_{1,1} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{w}_{1,2} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \qquad \vec{w}_{2,1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

gefunden haben, können wir die ganze Konstruktion beenden und die gefundenen Vektorketten zur gesuchten Transformationsmatrix X zusammentragen. Wir bilden sie, indem wir die bestimmten Vektoren in umgekehrter Reihenfolge, d.h. im Hasse-Digramm *von unten nach oben* als Spalten in die Matrix X einfüllen und erhalten schliesslich für die gesuchte Transformationsmatrix

$$X = X(\vec{w}_{1,2}, \vec{w}_{1,1}, \vec{w}_{2,1}) = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

von der wir oben bereits gesehen haben dass sie das Verlangte leistet.

Nachdem wir eine Jordanzerlegung (J, X) der Matrix A bestimmt haben, überzeugen wir uns nun als nächstes, dass die angegebenen Eigenschaften dieser Zerlegung genügen, um die Potenzen A^k und das Matrizenexponential e^{At} von A explizit berechnen zu können. Das folgt wieder daraus, dass die Potenzen von Jordanmatrizen J einfach explizit bestimmt werden können, da einerseits

$$J^k = (\mathrm{diag}(J_1, \ldots, J_m))^k = \mathrm{diag}(J_1^k, \ldots, J_m^k)$$

gilt und Potenzen von Jordanblöcken einfach bestimmt werden können.

Die dank $[\Lambda, N] = 0$ gültige binomische Fomel

$$(\Lambda + N)^k = \sum_{i=0}^{k} \binom{k}{i} \Lambda^{k-i} \cdot N^i = \Lambda^k + \binom{k}{1} \Lambda^{k-1} \cdot N + \binom{k}{2} \Lambda^{k-2} \cdot N^2 + \cdots + N^k$$

besteht nämlich in unserem Fall dank der Nilpotenz von N vom Grad 2 bloss aus den ersten beiden Summanden d.h. aus der Summe

$$J^k = (\Lambda + N)^k = \Lambda^k + k\Lambda^{k-1} \cdot N = \begin{pmatrix} 4^k & k4^{k-1} & 0 \\ 0 & 4^k & 0 \\ 0 & 0 & 4^k \end{pmatrix}$$

Damit und dank der gefundenen Faktorisierung der Jordanzerlegung

$$A \cdot X = X \cdot J, \qquad \text{bzw.} \quad A = X \cdot J \cdot X^{-1}$$

kann die gesuchte Potenz von A in der üblichen Form als Teleskopprodukt

$$A^k = X \cdot J^k \cdot X^{-1} = \begin{pmatrix} 4^k & k4^{k-1} & 2k4^{k-1} \\ 0 & 4^k - 2k4^{k-1} & -4k4^{k-1} \\ 0 & k4^{k-1} & 4^k + 2k4^{k-1} \end{pmatrix}$$

explizit beschrieben werden, wie der Leser durch Einsetzen in die Rekursionsgleichung der Matrizenpotenz überprüfen möge. Damit lässt sich dann der Zustand des Systems nach k Zeitschritten formelmässig durch

$$\begin{aligned} \vec{y}(k) = A^k \cdot \vec{a} &= \begin{pmatrix} 4^k & k4^{k-1} & 2k4^{k-1} \\ 0 & 4^k - 2k4^{k-1} & -4k4^{k-1} \\ 0 & k4^{k-1} & 4^k + 2k4^{k-1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} 4^k - 2k4^{k-1} \\ 4k4^{k-1} \\ -(4^k + 2k4^{k-1}) \end{pmatrix} \end{aligned}$$

explizit beschreiben.

Aus dieser expliziten Lösung des Anfangswertproblems entnehmen wir eine explizite Beschreibung der ursprünglichen drei Folgen. Es ist nämlich

$$\begin{cases} x(k) &= 4^k - 2k4^{k-1} \\ y(k) &= 4k4^{k-1} \\ z(k) &= -(4^k + 2k4^{k-1}) \end{cases}$$

wie wir soeben kontrolliert haben, aber wohl nur schwerlich erraten hätten.

Aus der charakteristischen Matrixgleichung $A^3 = 64E - 48A + 12A^2$ gewinnt man sofort eine rekursive Beschreibung dieser drei Folgen in Form einer linearen Differenzengleichung dritter Ordnung. Multiplizieren wir nämlich die charakteristische Matrizengleichung von rechts mit dem Zustandvektor $\vec{y}(k)$, erhalten wir für die Vektorfolge mit Hilfe ihrer Rekursionsgleichung erster Ordnung die Rekursionsgleichungen dritter Ordnung.

$$\begin{aligned} \vec{y}(k+3) = A^3 \cdot \vec{y}(k) &= (64E - 48A + 12A^2) \cdot \vec{y}(k) \\ &= 64\vec{y}(k) - 48\vec{y}(k+1) + 12\vec{y}(k+2) \end{aligned}$$

In Komponenten ausgeschrieben hat sie die Form

$$\begin{cases} x(k+3) = 64x(k) - 48x(k+1) + 12x(k+2) \\ y(k+3) = 64y(k) - 48y(k+1) + 12y(k+2) \\ z(k+3) = 64z(k) - 48z(k+1) + 12z(k+2) \end{cases}$$

Sie werden durch die selben drei angegebenen expliziten Folgen erfüllt, wie man ebenfalls durch Einsetzen nachprüfen kann. Man beachte, dass die drei Gleichungen dieses Systems von Rekursionsgleichungen dritter Ordnung eine gemeinsame Struktur haben und direkt aus dem charakteristischen Polynom $\chi_A(\lambda)$ der Matrix A abgelesen werden können.

Die je 3 notwendigen Anfangsbedingungen lauten hier

$$\begin{cases} x(0) = 1, & x(1) = 2, & x(2) = 0 \\ y(0) = 0, & y(1) = 4, & y(2) = 32 \\ z(0) = -1, & z(1) = -6, & z(2) = -32 \end{cases}$$

Im vorliegenden Beispiel der Matrix

$$A = \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix}$$

ist das charakteristische Polynom nicht minimal. Durch Einsetzen kann man leicht überprüfen, dass die Systemmatrix A sogar die polynomiale Gleichung

$$\mu_A(\lambda) = (\lambda - 4)^2 = 16 - 8\lambda + \lambda^2 = 0 \qquad \text{bzw.} \qquad A^2 = -16E + 8A$$

kleineren Grades erfüllt. Das Polynom μ_A teilt das charakteristische Polynom χ_A, da $\chi_A(\lambda) = \mu_A(\lambda) \cdot (\lambda - 4)$ gilt. Wir erwarten also, dass sich die drei Folgen sogar durch die zugehörige Rekursionsgleichung zweiter Ordnung

$$\vec{y}(k+2) = A^2 \cdot \vec{y}(k) = (-16E + 8A) \cdot \vec{y}(k) = -16\vec{y}(k) + 8\vec{y}(k+1)$$

beschreiben lassen, die in Komponenten ausgeschrieben die einheitliche Form

$$\left\{ \begin{array}{lcllll} x(k+2) & = & -16x(k) + 8x(k+1), & x(0) = 1 & x(1) = 2 \\ y(k+2) & = & -16y(k) + 8y(k+1), & y(0) = 0 & y(1) = 4 \\ z(k+2) & = & -16z(k) + 8z(k+1), & z(0) = -1 & z(1) = -6 \end{array} \right.$$

hat.

Bei kontinuierlicher statt diskreter Auffassung der Zeit würde man statt von einem Differenzengleichungssystem vom System

$$\left\{ \begin{array}{lcl} x'(t) & = & 4x(t) + y(t) + 2z(t) \\ y'(t) & = & 2y(t) - 4z(t) \\ z'(t) & = & y(t) + 6z(t) \end{array} \right.$$

homogener, linearer Differentialgleichungen mit der selben Systemmatrix A und den selben Anfangszustand $\vec{a}$ ausgehen und den Zustandsvektor $\vec{y}(t)$ mit Hilfe der drei gesuchten Funktionen $x(t)$, $y(t)$ und $z(t)$ definieren.

$$A = \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad \vec{a} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \in \mathbb{R}^3, \quad \vec{y}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} \in \mathbb{R}^3$$

Die vektorielle Differentialgleichung erster Ordnung

$$\vec{y}\,'(t) = A \cdot \vec{y}(t), \qquad \vec{y}(0) = \vec{a}, \qquad t \in \mathbb{R}$$

lässt sich mit Hilfe der oben angegebenen Faktorisierung der Jordanzerlegung ebenfalls leicht explizit lösen. Zunächst beachtet man, dass sich der Propagator e^{tJ} der Normalform J wegen der Summenzerlegung $J = \Lambda + N$ und dank des Kommutators $[\Lambda, N] = 0$ in zwei kommutierende[18] Summanden Λ und N und

[18] Genau so wenig, wie die binomische Formel für $(A+B)^k$ in der aus der Schule geläufigen Form gilt, wenn die beiden Matrizen A und B nicht kommutieren, gilt dann auch das aus der Schule bekannte Additionstheorem für e^{A+B} nicht! Das Matrizenprodukt $e^A \cdot e^B$ lässt sich im allgemeinen Fall in der Form $e^{Z(A,B)}$ ausrücken, wobei nach Baker-Campell-Hausdorff folgende Reihendarstellung gilt:

$$Z(A,B) = A + B + \frac{1}{2}[A,B] + \frac{1}{12}[A,[A,B]] - \frac{1}{12}[B,[A,B]] - \frac{1}{12}[B,[A,[A,B]]] + \cdots$$

Falls A und B kommutieren, verschwinden in dieser Reihe alle geschachtelten Kommutatoren.

wegen des damit gültigen Additionstheorems in der Form

$$e^{tJ} = e^{t(\Lambda+N)} = e^{t\Lambda} \cdot e^{tN}$$

faktorisieren lässt. Die beiden Faktoren lassen sich aber sofort bestimmen. Den Propagator einer Diagonalmatrix erhält man wegen der einfachen Arithmetik von Diagonalmatrizen in der Exponentialreihe

$$e^{t\Lambda} = E + t\Lambda + \frac{t^2}{2!}\Lambda^2 + \frac{t^3}{3!}\Lambda^3 + \cdots = \sum_{j=0}^{\infty} \frac{t^j}{j!}\Lambda^j$$

wie früher erwähnt, indem man die Exponentialfunktion auf die einzelnen Diagonalelemente anwendet und damit eine Digonalmatrix bildet. Es ist also

$$e^{t\Lambda} = \begin{pmatrix} e^{4t} & 0 & 0 \\ 0 & e^{4t} & 0 \\ 0 & 0 & e^{4t} \end{pmatrix}$$

Zur Berechnung des Propagators des nilpotenten Summanden N beachtet man, dass die Exponentialreihe wegen der Nilpotenz von N vom Grad 2 bloss aus den ersten beiden Summanden besteht und daher die einfache Gestalt

$$e^{tN} = E + tN = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + t \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

hat. Allgemein erhalten wir für das Exponential einer nilpotenten Matrix N des Nilpotenzgrades k die endliche Summendarstellung

$$e^{tN} = \sum_{j=0}^{k-1} \frac{1}{j!} t^j N^j$$

Setzen wir die Ergebnisse zusammen, erhalten wir für den Propagator der Normalform

$$e^{tJ} = e^{t\Lambda} \cdot e^{tN} = \begin{pmatrix} e^{4t} & te^{4t} & 0 \\ 0 & e^{4t} & 0 \\ 0 & 0 & e^{4t} \end{pmatrix}$$

Damit und mit der Faktorisierung

$$A = X \cdot J \cdot X^{-1}$$

kann der gesuchte Propagator von A schliesslich in der Form

$$e^{tA} = X \cdot e^{tJ} \cdot X^{-1} = \begin{pmatrix} e^{4t} & te^{4t} & 2te^{4t} \\ 0 & e^{4t} - 2te^{4t} & -4te^{4t} \\ 0 & te^{4t} & e^{4t} + 2te^{4t} \end{pmatrix}$$

explizit beschrieben werden, wie der Leser durch Einsetzen in die Differentialgleichung des Matrizenexponentials überprüfen möge. Damit lässt sich dann der

Zustand des Systems zur Zeit t formelmässig durch

$$\vec{y}(t) = e^{tA} \cdot \vec{a} = \begin{pmatrix} e^{4t} & te^{4t} & 2te^{4t} \\ 0 & e^{4t} - 2te^{4t} & -4te^{4t} \\ 0 & te^{4t} & e^{4t} + 2te^{4t} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

$$= \begin{pmatrix} e^{4t} - 2te^{4t} \\ 4te^{4t} \\ -(e^{4t} + 2te^{4t}) \end{pmatrix}$$

explizit beschreiben. Aus dieser Lösung des Anfangswertproblems entnehmen wir eine explizite Beschreibung der drei Komponentenfunktionen. Es ist nämlich

$$\begin{cases} x(t) = e^{4t} - 2te^{4t} \\ y(t) = 4te^{4t} \\ z(t) = -(e^{4t} + 2te^{4t}) \end{cases}$$

wie wir soeben kontrolliert haben. Man vergleiche diese Ergebnisse mit jenen des entsprechenden diskreten Problems bei der Berechnung der Matrizenpotenz.

Wie oben kann man das System von 3 Differentialgleichungen erster Ordnung in äquivalente Differentialgleichungen dritter Ordnung für die Koeffizientenfunktionen umformulieren. Leiten wir wir nämlich die Differentialgeichung erster Ordnung zweimal ab, geht sie durch Verwenden der charakteristischen Matrizengleichung in die Form

$$\begin{aligned} \vec{y}'''(t) = A^3 \cdot \vec{y}(t) &= (64E - 48A + 12A^2) \cdot \vec{y}(t) \\ &= 64\vec{y}(t) - 48\vec{y}'(t) + 12\vec{y}''(t) \end{aligned}$$

über. Sie hat in Komponenten ausgeschrieben die Form

$$\begin{cases} x'''(t) = 64x(t) - 48x'(t) + 12x''(t) \\ y'''(t) = 64y(t) - 48y'(t) + 12y''(t) \\ z'''(t) = 64z(t) - 48z'(t) + 12z''(t) \end{cases}$$

Sie werden durch die selben drei angegebenen expliziten Funktionen erfüllt, wie man ebenfalls durch Einsetzen nachprüfen kann. Man beachte, dass die drei Gleichungen dieses Systems von Differentialgleichungen dritter Ordnung eine gemeinsame Struktur haben und direkt aus dem charakteristischen Polynom χ_A der Matrix A abgelesen werden können. Mit Hilfe des Minimalpolynoms statt des charakteristischen Polynoms findet man für sie sogar ein System von Differentialgleichungen zweiter, statt dritter, Ordnung

$$\begin{cases} x''(t) = -16x(t) + 8x'(t), & x(0) = 1 \quad x'(1) = 2 \\ y''(t) = -16y(t) + 8y'(t), & y(0) = 0 \quad y'(1) = 4 \\ z''(t) = -16z(t) + 8z'(t), & z(0) = -1 \quad z'(1) = -6 \end{cases}$$

die man leicht durch Nachrechnen bestätigt.

Ein Vergleich der Rechnungen und Resultate im kontinuierlichen mit dem diskreten Fall zeigt, dass vollständig parallel vorgegangen wird. Um die lineare

Struktur des Problems nicht durch Analysis zu verschleiern, werden wir uns in Zukunft meistens auf den diskreten Fall beschränken. Dem eingeweihten Leser wird das Übertragen der Resultate kaum schwer fallen. ◯

Den an Hand des letzten Beispiels entwickelten Algorithmus zur Berechnung der Jordanschen Zerlegung (J, X) einer Matrix A studiert man ab besten an einigen typischen Beispielen, die allerdings unangenehm gross sein müssen, wenn man alle möglichen Phänomene erkennen möchte. Um also den Überblick nicht im Dschungel öder Rechnereien zu verlieren, überlassen wir einige von ihnen einer Maschine und geben nur die Resultate an, falls wir auf Grund der bisherigen Überlegungen davon ausgehen können, dass klar ist, was genau zu tun ist und die behaupteten Ergebnisse leicht kontrolliert werden können.

Beispiel. Die 8×8 Matrix

$$A = \begin{pmatrix} 3 & 3 & 0 & 0 & 0 & -1 & 0 & -2 \\ -3 & 4 & 1 & -1 & -1 & 0 & 1 & -1 \\ 0 & 6 & 3 & 0 & 0 & -2 & 0 & -4 \\ -2 & 4 & 0 & 1 & -1 & 0 & 2 & -5 \\ -3 & 2 & 1 & -1 & 2 & 0 & 1 & -2 \\ -1 & 1 & 0 & -1 & -1 & 3 & 1 & -1 \\ -5 & 10 & 1 & -3 & -2 & -1 & 6 & -10 \\ -3 & 2 & 1 & -1 & -1 & 0 & 1 & 1 \end{pmatrix}$$

hat das charakteristische Polynom

$$\chi(\lambda) = (\lambda - 3)^7 \cdot (\lambda - 2).$$

Offenbar hat also die Matrix A die beiden Eigenwerte des Spektrums

$$\sigma_A = \{2, 3\}.$$

Der eine Eigenwert $\lambda = 2$ ist leicht zu behandeln. Seine algebraische Vielfachheit ist 1 und daher muss auch seine geometrische Vielfachheit $h_1(2) = 1$ sein. Um eine Basis $\mathcal{B}_1(2)$ dieses Eigenraumes $H_1(2)$ zu finden, müssen wir zugehörige linear unabhängige Eigenvektoren bestimmen und dazu den Kern

$$\mathrm{Ker}(A - 2E)$$

berechnen. d.h. das homogene lineare Gleichungssystem

$$(A - 2E) \cdot \vec{x} = \vec{0}$$

lösen. Sein Lösungsraum wird etwa vom Eigenvektor

$$\vec{v}_{1,1}(2) = \begin{pmatrix} 2 \\ 0 \\ 4 \\ 2 \\ 1 \\ 0 \\ 6 \\ 2 \end{pmatrix}$$

aufgespannt. Weil die algebraische Vielfachheit dieses Eigenwertes 1 ist, muss $\tilde{s}(2) = 1$ sein. Weil wir in dieser Kette bereits Stabilität erreicht haben, ist es nicht notwendig, höhere Stufen dieses Eigenwertes zu untersuchen. d.h. wir können unsere Suche abbrechen.

Zum anderen Eigenwert $\lambda = 3$ gehört der Eigenraumes $H_1(3)$, der als Kern

$$\mathrm{Ker}(A - 3E)$$

erklärt ist. Um eine Basis $\mathcal{B}_1(3)$ dieses Raumes zu erhalten, müssen wir das homogene lineare Gleichungssystem

$$(A - 3E) \cdot \vec{x} = \vec{0}$$

lösen. Weil der Rang seiner Koeffizientenmatrix $\mathrm{Rang}(A - 3E) = 5$ ist, wird sein Lösungsraum von $h_1(3) = 8 - 5 = 3$ linear unabhängigen Eigenvektoren aufgespannt, die wir etwa in der Form

$$\vec{v}_{1,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_{1,2}(3) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_{1,3}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

wählen können.

Offenbar haben wir noch zuwenig linear unabhängige Vektoren, um damit eine invertierbare Matrix bilden zu können und müssen deshalb nun die nächste Stufe $s = 2$ untersuchen. Der zugehörige Hauptraum $H_2(3)$ ist als Kern

$$\mathrm{Ker}(A - 3E)^2,$$

d.h. als Lösungsmenge des homogenen linearen Gleichungssystems

$$(A - 3E)^2 \cdot \vec{x} = \vec{0}$$

erklärt. Weil seine Koeffizientenmatrix den Rang $\mathrm{Rang}(A - 3E)^2 = 2$ hat, ist sein Lösungsraum $h_2(3) = 8 - 2 = 6$-dimensional. Als Basis $\mathcal{B}_2(3)$ von $H_2(3)$ können wir etwa folgende 6 Vektoren

$$\vec{v}_{2,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{2,2}(3) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \vec{v}_{2,3}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\vec{v}_{2,4}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{2,5}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{2,6}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

benutzen.

Weil wir immer noch zu wenig linear unabhängie Vektoren haben, um damit eine invertierbare Matrix zu bilden, müssen wir als nächstes noch die nächste Stufe $s = 3$ untersuchen. Der zugehörige Hauptraum $H_3(3)$ ist als Kern

$$\mathrm{Ker}(A - 3E)^3,$$

d.h. als Lösungsmenge des homogenen linearen Gleichungssystems

$$(A - 3E)^3 \cdot \vec{x} = \vec{0}$$

erklärt. Weil seine Koeffizientenmatrix den Rang $\mathrm{Rang}(A - 3E)^3 = 1$ hat, ist sein Lösungsraum $h_3(3) = 8 - 1 = 7$-dimensional. Wegen $\tilde{s}(3) = 3$ können wir auf dieser Stufe die Suche abbrechen. Als Basis $\mathcal{B}_3(3)$ von $H_3(3)$ können wir etwa folgende 7 Vektoren

$$\vec{v}_{3,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{v}_{3,2}(3) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{v}_{3,3}(3) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{v}_{3,4}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\vec{v}_{3,5}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{3,6}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{3,7}(3) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

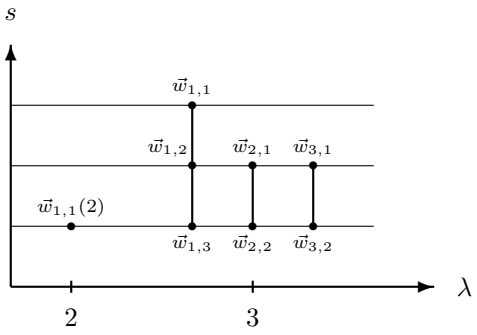
benutzen.

An dieser Stelle können wir nun schon Einiges über die Jordan-Zerlegung unserer Matrix A aussagen. Auf Grund der gefundenen Werte und der zugehörigen Tabellen

s	0	1	2	3	4	5	$\cdots$
$h_s(2)$	0	3	6	7	7	7	$\cdots$
$a_s(2)$	0	0	2	1	0	0	$\cdots$

s	0	1	2	3	$\cdots$
$h_s(2)$	0	1	1	1	$\cdots$
$a_s(2)$	0	1	0	0	$\cdots$

erkennen wir, dass die Kettenstruktur durch folgendes Hasse-Diagramm gegeben ist.



Die Jordansche Normalform J von A besteht also zum Eigenwert $\lambda = 3$ aus insgesamt $h_1(3) = 3$ Blöcken und zwar aus einem Block der Grösse 3 und aus zwei Blöcken der Grösse 2. Zusätzlich liefert der andere Eigenwert $\lambda = 2$ einen einzigen Block der Grösse 1. Deshalb muss sie folgende Gestalt haben:

$$J = \left(\begin{array}{ccc|cc|cc|c} 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{array}\right)$$

haben und ihr Minimalpolynom muss

$$\mu_A(\lambda) = (\lambda - 3)^3 \cdot (\lambda - 2)$$

sein. Was also noch fehlt, ist eine invertierbare Transformationsmatrix X, so dass die Faktorisierung

$$A \cdot X = X \cdot J$$

gilt.

Um zum Eigenwert $\lambda = 3$ eine Basis von Hauptvektorketten zu erhalten, wählen wir einen Vektor $\vec{w}_{1,1}(3)$ aus $H_3(3) = \mathrm{Ker}(A - 3E)^3$, der nicht im Teilraum $H_2(3) = \mathrm{Ker}(A-3E)^2$ liegt. Ein Blick auf die berechnete Basis von $H_3(3)$ zeigt, dass dafür etwa

$$\vec{w}_{1,1}(3) = \vec{v}_{3,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

in Frage kommt. Die damit angefangene Kette setzen wir nun nach unten fort, indem wir damit

$$\vec{w}_{1,2} = (A - 3E) \cdot \vec{w}_{1,1}(3) = \begin{pmatrix} 0 \\ -3 \\ 0 \\ -2 \\ -3 \\ -1 \\ -5 \\ -3 \end{pmatrix}, \quad \vec{w}_{1,3} = (A - 3E) \cdot \vec{w}_{1,2} = \begin{pmatrix} -2 \\ 0 \\ -4 \\ 0 \\ 0 \\ 0 \\ -2 \\ 0 \end{pmatrix}$$

berechnen. Für die beiden Vektoren $\vec{w}_{2,1}(3)$ und $\vec{w}_{3,1}(3)$ müssen wir nun zwei Vektoren aus $H_2(3) = \mathrm{Ker}(A-3E)^2$ so wählen, dass diese beiden Vektoren nicht im $H_1(3) = \mathrm{Ker}(A - 3E)$ liegen und zusammen mit dem Vektor $\vec{w}_{1,2}(3)$ linear unabhängig sind. Etwas rechnen zeigt, dass dafür

$$\vec{w}_{2,1}(3) = \vec{v}_{2,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{w}_{3,1}(3) = \vec{v}_{2,2}(3) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

in Frage kommen. Um die beiden damit angefangenen Ketten nach unten fortzusetzen, berechnen wir damit

$$\vec{w}_{2,2}(3) = (A - 3E) \cdot \vec{w}_{2,1}(3) = \begin{pmatrix} 0 \\ -1 \\ 0 \\ -2 \\ -1 \\ -1 \\ -3 \\ -1 \end{pmatrix}$$

und

$$\vec{w}_{3,2}(3) = (A - 3E) \cdot \vec{w}_{3,1}(3) = \begin{pmatrix} 1 \\ 0 \\ 2 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Zum anderen Eigenwert $\lambda = 2$ wählen wir den Vektor

$$\vec{w}_{1,1,}(2) = \vec{v}_{1,1}(2) = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Als Transformationsmatrix können wir damit also also die invertierbare Matrix

$$\begin{aligned} X &= \left(\vec{w}_{1,3}(3), \vec{w}_{1,2}(3), \vec{w}_{1,1}(3), \vec{w}_{2,2}(3), \vec{w}_{2,1}(3), \vec{w}_{3,2}(3), \vec{w}_{3,1}(3), \vec{w}_{1,1}(2)\right) \\ &= \begin{pmatrix} -2 & 0 & 1 & 0 & 1 & 1 & 0 & 2 \\ 0 & -3 & 0 & -1 & 0 & 0 & 1 & 0 \\ -4 & 0 & 0 & 0 & 2 & 2 & 0 & 4 \\ 0 & -2 & 0 & -2 & 0 & -1 & 0 & 2 \\ 0 & -3 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ -2 & -5 & 0 & -3 & 0 & 0 & 0 & 6 \\ 0 & -3 & 0 & -1 & 0 & 0 & 1 & 1 \end{pmatrix} \end{aligned}$$

verwenden. All das lässt sich mit folgendem Kode berechnen. ◯

Beispiel. Die 8×8 Matrix

$$A = \begin{pmatrix} 3 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 3 & 4 & 1 & -1 & -1 & 1 & -3 & 3 \\ -1 & 0 & 3 & 1 & 2 & -2 & 6 & -1 \\ 6 & 0 & 0 & 2 & 0 & 0 & 0 & 6 \\ 1 & -1 & 0 & 0 & 4 & 0 & 0 & 1 \\ 3 & -1 & -2 & 0 & 4 & 0 & 12 & 3 \\ 1 & 0 & -1 & 0 & 2 & -2 & 10 & 1 \\ 4 & -1 & 0 & -1 & 0 & 0 & 0 & 8 \end{pmatrix}$$

hat das charakteristische Polynom

$$\chi(\lambda) = (\lambda - 4)^6 \cdot (\lambda - 5)^2.$$

Offenbar hat also die Matrix A die beiden Eigenwerte des Spektrums

$$\sigma_A = \{4, 5\}.$$

Der erste Eigenwert $\lambda = 5$ hat die algebraische Vielfachheit 2. Um eine Basis $\mathcal{B}_1(5)$ des Eigenraumes $H_1(5)$ zu finden, müssen wir zugehörige linear unabhängige Eigenvektoren bestimmen und dazu den Kern

$$\mathrm{Ker}(A - 5E)$$

berechnen, d.h. das homogene lineare Gleichungssystem

$$(A - 5E) \cdot \vec{x} = \vec{0}$$

lösen. Weil der Rang seiner Koeffizientenmatrix $\mathrm{Rk}(A-5E)=7$ ist, beträgt die geometrische Vielfachheit dieses Eigenwertes $h_1(5)=8-7=1$ und der Eigenraum $H_1(5)$ wird etwa vom Eigenvektor

$$\vec{v}_{1,1}(5)=\begin{pmatrix}0\\0\\1\\0\\0\\2\\1\\0\end{pmatrix}$$

aufgespannt. Um die nächste Stufe $s=2$ dieses Eigenwertes zu untersuchen, benötigen wird den Hauptraum $H_2(5)$. Es ist definitionsgemäss der Kern

$$\mathrm{Ker}(A-5E)^2,$$

d.h. die Lösungsmenge des homogenen linearen Gleichungssystems

$$(A-5E)^2\cdot\vec{x}=\vec{0}.$$

Weil seine Koeffizientenmatrix den Rang $\mathrm{Rang}(A-5E)^2=6$ hat, ist sein Lösungsraum $h_2(3)=8-6=2$-dimensional. Als Basis $\mathcal{B}_2(5)$ von $H_2(5)$ können wir etwa folgende 2 Vektoren

$$\vec{v}_{2,1}(5)=\begin{pmatrix}0\\1\\0\\2\\0\\0\\0\\1\end{pmatrix},\qquad \vec{v}_{2,2}(5)=\begin{pmatrix}0\\0\\1\\0\\0\\2\\1\\0\end{pmatrix}$$

benutzen. Weil nun die algebraische Vielfachheit dieses Eigenwertes errreicht ist, gilt $\tilde{s}(5)=2$ und wir können unsere Suche abbrechen.

Zum anderen Eigenwert $\lambda=4$ gehört der Eigenraumes $H_1(4)$, der als Kern

$$\mathrm{Ker}(A-4E)$$

erklärt ist. Um eine Basis $\mathcal{B}_1(4)$ dieses Raumes zu erhalten, müssen wir das homogene lineare Gleichungssystem

$$(A-4E)\cdot\vec{x}=\vec{0}$$

lösen. Weil der Rang seiner Koeffizientenmatrix $\mathrm{Rang}(A-4E)=5$ ist, wird sein Lösungsraum von $h_1(4)=8-5=3$ linear unabhängigen Eigenvektoren

aufgespannt, die wir etwa in der Form

$$\vec{v}_{1,1}(4) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \qquad \vec{v}_{1,2}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_{1,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

wählen können.

Offenbar haben wir noch zuwenig linear unabhängige Vektoren, um damit eine invertierbare Matrix bilden zu können und müssen deshalb nun die nächste Stufe $s = 2$ untersuchen. Der zugehörige Hauptraum $H_2(4)$ ist als Kern

$$\mathrm{Ker}(A - 4E)^2,$$

d.h. als Lösungsmenge des homogenen linearen Gleichungssystems

$$(A - 4E)^2 \cdot \vec{x} = \vec{0}$$

erklärt. Weil seine Koeffizientenmatrix den Rang $\mathrm{Rk}(A - 4E)^2 = 4$ hat, ist sein Lösungsraum $h_2(4) = 8 - 4 = 4$-dimensional. Als Basis $\mathcal{B}_2(4)$ von $H_2(4)$ können wir etwa folgende 4 Vektoren

$$\vec{v}_{2,1}(4) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \vec{v}_{2,2}(4) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \vec{v}_{2,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \vec{v}_{2,4}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

benutzen.

Weil wir immer noch zu wenig linear unabhängie Vektoren haben, um damit eine invertierbare Matrix zu bilden, müssen wir als nächstes die nächste Stufe $s = 3$ untersuchen. Der zugehörige Hauptraum $H_3(4)$ ist als Kern

$$\mathrm{Ker}(A - 4E)^3,$$

d.h. als Lösungsmenge des homogenen linearen Gleichungssystems

$$(A - 4E)^3 \cdot \vec{x} = \vec{0}$$

erklärt. Weil seine Koeffizientenmatrix den Rang $\mathrm{Rang}(A - 4E)^3 = 3$ hat, ist sein Lösungsraum $h_3(4) = 8 - 3 = 5$-dimensional. Als Basis $\mathcal{B}_3(4)$ von $H_3(4)$

können wir etwa folgende 5 Vektoren

$$\vec{v}_{3,1}(4) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \quad \vec{v}_{3,2}(4) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{v}_{3,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 6 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\vec{v}_{3,4}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \qquad \vec{v}_{3,5}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

benutzen.

Noch einmal haben wir zu wenig linear unabhängie Vektoren, um damit eine invertierbare Matrix zu bilden. Deshalb müssen wir als nächstes die nächste Stufe $s = 4$ untersuchen. Der zugehörige Hauptraum $H_4(4)$ ist als Kern

$$\mathrm{Ker}(A - 4E)^4,$$

d.h. als Lösungsmenge des homogenen linearen Gleichungssystems

$$(A - 4E)^4 \cdot \vec{x} = \vec{0}$$

erklärt. Weil seine Koeffizientenmatrix den Rang $\mathrm{Rk}(A - 4E)^4 = 2$ hat, ist sein Lösungsraum $h_4(4) = 8 - 2 = 6$-dimensional und wir können unsere Suche abbrechen, da $\tilde{s}(4) = 6$ ist. Als Basis $\mathcal{B}_4(4)$ von $H_4(4)$ können wir etwa folgende 6 Vektoren

$$\vec{v}_{4,1}(4) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \quad \vec{v}_{4,2}(4) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{v}_{4,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\vec{v}_{4,4}(4) = \begin{pmatrix} 0 \\ 0 \\ 6 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_{4,5}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \qquad \vec{v}_{4,6}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

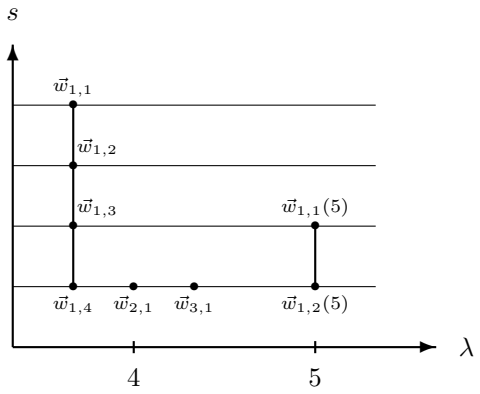
benutzen.

An dieser Stelle können wir nun schon Einiges über die Jordan-Zerlegung unserer Matrix A aussagen. Auf Grund der gefundenen Werte und der zugehörigen Tabellen

s	0	1	2	3	4	$\cdots$
$h_s(5)$	0	1	2	2	2	$\cdots$
$a_s(5)$	0	0	1	0	0	$\cdots$

s	0	1	2	3	4	5	6	$\cdots$
$h_s(4)$	0	3	4	5	6	6	6	$\cdots$
$a_s(4)$	0	2	0	0	1	0	0	$\cdots$

erkennen wir, dass die Kettenstruktur durch folgendes Hasse-Diagramm gegeben ist.



Die Jordansche Normalform J von A besteht also zum Eigenwert $\lambda = 4$ aus insgesamt $h_1(4) = 3$ Blöcken und zwar aus einem Block der Grösse 4 und aus zwei Blöcken der Grösse 1. Zusätzlich liefert der andere Eigenwert $\lambda = 5$ einen einzigen Block der Grösse 2. Deshalb muss sie folgende Gestalt haben:

$$J = \left(\begin{array}{cccc|c|c|cc} 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \end{array}\right)$$

haben und ihr Minimalpolynom muss

$$\mu_A(\lambda) = (\lambda - 4)^4 \cdot (\lambda - 5)^2$$

sein. Was also noch fehlt, ist eine invertierbare Transformationsmatrix X, so dass die Faktorisierung

$$A \cdot X = X \cdot J$$

gilt.

Um zum Eigenwert $\lambda = 4$ eine Basis von Hauptvektorketten zu erhalten, wählen wir einen Vektor $\vec{w}_{1,1}(4)$ aus $\mathrm{Ker}(A-4E)^4$, der nicht in $\mathrm{Ker}(A-4E)^3$ liegt. Ein

Blick auf die berechnete Basis von $H_4(3)$ zeigt, dass dafür etwa

$$\vec{w}_{1,1}(4) = \vec{v}_{4,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

in Frage kommt. Die damit angefangene Kette setzen wir nun nach unten fort, indem wir damit

$$\vec{w}_{1,2}(4) = (A-4E)\cdot\vec{w}_{1,1}(4) = \begin{pmatrix} -1 \\ 0 \\ 2 \\ 0 \\ 1 \\ 3 \\ 1 \\ 1 \end{pmatrix}, \quad \vec{w}_{1,3}(4) = (A-4E)\cdot\vec{w}_{1,2}(4) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

und

$$\vec{w}_{1,4}(4) = (A-4E)\cdot\vec{w}_{1,3} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ 0 \\ -1 \end{pmatrix}$$

berechnen. Für die beiden Vektoren $\vec{w}_{2,1}(4)$ und $\vec{w}_{3,1}(4)$ müssen wir nun zwei Vektoren aus dem Raum $H_1(4) = \mathrm{Ker}(A-4E)$ so wählen, dass diese beiden Vektoren zusammen mit dem Vektor $\vec{w}_{1,1}(4)$ eine Basis von $H_1(4) = \mathrm{Ker}(A-4E)$ ist. Etwas rechnen zeigt, dass dafür

$$\vec{w}_{2,1}(4) = \vec{v}_{1,1}(4) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \qquad \vec{w}_{3,1}(4) = \vec{v}_{1,3}(4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

in Frage kommen.

Zum anderen Eigenwert $\lambda = 5$ wählen wir einen Vektor $\vec{w}_{1,1}(5)$ aus $H_2(5)$ der

nicht in $H_1(5) = \mathrm{Ker}(A - 5E)$ ist. Dafür kommt etwas

$$\vec{w}_{1,1}(5) = \vec{v}_{2,1}(5) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

in Frage. Die damit angefangene neue Kette wird durch den Vektor

$$\vec{w}_{1,2}(5) = (A - 5E) \cdot \vec{w}_{1,1}(5) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

fortgesetzt. Als Transformationsmatrix können wir damit also also die invertierbare Matrix

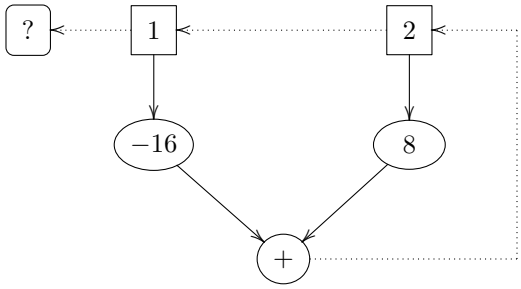
$$\begin{aligned} X &= \left(\vec{w}_{1,4}(4), \vec{w}_{1,3}(4), \vec{w}_{1,2}(4), \vec{w}_{1,1}(4), \vec{w}_{2,1}(4), \vec{w}_{3,1}(4), \vec{w}_{1,2}(5), \vec{w}_{1,1}(5)\right) \\ &= \begin{pmatrix} 1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 2 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 3 & 0 & 0 & 3 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ -1 & 0 & 1 & 1 & -1 & 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

verwenden. ◯

Die in einem früheren Beispiel angetroffenen drei Rekursionsgleichungen höher Ordnung der drei Komponentenfolgen

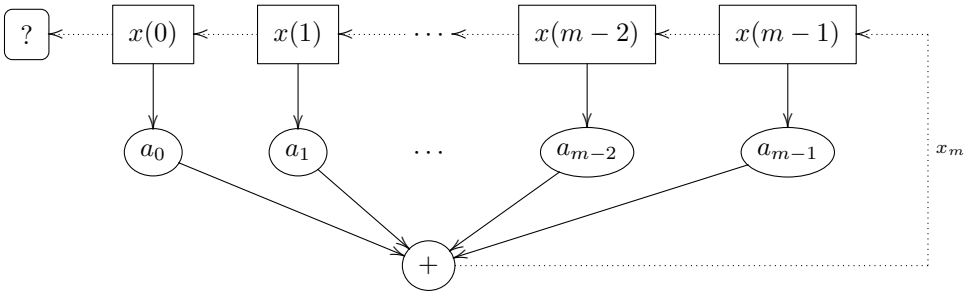
$$\left\{ \begin{array}{lcll} x(k+2) & = & -16x(k) + 8x(k+1), & x(0) = 1 \quad x(1) = 2 \\ y(k+2) & = & -16y(k) + 8y(k+1), & y(0) = 0 \quad y(1) = 4 \\ z(k+2) & = & -16z(k) + 8z(k+1), & z(0) = -1 \quad z(1) = -6 \end{array} \right.$$

lassen sich mit Hilfe eines minimalen linearen *Schieberegisters* in der Form



realisieren. Man beachte, dass die drei Rekursionsgleichungen für die einzelnen Komponenten und damit die zugehörigen Schieberegister in ihrem Aufbau übereinstimmen. Sie unterscheiden sich einzig in den Anfangswerten. Dieses Schieberegister lässt sich dank der Minimalistät des verwendeten Polynoms nicht mehr weiter verkleinern. Sonst würde es drei geometrische Folgen produzieren und der Quotient aufeinanderfolgender Glieder müsste konstant sein.

Techniker verwenden im Umgang mit solche rekursiven Folgen ein handfestes Bild und verstehen unter einem *linearen rückgekoppelten Schieberegister* der Länge m einen Schaltkreis der Bauart



Dabei wird jede der m quadratischen Speicherzellen in der oberen Zeile von links nach rechts mit einem der m Anfangswerte $x(0), x(1), \cdots, x(m-1)$ gefüllt (Initialisierung). Beim ersten durch eine externe Uhr gesteuerten Takt wird mit den fest gewählten Koeffizienten a_j die Linearkombination

$$x(m) = a_0\ x(0) + a_1\ x(1) + \cdots + a_{m-2}\ x(m-2) + a_{m-1}\ x(m-1) = \sum_{k=0}^{m-1} a_k x(k)$$

bestimmt und jeder Zelleninhalt um eine Position nach links verschoben. Der Wert in der Zelle ganz links wird in Pfeilrichtung ausgegeben und der Wert in der Zelle ganz rechts wird durch die berechnete Linearkombination ersetzt. (Rekursion)

Die eindeutig bestimmte Ausgabe eines solchen Schieberegisters der Länge m nach k Takten kann also durch die homogene lineare Differenzengleichung m. Ordnung mit konstanten Koeffizienten

$$x(k+m) = a_0\ x(k) + a_1\ x(k+1) + \cdots + a_{m-2}\ x(k+m-2) + a_{m-1}\ x(k+m-1)$$

beschrieben werden. Dieses Schieberegister bzw. die zugehörigen Differenzengleichung lässt sich mit Hilfe des *charakteristischen Polynoms*

$$\chi(\lambda) = a_0 + a_1\lambda + \cdots + a_{m-2}\lambda^{m-2} + a_{m-1}\lambda^{m-1} - \lambda^m = \sum_{k=0}^{m-1} a_k\lambda^k - \lambda^m$$

bzw. die zugehörige charakteristische Gleichung

$$\lambda^k = \sum_{k=0}^{m-1} a_k\lambda^k$$

vom Grad m knapp beschrieben. Techniker verwenden statt des charakteristischen Polynoms oft das gleichwertige *Rückkopplungspolynom*

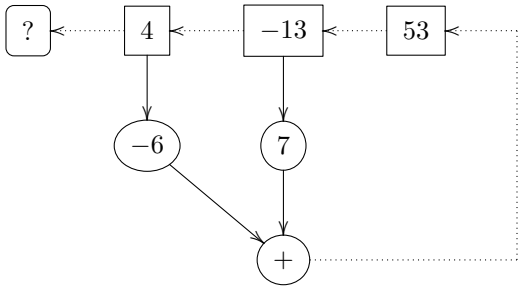
$$\rho(x) = -\chi\Big(\frac{1}{x}\Big) \cdot x^m = 1 - a_{m-1}x - a_{m-2}x^2 - \cdots - a_1x^{m-1} - a_0x^m$$

in dem also die Koeffizienten im Wesentlichen in umgekehrter Reihenfolge und mit umgekehrten Vorzeichen auftreten.

Lineare Schieberegister liefern bequeme dynamische Systeme (Maschinen), weil:

1. Sie einfach praktisch gebaut werden können.
2. Ihr Langzeitverhalten gut untersucht werden kann.
3. Ihre Ausgabe gut statistisch untersucht werden kann.
4. Sie eine mathematische Struktur haben, die mit Hilfe der Methoden der linearen und der kommutativen Algebra untersucht werden kann.

Beispiel. Im Schieberegister



sind nicht alle Zellen angezapft. Es wird durch die lineare Differenzengleichung

$$x(k+3) = -6x(k) + 7x(k+1)$$

dritter Ordnung mit dem charakteristischen Polynom

$$\chi(\lambda) = -6 + 7\lambda - \lambda^3 = -(\lambda - 1) \cdot (\lambda - 2) \cdot (\lambda + 3)$$

beschrieben.

Die Anfangsbedingung $x(0) = 4, x(1) = -13, x(2) = 53$ liefert die Folge mit den Werten

k	0	1	2	3	4	5	$\cdots$
$x(k)$	4	-13	53	-115	449	-1123	$\cdots$

Wie man durch Einsetzen in die Rekursionsgleichung sieht, liefern die drei Folgen $x_1(k) = 1^k$, $x_2(k) = 2^k$ und $x_3(k) = (-3)^k$ Lösungen dieser Differenzengleichungen, die allerdings zu den unterschiedlichen Anfangsbedingungen

$$\begin{array}{lll} x_1(0) = 1 & x_1(1) = 1 & x_1(2) = 1 \\ x_2(0) = 1 & x_2(1) = 2 & x_2(2) = 4 \\ x_3(0) = 1 & x_3(1) = -3 & x_3(2) = 9 \end{array}$$

gehören. Die drei Basislösungen können sofort aus den Nullstellen des charakteristischen Polynoms gewonnen werden. Aus den drei Basislösungen kann jede beliebige Lösung der Rekursionsgleichung als Linearkombination in der Form

$$x(k) = a \cdot x_1(k) + b \cdot x_2(k) + c \cdot x_3(k) = a \cdot 1^k + b \cdot 2^k + c \cdot (-3)^k$$

zusammengebaut werden.

Soll zum Beispiel jene Lösung linear kombiniert werden, die zur ursprünglichen Anfangsbedingung

$$x(0) = 4, \quad x(1) = -13, \quad x(2) = 53$$

gehört, so ist für die Koeffizienten das lineare Gleichungssystem

$$\left\{ \begin{array}{rrrr} a+ & b+ & c= & 4 \\ a+ & 2b- & 3c= & -13 \\ a+ & 4b+ & 9c= & 53 \end{array} \right. \qquad \left(\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 1 & 2 & -3 & -13 \\ 1 & 4 & 9 & 53 \end{array} \right)$$

zu lösen. Um nicht mit so grossen (sic!) Zahlen rechnen zu müssen, lösen wir das System gleich allgemeiner mit beliebiger rechter Seite und gehen dazu von der Blockmatrix

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & -3 & 0 & 1 & 0 \\ 1 & 4 & 9 & 0 & 0 & 1 \end{array} \right)$$

aus. Im Eliminationsverfahren addieren wir zunächst das (-1)-fachen der ersten zur zweiten und dritten Zeile und erhalten

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & -4 & -1 & 1 & 0 \\ 0 & 3 & 8 & -1 & 0 & 1 \end{array} \right)$$

Addition des (-3)-fachen der zweiten Zeile zur dritten liefert die Stufenform

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & -4 & -1 & 1 & 0 \\ 0 & 0 & 20 & 2 & -3 & 1 \end{array} \right)$$

Addition der dritten Zeile zum 5-fachen der zweiten und zum (-20)-fachen der ersten Zeile liefert

$$\left(\begin{array}{ccc|ccc} -20 & -20 & 0 & -18 & -3 & 1 \\ 0 & 5 & 0 & -3 & 2 & 1 \\ 0 & 0 & 20 & 2 & -3 & 1 \end{array} \right)$$

Addition des 4-fachen der zweiten Zeile liefert die reduzierte Stufenform

$$\left(\begin{array}{ccc|ccc} -20 & 0 & 0 & -30 & 5 & 5 \\ 0 & 5 & 0 & -3 & 2 & 1 \\ 0 & 0 & 20 & 2 & -3 & 1 \end{array} \right)$$

In anderen Worten formuliert, haben wir hier die Inverse der Koeffizientenmatrix X berechnet und die beiden inversen Matrizen

$$X = \left(\begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 1 & 4 & 9 \end{array} \right), \qquad X^{-1} = \frac{1}{20} \left(\begin{array}{ccc} 30 & -5 & -5 \\ -12 & 8 & 4 \\ 2 & -3 & 1 \end{array} \right)$$

erhalten. Es scheint also, als ob unser System durch eine Matrix A beschreiben werden kann, die auf Grund der Diagonalisierung

$$A \cdot X = X \cdot \Lambda, \qquad A = X \cdot \Lambda \cdot X^{-1} = \left(\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & 7 & 0 \end{array} \right)$$

hat, wobei die Diagonalmatrix

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$

die drei gefundenen Eigenwerte auf der Diagonalen und X die zugehörigen Eigenvektoren als Spalten hat.

Man beachte, dass die Matrix X und A auf Grund der Fragestellung eine ganz spezielle Gestalt haben. Die Systemmatrix A ist eine sogn. *Begleitermatrix* und X ist eine sogn. *Vandermonde-Matrix*. Solche Matrizen kommen also kommen in den Anwendungen oft vor und es scheint, dass die Matrix X invertierbar sein muss, was wir ja am Beispiel überprüft haben. Wie unser Beispiel ferner zeigt, diagonalisieren Vandermonde-Matrizen Begleitermatrizen.

Aus der gefundenen Lösung des linearen Gleichungssystems entnehmen wir insbesondere für das konkrete numerische Problem die eindeutig bestimmte Lösung $a = -4$, $b = 3$ und $c = 5$. Daher hat die gesuchte Folge die explizite Beschreibung

$$x(k) = -4 \cdot x_1(k) + 3 \cdot x_2(k) + 5 \cdot x_3(k) = -4 + 3 \cdot 2^k + 5 \cdot (-3)^k$$

wie man leicht durch Einsetzen überprüft. Offenbar haben wir eine explizite Beschreibung der durch das Schieberegister erzeugten Folge gefunden. Man beachte den Einfluss der drei unterschiedlichen Eigenwerte auf das Verhalten der Folge. Der Eigenwert $\lambda_1 = 1$ liefert eine additive Konstante, der Eigenwert $\lambda_2 = 2$ führt zu einem exponentiellen Wachstum und der betragsmässig grösste Eigenwert $\lambda_3 = -3$ dominiert das Verhalten und führt insbesondere zu einer Oszillation und einem betragsmässigen exponentiellen Wachstum. En passant haben wir wir das asymptotische Verhalten

$$x(k) \sim 5 \cdot (-3)^k$$

der Folge geklärt. ◯

Nicht nur liefert jedes System linearer Differenzengleichungen eine lineare Differenzengleichung höherer Ordnung. Umgekehrt kann — und soll aus mathematischen Gründen — jede Differenzengleichung (bzw. Differentialgleichung) höherer Ordnung als System von Differenzengleichungen (Differentialgleichungen) erster Ordnung aufgefasst werden!

Beispiel. Im Fall der im letzten Beispiel benutzten Rekursion

$$x(k+3) = -6\ x(k) + 7\ x(k+1)$$

der Ordnung 3 kodiert man den Zustand des zugehörigen Schieberegisters durch den 3-komponentigen Vektor

$$\vec{y}(k) = \begin{pmatrix} x(k) \\ x(k+1) \\ x(k+2) \end{pmatrix} \in \mathbb{R}^3$$

Man führt also die drei zeitverschobenen Zustandsvariablen

$$y_1(k) = x(k), \qquad y_2(k) = x(k+1), \qquad y_3(k) = x(k+2)$$

ein und drückt alle diese Variablen nun mit Hilfe dieser Definitionen und der Rekursionsgleichung als Linearkombination der neuen Zustandsvariablen aus.

Im Beispiel ergibt sich das System von Rekursionsgleichungen erster Ordnung

$$\begin{cases} y_1(k+1) = & & y_2(k) \\ y_2(k+1) = & & & y_3(k) \\ y_3(k+1) = -6y_1(k) & + 7y_2(k) \end{cases}$$

Dieses dynamische System kann mit Hilfe der Systemmatrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & 7 & 0 \end{pmatrix}$$

in matrizieller Form $\vec{y}(k+1) = A \cdot \vec{y}(k)$ beschrieben werden, die wir oben aus der Diagonalisierung $A = X \cdot \Lambda \cdot X^{-1}$ erhalten haben. Kodieren wir noch den seinerzeit verwendeten Anfangszustand durch den Vektor

$$\vec{a} = \begin{pmatrix} 4 \\ -13 \\ 53 \end{pmatrix},$$

wird die Folge der Zustände des Schieberegisters durch die lineare Rekusion

$$\vec{y}(k+1) = A \cdot \vec{y}(k), \qquad \vec{y}(0) = \vec{a}$$

beschrieben und das beschriebene Verdoppelungsverfahren zur Berechnung der Potenzen von A liefert ein effizientes Verfahren zur numerischen Berechnung der Zustände des zugehörenden Schieberegisters.

Die gefundene *Begleitermatrix* des charakteristischen Polynoms

$$\chi(\lambda) = -6 + 7\lambda - \lambda^3$$

hat in der oberen Nebendiagonale lauter 1 und in der untersten Zeile die Koeffizienten des charakteristischen Polynoms in aufsteigender Reihenfolge. Wir erwarten selbstverständlich, dass das charakteristische Polynom einer Begleitermatrix bis auf ein Vorzeichen gerade das Polynom, mit dem sie gebildet wurde.

Zur Bestätigung dieser Vermutung und und zur Übung berechnen wir nun ihr charakteristisches Poynom, indem wir die Eigenwertgleichung

$$A \cdot \vec{x} = \lambda\vec{x}, \qquad (A - \lambda E) \cdot \vec{x} = \vec{0}$$

mit der Koeffizientenmatrix

$$A - \lambda E = \begin{pmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \\ -6 & 7 & -\lambda \end{pmatrix}$$

lösen. Um in jedem Fall eine umkehrbare Operation durchzuführen vertauschen wir die erste und die dritte Zeile.

$$\begin{pmatrix} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ -\lambda & 1 & 0 \end{pmatrix}$$

In der entstandenen äquivalenten Matrix addieren nun das $(-\lambda)$-fache der ersten Zeile zum 6-fachen der dritten Zeile und erhalten

$$\begin{pmatrix} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ 0 & 6-7\lambda & \lambda^2 \end{pmatrix}$$

Um nun weiterhin unabhängig von λ Äquivalenzumformungen durchführen zu können, sorgen wir durch Wechselwegnahme dafür, dass an der gewünschten Stelle zunächst eine Konstante entsteht, die wir dann zu einer 0 machen können. Dazu addieren wir zunächst das (-7)-fache der zweiten Zeile zur dritten und erhalten die Matrix

$$\begin{pmatrix} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ 0 & 6 & \lambda^2-7 \end{pmatrix}$$

Sie hat an der fraglichen Stelle die gewünschte Konstante. Durch Vertauschen der letzten beiden Zeilen erhalten wir die äquivalente Matrix

$$\begin{pmatrix} -6 & 7 & -\lambda \\ 0 & 6 & \lambda^2-7 \\ 0 & -\lambda & 1 \end{pmatrix}$$

in der wir schliesslich das λ-fache der zweiten Zeile zum 6-fachen der dritten Zeile addieren und die Stufenform

$$S(\lambda) = \begin{pmatrix} -6 & 7 & -\lambda \\ 0 & 6 & \lambda^2-7 \\ 0 & 0 & \chi(\lambda) \end{pmatrix}$$

in der unten rechts tatsächlich das erwartete charakteristische Polynom

$$\chi_A(\lambda) = \det(A-\lambda E) = -6+7\lambda-\lambda^3$$

entstanden ist. Die Begleitermatrix A von $\chi_A(\lambda)$ ist nach dem Satz von Cayley-Hamilton Nullstelle des charakteristischen Polynoms, d.h. es ist

$$\chi_A(A) = -6E+7A-A^3 = 0,$$

wie man leicht bestätigt. Offenbar ist es also leicht, matrizielle Lösungen von Polynomgleichungen zu erhalten und die Matrizenrechnung steht in enger Beziehung zur Theorie der Polynome. Jedem normierten Polynom lässt sich mit Hilfe der Begleitermatrix auf kanonische Art eine quadratische Matrix zuordnen, deren Eigenvektoren sich mit Hilfe der Nullstellen des Polynoms leicht bestimmen lassen. Allgemein gilt nämlich:

Definition. Zum Polynom

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-2}x^{n-2} + a_{n-1}x^{n-1} - x^n = \sum_{k=0}^{n-1} a_kx^k - x^n$$

vom Grad n gehört die *Begleitermatrix*

$$B(p) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-2} & a_{n-1} \end{pmatrix} \in \mathbb{R}^{n,n}$$

Ihr charakteristisches Polynom ist

$$\chi_{B(p)}(\lambda) = p(\lambda) = \prod_{k=1}^{n}(\lambda - \lambda_k).$$

Weil jeder Eigenwert λ_k von $B(p)$ die charakteristische Gleichung

$$\lambda_k^n = \sum_{j=0}^{n-1} a_j \lambda_k^j$$

erfüllt, ist der Vektor

$$\vec{v}(k) = \begin{pmatrix} 1 \\ \lambda_k \\ \lambda_k^2 \\ \cdots \\ \lambda_k^{n-2} \\ \lambda_k^{n-1} \end{pmatrix}$$

ein Eigenvektor von $B(p)$, der zum Eigenwert λ_k gehört. Daher wird die Begleitermatrix durch die Vandermonde-Matrix

$$V(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{n-1}, \lambda_n) = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_{n-1} & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \cdots & \lambda_{n-1}^2 & \lambda_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_1^{n-2} & \lambda_2^{n-2} & \lambda_3^{n-2} & \cdots & \lambda_{n-1}^{n-2} & \lambda_n^{n-2} \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \lambda_3^{n-1} & \cdots & \lambda_{n-1}^{n-1} & \lambda_n^{n-1} \end{pmatrix}$$

diagonalisiert, d.h. es ist

$$B(p) \cdot V(\lambda_1, \lambda_2, \ldots, \lambda_n) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \cdot V(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Dabei bezeichnen $\lambda_1, \lambda_2, \ldots \lambda_k, \ldots, \lambda_n$ die Nullstellen von $p(\lambda)$ und für das typische Element der Vandermonde-Matrix gilt

$$V(\lambda_1, \lambda_2, \ldots, \lambda_n)_{j,k} = \lambda_k^{j-1}$$

Aus der Begleitermatrix $B(p)$ des Polynms $p(x)$ erhält man also umgekehrt das Polynom $p(x)$ zurück, mit dem man sie gebildet hat. Die enge Beziehung zwischen Polynomen und Matrizen wird noch enger, wenn man bedenkt, dass mit dem Satz von Cayley-Hamilton jede Matrix Nullstelle ihres eigenen charakteristischen Polynoms ist. Aus diesen beiden Sätzen zusammen folgt die bemerkenswerte Tatsache, dass jedes Polynom $p(x)$ sein Begleitermatrix $B(p)$ als Matrixlösung hat. Insbesondere ist jede Polynomgleichung mit Matrizen lösbar und daher bieten sich Matrizen als verallgemeinerte Zahlen an.

Das betrachtete dynamische System ist reversibel. Durch Zeitumkehr kann jeder beliebige Zustand aus einem gewissen Anfangszustand erhalten werden. Um den inversen Prozess zu beschreiben, benötigen wir die Inverse

$$A^{-1} = \frac{1}{6}\begin{pmatrix} 7 & 0 & -1 \\ 6 & 0 & 0 \\ 0 & 6 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

Um etwa obigen Zustand $\vec{a}$ in zwei Schritten zu erreichen, muss man das System im Zustand

$$(A^{-1})^2 \cdot \vec{a} = A^{-2} \cdot \vec{a} = \frac{1}{36}\begin{pmatrix} 49 & -6 & -7 \\ 42 & 0 & -6 \\ 36 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ -13 \\ 53 \end{pmatrix} = \frac{1}{36}\begin{pmatrix} -97 \\ -150 \\ 144 \end{pmatrix}$$

starten, wie man leicht berechnet und dann kontrolliert.

Weil das charakteristische Polynom $\chi_A(\lambda) = (\lambda-1)\cdot(\lambda-2)\cdot(\lambda+3)$ die drei Nullstellen $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = -3$ hat, sind hier die Eigenwerte sogar ganzzahlig und als zugehörige Eigenvektoren lassen sich die Anfangsbedingungen der drei oben angegebenen Basislösungen $x_1(k) = 1^k$, $x_2(k) = 2^k$ und $x_3(k) = (-3)^k$ d.h. die Vektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} 1 \\ -3 \\ 9 \end{pmatrix}$$

verwenden, wie man leicht an Hand der Eigenwertgleichung $A \cdot \vec{v}_j = \lambda_j \vec{v}$ und der übersichtlichen Struktur von Begleitermatrizen bestätigt. Eigenvektoren von Begleitermatrizen lassen sich also einfach bestimmen.

Mit Hilfe dieses Eigensystems lassen sich nun die Transformationsmatrizen

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 1 & 4 & 9 \end{pmatrix}, \qquad X^{-1} = \frac{1}{20}\begin{pmatrix} 30 & -5 & -5 \\ -12 & 8 & 4 \\ 2 & -3 & 1 \end{pmatrix}$$

bestimmen, die wir bereits bei einer früherer Glegenheit berechnet haben. Mit der Diagonalmatrix

$$\Lambda = X^{-1} \cdot A \cdot X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$

erhalten wir die Diagonalisierung $A = X \cdot \Lambda \cdot X^{-1}$ und für die Matrizenpotenzen $A^k = X \cdot \Lambda^k \cdot X^{-1}$ die explizite Beschreibung

$$\frac{1}{20}\begin{pmatrix} 30-12\,2^k+\ 2\,(-3)^k & -5+\ 8\,2^k-\ 3\,(-3)^k & -5+\ 4\,2^k+\ (-3)^k \\ 30-24\,2^k-\ 6\,(-3)^k & -5+16\,2^k+\ 9\,(-3)^k & -5+\ 8\,2^k-3\,(-3)^k \\ 30-48\,2^k+18\,(-3)^k & -5+32\,2^k-27\,(-3)^k & -5+16\,2^k+9\,(-3)^k \end{pmatrix}$$

Für für $k = 1$ ergibt sich die Begleitermatrix A und für $k = -1$ die oben bestimme Inverse A^{-1}.

Selbstverständlich liefert die erste Komponente des Zustandes $\vec{y}(k) = A^k \cdot \vec{a}$ die seinerzeit bereits bestimmte explizite Beschreibung der Folge

$$x(k) = -4 + 3 \cdot 2^k + 5 \cdot (-3)^k$$

Obwohl also beide Sichtweisen — via Systeme erster Ordnung bzw. via Rekursionsgleichungen höherer Ordnung — im Prinzip die selben Resultate liefern, empfiehlt es sich aus konzeptionellen Gründen, Differenzengleichungen höherer Ordnung als Systeme erster Ordnung in höheren Dimensionen aufzufassen, weil man dann auf die geometrische Intuition zurückgreifen kann. Die analoge Bemerkung gilt für die entsprechenden Differentialgleichungen höherer Ordnung. Insbesondere werden auf diese Weise viele Mehrspurigkeiten der Ingenieur-Mathematik im Umkreis von Z- und Laplace-Transformation völlig überflüssig und können als Spezialfälle von Fragestellungen aus der Matrizenrechnung verstanden und behandelt werden, um die man sowieso für seriöse Anwendungen nicht herumkommt. ◯

Das Eliminationsverfahren liefert also einen effektiven Algorithmus zur Berechnung des charakteristischen Polynoms einer Matrix A. Das Problem, dieses Polynom zu faktorisieren und so die Eigenwerte von A zu bestimmen, ist im allgemeinen nicht linear, ja in der Regel nicht einmal rational und daher viel anspruchsvoller! Dazu müssen wir insbesondere zuerst genügend geeignete Zahlen zur Verfügung haben, die als Lösungen in Frage kommen und dann müssen Matrizenalgebra und Vektorgeometrie zur kommutativen Algebra und zur algebraischen Geometrie verallgemeinert werden. In der Numerik werden subtile Methoden besprochen, um Eigenwertgleichungen näherungsweise zu lösen.

Oft sind die Parameter eines Schieberegisters der Länge m nicht zum Vorneherein bekannt, sondern sollen so bestimmt werden, dass es ein vorgeschriebenes Verhalten hat. Weil dazu die m Anfangswerte und die m Koeffizienten bestimmt werden müssen, sollten Daten der Länge $2m$ zur Verfügung stehen.

Beispiel. Um das kleinste Schieberegister zu finden, das die Folge mit dem Anfangsstück der Länge $2m = 6$

k	0	1	2	3	4	5
$x(k)$	5	0	30	-30	210	-390

produziert, machen wir für die Rekursionsgleichung der Ordnung $\lceil \frac{2m}{2} \rceil = 3$ den Ansatz

$$x(k+3) = \alpha \cdot x(k) + \beta \cdot x(k+1) + \gamma \cdot x(k+2)$$

und haben die unbekannten Koeffizienten zu bestimmen. Einsetzen der bekannten Werte liefert für $k = 0, 1, 2$ die Gleichungen des folgenden linearen Gleichungssystems

$$\left\{ \begin{array}{rrrr} 5\alpha & + \ 0\beta + & 30\gamma = & -30 \\ 0\alpha & + 30\beta - & 30\gamma = & 210 \\ 30\alpha & - 30\beta + & 210\gamma = & -390 \end{array} \right. \qquad \left(\begin{array}{rrr|r} 5 & 0 & 30 & -30 \\ 0 & 30 & -30 & 210 \\ 30 & -30 & 210 & -390 \end{array} \right)$$

mit der zugehörigen erweiterten Matrix rechts daneben. Um möglichst kleine Zahlen zu erhalten, dividieren wir die erste Zeile durch 5 und die beiden anderen durch 30. Dann lautet die erweiterte Matrix des Systems

$$\left(\begin{array}{rrr|r} 1 & 0 & 6 & -6 \\ 0 & 1 & -1 & 7 \\ 1 & -1 & 7 & -13 \end{array} \right)$$

Addieren wir das (-1)-fache der ersten Zeile zur dritten, erhalten wir die Matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & 6 & -6 \\ 0 & 1 & -1 & 7 \\ 0 & -1 & 1 & -7 \end{array}\right), \qquad \left(\begin{array}{ccc|c} 1 & 0 & 6 & -6 \\ 0 & 1 & -1 & 7 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

in der man erkennt, dass die dritte Zeile überflüssig ist, weil sie das negative der zweiten Zeile ist. Das Gleichungssystem kann also mit Hilfe der rechts stehenden reduzierten Stufenform gelöst werden, die durch Addition der zweiten zur dritten Zeile entsteht. Daraus lesen wir die Lösung in vektorieller Form

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} -6 \\ 7 \\ 0 \end{pmatrix} + t \begin{pmatrix} -6 \\ 1 \\ 1 \end{pmatrix}$$

ab. Die Lösung des linearen Gleichungssystems, die geometrisch die Schnittgerade der drei durch die ursprünglichen linearen Gleichungen beschriebenen Ebenen beschreibt, ist nicht eindeutig bestimmt, sondern jedes $t \in \mathbb{Z}$ in den Formeln $\alpha = -6 - 6t$, $\beta = 7 + t$ und $\gamma = t$ liefert eine Lösung und damit eine Rekursionsgleichung für die gegebene Liste. Für ganzzahlige $t \in \mathbb{Z}$ sind die Koeffizienten der Rekursion sogar ganzzahlig.

Für die Wahl $t = 0$ ist $\alpha = -6$, $\beta = 7$ und $\gamma = 0$ und daher kann das gegebene Anfangsstück der Folge durch die bereits im letzten Beispiel benutzte Rekursion

$$x(k+3) = -6 \cdot x(k) + 7 \cdot x(k+1), \qquad x(0) = 5, \quad x(1) = 0, \quad x(2) = 30$$

dritter Ordnung und das zugehörige Schieberegister — allerdings mit neuen Anfangswerten — erzeugt werden. Die damit erzeugten Folgenglieder sind dann $x(6) = 1'650$ und $x(7) = -3'990$.

Für die Wahl $t = -1$ erhalten wir hingegegen $\alpha = 0$, $\beta = 6$ und $\gamma = -1$ mit der zugehörigen Rekursion

$$x(k+3) = 6 \cdot x(k+1) - x(k+2), \qquad x(0) = 5, \quad x(1) = 0, \quad x(2) = 30$$

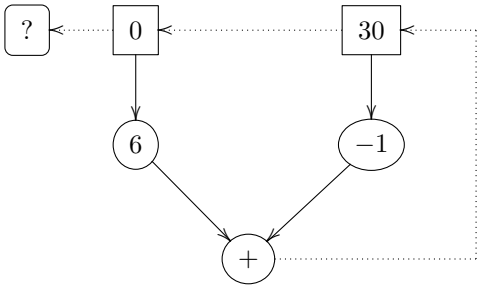
Man beachte, dass es sich nach dem Zeitshift $\tilde{x}(k) = x(k+1)$ um eine Rekursion zweiter Ordnung handelt, die in der Form

$$\tilde{x}(k+2) = 6 \cdot \tilde{x}(k) - \tilde{x}(k+1), \qquad \tilde{x}(0) = x(1) = 0, \quad \tilde{x}(1) = x(2) = 30$$

geschrieben werden kann und das charakteristische Polynom

$$\chi(\lambda) = 6 - \lambda - \lambda^2 = -(\lambda - 2)(\lambda + 3)$$

zweiter Ordnung hat, zu dem das Schieberegister



gehört. Es produziert tatsächlich das Anfangsstück der gegebenen Folge.

k	0	1	2	3	4
$\tilde{x}(k)$	0	30	-30	210	-390

Man könnte hier sogar die selben Anfangswerte $\hat{x}(0) = 5$ und $\hat{x}(1) = 0$ wie für die ursprüngliche Folge $x(n)$ wählen, weil tatsächlich $\hat{x}(2) = 30$ gilt. Die Folge kann also durch die Rekursion

$$\hat{x}(k+2) = 6 \cdot \hat{x}(k) - \hat{x}(k+1), \qquad \hat{x}(0) = 5, \quad \hat{x}(1) = 0$$

beschrieben werden.

Aus diesen endlich vielen Werten kann man aber noch nicht schliessen, dass die beiden Folgen $x(k)$ und $\tilde{x}(k)$, die ja durch völlig unterschiedliche Rekursionen

$$x(k+3) = -6x(k) + 7x(k+1), \qquad \text{und} \qquad \hat{x}(k+2) = 6\hat{x}(k) - \hat{x}(k+1)$$

beschrieben sind, für alle Werte von k den selben Wert liefern. Diesen Sachverhalt kann man mit Hilfe von vollständiger Induktion beweisen. Weil dieses mathematische Werkzeug nicht mehr allen Schülern zur Verfügung steht, wählen wir einen anderen Weg zur Begründung.

Dass die Folge $x(k)$ sogar mit einer Rekursionsgleichung kleinerer Ordnung beschrieben werden kann als ursprünglich erwartet, hängt damit zusammen, dass das lineare Gleichungssystem nicht nur eine einzige Lösung hat. Man beachte, dass das charakteristische Polynom der minimalen Rekursionsgleichung zweiter Ordnung ein Teiler des charakteristischen Polynoms der im letzten Beispiel benutzten Rekursion dritter Ordnung ist. Die zugehörige Folge wird daher durch eine explizite Formel der Art

$$\hat{x}(k) = b \cdot 2^k + c \cdot (-3)^k$$

beschrieben. Zur Bestimmung der beiden Koeffizienten müssen wir das lineare Gleichungssystem

$$\begin{cases} b + \ c = 5 \\ 2b - 3c = 0 \end{cases}$$

lösen. Seine eindeutige Lösung ist $b = 3$ und $c = 2$. Daher hat also die Fortsetzung des gegebenen Anfangsstücks die explizite Beschreibung

$$\hat{x}(k) = 3 \cdot 2^k + 2 \cdot (-3)^k = x(k)$$

Nachdem wir explizite Beschreibungen der beiden Folgen $x(k)$ und $\hat{x}(k)$ zur Verfügung haben, ist es nun leicht einzusehen, dass die beiden Folgen für alle k übereinstimmen müssen. Dazu kontrolliert man, dass die Folge $\hat{x}(k)$ die Rekursionsgleichung und die Anfangsbedingungen der anderen Folge $x(k)$ erfüllt und umgekehrt. Selbstverständlich hat die geshiftete Folge die explizite Darstellung

$$\tilde{x}(k) = x(k+1) = 3 \cdot 2^{k+1} + 2 \cdot (-3)^{k+1} = 6 \cdot 2^k - 6 \cdot (-3)^k$$

Das gefundene Schieberegister für die Folge $x(k) = \hat{x}(k)$ ist minimal, weil ein kleineres mit einer einzigen Zelle eine geometrische Folge liefern müsste. In unserem Beispiel hat also die ursprünglich gegebene Liste die *lineare Komplexität*

2, weil das kleinste Schieberegister, dass das gegebene Anfangsstück produzieren kann, die Länge $m = 2$ hat. ◯

Beispiel. Soll im Gegensatz dazu das kleinste Schieberegister bestimmt werden, das das bereits angetroffene Anfangsstück der Länge $2m$

k	0	1	2	3	4	5
$x(k)$	4	-13	53	-115	449	-1123

produziert, so liefert der selbe Ansatz wie oben für eine Rekursion der Ordnung $\lceil \frac{2m}{2} \rceil = 3$ das lineare Gleichungssystem

$$\left\{\begin{array}{rrrr} 4\alpha - & 13\beta + & 53\gamma = & -115 \\ -13\alpha + & 53\beta - & 115\gamma = & 449 \\ 53\alpha - & 115\beta + & 449\gamma = & -1123 \end{array}\right. \qquad \left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ -13 & 53 & -115 & 449 \\ 53 & -115 & 449 & -1123 \end{array}\right)$$

mit der daneben stehenden erweiterte Matrix zu lösen. Addition des 13-fachen der ersten Zeile zum 4-fachen der zweiten und des 53-fachen der ersten zum (-4)-fachen der dritten Zeile liefert

$$\left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ 0 & 43 & 229 & 301 \\ 0 & -229 & 1013 & -1603 \end{array}\right)$$

Addition des 229-fachen der zweiten Zeile zum 43-fachen der dritten ergibt

$$\left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ 0 & 43 & 229 & 301 \\ 0 & 0 & 96000 & 0 \end{array}\right), \qquad \left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ 0 & 43 & 229 & 301 \\ 0 & 0 & 1 & 0 \end{array}\right)$$

in der wir die dritte Zeile durch $96'000$ dividiert haben, um mit möglichst kleinen Zahlen weiterrechnen zu können. Addition des (-229)-fachen der dritten Zeile zur zweiten und des (-53)-fachen der dritten Zeile zur ersten liefert

$$\left(\begin{array}{rrr|r} 4 & -13 & 0 & -115 \\ 0 & 43 & 0 & 301 \\ 0 & 0 & 1 & 0 \end{array}\right), \qquad \left(\begin{array}{rrr|r} 4 & -13 & 0 & -115 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & 0 \end{array}\right)$$

in der wir die zweite Zeile durch 43 dividiert haben. Addition des 13-fachen der zweiten Zeile zur ersten und Division der entstehenden ersten Zeile durch 4 ergibt

$$\left(\begin{array}{rrr|r} 4 & 0 & 0 & -24 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & 0 \end{array}\right), \qquad \left(\begin{array}{rrr|r} 1 & 0 & 0 & -6 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & 0 \end{array}\right)$$

die erweiterte Matrix in reduzierter Stufenform. Daraus lesen wir ab, dass die gesuchte Lösung des linearen Gleichungssystems im Gegensatz zum letzten Beispiel eindeutig bestimmt und von der Form $\alpha = -6$, $\beta = 7$ und $\gamma = 0$ ist. Daher kann dieses Anfangsstück nur durch die bereits im letzten Beispiel benutzte Rekursion

$$x(k+3) = -6 \cdot x(k) + 7 \cdot x(k+1)$$

dritter Ordnung erzeugt werden und hat damit die lineare Komplexität 3. ◯

Beispiel. Um das kleinste Schieberegister zu finden, das die Teilliste der ungeraden Länge $n = 5$

k	0	1	2	3	4
$x(k)$	4	-13	53	-115	449

produziert, machen wir auch diesmal für die Rekursionsgleichung der Ordnung $\lceil \frac{n}{2} \rceil = 3$ den selben Ansatz

$$x(k+3) = \alpha \cdot x(k) + \beta \cdot x(k+1) + \gamma \cdot x(k+2)$$

und haben die unbekannten Koeffizienten zu bestimmen. Einsetzen der bekannten Werte liefert für $k = 0, 1$ die Gleichungen des folgenden linearen Gleichungssystems

$$\left\{ \begin{array}{rrrr} 4\alpha - 13\beta + & 53\gamma = & -115 \\ -13\alpha + 53\beta - 115\gamma = & & 449 \end{array} \right. \qquad \left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ -13 & 53 & -115 & 449 \end{array} \right)$$

mit der zugehörigen erweiterten Matrix. Addition des 13-fachen der ersten Zeile zum 4-fachen der zweiten macht daraus

$$\left(\begin{array}{rrr|r} 4 & -13 & 53 & -115 \\ 0 & 43 & 229 & 301 \end{array} \right)$$

Addition des 13-fachen der zweiten Zeile zum 43-fachen der ersten ergibt die Stufenform

$$\left(\begin{array}{rrr|r} 172 & 0 & 5256 & -1032 \\ 0 & 43 & 229 & 301 \end{array} \right), \qquad \left(\begin{array}{rrr|r} 43 & 0 & 1314 & -258 \\ 0 & 43 & 229 & 301 \end{array} \right)$$

deren erste Zeile wir noch durch den grössten gemeinsamen Teiler 4 dividiert haben. Daraus lesen wir ab, die Lösung dieses Systems in vektorieller Form

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} -6 \\ 7 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1314 \\ -229 \\ 43 \end{pmatrix}$$

ab und erkennen, dass sie nicht eindeutig bestimmt ist. Der Wert $t = 0$ liefert $\alpha = -6$, $\beta = 7$ und $\gamma = 0$ und damit die im letzten Beispiel benutzte Fortsetzung mit der Rekursionsgleichung

$$x(k+3) = -6 \cdot x(k) + 7 \cdot x(k+1)$$

dritter Ordnung.

Soll das gegebenen Anfangsstück durch eine Rekursionsgleichung zweiter Ordnung beschrieben werden, müssen wir dafür sorgen, dass $\alpha = 0$ wird. Das erreichen wir durch die Wahl $t = -\frac{1}{219}$ bzw. mit den Koeffizienten $\beta = \frac{1762}{219}$ und $\gamma = -\frac{43}{219}$. Die zugehörige Rekursion zweiter Ordnung lautet diesmal

$$x(k+3) = \frac{1762}{219} \cdot x(k+1) - \frac{43}{219} \cdot x(k+2), \quad x(1) = -13, \quad x(2) = 53$$

bzw. nach der Substitution $\tilde{x}(k) = x(k+1)$ (Zeitshift)

$$\tilde{x}(k+2) = \frac{1762}{219} \cdot \tilde{x}(k) - \frac{43}{219} \cdot \tilde{x}(k+1), \qquad \tilde{x}(0) = -13, \quad \tilde{x}(1) = 53$$

Man beachte, dass diesmal die Koeffizienten der kürzeren Rekursionsformel nicht ganzzahlig, sondern rational sind und die verschobenen Anfangswerte gewählt werden müssen. Damit erhält man $\tilde{x}(4) = -\frac{73979}{73}$. ◯

Weil die Aufgabe, ein (minimales) Schieberegister zu konstruieren, das eine gegebenen Liste von Werten fortsetzt, eine grosse praktische Rolle spielt, sieht man sich nach einer Methode um, dieses Problem so effizient wie möglich zu behandeln. Dass diese Aufgabe nicht beliebige lineare Gleichungssysteme liefert, zeigt schon ein Blick auf die beiden im Beispiel angegebenen erweiterten Matrizen. Sie haben offenbar eine spezielle Gestalt, die von der Art der konkreten Aufgabe herrührt: aufeinanderfolgende Zeilen gehen aus einem Vektor durch einen Linksverschiebung auseinander hervor. Es ist zu erwarten, dass diese spezielle Form des Problems auch speziell effiziente Methoden ermöglicht. Tatsächlich wird der Informatiker die Daten eines solchen Gleichungssystems nicht als rechteckige erweiterte 2-dimensionale Matrix, sondern in der Form der ursprünglichen 1-dimensionalen Liste abspeichern, um Speicherplatz zu sparen. Eine genaue Analyse des Problems liefert in gewissen, für die Praxis relevanten Fällen einen Algorithmus, der die Lösung des zugehörigen linearen Gleichungssystems und die lineare Komplexität der Liste auf effizientere Art liefert, als der Eliminationsalgorithmus, der zur Lösung beliebiger lineare Gleichungssysteme konstruiert ist. Der Algorithmus von Berlekamp-Massey hat die Komplexität der Ordnung $O(n^2)$ im Vergleich zur Ordnung $O(n^3)$ des oben verwendeten Eliminationsalgorithmus. Er dient zur effzienten Dekodierung der sogn. BCH-Kodes.

Beispiel. Falls ein lineares, homogenes autonomes Differenzengleichungssystem mit der Rekursionsgleichung $\vec{y}(k+1) = A \cdot \vec{y}(k)$ die Vektoren

$$\vec{y}(3) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{y}(4) = \begin{pmatrix} -1 \\ 3 \\ -1 \end{pmatrix}, \quad \vec{y}(5) = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}, \quad \vec{y}(6) = \begin{pmatrix} -1 \\ 3 \\ -7 \end{pmatrix}$$

der Vektorfolge $k \mapsto \vec{y}(k)$ produziert, so geht es in den Anwendungen oft darum, die unbekannte Systemmatrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

zu bestimmen, die diese Datenvektoren produziert. Die gegebenen Vektoren erfüllen also definitionsgemäss die Rekursionsgleichung für $k = 3, 4, 5$. Diese 3 Vektorgleichungen liefern, komponentenweise ausgeschrieben, die $3^2 = 9$ linearen Gleichungen

$$\begin{cases} 1a_{11} + 0a_{12} + 1a_{13} = -1 \\ -1a_{11} + 3a_{12} - 1a_{13} = 1 \\ 1a_{11} + 0a_{12} - 5a_{13} = -1 \end{cases}$$

$$\begin{cases} 1a_{21} + 0a_{22} + 1a_{23} = 3 \\ -1a_{21} + 3a_{22} - 1a_{23} = 0 \\ 1a_{21} + 0a_{22} - 5a_{23} = 3 \end{cases} \qquad \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & -1 & 3 & -1 \\ -1 & 3 & -1 & 1 & 0 & -5 \\ 1 & 0 & -5 & -1 & 3 & -7 \end{array}\right)$$

$$\begin{cases} 1a_{31} + 0a_{32} + 1a_{33} = -1 \\ -1a_{31} + 3a_{32} - 1a_{33} = -5 \\ 1a_{31} + 0a_{32} - 5a_{33} = -7 \end{cases}$$

für die unbekannten Koeffizienten von A. Ein Blick auf dieses System zeigt, dass es nur schwach gekoppelt ist und man sich deshalb einiges an Rechenarbeit sparen kann — und soll! Tatsächlich lassen sich die 9 Gleichungen in 3 Teilsysteme von je 3 Gleichungen gruppieren, deren Koeffizientenmatrizen übereinstimmen und aus den ersten drei gegebenen Vektoren gebildet werden können. Die Teilsysteme unterscheiden sich also nur in ihren Konstantenvektoren, die jeweils aus den gegebenen Vektoren gebildet werden und sie können daher in Form der danebenstehenden Blockmatrix zusammengefasst werden.

Zur systematischen Beschreibung der Systeme können wir auch je 3 aufeinanderfolgende Vektoren der Vektorfolge zu einer 3×3 Matrix mit diesen drei Vektoren als Spalten zusammenfassen. Ferner definieren mit Hilfe aufeinanderfolgender Zustandsvektoren die assoziierte Matrizenfolge

$$M(k) = \Big(\vec{y}(k), \vec{y}(k+1), \vec{y}(k+2)\Big), \qquad k \in \mathbb{N}$$

Diese Matrizen erfüllen die Rekursion

$$M(k+1) = A \cdot M(k)$$

weil ihre Spalten definitionsgemäss der Rekursion $\vec{y}(k+1) = A \cdot \vec{y}(k)$ genügen. Daher gilt die explizite Beschreibung

$$M(k) = A^k \cdot M(0), \quad k \geq 0$$

In unserem Fall ist

$$M(3) = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 3 & 0 \\ 1 & -1 & -5 \end{pmatrix} \quad \text{und} \quad M(4) = \begin{pmatrix} -1 & 1 & -1 \\ 3 & 0 & 3 \\ -1 & -5 & -7 \end{pmatrix}$$

Die gemeinsame Koeffizientmatrix der drei Teilsysteme ist die Transponierte $M^T(3)$. Auf den rechten Seiten des j-ten Teilsystems kommen gerade die j-ten Spalten von $M^T(4)$ vor. Die Lösungen des j-ten Systems bilden die j-te Zeile der gesuchten Matrix A. Formal können wir die gesuchte Matrix aus der rekursiven Beziehung $M(4) = A \cdot M(3)$ durch Auflösen nach A in der Form

$$A = M(4) \cdot \Big(M(3)\Big)^{-1}$$

bestimmen, falls die Inverse von $M(3)$ existiert.

Um nicht von dieser Zusatzbedingung abhängig zu sein, lösen wir das entstandene lineare Gleichungssystem algorithmisch. Dazu wenden wir den Eliminationsalgorithmus auf die gefundene Blockmatrix

$$\big(M^T(3), M^T(4)\big) = \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & -1 & 3 & -1 \\ -1 & 3 & -1 & 1 & 0 & -5 \\ 1 & 0 & -5 & -1 & 3 & -7 \end{array}\right)$$

an. Im linken Block stehen also die ersten 3 gegebenen Vektoren und im rechten Block stehen die letzten 3 Vektoren, je zeilenweise von oben nach unten angeordnet. Im Spezialfall, wo die Matrix $M(3)$ invertierbar ist und im linken Block am Schluss also die Einheitsmatrix steht, finden wir im rechten Block die Transponierte von A.

Addition der ersten Zeile zur zweiten Zeile und des (-1)-fachen der ersten Zeile zur dritten Zeile liefert

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & -1 & 3 & -1 \\ 0 & 3 & 0 & 0 & 3 & -6 \\ 0 & 0 & -6 & 0 & 0 & -6 \end{array}\right) \quad \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & -1 & 3 & -1 \\ 0 & 1 & 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right)$$

im linken Block bereits eine obere Dreiecksmatrix, deren zweite Zeile wir noch durch 3 und deren dritte Zeile wir durch (-6) dividiert haben, um mit möglichst kleinen Zahlen weiterrechnen zu können. Addition des (-1)-fachen der dritten Zeile zur ersten Zeile

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 3 & -2 \\ 0 & 1 & 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right)$$

zeigt, dass in unserem Fall $M(3)$ invertierbar ist und liefert damit durch Transposition des rechten Blockes die gesuchte Erzeugermatrix

$$A = \left(\begin{array}{ccc} -1 & 0 & 0 \\ 3 & 1 & 0 \\ -2 & -2 & 1 \end{array}\right)$$

des Prozesses. Um ein zugehöriges Schieberegister zu konstruieren, benötigen wir das charakteristische Polynom $\chi_A(\lambda) = -1 + \lambda + \lambda^2 - \lambda^3 = -(\lambda - 1)^2(\lambda + 1)$. Um den Vektor $\vec{y}(2)$ zu berechnen, beachten wir, dass die Rekursionsgleichung das lineare Gleichungssystem $A \cdot \vec{y}(2) = \vec{y}(3)$ liefert, das durch die Blockmatrix

$$\left(\begin{array}{ccc|c} -1 & 0 & 0 & 1 \\ 3 & 1 & 0 & 0 \\ -2 & -2 & 1 & 1 \end{array}\right)$$

kodiert wird und durch nochmaliges Anwenden des Eliminationsalgorithmus gelöst wird. Addition des 3-fachen der ersten Zeile zur zweiten und des (-2)-fachen der ersten zur dritten liefert

$$\left(\begin{array}{ccc|c} -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & -2 & 1 & -1 \end{array}\right)$$

Als nächstes addieren wir das 2-fache der zweiten Zeile und erhalten die Matrix

$$\left(\begin{array}{ccc|c} -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 5 \end{array}\right) \quad \left(\begin{array}{ccc|c} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 5 \end{array}\right)$$

deren erste Zeile wir noch mit (-1) multipliziert haben. Den gesuchten Vektor

$$\vec{y}(2) = \left(\begin{array}{c} -1 \\ 3 \\ 5 \end{array}\right)$$

lesen wir im rechten Block ab. Natürlich ist das vorliegende System nur deshalb reversibel, weil seine Systemmatrix A invertierbar ist, wie wir im linken Block erkennen. Tatsächlich ist $\vec{y}(2) = A^{-1} \cdot \vec{y}(3)$. ◯

Wir wollen mit den folgenden Beispielen die entwickelten Methoden an einigen typischen Beispielen zusammenfassen und auf einige zusätzliche Aspekte hinweisen.

Beispiel. Der Italiener Leonardo de Pisa, bekannt unter dem Übernamen Fibonacci[19], untersuchte die Nachkommenschaft eines Kaninchenpaares, die ausserordentlich gross ist, wie sich vielleicht schon herumgesprochen hat. Zur qualitativen Untersuchung der Anzahl $f(k)$ Kaninchenpaare, die am Anfang der k-ten Zeiteinheit leben, machen wir die folgenden Modellannahmen:

1. Am Anfang der ersten Zeiteinheit lebt ein einziges, neugeborenes Kaninchenpaar ($\ominus$).

2. Jedes neugeborene Kaninchenpaar wird nach der Pubertät von einer Zeiteinheit geschlechtsreif ($\oplus$).

3. Jedes geschlechtsreife Paar bringt nach jeder Zeiteinheit ein weiteres Paar zur Welt.

4. Kaninchen leben ewig.

Unter diesen Annahmen wird das ursprünglich vorhandene Paar[20] am Anfang der zweiten Zeiteinheit gebärfähig und gebiert am Anfang der dritten Zeiteinheit ein weiteres Paar. Auch am Anfang der vierten Zeiteinheit bringt das ursprünglich vorhandene Paar ein neues Paar zur Welt, wie der folgende Stammbaum zeigt.

Aus dem Fibonacci-Baum entnehmen wir folgende numerische Information:

k	1	2	3	4	5	6	$\cdots$
$f(k)$	$1\ominus$	$1\oplus$	$1\ominus + 1\oplus$	$1\ominus + 2\oplus$	$2\ominus + 3\oplus$	$3\ominus + 5\oplus$	$\cdots$

Bezeichnen wir mit $y_1(k)$ die Anzahl juveniler ($\ominus$) und mit $y_2(k)$ die Anzahl erwachsener ($\oplus$) Paare am Anfang der k-ten Zeiteinheit, erhalten wir für die gesuchten Zahlen

$$f(k) = y_1(k) + y_2(k)$$

für kleine k die folgenden numerischen Werte:

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\cdots$
$y_1(k)$	1	0	1	1	2	3	5	8	13	21	34	55	89	144	$\cdots$
$y_2(k)$	0	1	1	2	3	5	8	13	21	34	55	89	144	233	$\cdots$
$f(k)$	1	1	2	3	5	8	13	21	34	55	89	144	233	377	$\cdots$

Diese Folge hat es, neben den Werbereien der Krämer, als Warnung an die explodierende Menschheit, sich nicht weiter wie Karnickel zu vermehren, bis in die Halle des Zürcher Hauptbahnhofs geschafft.

Zur Berechnung weiterer Werte dieser Folge benötigen wir eine Einsicht in die Struktur dieses Problems, die sich als Wachstumsgesetz in Form einer Rekursionsformel für die gesuchte Folge manifestieren wird.

[19] Verballhornung von filius Bonacci; 1170 – 1240.

[20] Wer will, kann die Männchen unterdrücken und sich auf die Weibchen beschränken.

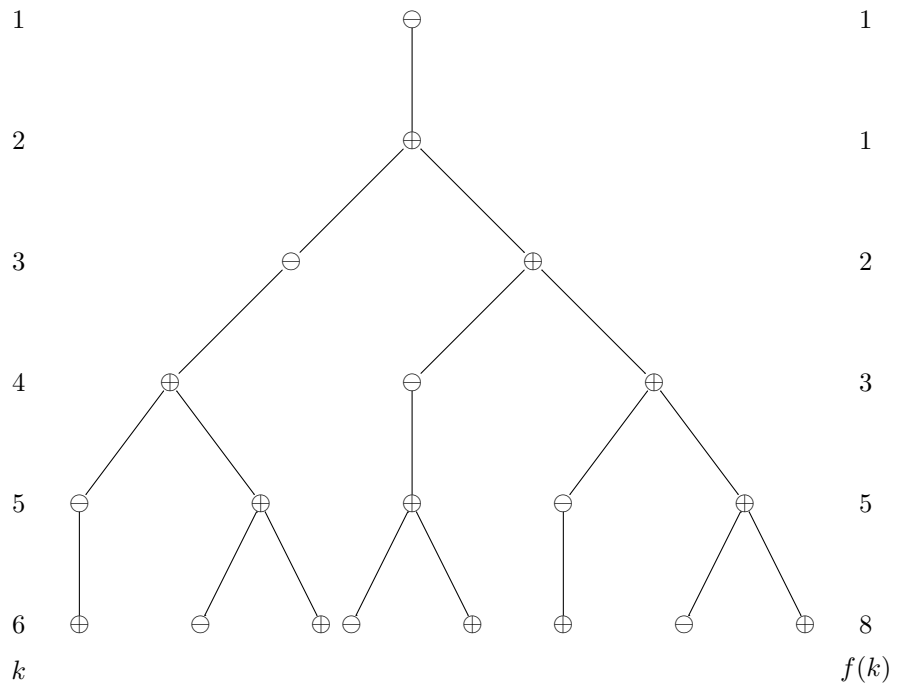


Abbildung 2.6: Stammbaum einer Kaninchen-Population.

Satz. Die Fibonacci-Zahlen erfüllen die berühmte lineare Rekursion

$$f(k+2) = f(k) + f(k+1), \quad f(0) = 0, f(1) = 1$$

Es handelt sich zweifelsfrei um die einfachste nicht triviale lineare, homogene Rekursionsgleichung mit konstanten Koeffizienten, die deshalb für didaktische Zwecke besonders gnstig ist.

Beweis. Ein Blick in den Kaninchenstall am Anfang der $(k+1)$-ten Zeiteinheit zeigt, dass dann definitionsgemäss genau $f(k+1)$ Kaninchenpaare vorhanden sind. Von diesen sind genau $f(k)$ geschlechtsreif, nämlich diejenigen, die schon am Anfang der k-ten Zeiteinheit gelebt haben. Genau sie werden also am Anfang der nächsten Zeiteinheit ein junges Paar zur Welt bringen. Die Anzahl $f(k+2)$ am Anfang der $(k+2)$-ten Zeiteinheit setzt sich aus der Anzahl $f(k)$ Paare zusammen, die am Anfang der $(k+2)$-ten Zeiteinheit geboren werden, zuzüglich zu den $f(k+1)$ überlebenden Paaren der vorherigen Zeiteinheit. Aus dieser Überlegung ergibt sich die behautptete Rekursionsformel, die als mathematisches Modell dieser Geschichte aufgefasst werden kann.

Man kann mit Hilfe der gegebenen Information auch gleich ein lineares Gleichungungssystem erster Ordnung für die beiden Folgen $y_1(k)$ und $y_2(k)$ angeben. Auf Grund der Modellannahmen gilt:

$$\left\{ \begin{array}{rcl} y_1(k+1) & = & y_2(k) \\ y_2(k+1) & = & y_1(k) + y_2(k) \end{array} \right.$$

Daraus ergibt sich für $f(k) = y_1(k) + y_2(k)$ durch Einsetzen

$$\begin{aligned} f(k+2) &= y_1(k+2) + y_2(k+2) = y_2(k+1) + y_1(k+1) + y_2(k+1) \\ &= f(k+1) + y_1(k) + y_2(k) = f(k+1) + f(k) \end{aligned}$$

die behauptete Rekursionsgleichung zweiter Ordnung. □

Dieses Wachstumsgesetz besagt, dass am Anfang einer Zeiteinheit gerade so viele Kaninchenpaare vorhanden sind, wie in den beiden vorangehenden Zeiteinheiten zusammen und ermöglicht es im Prinzip, beliebige Werte unserer Folge zu berechnen. Die gefundene Rekursion gilt auch für $k = 0$, falls man die Anfangsbedingung $f(0) = 0$ benutzt. Es ist also bequem anzunehmen, dass zur Zeit $k = 0$ keine Kaninchenpaare vorhanden waren.

Je nach beabsichigter Anwendung wird die Fibonacci-Folge gelegentlich durch andere Anfangsbedingungen festgelegt. Oft benutzt man etwa die Anfangsbedingung $x(0) = 1$ und $x(1) = 1$ bzw. $y(0) = 1$ und $y(1) = 2$ und erhält dann selbstverständlich die verschobene Fibbonacci- Folge $x(n) = f(n+1)$ bzw. $y(n) = f(n+2)$. Im Baum wird dabei einfach die Wurzel nach unten verschoben.

In der Kombinatorik interessiert man sich beispielsweise für die Anzahl Möglichkeiten, eine total geordnete n-elementige Menge in Blöcke der Grösse 1 oder 2 zu zerlegen. Beispielsweise erhält man für $n = 4$ die folgenden $x(4) = f(5) = 5$ Blockstrukturen

$$\bullet\,|\,\bullet\,|\,\bullet\,|\,\bullet, \qquad \bullet\bullet\,|\,\bullet\,|\,\bullet, \qquad \bullet\,|\,\bullet\,|\,\bullet\bullet, \qquad \bullet\,|\,\bullet\bullet\,|\,\bullet, \qquad \bullet\bullet\,|\,\bullet\bullet$$

Diese $x_n = f(n+1)$ Blockstrukturen entsprechen eindeutig der Anzahl Möglichkeiten, die natürliche Zahl n als Summe von 1 und 2 zu zerlegen, wobei wir verschieden Reihenfolgen der Summation berücksichtigen. Beispielsweise entsprechen für $n = 4$ obige Blockstrukturen den Summenzerlegungen

$$1+1+1+1, \qquad 2+1+1, \qquad 1+1+2, \qquad 1+2+1, \qquad 2+2$$

In der Kombinatorik spielen die Fibonacci-Zahlen auch deshalb eine Rolle, weil die Anzahl Teilmengen der k-elementigen Menge $\{1, 2, 3, \ldots, k\}$, die kein Paar benachbarter Zahlen enthalten, durch $y(k) = f(k+2)$ abgezählt werden können. Beispielsweise hat die 3-elementige Menge $\{1, 2, 3\}$ die folgenden $2^3 = 8$ Teilmengen:

$$\emptyset, \quad \{1\}, \quad \{2\}, \quad \{3\}, \quad \underline{\{1,2\}}, \quad \{1,3\}, \quad \underline{\{2,3\}}, \quad \underline{\{1,2,3\}}$$

Unter ihnen erfüllen die unterstrichenen 3 Teilmengen die Zusatzbedingung nicht, so dass also insgesamt $y(3) = f(5) = 5$ Teilmengen mit dieser Eigenschaft existieren. Von den zugehörigen $2^3 = 8$ Binärzahlen der Länge $k = 3$

$$000, \quad 100, \quad 010, \quad 001, \quad \underline{110}, \quad 101, \quad \underline{011}, \quad \underline{111}$$

sind die unterstrichenen 3 durch die Zusatzeigenschaft charakterisiert, dass in ihnen 2 aufeinanderfolgende 1 vorkommen. Daher liefert die verschobene Fibonacci-Zahl $y(3) = f(5) = 5$ die Anzahl Binärzahlen der Länge $k = 3$, in denen keine zwei aufeinanderfolgende 1 vorkommen[21].

[21] Allgemeiner kann man zeigen, dass die Anzahl Binärzahlen der Länge k, in denen keine t

Als weitere möglich Anwendung der Fibonacci-Folge erwähnen wir die Wörter der Sprache $\mathcal{F}$ über dem Alphabet $\{A, B\}$, die wir durch simultanes Anwenden folgender grammatikalischer Transformationsregeln erklären:

$$\begin{cases} A \rightsquigarrow B \\ B \rightsquigarrow AB \end{cases}$$

Beginnen wir nun auf das einbuchstabige Anfangswort $w_0 = A$ diese Substituionsregeln anzuwenden, generieren wir der Reihe nach die Wörter der Sprache $\mathcal{F} \subset \{\mathcal{A}, \mathcal{B}\}^*$:

k	w_k	k	w_k
1	A	5	$ABBAB$
2	B	6	$BABABBAB$
3	AB	7	$ABBABBABABBAB$
4	BAB	8	$BABABBABABBABBABABBAB$

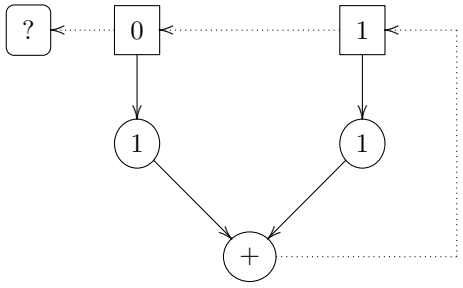
Offensichtlich besteht das Wort w_k aus $f(k)$ Buchstaben, nämlich $y_1(k)$ viele A's und $y_2(k)$ vielen B's. Man beachte, dass das Wort w_k im Fibonacci-Baum in der Zeile k auftritt, falls man den Buchstaben A durch das Symbol $\ominus$ und den Buchstaben B durch das Symbol $\oplus$ ersetzt.

Damit ist die Modellierung bzw. die Motivation für die zu betrachtete Folge abgeschlossen und wir können uns nun mit Mathematik befassen.

Man beachte, dass die gefundene Rekursionsgleichung linear homogen mit konstanten Koeffizienten ist und die zugehörige charakteristische Gleichung

$$\lambda^2 = \lambda + 1, \qquad 1 + \lambda - \lambda^2 = 0$$

zweiter Ordnung lautet. Daher kann die Fibonacci-Folge durch das simple Schieberegister



produziert werden. Obwohl in diesem Beispiel die Rekursionsgleichung bloss eine einzige Addition enthält, dürfte es nicht ohne weiteres möglich sein, für die Zahlenfolge $f(k)$ eine explizite Formel zu erraten, die für theoretische Zwecke nützlich wäre. Dazu benötigen wir die Eigenwerte. Quadratische Ergänzung liefert zur charakteristischen Gleichung die Scheitelform

$$\left(\lambda - \frac{1}{2}\right)^2 = \frac{1}{4} + 1 = \frac{5}{4}$$

aufeinanderfolgende 1 vorkommen, durch die Fibonacci-Folge t-ter Ordnung $f^{[t]}(n)$ gegeben ist, die durch die Rekursion

$$f^{[t]}(n+t) = f^{[t]}(n+t-1) + f^{[t]}(n+t-2) + \cdots + f^{[t]}(n)$$

der Ordnung t gegeben ist. Deshalb beträgt die Wahrscheinlichkeit, dass eine 100-stellige Binärzahl keine 6 aufeinanderfolgende 1 (oder aus Symmetriegründen 0) hat, etwa 45% und für eine 200-stellige Binärzahl beträgt sie etwa 3%.

aus der die beiden gesuchten irrationalen Eigenwerte

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \approx 1.618\ldots, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2} \approx -0.618\ldots$$

sofort abgelesen werden können. Für sie gilt nach dem Ausmultiplizieren des faktorisierten charakteristischen Polynoms

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2) + \lambda_1\lambda_2$$

durch Koeffizientenvergleich die Beziehungen $\lambda_1 + \lambda_2 = 1$ und $\lambda_1 \cdot \lambda_2 = -1$.
Mit diesen Eigenwerten erhalten wir sofort die beiden speziellen Basislösungen

$$f_1(k) = \lambda_1^k, \qquad f_2(k) = \lambda_2^k$$

unserer Differenzengleichung, wie man durch Einsetzten leicht bestätigt. Es sind die beiden einzigen nicht trivialen Lösungen, die als Exponentialfunktion beschrieben werden können, wie man mit Hilfe des Exponentialansatzes $f(k) = \lambda^k$ leicht bestätigt. Die beiden Basislösungen gehören zu den Anfangsbedingungen

$$\begin{aligned} f_1(0) &= 1, \quad f_1(1) = \lambda_1 \\ f_2(0) &= 1, \quad f_2(1) = \lambda_2 \end{aligned}$$

die noch nicht mit jenen der Fibonacci-Folge übereinstimmen. Um aus diesen beiden Basislösungen die gesuchte Folge linear kombinieren zu können, machen wir wegen der Linearität der Rekursionsgleichung für ihre allgemeine Lösung den Ansatz

$$f(k) = c_1 f_1(k) + c_2 f_2(k) = c_1\lambda_1^k + c_2\lambda_2^k$$

und müssen nun die Koeffizienten c_1, c_2 so bestimmen, dass die gewünschten Anfangsbedingungen $f(0) = 0$ und $f(1) = 1$ erfüllt sind. Der Ansatz liefert für $k = 0$ und $k = 1$ das lineare Gleichungssystem

$$\begin{cases} k=0: & c_1 + \quad c_2 = 0 \\ k=1: & \lambda_1 c_1 + \lambda_2 c_2 = 1 \end{cases}$$

mit der eindeutigen Lösung

$$c_1 = \frac{1}{\lambda_1 - \lambda_2} = \frac{1}{\sqrt{5}} = \frac{\sqrt{5}}{5}, \qquad c_2 = \frac{1}{\lambda_2 - \lambda_1} = -\frac{1}{\sqrt{5}} = -\frac{\sqrt{5}}{5}$$

Daher hat die Fibonacci-Folge die explizite Beschreibung

$$f(k) = \frac{\sqrt{5}}{5} \cdot \lambda_1^k - \frac{\sqrt{5}}{5} \cdot \lambda_2^k = \frac{\sqrt{5}}{5} \cdot \Big(\frac{1+\sqrt{5}}{2}\Big)^k - \frac{\sqrt{5}}{5} \cdot \Big(\frac{1-\sqrt{5}}{2}\Big)^k$$

wie man nun nachträglich durch Einsetzen in die Rekursionsgleichung verifiziert. Das Erstaunliche an dieser expliziten Formel für die Fibonacci-Folge ist, dass sich für jede natürliche Zahl k die irrationalen Terme gegenseitig so aufheben, dass am Schluss ein ganzzahliger Wert entsteht!

Um nun mit dieser Formel das Wachstum der Kaninchenpopulation numerisch zu untersuchen, berechnen wir einige konkrete Werte. In der folgenden Tabelle listen wir die beiden Summanden

$$e(k) = \frac{\sqrt{5}}{5} \cdot \lambda_1^k = \frac{\sqrt{5}}{5} \cdot \Big(\frac{1+\sqrt{5}}{2}\Big)^k, \qquad z(k) = -\frac{\sqrt{5}}{5} \cdot \lambda_2^k = -\frac{\sqrt{5}}{5} \cdot \Big(\frac{1-\sqrt{5}}{2}\Big)^k$$

der gefundenen Formel separat auf und erhalten die numerischen Werte:

k	$e(k)$	$z(k)$	$f(k) = e(k) + z(k)$
0	$0.447\ldots$	$-0.447\ldots$	0
1	$0.723\ldots$	$0.276\ldots$	1
2	$1.170\ldots$	$-0.170\ldots$	1
3	$1.894\ldots$	$0.105\ldots$	2
4	$3.065\ldots$	$-0.065\ldots$	3
5	$4.959\ldots$	$0.040\ldots$	5
6	$8.024\ldots$	$-0.024\ldots$	8
7	$12.984\ldots$	$0.015\ldots$	13
8	$21.009\ldots$	$-0.009\ldots$	21
9	$33.994\ldots$	$0.005\ldots$	34

Offenbar liegt der erste Summand $e(k)$ in der Nähe der Zahl $f(k)$ und der zweite Summand $z(k)$ korrigiert die Abweichung. Er oszilliert um 0 und wird betragsmässig sehr rasch klein. Daher gilt also die Näherungsformel

$$f(k) \sim e(k), \qquad \lim_{k\to\infty} z(k) = 0$$

Das Wachstumsverhalten wird durch den ersten Summanden bestimmt und die Kaninchenpopulation wächst exponentiell. Das typische explosive Verhalten der Population entnimmt man dem Graphen der Fibonacci-Folge.

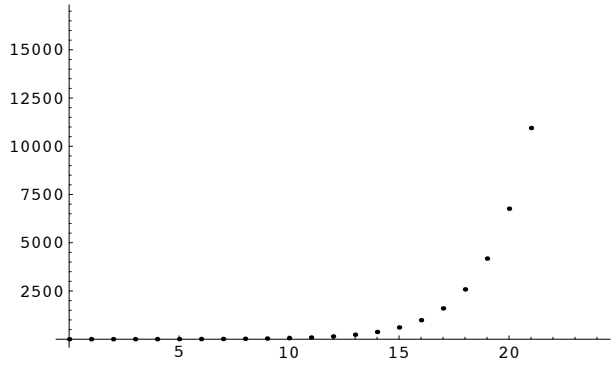


Abbildung 2.7: Graph der Fibonacci-Folge $f(k)$.

Nach zwei Jahren sind also $f(24) = 46'368$ Kaninchenpaare vorhanden.

Um die Qualität der Approximation beurteilen zu können, vergleicht man den exakten ganzzahligen Wert $f(100) = 354'224'848'179'261'915'075$ mit der Näherung

$$e(100) = 3.54224848179261915075 \cdot 10^{20}$$

Der Fehler ist von der Grössenordnung $z(100) = 5.6 \cdot 10^{-22}$ der für die meisten praktischen Zwecke vernachlässigt werden kann.

Auf Grund dieser Näherung gilt sogar

$$f(k) = \left\lfloor e(k) + \frac{1}{2} \right\rfloor$$

Daher lässt sich der ganzzahlige Wert $f(k)$ exakt berechnen, falls ein genügend genauer Wert der irrationalen Zahl $\sqrt{5}$ zur Verfügung steht.
Betrachten wir die Fibonacci-Quotienten

$$p(k) = \frac{f(k)}{f(k+1)}$$

Diese Zahl kann wegen $p(k) = \frac{y_2(k)}{f(k)}$ im ursprünglichen Modell als Anteil (Wahrscheinlichkeit) der geschlechtsreifen Paare m k-ten Monat interpretiert werden. Entsprechend kann $1 - p(k) = \frac{y_1(k)}{f(k)}$ als Anteil der juvenilen Paare interpretiert werden. Die numerischen Werte

k	$f(k)$	$f(k+1)$	$p(k)$	$1-p(k)$	$r(k)$
0	0	1	0	1	–
1	1	1	1	0	1
2	1	2	$\frac{1}{2} = 0.5$	$\frac{1}{2} = 0.5$	2
3	2	3	$\frac{2}{3} = 0.666$	$\frac{1}{3} = 0.333$	$\frac{3}{2} = 1.5$
4	3	5	$\frac{3}{5} = 0.6$	$\frac{2}{5} = 0.4$	$\frac{5}{3} = 1.666$
5	5	8	$\frac{5}{8} = 0.625$	$\frac{3}{8} = 0.375$	$\frac{8}{5} = 1.6$
6	8	13	$\frac{8}{13} = 0.615$	$\frac{5}{13} = 0.385$	$\frac{13}{8} = 1.625$
7	13	21	$\frac{13}{21} = 0.619$	$\frac{8}{21} = 0.381$	$\frac{21}{13} = 1.615$
8	21	34	$\frac{21}{34} = 0.617$	$\frac{13}{34} = 0.383$	$\frac{34}{21} = 1.619$
9	34	55	$\frac{34}{55} = 0.618$	$\frac{21}{55} = 0.382$	$\frac{55}{34} = 1.617$

suggerieren, dass diese Anteile am Anfang stark schwanken und sich dann schnell einem Grenzwert nähern. Dieses Verhalten erkennt man besonders anschaulich am Histogramm von $p(k)$.

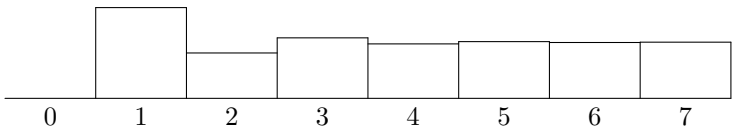


Abbildung 2.8: Histogramm des Anteils geschlechtsreifen Paare.

Um diese Vermutung zu bestätigen, brauchen wir zunächst einen Kandidaten für den gesuchten Grenzwert. Dazu beschreiben wir die Folge der Anteile mit Hilfe einer Rekursion. Definitionsgemäss ist

$$p(k+1) = \frac{f(k+1)}{f(k+2)} = \frac{f(k+1)}{f(k+1)+f(k)} = \frac{1}{1+\frac{f(k)}{f(k+1)}} = \frac{1}{1+p(k)}, \quad k \geq 0$$

Ein Blick auf diese Rekursion zeigt, dass sie nicht linear ist.
Falls die Folge $p(k)$ tatsächlich gegen einen Grenzwert x konvergiert, so muss dieser Grenzwert die Fixpunktgleichung

$$x = \frac{1}{1+x}$$

erfüllen, die nach dem Umstellen zur quadratischen Gleichung

$$x^2 + x = 1, \qquad \text{bzw.} \qquad \left(x + \frac{1}{2}\right)^2 = 1 + \frac{1}{4} = \frac{5}{4}$$

wird. Sie hat die beiden reellen Lösungen

$$x_1 = \frac{-1+\sqrt{5}}{2} = -\lambda_2 \approx 0.618\ldots \qquad x_2 = \frac{-1-\sqrt{5}}{2} = -\lambda_1 \approx -1.618$$

Es gilt $x_1 + x_2 = x_1 \cdot x_2 = -1$.
Die zweite Lösung ist für unsere Zwecke auf Grund des Vorzeichens unbrauchbar. Daher kommt also für den gesuchten Grenzwert höchstens

$$x_1 = \tau = \frac{-1+\sqrt{5}}{2} \approx 0.618\cdots$$

in Frage. Diese Zahl erfüllt definitionsgemäss die Fixpunktgleichung

$$\tau = \frac{1}{1+\tau}, \qquad \tau^2 + \tau - 1 = 0$$

Auch die oben gefundene asymptotische Beschreibung der Fibonacci-Zahlen legt diesen Grenzwert

$$p(k) = \frac{f(k)}{f(k+1)} \sim \frac{e(k)}{e(k+1)} = \frac{\lambda_1^k}{\lambda_1^{k+1}} = \frac{1}{\lambda_1} = -\lambda_2 = \tau$$

nahe. Um nun zu zeigen, dass die Folge $p(k)$ tatsächlich diesem Grenzwert τ zustrebt, berechnen wir zunächst die Abweichung $|\tau - p(k)|$. Auf Grund der Rekursion von $p(k)$ und der Fixpunkteigenschaft von τ ist zunächst

$$\tau - p(k) = \tau - \frac{1}{1+p(k-1)} = \frac{1}{1+\tau} - \frac{1}{1+p(k-1)} = \frac{p(k-1)-\tau}{(1+\tau)\cdot(1+p(k-1))}$$

Da $0 < \tau$ und $0 \le p(k-1)$ für $k \ge 1$ ist, gilt $1 < 1+\tau$ und $1 \le 1+p(k-1)$. Daher sind die beiden Faktoren des Nenners strikt positiv und wir erhalten für die Abweichung die Abschätzung ($k \ge 1$)

$$|\tau - p(k)| = \frac{|p(k-1)-\tau|}{(1+\tau)\cdot(1+p(k-1))} \le \frac{|p(k-1)-\tau|}{1+\tau} = \frac{|\tau - p(k-1)|}{1+\tau}$$

Wenden wir diese Abschätzung rekursiv an, erhalten wir die gesuchte Abschätzung

$$|\tau - p(k)| \le \frac{|\tau - p(k-1)|}{1+\tau} \le \frac{|\tau - p(k-2)|}{(1+\tau)^2} \le \cdots \le \frac{|\tau - p(2)|}{(1+\tau)^{k-2}} \le \frac{|\tau - p(1)|}{(1+\tau)^{k-1}}$$

Da im Nenner $(1+\tau) > 1$ ist, wird diese Abweichung mit wachsendem k beliebig klein und der Wert $p(k)$ kommt der Zahl τ tatsächlich beliebig nahe. Diese Tatsache beschreiben wir durch den Grenzwert

$$\lim_{k\to\infty} p(k) = \tau$$

Insbesondere konvergiert also die Folge $p(k)$ gegen den *goldenen Schnitt* τ.

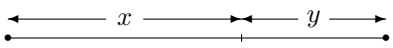


Abbildung 2.9: Geometrische Definition des goldenen Schnittes.

Diese irrationale Zahl τ taucht in der Mathematik und in den Anwendungen in diversen Fragestellungen oft und unerwartet auf. Sie spielt aus ästhetischen Gründen eine Rolle in der Architektur und in der Kunst und in der Biologie taucht sie beim Wachstumsverhalten vieler Populationen auf. Geometrisch sagt man, eine Strecke werde im goldenen Schnitt geteilt, wenn sich die ganze Strecke zum grösseren Abschnitt (Major) x gleich verhält, wie der grössere Abschnitt x zum kleineren y (Minor).

Wählen wir die Länge der Strecke als Einheit, so gilt definitionsgemäss die Proportion $1 : x :: x : y$ oder in moderner Sprache mit Brüchen geschrieben

$$\frac{1}{x} = \frac{x}{y}, \qquad x + y = 1$$

Daraus ergibt sich aber sofort durch Einsetzen und Umformen

$$\frac{1}{x} = \frac{x}{1-x}, \qquad 1 - x = x^2$$

dass x die charakteristische quadratische Gleichung des goldenen Schnittes erfüllt. Weil der goldene Schnitt als Grenzwert aufeinanderfolgender Glieder der Fibonacci-Folge entsteht, hat er die regelmässige Kettenbruchdarstellung

$$\tau = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}} = [0; 1, 1, 1, 1, 1, \ldots]$$

aus der folgt, dass es sich um die irrationalste aller reellen Zahlen handelt, die sich nur sehr langsam durch die Quotienten aufeinanderfolgender Fibonacci-Zahlen rational approximiert werden kann. Auf Grund dieser Kettenbruchentwicklung vermutet man, dass der goldene Schnitt eine interessante Selbstähnlichkeit hat. Um sie geometrisch leicht zu erkennen, geht man von einem goldenen Rechteck aus, dessen Seitenverhältnis $y : x$ der goldene Schnitt ist.

Um mit Hilfe von Zirkel und Lineal zu einer Strecke y eine Strecke $x = \overline{AX}$ zu konstruieren, so dass y Minor und x Mayor sind, konstruiert man das Quadrat AY über der Strecke y und beachtet, dass für den Mittelpunkt M von $\overline{AB}$ die Bedingung $d(M, X) = d(M, Y)$ gilt.

Schneidet man umgekehrt einem solchen goldenen Rechteck das grösstmögliche enthaltene Quadrat AY ab, bleibt ein Restrechteck XY übrig, dessen Seitenverhältnis wiederum der goldene Schnitt ist. Durch Wiederholen dieser Abschneideprozedur findet man eine unendliche Folge ineinandergeschachtelter goldener Rechtecke. Mit Hilfe dieser Selbstähnlichkeit kann man Quasikristalle und chaotische Systeme konstruieren, die auf dem goldenen Schnitt beruhen.

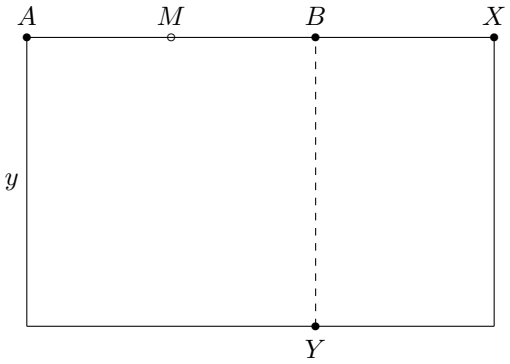


Abbildung 2.10: Geometrische Konstruktion des goldenen Schnittes.

Wir erwarten nun, dass die Folge der reziproken Fibonacci-Quotienten

$$r(k) = \frac{f(k+1)}{f(k)}, \quad k \geq 1$$

die der Rekursion

$$r(k+1) = \frac{f(k+2)}{f(k+1)} = \frac{f(k+1)+f(k)}{f(k+1)} = 1 + \frac{f(k)}{f(k+1)} = 1 + \frac{1}{r(k)}, \quad k \geq 1$$

genügt, gegen den reziproken Grenzwert

$$\lim_{k\to\infty} r(k) = \frac{1}{\tau} = \frac{2}{-1+\sqrt{5}} = \frac{1+\sqrt{5}}{2} = \lambda_1$$

d.h. gegen den Eigenwert λ_1 konvergiert, der die Fixpunkteigenschaft erfüllt.

$$\lambda_1 = 1 + \frac{1}{\lambda_1}$$

Man erhält sie durch Divison der charakteristischen Gleichung $\lambda^2 = \lambda + 1$ durch λ. Um auch diese Vermutung zu bestätigen, berechnen wir diesmal die Abweichung $|\lambda_1 - r(k)|$. Auf Grund der Rekursion von $r(k)$ und der Fixpunkteigenschaft von λ_1 gilt zunächst für den Unterschied

$$\lambda_1 - r(k) = 1 + \frac{1}{\lambda_1} - \left(1 + \frac{1}{r(k-1)}\right) = \frac{1}{\lambda_1} - \frac{1}{r(k-1)} = \frac{r(k-1)-\lambda_1}{\lambda_1 \cdot r(k-1)}$$

Da $0 < \lambda_1$ und $1 \leq r(k-1)$ für $k \geq 2$ ist, erhalten für die Abweichung die Abschätzung

$$|\lambda_1 - r(k)| = \frac{|r(k-1)-\lambda_1|}{\lambda_1 \cdot r(k-1)} \leq \frac{|r(k-1)-\lambda_1|}{\lambda_1}, \quad k \geq 2$$

Wenden wir diese Abschätzung rekursiv an, erhalten wir die gesuchte Abschätzung

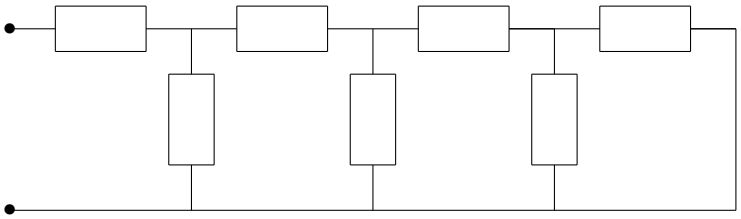
$$|\lambda_1 - r(k)| \leq \frac{|\lambda_1 - r(k-1)|}{\lambda_1} \leq \frac{|\lambda_1 - r(k-2)|}{\lambda_1^2} \leq \cdots \leq \frac{|\lambda_1 - r(2)|}{\lambda_1^{k-2}}$$

Da im Nenner $\lambda_1 > 1$ ist, wird diese Abweichung mit wachsendem k beliebig klein und der Wert $r(k)$ kommt der Zahl λ_1 tatsächlich beliebig nahe. Diese Tatsache zeigt den Grenzwert

$$\lim_{k\to\infty} r(k) = \lambda_1$$

Diesen Sachverhalt hat zwar schon Kepler vermutet. Er konnte aber erst etwa 100 Jahre später exakt begründet werden, als die besprochenen Zusammenhänge klar wurden.

Dieser Grenzwert und seine approximierende Folge lässt sich durch das elektrische Netzwerk



physikalisch interpretieren, in dem alles Einheitswiderstände sind. Bauen wir an dieses Netzwerk, wie bei einer Leiter, weitere Sprossen an, indem wir ganz rechts abwechslungsweise einen zusätzlichen vertikalen und dann einen weiteren horizontalen Widerstand einbauen, erhalten wir eine Folge von Netzwerken der selben Art mit dem Gesamtwiderstand R_k. Natürlich gilt $R_1 = 1$ und die Rekursionsgleichung

$$R_{k+1} = 1 + \frac{1}{R_k}$$

Daher hat das zugehörige „unendliche“ Netzwerk den Widerstand $R = \lambda_1$.

Die effiziente exakte numerische Berechnung der Werte f_k hat es in sich. Selbstverständlich kann man dazu die Rekursionsformel heranziehen und damit im Prinzip ein einfaches rekursives Programm schreiben. Experimentiert man damit, so stellt man fest, dass dieser Algorithmus schon für mässig grosse k sehr langsam ist und der Aufwand zur Berechnung immer grösser wird. Um die Laufzeit T_k dieses rekursiven Programms abzuschätzen, wählen wir die Zeiteinheit so, dass $T_1 = 1$ gilt. Auf Grund der Rekursionsformel gilt für die Laufzeit

$$T_{k+2} \geq T_k + T_{k+1}$$

Das ist analog zur Rekursionsformel für die Fibonacci-Zahlen, nur ist hier das Gleichheitszeichen durch ein Ungleichheitszeichen ersetzt. Die Komplexität des rekursiven Algorithmus zur Berechnung der Fibanocci-Folge wächst also mindestens so stark wie die Fibonacci-Folge — also exponentiell! Der Algorithmus zur rekursiven Berechnung der Fibonacci-Zahlen, der in vielen Programmier-Lehrbüchern als Musterbeispiel für rekursives Programmieren dargestellt wir, beweist seine eigene Unbrauchbarkeit!

Um abzuschätzen, wie lange eine solche Berechnung dauert, nehmen wir an, wir hätten einen Supercomputer, für den die oben erwähnte Zeiteinheit eine Nano-Sekunde, d.h. 10^{-9} Sekunden ist. Dann ist mit dem früher bestimmten Wert $T_{100} \geq 3.54 \cdot 10^{20} \cdot 10^{-9} = 3.54 \cdot 10^{11}$ Sekunden. Da ein Jahr etwa $3.15 \cdot 10^7$ Sekunden hat, kommt man auf einen Zeitbedarf von über $10'000$ Jahren.

Das Problem mit dem rekursiven Algorithmus liegt offenbar darin, dass die beiden Funktionsaufrufe unabhängig voneinander durchgeführt werden und der zweite nicht von den Zwischenergebnissen des ersten profitiert, so dass viele Berechnungen mehrfach durchgeführt werden. Es ist zwar in diesem Beispiel leicht möglich, den rekursiven in einen iterativen Algorithmus umzuschreiben, der dann eine befriedigendere Laufzeit hat. Einen noch viel effizienteren Algorithmus findet man allerdings, wenn man das besprochende Verdoppelungsverfahren verwendet.

Dazu bestimmen wir das zur Fibonacci-Folge gehörige System von linearen Differenzengleichungen. Wir definieren also mit den beiden Zustandsvariablen $y_1(k) = f(k)$ und $y_2(k) = f(k+1)$ in bekannter Manier den Zustandsvektor

$$\vec{y}(k) = \begin{pmatrix} f(k) \\ f(k+1) \end{pmatrix}$$

Die Rekursionsgleichung nimmt damit die Form

$$\begin{cases} y_1(k+1) = & y_2(k) \\ y_2(k+1) = y_1(k) + & y_2(k) \end{cases}$$

an und kann mit Hilfe der Systemmatrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{R}^{2,2}$$

in matrizieller Form $\vec{y}(k+1) = A \cdot \vec{y}(k)$ beschrieben werden. Die Matrix A kann als Begleitermatrix des charakteristischen Polynoms

$$\chi(\lambda) = 1 + \lambda - \lambda^2$$

aufgefasst werden. Es hat tatsächlich die Matrix A als Nullstelle. Der Zustand $\vec{y}(k)$ kann mit Hilfe der Matrizenpotenz A^k aus dem Anfangszustand

$$\vec{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \mathbb{R}^2$$

direkt in der Form $\vec{y}(k) = A^k \cdot \vec{a}$ berechnet werden. Der gesuchte Algorithmus zur Berechnung der Fibonacci-Folge läuft also darauf hinaus, Matrizenpotenzen zu berechnen, was mit dem Verdoppelungsverfahren effizient möglich ist.

Zur Berechnung der Potenz A^{100} benötigen wir die Binärdarstellung des Exponenten. Es ist

$$100 = 64 + 32 + 4 = 2^6 + 2^5 + 2^2 = (1100100)_2$$

Nun berechnen wir die sukzessiven Quadrate A^{2^j} für $1 \leq j \leq \lfloor \log_2(100) \rfloor = 6$:

$$\begin{aligned} A^2 &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \\ A^4 = A^{2^2} = \left(A^2\right)^2 &= \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix} \\ A^8 = A^{2^3} = \left(A^{2^2}\right)^2 &= \begin{pmatrix} 13 & 21 \\ 21 & 34 \end{pmatrix} \end{aligned}$$

$$A^{16} = A^{2^4} = \left(A^{2^3}\right)^2 = \begin{pmatrix} 610 & 987 \\ 987 & 1597 \end{pmatrix}$$

$$A^{32} = A^{2^5} = \left(A^{2^4}\right)^2 = \begin{pmatrix} 1'346'269 & 2'178'309 \\ 2'178'309 & 3'524'578 \end{pmatrix}$$

$$A^{64} = A^{2^6} = \left(A^{2^5}\right)^2 = \begin{pmatrix} 6'557'470'319'842 & 10'610'209'857'723 \\ 10'610'209'857'723 & 17'167'680'177'565 \end{pmatrix}$$

Daraus erhalten wir für die gesuchte Matrix $A^{100} = A^{2^2} \cdot A^{2^5} \cdot A^{2^6}$ den Wert

$$\begin{pmatrix} 218'922'995'834'555'169'026 & 354'224'848'179'261'915'075 \\ 354'224'848'179'261'915'075 & 573'147'844'013'817'084'101 \end{pmatrix}$$

Eine Rechnung, die auf einem PC keine messbare Zeit erfordert. Weil die zu multiplizierenden und addierenden Zahlen immer grösser werden, wächst die Komplexität[22] schneller als $O\big(\log_2(k)\big)$. Bei der Verwendung endlicher Körper in der Informatik fallen die Kosten der Grundoperationen allerdings kaum ins Gewicht.

Zur Berechnung von $f(100)$ berechnen wir den Zustand $\vec{y}(100) = A^{100} \cdot \vec{a}$ und erhalten wegen der Anfangsbedingung $\vec{a} = \vec{e}_2$ den zweiten Spaltenvektor

$$\vec{y}(100) = \begin{pmatrix} 354'224'848'179'261'915'075 \\ 573'147'844'013'817'084'101 \end{pmatrix}$$

Daraus lesen wir als erste Komponente den bereits früher angegebenen Wert $f(100) = 354'224'848'179'261'915'075$ ab. Man beachte, dass wegen der speziellen Form des Anfangszustandes und der Symmetrie der Systemmmatrix der Wert $f(100)$ in der Nebendiagonalen von A^{100} steht. Mit diesem Algorithmus erfordert auch die numerische Berechnung des exakten Wertes von

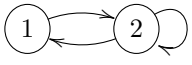
$$\begin{aligned} f(1000) \;=\; & 43'466'557'686'937'456'435'688'527'675'040'625'802'564'660'517' \\ & 371'780'402'481'729'089'536'555'417'949'051'890'403'879'840'079' \\ & 255'169'295'922'593'080'322'634'775'209'689'623'239'873'322'471' \\ & 161'642'996'440'906'533'187'938'298'969'649'928'516'003'704'476' \\ & 137'795'166'849'228'875 \approx 4.3466 \cdot 10^{208} \end{aligned}$$

kaum messbare Zeit, obwohl der Autor diese Zahl lieber nicht an die Wandtafel schreibt und schon gar keine Ahnung habt, wie sie heisst, weil Römer und wohl auch Kaninchenzüchter selten mit solch riesigen Zahlen handieren.

Interpretieren wir die Systemmatrix der Fibonacci-Folge

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

als Adjazenzmatrix des Fibonacci-Graphen



[22] Die Addition k-stelliger Dezimalzahlen aus der Schule hat die Komplexität $O(k)$. Der gewöhnliche Multiplikations-Algorithmus für k-stellige Dezimalzahlen hat die Komplexität $O\big(k^2\big)$. Schönhage und Strassen zeigten mit Hilfe von FFT, dass die Multiplikation von k-Bit-Zahlen eine asymptotische Komplexität $O\Big(k \log_2(k) \log_2(log_2(k))\Big)$ hat.

d.h. beschreibt a_{ji} die Anzahl Kanten vom Knoten i zum Knoten j, so beschreiben die Potenzen A^k die Anzahl Weg der Länge k in diesem Graphen. Aus der Matrix

$$A^5 = \begin{pmatrix} 3 & 5 \\ \mathbf{5} & 8 \end{pmatrix}$$

lesen wir ab, dass es insgesamt 5 Wege der Länge 5 vom Knoten 1 zum Knoten 2 geben sollte. Sie werden explizit durch die Listen

$$\begin{array}{ll} 1 \to 2 \to 2 \to 2 \to 2 \to 2, & 1 \to 2 \to 2 \to 2 \to 1 \to 2, \\ 1 \to 2 \to 2 \to 1 \to 2 \to 2, & 1 \to 2 \to 1 \to 2 \to 2 \to 2 \\ 1 \to 2 \to 1 \to 2 \to 1 \to 2. & \end{array}$$

beschrieben. Entsprechend sind die 3 Schleifen der Länge 5 des Knotens 1 gegeben durch

$$\begin{array}{ll} 1 \to 2 \to 2 \to 2 \to 2 \to 1, & 1 \to 2 \to 2 \to 1 \to 2 \to 1, \\ 1 \to 2 \to 1 \to 2 \to 2 \to 1. & \end{array}$$

Diese Wege vom Anfanszustand 1 aus im Fibonacci-Graphen lassen sich leicht als Wege im zugehörigen Fibonacci-Baum verfolgen, in dem die eindeutig bestimmten Wege von der Wurzel zu einem der Knoten bijektiv den Wegen im Graphen entsprechen. In der folgenden Figur lassen sich also sämtliche Wege der Länge höchstens 5 verfolgen.

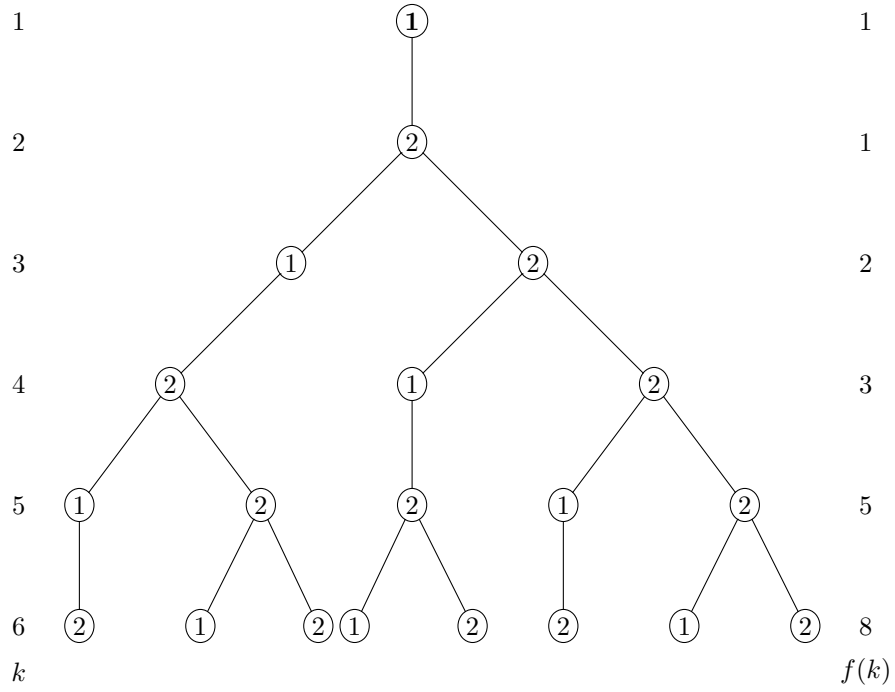


Abbildung 2.11: Die Wege vom Anfangszustand **1** im Fibonacci-Automaten.

Dieser Baum ist uns, mit einer etwas anderen Bezeichnung der Knoten, d.h. $\ominus$ statt 1 und $\oplus$ statt 2 als Stammbaum der Kaninchen-Population bereits bekannt.

Will man zur effizienten Berechnung von $f(k)$ den Weg über Matrizen vermeiden, und direkt mit den Fibonacci-Zahlen arbeiten, muss man den mathematischen Gehalt des Verdoppelungsverfahrens auf diese Skalare übertragen und benötigt eine Verdoppelungsformel, mit der man dann die Fibonacci-Folge effizient berechnen kann. Um sie zu finden, gehen wir von der Vektorfolge zur assozierten Matrizenfolge über und definieren mit Hilfe der Zustandsvektoren die assoziierte Matrizenfolge

$$F(k) = \Big(\vec{y}(k), \vec{y}(k+1)\Big) = \begin{pmatrix} f(k) & f(k+1) \\ f(k+1) & f(k+2) \end{pmatrix}$$

Wir beachten, dass diese Matrizen die Rekursionsgleichung

$$F(k+1) = A \cdot F(k), \qquad k \in \mathbb{N}$$

erfüllt, weil ihre Spalten definitionsgemäss der Rekursion $\vec{y}(k+1) = A \cdot \vec{y}(k)$ genügen. Daher hat diese Matrizenfolge die explizite Beschreibung

$$F(k) = A^k \cdot F(0), \quad k \geq 0$$

Die hinter dem Verdoppelungsalgorithmus stehende zentrale Verdoppelungsformel

$$A^{2k} = A^k \cdot A^k = \Big(A^k\Big)^2$$

übertragen wir nun auf die Matrizenfolge. Zunächst ist

$$F(2k) = A^{2k} \cdot F(0) = \big(A^k \cdot A^k\big) \cdot F(0) = A^k \cdot \big(A^k \cdot F(0)\big) = A^k \cdot F(k)$$

Aus der expliziten Beschreibung $F(k) = A^k \cdot F(0)$ erhalten wir durch Auflösen nach A^k die Beziehung

$$A^k = F(k) \cdot F^{-1}(0)$$

die wir oben einsetzen und schliesslich die Verdoppelungsformel

$$F(2k) = F(k) \cdot F^{-1}(0) \cdot F(k)$$

für die assoziierte Matrizenfolge erhalten. Im konkreten Beispiel ist

$$F(0) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = A, \qquad F^{-1}(0) = \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}$$

und die gefundene Verdoppelungsformel $F(2k) = F(n) \cdot F^{-1}(0) \cdot F(k)$ nimmt in Komponenten die Gestalt

$$\begin{pmatrix} f(2k) & f(2k+1) \\ f(2k+1) & f(2k+2) \end{pmatrix}$$
$$= \begin{pmatrix} f(k) & f(k+1) \\ f(k+1) & f(k+2) \end{pmatrix} \cdot \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} f(k) & f(k+1) \\ f(k+1) & f(k+2) \end{pmatrix}$$

an. Koeffizientenvergleich der ersten Spalten und Verwenden der Rekursion der Fibonacci-Folge $f(k+2) = f(k)+f(k+1)$ liefert die gesuchten Verdoppelungsformeln

$$\left\{ \begin{array}{lcl} f(2k) & = & 2f(k)f(k+1) - f^2(k) \\ f(2k+1) & = & f^2(k+1) - f(k)f(k+1) + f(k)f(k+2) \end{array} \right.$$

Sie nehmen mit der Rekursion der Fibonacci-Folge $f(k+2) = f(k) + f(k+1)$ die Form

$$\left\{ \begin{array}{lcl} f(2k) & = & 2f(k)f(k+1) - f^2(k) \\ f(2k+1) & = & f^2(k+1) + f^2(k) \end{array} \right.$$

an.

Damit kann aus dem Paar $\big(f(k), f(k+1)\big)$ direkt das Paar $\big(f(2k), f(2k+1)\big)$ berechnet werden, dessen erste Komponente die doppelte Nummer hat. Mit Hilfe der Rekursionsgleichung kann daraus das Paar $(f(2k+1), f(2k+2))$ bestimmt werden, dessen erste Komponente eine ungerade Nummer hat.

Diese Verdoppelungsformeln können nun zur effizienten numerischen Berechnung der Fibonacci-Zahlen $f(k)$ herangezogen werden, ohne dass Matrizen in Erscheinung treten. Dazu geht man wie früher bei der Besprechung des Verdoppelungsverfahrens für Matrizen von der Binärdarstellung

$$k = \big(b_n b_{n-1} \ldots b_1 b_0\big)_2 = \sum_{j=0}^{n} b_j 2^j, \quad b_j \in \mathbb{Z}_2, b_n = 1$$

der Nummer aus und definiert die approximierende Folge

$$k_m = \big(b_n b_{n-1} \ldots b_{m+1} b_m\big)_2 = \sum_{j=m}^{n} b_j 2^j, \qquad 0 \leq m \leq n$$

Es ist $k_0 = k$ und $k_n = 1$ sowie

$$k_m = \left\{ \begin{array}{ll} 2k_{m+1} + 1 & \text{falls } b_m = 1 \\ 2k_{m+1} & \text{falls } b_m = 0 \end{array} \right.$$

Daher kann man mit Hilfe der gefundenen Verdoppelungsformeln aus dem Paar $\big(f(2k_{m+1}), f(2k_{m+1}+1)\big)$ das Paar $\big(f(k_m), f(k_m+1)\big)$ und bei Bedarf mit Hilfe der Rekursion das Paar $\big(f(k_m+1), f(k_m+2)\big)$ berechnen und erhält durch absteigende Iteration über $m = n, \ldots, 0$ in $n+1$ Schritten das gesuchte Resultat $f(k_0) = f(k)$.

Um zum Beispiel $f(100)$ zu berechnen, gehen wir früher von der Binärdarstellung

$$100 = 64 + 32 + 4 = (1100100)_2$$

aus. Hier ist $n = 6$ und die approximierende Folge war

$k_6 = 1$	$k_3 = (1100)_2 = 2k_4 = 12$
$k_5 = (11)_2 = 2k_6 + 1 = 3$	$k_2 = (11001)_2 = 2k_3 + 1 = 25$
$k_4 = (110)_2 = 2k_5 = 6$	$k_1 = (110010)_2 = 2k_2 = 50$
	$k_0 = (1100100)_2 = 2k_1 = 100$

Wir berechnen aus dem Paar $(f(0), f(1)) = (0, 1)$ mit Hilfe der Verdoppelungsformeln die Paare der Tabelle

m	b_m	$(f(k_m), f(k_{m+1}))$
6	$b_6 = 1$	$(f(0), f(1)) = (0, 1)$ $(f(1), f(2)) = (1, 1)$
5	$b_5 = 1$	$(f(2), f(3)) = (2, 3)$ $(f(3), f(4)) = (3, 5)$
4	$b_4 = 0$	$(f(6), f(7)) = (8, 13)$
3	$b_3 = 0$	$(f(12), f(13)) = (144, 233)$
2	$b_2 = 1$	$(f(24), f(25)) = (46368, 75025)$ $(f(25), f(26)) = (75025, 121393)$
1	$b_1 = 0$	$(f(50), f(51)) = (12586269025, 20365011074)$
0	$b_0 = 0$	$(f(100), f(101)) = (354224848179261915075, ?)$

und erhalten natürlich den selben Wert $f(100)$ wie oben. Diese Tabelle lässt sich durch das folgende Programm-Fragement realisieren, das man mit jenem für die effizienten Berechnung von Matrizenpotenzen vergleiche.

```
function fib(k:integer):integer;
var m, x, y, xx, t: integer;
begin
    if k ≤ 1 then return k; end;
    x := 0; y:=1;
    for m:=bitlength(k)-1 to 0 by −1 do
        xx := x ∗ x;
        x := 2 ∗ x ∗ y − xx;
        y := xx + y ∗ y;
        if bittest(k, m) then
          t := x;
          x := x + y;
          y := t;
        end;
    end;
    return x;
end.
```

Die Variablen `x` und `y` enthalten also nach jedem Schleifendurchgang mit dem Index m die Werte $f(k_m)$ und $f(k_m + 1)$. Natürlich braucht dieses Programm zur Berechnung von $f(k)$ nur $O\big(\log_2(k)\big)$ viele Multiplikation und Additionen. Theoretische Einsichten haben einen effizienten Algorithmus geliefert.

Um schliesslich die Matrizenpotenz A^k durch eine explizite Formel zu beschreiben, benötigen wir die bereits bestimmten Eigenwerte

$$\lambda_1 = \frac{1+\sqrt{5}}{2}, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2}$$

von A. Weil A als Begleitermatrix des charakteristischen Polynoms aufgefasst werden kann, ergeben sich die beiden zugehörigen Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix}$$

sofort aus den oben angegebenen Anfangsbedingungen der beiden Basislösungen $x_1(k) = \lambda_1^k$ und $x_2(k) = \lambda_2^k$. Wie man leicht kontrolliert, erfüllen sie nämlich die Eigenwertgleichung $A \cdot \vec{v}_j = \lambda_j \vec{v}_j, j = 1, 2$.

Mit Hilfe der Eigenvektoren bilden wir nun die invertierbare Matrix

$$X = (\vec{v}_1, \vec{v}_2) = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix}, \qquad X^{-1} = \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix}$$

Koordinatentransformation in die Eigenbasis liefert die Diagonalmatrix

$$\Lambda = X^{-1} \cdot A \cdot X = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

mit den Eigenwerten in der Diagonalen. Durch Auflösen ergibt sich daraus die Diagonalisierung der Systemmatrix

$$A = X \cdot \Lambda \cdot X^{-1}$$

und schliesslich die Matrizenpotenz als Teleskopprodukt

$$A^k = X \cdot \Lambda^k \cdot X^{-1} = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1\lambda_2^k - \lambda_2\lambda_1^k & \lambda_1^k - \lambda_2^k \\ \lambda_1\lambda_2(\lambda_2^k - \lambda_1^k) & \lambda_1^{k+1} - \lambda_2^{k+1} \end{pmatrix}$$

Durch Multiplikation mit dem Anfangszustand erhalten wir

$$\begin{aligned} \vec{y}(k) = A^k \cdot \vec{a} &= \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1\lambda_2^k - \lambda_2\lambda_1^k & \lambda_1^k - \lambda_2^k \\ \lambda_1\lambda_2(\lambda_2^k - \lambda_1^k) & \lambda_1^{k+1} - \lambda_2^{k+1} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1^k - \lambda_2^k \\ \lambda_1^{k+1} - \lambda_2^{k+1} \end{pmatrix} \end{aligned}$$

Die erste Komponente des Zustandes $\vec{y}(k)$ liefert definitionsgemäss

$$f(k) = \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2} = \frac{1}{\sqrt{5}} \cdot \lambda_1^k - \frac{1}{\sqrt{5}} \cdot \lambda_2^k = \frac{\sqrt{5}}{5} \cdot \lambda_1^k - \frac{\sqrt{5}}{5} \cdot \lambda_2^k$$

was wegen $\lambda_1 - \lambda_2 = \sqrt{5}$ mit der oben bereits hergeleiteten expliziten Formel übereinstimmt.

Aus dem asymptotischen Verhalten der Eigenwerte

$$\lim_{k\to\infty} \lambda_1^k = \infty, \qquad \lim_{k\to\infty} \lambda_2^k = 0$$

erhalten wir für die Matrizenpotenz mit $\lambda_1 \cdot \lambda_2 = -1$ asymptotisch

$$A^k \sim \frac{\lambda_1^k}{\lambda_1 - \lambda_2} \begin{pmatrix} -\lambda_2 & 1 \\ 1 & \lambda_1 \end{pmatrix}$$

In unserem numerischen Beispiel erhalten wir die Näherung

$$A^{100} \sim \begin{pmatrix} 2.18922995834555169026 \cdot 10^{20} & 3.54224848179261915075 \cdot 10^{20} \\ 3.54224848179261915075 \cdot 10^{20} & 5.73147844013817084101 \cdot 10^{20} \end{pmatrix}$$

die man mit dem exakten Wert vergleiche. Die explizite Formel nimmt asymptotisch die Näherungsform

$$f(k) \sim \frac{\lambda_1^k}{\lambda_1 - \lambda_2} = e(k)$$

an, die wir auch bereits kennen. Die Explosion der Population ist eine Konsequenz des Umstands, dass für den betragsgrössten Eigenwert $|\lambda_1| > 1$ gilt. Für den Fibonacci-Quotienten bzw. den reziproken Fibonacci-Quotienten erhalten wir asymptotisch

$$p(k) = \frac{f(k)}{f(k+1)} \sim \frac{1}{\lambda_1} = \tau, \qquad r(k) = \frac{f(k+1)}{f(k)} \sim \lambda_1$$

was wir auch bereits wissen.

Obwohl theoretische Physiker immer mehr überzeugt sind, dass Zeitintervalle, die kürzer, als die sogn. Planck-Zeit $t_P = \sqrt{\frac{\hbar G}{c^5}} \approx 5.391 \cdot 10^{-44}$ [s] sind, keinen Sinn machen, glauben die Analytiker statt an einen diskreten an einen kontinuierlichen Zeitverlauf und würden sich daher statt für die Lösung der diskreten Rekursionsgleichung $f(k+2) = f(k+1) + f(k)$ für die Lösung der linearen, homogenen autonomen Fibonacci-Differentialgleichung

$$f'' = f' + f, \qquad f'' - f' - f = 0, \qquad f(0) = 0, f'(0) = 1$$

interessieren. Um die Lösung dieses Anfangswertproblems zu finden, würden sie analog zum diskreten Fall vorgehen und zunächst für die gesuchte Lösungsfunktion $t \mapsto f(t)$ den *Exponentialansatz*

$$f(t) = e^{\lambda t}$$

machen. Setzt man ihn in die Differentialgleichung ein, stellt man fest, dass λ eine Nullstelle des selben charakteristischen Polynoms

$$\chi(\lambda) = 1 + \lambda - \lambda^2$$

sein muss, das wir bereits im diskreten Fall benutzt haben. Daher kommen auch im kontinuierlichen Problem als mögliche Eigenwerte nur die beiden uns bereits bekannten irrationalen Zahlen

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \approx 1.618\ldots, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2} \approx -0.618$$

in Frage. Die beiden Basislösungen $f_1(t) = e^{\lambda_1 t}$ und $f_2(t) = e^{\lambda_2 t}$ erfüllen also nach Konstruktion die Differentialgleichung und es sind die beiden einzigen Lösungen, die in Form einer Exponentialfunktion geschrieben werden können. Sie erfüllen die Anfangsbedingungen

$$\begin{aligned} f_1(0) &= 1, \quad f_1'(0) = \lambda_1 \\ f_2(0) &= 1, \quad f_2'(0) = \lambda_2 \end{aligned}$$

die noch nicht mit jenen des Anfangswertproblems übereinstimmen. Deshalb kombinieren wir die gefundenen Basislösungen linear und machen für die gesuchte allgemeine Lösung den Ansatz

$$f(t) = c_1 f_1(t) + c_2 f_2(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

Dabei handelt es sich wegen der Linearität der Differentialgleichung für eine beliebige Wahl der beiden Konstanten c_1 und c_2 um eine Lösung der Differentialgleichung. Diese Konstanten sollen nun so bestimmt werden, dass auch die gewünschte Anfangsbedingung $f(0) = 0$ und $f'(0) = 1$ erfüllt ist. Einsetzen liefert das selbe lineare Gleichungssystem

$$\begin{cases} c_1 + c_2 = 0 \\ \lambda_1 c_1 + \lambda_2 c_2 = 1 \end{cases}$$

wie im diskreten Fall. Es besitzt die uns bereits bekannte eindeutige Lösung

$$c_1 = \frac{1}{\lambda_1 - \lambda_2} = \frac{1}{\sqrt{5}} = \frac{\sqrt{5}}{5}, \qquad c_2 = \frac{1}{\lambda_2 - \lambda_1} = -\frac{1}{\sqrt{5}} = -\frac{\sqrt{5}}{5}$$

Daher hat das Fibonacci-Anfangswertproblem die Lösung

$$f(t) = \frac{\sqrt{5}}{5} \cdot e^{\lambda_1 t} - \frac{\sqrt{5}}{5} \cdot e^{\lambda_2 t}$$

Diese Lösungsfunktion bzw. ihr zugehöriger Graph zeigt das bereits bekannte exponentielle Wachstumsverhalten.

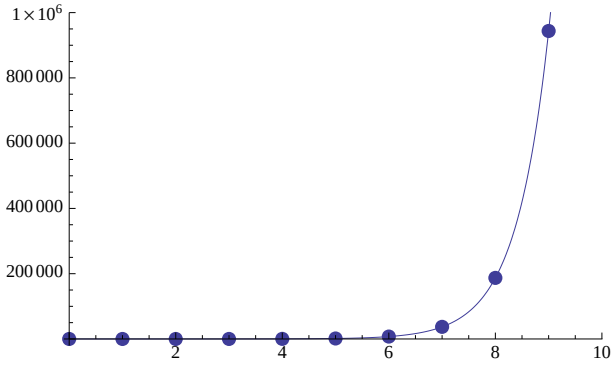


Abbildung 2.12: Der Graph der Fibonacci-Funktion $f(t)$.

Im Umgang mit Differentialgleichungen ist es zweckmässig, Differentialgleichungen höherer Ordnung durch ein System von Differentialgleichungen erster Ordnung zu ersetzen. Dazu definieren wir im vorliegenden Beispiel die beiden Zustandsvariablen $y_1(t) = f(t)$ und $y_2(t) = f'(t)$ und erklären damit den Zustandsvektor

$$\vec{y}(t) = \begin{pmatrix} f(t) \\ f'(t) \end{pmatrix} \in \mathbb{R}^2$$

Die Differentialgleichung nimmt damit die Form des Systems

$$\begin{cases} y_1'(t) = y_2(t) \\ y_2'(t) = y_1(t) + y_2(t) \end{cases}$$

an und kann mit Hilfe der Systemmatrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{R}^{2,2}$$

in matrizieller Form $\vec{y}\,'(t) = A \cdot \vec{y}(t)$ beschrieben werden. Für den Anfangszustand erhalten wir

$$\vec{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \mathbb{R}^2$$

Die Matrix A ist uns aus dem diskreten Fall bereits bekannt und beschreibt das zugehörige Vektorfeld

$$\vec{y} \mapsto A \cdot \vec{y}$$

Jeder Punkt $\vec{y}$ des Phasenraumes $\mathbb{R}^2$ beschreibt also einen Zustand des Systems. Durch die vielen kleinen Pfeilchen der Figur wird geometrisch das momentane Änderungsverhalten der Zustände beschrieben.

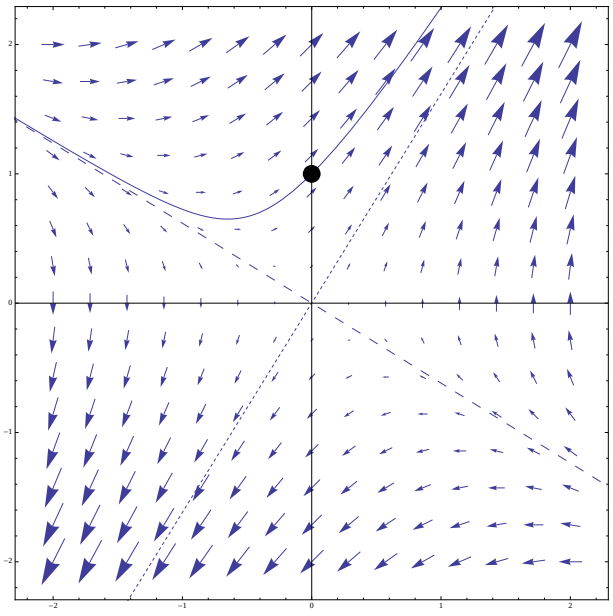


Abbildung 2.13: Vektorfeld der linearen Differentialgleichung $\vec{y}\,'(t) = A \cdot \vec{y}(t)$ mit der Integralkurve durch den Anfangszustand $\vec{y}(0) = \vec{a}$.

Durch die Differentialgleichung wird also zu jedem Punkt $\vec{y} \in \mathbb{R}^2$ ein Vektor $A \cdot \vec{y}$ festgelegt; anschaulich beschreibt dieses Vektorfeld eine Strömung im Phasenraum. Die Lösung der Differentialgleichung $t \mapsto \vec{y}(t)$ beschreibt anschaulich die eingezeichnete Bahn eines Korks, der im fett markierten Anfangszustand $\vec{a}$ in diese Strömung geworfen und dann von ihr mitgenommen wird. In unserem Beispiel erkennt man deutlich die beiden gestrichelt gezeichneten Richtungen, die von den beiden Eigenvektoren aufgespannt werden. Es sind die beiden einzigen geradlinigen Bahnkurven durch den Ursprung. Dass ihre beiden Richtungen orthogonal zueinander sind, hängt damit zusammen, dass in diesem Beispiel die Systemmatrix A symmetrisch ist. Offensichtlich nähert sich dieses System asymptotisch diesen beiden geraden Bahnkurven.

Wie wir gesehen haben, erhält man die Matrix A direkt als Begleitermatrix des charakteristischen Polynoms

$$\chi(\lambda) = 1 + \lambda - \lambda^2$$

Auf Grund des Satzes von Cayley-Hamilton gilt in der Tat $\chi_A(A) = 0$. Der Zustand des Systems $\vec{y}(t)$ zur Zeit t kann formal mit Hilfe des Propagators e^{At} aus dem Anfangszustand $\vec{a}$ direkt in der Form $\vec{y}(t) = e^{At} \cdot \vec{a}$ bestimmt werden. Der gesuchte Algorithmus zur Berechnung der Lösung des Fibonacci-Anfangswertproblems läuft also darauf hinaus, Propagatoren zu berechnen. Das ist mit Hilfe der Exponentialreihe

$$e^A = E + \frac{1}{1!}A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \frac{1}{4!}A^4 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}A^k$$

numerisch effizient möglich, weil diese Reihe für jede Matrix A sehr schnell konvergiert. Im vorliegenden Fall liefern die Matrizenpotenzen von A

$$\begin{aligned} A^2 &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \quad A^4 = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix} \\ A^5 &= \begin{pmatrix} 3 & 5 \\ 5 & 8 \end{pmatrix}, \quad A^6 = \begin{pmatrix} 5 & 8 \\ 8 & 13 \end{pmatrix}, \quad A^7 = \begin{pmatrix} 8 & 13 \\ 13 & 21 \end{pmatrix} \end{aligned}$$

mit den Fibonacci-Zahlen als Elemente für das Matrizenexponential den approximativen Wert siebter Ordnung

$$e^A \approx \sum_{k=0}^{7} \frac{1}{k!}A^k = \begin{pmatrix} \frac{8989}{5040} & \frac{3383}{1680} \\ \frac{3383}{1680} & \frac{1367}{360} \end{pmatrix} = \begin{pmatrix} 1.78353\ldots & 2.01369\ldots \\ 2.01369\ldots & 3.79722\ldots \end{pmatrix}$$

und damit für den Zustand des System zur Zeit $t = 1$

$$\vec{y}(1) = e^A \cdot \vec{a} \approx \begin{pmatrix} 2.01369\ldots \\ 3.79722\ldots \end{pmatrix}$$

Diese numerischen Werte machen sich neben den Funktionswerten der bereits bestimmten Lösung des Anfangswertproblems

$$\begin{aligned} f(1) &= \frac{\sqrt{5}}{5}e^{\lambda_1} - \frac{\sqrt{5}}{5}e^{\lambda_2} \approx 2.01432\ldots \\ f'(1) &= \frac{\sqrt{5}}{5}\lambda_1 e^{\lambda_1} - \frac{\sqrt{5}}{5}\lambda_2 e^{\lambda_2} \approx 3.79825\ldots \end{aligned}$$

gar nicht schlecht und können bei Bedarf durch Summation weiterer Terme der Reihe beliebig verbessert werden.

Um die Differentialgleichung durch eine elementare Formel zu lösen, müssen wir den Propagator e^{At} symbolisch berechnen, was mit der besprochenen Diagonalisierung der Systemmatrix A leicht gelingt. Aus ihren bekannten Eigenwerten

$$\lambda_1 = \frac{1 + \sqrt{5}}{2}, \qquad \lambda_2 = \frac{1 - \sqrt{5}}{2}$$

und den beiden zugehörigen Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix}$$

bilden wir in geläufiger Weise die Transformationsmatrix

$$X = (\vec{v}_1, \vec{v}_2) = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix}, \qquad X^{-1} = \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix}$$

Transformation in die Eigenbasis liefert die Diagonalmatrix

$$\Lambda = X^{-1} \cdot A \cdot X = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

mit den Eigenwerten in der Diagonalen. Durch Auflösen ergibt sich daraus die Diagonalisierung

$$A = X \cdot \Lambda \cdot X^{-1}$$

Für den Propagator der Diagonalmatrix Λ erhalten wir sofort

$$e^{\Lambda t} = \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}$$

und daher durch Transformation in die Eigenbasis den gesuchten Propagator

$$\begin{aligned} e^{At} &= X \cdot e^{\Lambda t} \cdot X^{-1} = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \cdot \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} -\lambda_2 & 1 \\ \lambda_1 & -1 \end{pmatrix} \\ &= \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 e^{\lambda_2 t} - \lambda_2 e^{\lambda_1 t} & e^{\lambda_1 t} - e^{\lambda_2 t} \\ \lambda_1 \lambda_2 (e^{\lambda_2 t} - e^{\lambda_1 t}) & \lambda_1 e^{\lambda_1 t} - \lambda_2 e^{\lambda_2 t} \end{pmatrix} \end{aligned}$$

Seine Eigenwerte sind also $e^{\lambda_1 t}$ und $e^{\lambda_2 t}$ und er erfüllt wie erwartet die Differentialgleichung $(e^{At})' = A \cdot e^{At}$ sowie die Anfangsbedingung $e^{A0} = E$. Für $t = 1$ erhalten wir die Matrix

$$e^A = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1} & e^{\lambda_1} - e^{\lambda_2} \\ \lambda_1 \lambda_2 (e^{\lambda_2} - e^{\lambda_1}) & \lambda_1 e^{\lambda_1} - \lambda_2 e^{\lambda_2} \end{pmatrix} \approx \begin{pmatrix} 1.78392 & 2.01432 \\ 2.01432 & 3.79825 \end{pmatrix}$$

die man mit der oben durch Summieren der Exponentialreihe berechneten Näherung vergleiche.

Durch Multiplikation mit dem Anfangszustand erhalten wir die Lösung

$$\begin{aligned} \vec{y}(t) = e^{At} \cdot \vec{a} &= \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 e^{\lambda_2 t} - \lambda_2 e^{\lambda_1 t} & e^{\lambda_1 t} - e^{\lambda_2 t} \\ \lambda_1 \lambda_2 (e^{\lambda_2 t} - e^{\lambda_1 t}) & \lambda_1 e^{\lambda_1 t} - \lambda_2 e^{\lambda_2 t} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} e^{\lambda_1 t} - e^{\lambda_2 t} \\ \lambda_1 e^{\lambda_1 t} - \lambda_2 e^{\lambda_2 t} \end{pmatrix} \end{aligned}$$

Die erste Komponente des Zustandes $\vec{y}(t)$ liefert wegen $\lambda_1 - \lambda_2 = \sqrt{5}$ die früher bereits bestimmte Lösungsfunktion der Differentialgleichung

$$f(t) = \frac{1}{\lambda_1 - \lambda_2} \cdot \left(e^{\lambda_1 t} - e^{\lambda_2 t} \right) = \frac{\sqrt{5}}{5} \cdot e^{\lambda_1 t} - \frac{\sqrt{5}}{5} \cdot e^{\lambda_2 t}$$

und die zweite Komponente ist definitionsgemäss ihre Ableitung $f'(t)$.

Aus dem asymptotischen Verhalten der Eigenwerte

$$\lim_{k \to \infty} \lambda_1^k = \infty, \qquad \lim_{k \to \infty} \lambda_2^k = 0$$

erhalten wir für den Propagator mit $\lambda_1 \cdot \lambda_2 = -1$ asymptotisch

$$e^{At} \sim \frac{e^{\lambda_1 t}}{\lambda_1 - \lambda_2} \begin{pmatrix} -\lambda_2 & 1 \\ 1 & \lambda_1 \end{pmatrix}$$

Ein detailierter Vergleich der Resultate im diskreten bzw. im kontinuierlichen Fall bei der Differenzen. bzw. der Differentialgleichung zeigt, dass in beiden Fällen völlig analog vorgegangen wird und sich die entsprechenden Ergebnisse direkt übertragen lassen. Um die Analogie weiter zu vertiefen, wollen wir uns noch überlegen, wie wir die beiden Fälle durch Diskretisieren ineinander überführen können. Dazu Unterteilen wir die als kontinuierlich aufgefasste Zeit so in k diskrete Zeitschritte der Länge Δt, dass $t = k \cdot \Delta t$ gilt. Um für alle $k \in \mathbb{N}$ die Diskretisierungsbedingung

$$e^{At} = e^{A \cdot k \cdot \Delta t} = \left(e^{A \cdot \Delta t}\right)^k = C^k$$

zu erfüllen, müssen wir als Systemmatrix des zugehörigen diskreten Systems

$$C = e^{A \cdot \Delta t}$$

wählen, damit die Lösung des diskreten Problems auf der Lösungsfunktion des kontinuierlichen Problems liegt.

Im Fall der Fibonacci-Differentialgleichung müssen wir bei einer gewählten Schrittweite von $\Delta t = 1$ für das zugehörige diskrete System die Matrix

$$C \quad = \quad e^A = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1} & e^{\lambda_1} - e^{\lambda_2} \\ \lambda_1 \lambda_2 (e^{\lambda_2} - e^{\lambda_1}) & \lambda_1 e^{\lambda_1} - \lambda_2 e^{\lambda_2} \end{pmatrix}$$

wählen. Ein Blick auf den oben gezeichneten Graphen der Lösung $f(t)$ der Fibonacci-Differentialgleichung zeigt, dass die fett gezeichneten Punkte des diskreten Systems $\vec{y}(k) = C^k \cdot \vec{a}$ tatsächlich für alle diskreten Zeiten k auf diesem Graphen liegen.

Linearisiert man die Exponentialmatrix C, d.h. bricht die Exponentialreihe nach dem linearen Term ab, erhält man die Näherung

$$C = e^{A \cdot \Delta t} \approx E + A \cdot \Delta t$$

Für die Differentialgleichung $\vec{y}\,'(t) = A \cdot \vec{y}(t)$ erhalten wir die Näherungslösung

$$\vec{y}(k) = \left(E + A \cdot \Delta t\right)^k \cdot \vec{a} \qquad k \geq \mathbb{N}$$

von der wir erwarten, dass sie für festes t um so genauer wird, je kleiner Δt bzw. je grösser k wird. Dieses Verhalten bestätigt man an Hand folgender Figur.

Die Punkte der Folge $(k \cdot \Delta t, \vec{y}_1(k))$ liegen am Anfang recht gut in der Nähe der gesuchten Kurve und entfernen sich mit wachsendem k immer weiter von ihr. Der Näherungsfehler an der Stelle $t = k \cdot \Delta t$ ist proportional zu Δt. Er verkleinert sich also, falls wir eine kleinere Schrittweite Δt wählen. Für die Figur sind Schrittweiten $\Delta t = 0.1$, $\Delta t = 0.05$ und $\Delta t = 0.01$ benutzt worden. Im dritten Fall mussten zehnmal so viele Punkte berechnet werden, die dann natürlich nur ein zehntel so weit voneinander entfernt sind. Falls wir einen festen

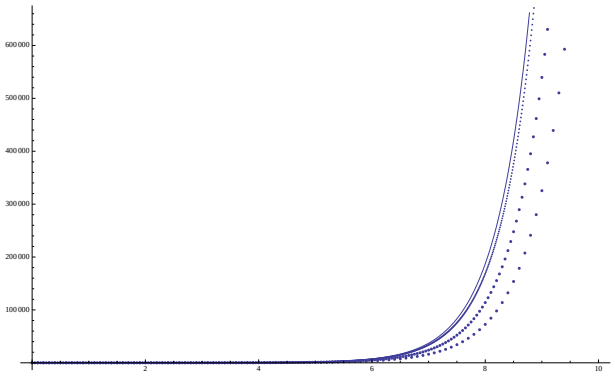


Abbildung 2.14: Lösungskurve der Differentialgleichung $f'' = f' + f$ mit den Näherungen, die zu $\Delta t = 0.1$, $\Delta t = 0.05$ und $\Delta = 0.01$ gehören.

Zeitpunkt t wählen, so konvergieren die numerischen Approximationen jedoch offensichtlich gegen den korrekten Wert $f(t)$, falls $\Delta t = \frac{t}{k}$ beliebig klein, d.h. bei festem t die Nummer beliebig gross wird. In der Tat existiert der etwas heikle Grenzwert und für den Propagator gilt die Darstellung als Grenzwert

$$\lim_{k\to\infty} \left(E + A \cdot \frac{t}{k}\right)^k = e^{At}$$

die auf Euler zurückgeht und in der Schule unter der Rubrik stetige Verzinsung abgehandelt wurde. Sie zeigt, dass die Lösung autonomer linearer Differentialgleichungen näherungsweise durch Matrizenpotenzen berechnet werden kann, die duch die Rekursion

$$\vec{y}(k+1) = (E + \Delta t \cdot A) \cdot \vec{y}(k), \qquad \vec{y}(0) = \vec{a}$$

beschreiben werden kann. Im Gegensatz zur Potenzreihen-Darstellung

$$e^{At} = E + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k$$

ist sie jedoch zur numerischen Berechnung des Propagators ungeeignet, weil die Folge sehr langsam konvergiert. In der numerischen Mathematik verwendet man statt der Linearisierung von C oft die Näherung vierter Ordnung

$$C \approx E + \Delta t \cdot A + \frac{(\Delta t)^2}{2!}A^2 + \frac{(\Delta t)^3}{3!}A^3 + \frac{(\Delta t)^4}{4!}A^4$$

und berechnet durch Potenzieren die Vektorfolge

$$\vec{y}(k+1) = \left(E + \Delta t \cdot A + \frac{(\Delta t)^2}{2!}A^2 + \frac{(\Delta t)^3}{3!}A^3 + \frac{(\Delta t)^4}{4!}A^4\right) \cdot \vec{y}(k), \qquad \vec{y}(0) = \vec{a}$$

die als Runge-Kutta-Verfahren vierter Ordnung bekannt ist. Sein Fehler ist proportional zu $(\Delta t)^4$. Eine Halbierung der Schrittweite berwirkt also, dass die Fehler auf einen 16-tel reduziert werden. ◯

Nachdem die bisher besprochenen Beispiele linearer Vektorfolgen die *geometrischen Folgen* vom skalaren Fall auf höhere Dimensionen verallgemeinern, wollen wir nun auch die *arithmetischen Folgen* verallgemeinern.

Geometrische Folgen werden explizit durch Exponentialfunktionen und beim Vorliegen komplexer Eigenwerte durch Kreisfunktionen beschrieben. Im Gegensatz dazu benötigt man zur expliziten Beschreibung arithmetischer Folgen die Polynome.

Diese drei Klassen reeller Funktionen entsprechen im kontinuierlichen Fall den drei Prototypen linearer, autonomer, homogener Differentialgleichungen

$$f' = af, \qquad f'' = -\omega f, \qquad f^{(k+1)} = 0$$

wobei in jedem Fall nach t abzuleiten ist. Ihre allgemeinen Lösungen sind

$$f(t) = ce^{at}, \quad f(t) = a\cos(\omega t) + b\sin(\omega t) = C\cos(\omega t + \varphi), \quad f(t) = \sum_{j=0}^{k} a_j t^j$$

Die freien Konstanten dienen zum Anpassen an die Anfangsbedingungen.

Wir besprechen nun die entsprechenden diskreten Versionen dieser fundamentalen Funktionsklassen und ihrer zugehörigen Differentialgleichungen bzw. Systemen von Differenzialgleichungen. Die zugehörigen Differenzengleichungen bzw. Systeme von linearen, autonomen, homogenen Differenzengleichungen entstehen beispielsweise beim Diskretieren der zugehörigen Differentialgleichungen und sind zur numerischen Berechnung von Näherungslösungen erforderlich. Die zur vollständigen Behandlung solcher linearer autonomer Systeme erforderliche Jordan'sche Normalform werden wir später, mit etwas mehr Erfahrung, zusammen mit den Hauptvektoren, behandeln.

Während bei der geometrischen Folge, die durch die Rekursionsgleichung

$$x_{n+1} = qx_n$$

definiert ist, der Quotient aufeinanderfolgender Glieder konstant ist, und die Differenz

$$\Delta(x_n) = x_{n+1} - x_n = qx_n - x_n = (q-1)x_n = ax_n$$

proportional zum Momentanzustand ist, ist die Differenz bei den arithmetischen Folgen konstant und daher verschwindet für sie die zweite Differenz, d.h. die Differenz der Differenz. Für eine arithmetische Folge x_n gilt also

$$\Delta(x_n) = x_{n+1} - x_n = d, \qquad \Delta^2(x_n) = 0$$

Das typische Verhalten einer arithmetischen Folge erkennt man in der folgenden Figur an Hand der Folge der geraden Zahlen $x_n = 2n$.

Diese arithmetische Folge mit der Differenz $d = 2$ wird in der Schule meistens als 2-er Reihe bezeichnet, weil es sich um die Summenfolge der konstanten Folge $c_n = 2$ handelt, da $\Delta(x_n) = c_n$ gilt.

Beispiel. Physiker würden in diesem Zusammenhang von einer *gleichförmigen Bewegung* des Punktes reden, weil er in gleichen Zeitetappen der Länge T die

selbe Distanz d zurücklegt. Den Quotienten $v = \frac{d}{T}$ würden sie als *Geschwindigkeit* bezeichnen. Um die Geschwindigkeit zu messen, müssen also eine Länge und eine Zeitdauer gemessen werden.

Gleichförmige Bewegungen stehen in engem Zusammenhang mit harmonischen Schwingungen mit einer festen Frequenz f.

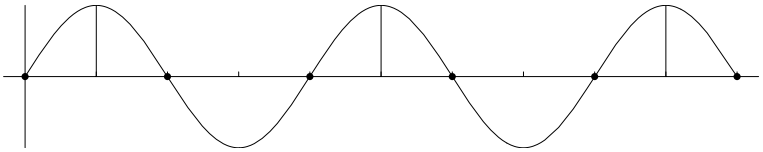


Abbildung 2.15: Aufzeichnung einer harmonischen Schwingung.

Die feste Zeitdifferenz zwischen zwei aufeinanderfolgenden Maxima beträgt $T = \frac{1}{f}$ und wird als Periodendauer bezeichnet. Die Zeitdifferenz zwischen aufeinanderfolgender Nulldurchgängen, die auch eine arithmetische Folge bilden, beträgt die halbe Periodendauer $\frac{1}{2f}$.

Bezeichnet x_n den in den ersten n Etappen insgesamt zurückgelegten Weg, so gilt für eine gleichförmige Bewegung definitionsgemäss

$$\Delta(x_n) = x_{n+1} - x_n = d$$

die sich auch als Rekursionsgleichung erster Ordnung

$$x_{n+1} = x_n + d$$

schreiben lässt. Definitionsgemäss gilt die Anfangsbedingung $x_1 = d$.

Weil diese Rekursion nicht homogen ist, lässt sie sich nicht direkt durch ein Schieberegister realisieren. Zu diesem Zweck bilden wir ihre zweite Differenz, d.h. die Differenz der Differenz und erhalten die Differenzengleichung zweiter Ordnung

$$\begin{aligned} \Delta^2(x_n) &= \Delta(x_{n+1}) - \Delta(x_n) = \big(x_{n+2} - x_{n+1}\big) - \big(x_{n+1} - x_n\big) \\ &= x_{n+2} - 2x_{n+1} + x_n = 0 \end{aligned}$$

die als lineare, autonome homogene Rekursionsgleichung zweiter Ordnung

$$x_{n+2} = 2x_{n+1} - x_n, \qquad x_1 = d, x_2 = 2d$$

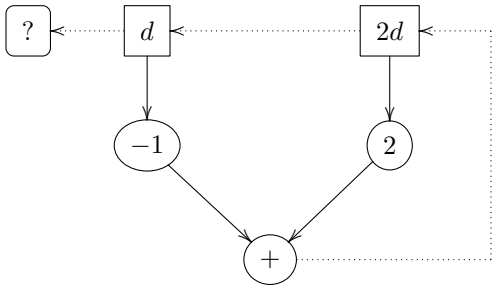
aufgefasst werden kann. Diese lineare Rekursionsgleichung zweiter Ordnung lässt sich mit dem charakteristischen Polynom

$$\chi(\lambda) = 1 - 2\lambda + \lambda^2 = (1 - \lambda)^2$$

und der Begleitermatrix

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$$

durch das lineare Schieberegister



erzeugen. Für $d = 1$ liefert dieses Schiebregister die Folge der natürlichen Zahlen $x_n = n$, mit der das Unternehmen Mathematik in unserer Kindheit angefangen hat. Die 2-er Reihe aus der Schule ergibt sich für $d = 2$.

Man beachte, dass A den doppelten Eigenwert $\lambda = 1$ hat. Zur Bestimmung des Eigensystems gehen wir deshalb von der Matrix $A - E$ aus und suchen nun möglichst viele unabhängige nichttriviale Lösungen für das lineare Gleichungssystem $(A-E)\cdot\vec{v} = \vec{0}$ bzw. für die Fixpunktgleichung $A\cdot\vec{v} = \vec{v}$. Seine erweiterte Matrix lautet

$$\left(\begin{array}{cc|c} -1 & 1 & 0 \\ -1 & 1 & 0 \end{array}\right)$$

Weil die beiden Zeilen übereinstimmen, lautet seine allgemeine Lösung

$$\vec{v} = t\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Offenbar ist es in diesem Beispiel nicht möglich, zwei unabhängige Eigenvektoren zu finden. Wegen dieser Entartung ist die Matrix A nicht diagonalisierbar. In diesem Beispiel ist es aber auch sonst leicht, für die Matrizenpotenz eine explizite Formel zu finden. Etwas Probieren liefert die Vermutung

$$A^k = \begin{pmatrix} 1-k & k \\ -k & k+1 \end{pmatrix}$$

die man durch Einsetzen in die Rekursionsgleichung der Potenz bestätigt. Für solche entarteten Sonderfälle muss man also die Eigenwerttheorie verallgemeinern und wird dabei statt zur Diagonalsierung zur allgemeineren Jordan'schen Normalform $J(A)$ geführt für die man neben den Eigenvektoren noch die sogn. Hauptvektoren benötigt.

Für die arithmetische Folge x_n lässt sich auch ohne Diagonalisierung leicht eine explizite Beschreibung finden. Mit dem Anfangsglied x_1 berechnet man mit Hilfe der Rekursion die ersten paar Folgenglieder und vermutet die Beziehung

$$x_n = x_1 + (n-1)d$$

die man durch Einsetzen in die Rekursionsgleichung leicht bestätigt.

Beim der gleichförmigen Bewegung gilt auf Grund der Anfangsbedingung $x_1 = d$ und wir erhalten für den in den ersten n-ten Etappen zurückgelegten Weg die explizite Beschreibung

$$x_n = n \cdot d$$

Charakteristisch für die gleichförmige Bewegung ist die Tatsache, dass die sich ergebene Geschwindigkeit $v = \frac{d}{T}$ unabhängig von der Länge des Zeitintervalls T ist. Deshalb formuliert man dieses Resultate heute in kontinuierlicher, statt in diskreter Sprache. Um die Werte für einen festen Zeitpunkt $t \in \mathbb{R}_{\geq}$ zu erhalten, teilt man die Zeit so in n gleich lange Intervalle der Länge T ein, dass $t = n \cdot T$ bzw. $n = \frac{t}{T}$ gilt.

Damit erhält man für den mit Hilfe von t ausgedrückten Gesamtweg die von T unabhängige Beziehung

$$s(t) = n \cdot d = n \cdot v \cdot T = \frac{t}{T} \cdot v \cdot T = v \cdot t$$

Der Graph dieser reellen Funktion ist eine Gerade durch den Ursprung mit der Steigung v. ◯

Entsprechend verschwinden bei den höheren arithmetischen Folgen höhere Differenzen. Sie werden durch folgende Differenzengleichungen beschrieben.

Definition. Unter einer arithmetischen Folge k-ter Ordnung verstehen wir eine Folge x_n, die der Differenzengleichung $\Delta^{k+1}(x_n) = 0$ genügt.

Das folgende Beispiel spielt in der Geschichte der Physik eine zentrale Rolle, weil es den historisch ersten Versuch beschreibt, physikalische Gesetze auf Grund von experimentellen Beobachtungen und Messungen zu beschreiben.

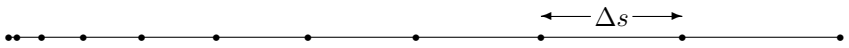
Beispiel. Galilei[23] hat sich 1590 mit der Frage beschäftigt, welchem Gesetz die Distanzen und die Zeiten beim freien Fall in der Nähe der Erdoberfläche gehorchen, falls man den Luftwiderstand vernachlässigen kann (ideales Fallgesetz). Dazu hat er Experimente durchgeführt[24], die man heute im Hörsaal wiederholt.

Der Autor erinnert sich an ein Experiment, bei dem ein Holzstab, der mit weissem Papier umwickelt und an einem Galgen aufgehängt war und durch einen Auslöser freigegeben wurde. Der Stab fiel dann frei durch einen feinen, in einer Ebene kreisenden Tintenstrahl zu Boden. Der Strahl spritzte aus der seitlichen Düse eines sich mit der festen Frequenz f gleichförmig drehenden Tintenfasses. Die qualitative Auswertung des Experimentes ergab auf dem Papier eine Folge von Tintenspritzern, die zwar in konstanten Zeitintervallen der Länge $T = \frac{1}{f}$

[23] 1564 – 1642.

[24] Aus heutiger Sicht ist man sicher, dass Galilei die meisten Fall-Experimente gar nicht wirklich durchgeführt, sondern die Ergebnisse seinen theoretischen Vorstellungen angepasst hat. Bemerkenswert ist der berühmte Fallversuch vom schiefen Turm zu Pisa, mit dem er vorgab zu beweisen, dass die Fallgeschwindigkeit eines Gegenstandes von seinem Gewicht unabhängig ist und der die Ansichten des Aristoteles widerlegen sollte, wonach die Geschwindigkeit sich proportional zum Gewicht verhält. Hätte er dieses Experiment wirklich durchgeführt, wäre ihm aufgefallen, dass die gleich grosse Holz- und Eisenkugel, von denen bei ihm die Rede ist, die Erde gar nicht gleichzeitig erreichen. Die „Wiederholung“ der Experimente ergab 1978, dass die Eisenkugel etwas schneller als die Holzkugel fiel. Der Geschwindigkeitsunterschied war zwar nicht proportional zum Gewichtsunterschied, wie Aristoteles angenommen hatte. Galilei's Vertrauen in seine Theorie war offenbar so stark, dass er die physikalischen Gesetzmässigkeiten zunächst theoretisch herleitete und die Experimente — wenn schon — erst danach durchführte, um seine Hypothesen zu bestätigen und durch dieses Frisieren der Daten nach dem Standard heutiger “political correctness” kurzerhand betrog. Nichts ist halt so praktisch, wie eine gute Theorie.

aufgespritzt wurden, deren Abstände Δs aber dauernd zunahmen[25].



Bei Wiederholung des Versuches mit einem Stab aus anderem Material, etwa mit einem Eisenrohr, statt mit einem Holzstab, ergaben sich die selben Werte.

Eine Variante dieses Experimentes findet man im klassischen Lehrbuch der Mechanik von Mach beschrieben.

> Eine berusste Glasscheibe falle frei vertikal nach unten, während ein horizontal schwingender Stift, der beim ersten Durchgang durch seine Gleichgewichtslage die Fallbewegung auslöst, eine Kurve auf die Glasscheibe zeichnet. Wegen der konstanten Schwingungsdauer des Stiftes und der zunehmenden Fallgeschwindigkeit der Glasscheibe werden die vom Stift aufgezeichneten Wellen immer länger.

Das Experiment liefert die folgende typische Messkurve.

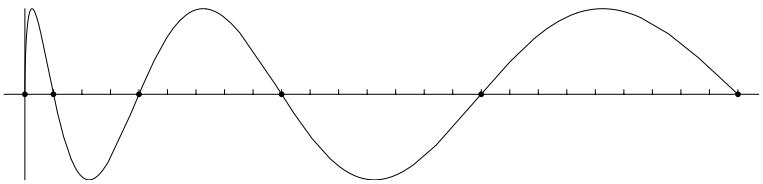


Abbildung 2.16: Aufzeichnung des freien Falls.

Offenbar steckt hinter den Abständen der Nullstellen dieser Kurve, die in gleichen Zeitintervallen durchlaufen wurden, ein einfaches Muster, das sich wie folgt in Worte fassen lässt:

1. Bei einer idealen Fallbewegung wächst in den einzelnen Zeitintervallen der zurückgelegte Weg von einem Zeitintervall zum nächsten um den selben Wert d.

2. Im ersten Zeitintervall beträgt der zurückgelegte Weg $\frac{d}{2}$.

Zur quantitativen Auswertung dieser beiden Beobachtungen benutzen wir moderne Einheiten. Wir nehmen an, der Stift schwinge harmonisch mit der Frequenz $f = 25$ [Hz]. Daher beträgt die feste Zeitdifferenz zwischen zwei aufeinanderfolgenden Nulldurchgängen $T = \frac{1}{50}$ [s]. Für die in der n-ten Zeitetappe der Länge T zurückgelegten Distanzen x_n [cm] erhält man (bis auf die unvermeid-

[25] Diese Experiment würde man wohl heute, um die Sauerei zu vermeiden, mit Hilfe von Lichtschranken durchführen, wobei Lichtstrahlen auf Photozellen gerichtet und durch den vorbeifliegenden Stab unterbrochen und wieder frei gegebenen werden und dabei elektronische Uhren steuern, an denen dann die benötigten Zeiten abgelesen werden können.

lichen Messfehler) die Werte folgender Tabelle:

n	x_n		Δx_n		s_n	
1	0.1962	$1\cdot\frac{d}{2}$	0.3924	d	0.1962	$1\cdot\frac{d}{2}$
2	0.5886	$3\cdot\frac{d}{2}$	0.3924	d	0.7848	$4\cdot\frac{d}{2}$
3	0.981	$5\cdot\frac{d}{2}$	0.3924	d	1.7658	$9\cdot\frac{d}{2}$
4	1.3734	$7\cdot\frac{d}{2}$	0.3924	d	3.1392	$16\cdot\frac{d}{2}$
5	1.7658	$9\cdot\frac{d}{2}$	0.3924	d	4.905	$25\cdot\frac{d}{2}$
6	2.1582	$11\cdot\frac{d}{2}$	0.3924	d	7.0632	$36\cdot\frac{d}{2}$
7	2.5506	$13\cdot\frac{d}{2}$	0.3924	d	9.6138	$49\cdot\frac{d}{2}$
8	2.943	$15\cdot\frac{d}{2}$	0.3924	d	12.5568	$64\cdot\frac{d}{2}$
9	3.3354	$17\cdot\frac{d}{2}$	0.3924	d	15.8922	$81\cdot\frac{d}{2}$
10	3.7278	$19\cdot\frac{d}{2}$			19.62	$100\cdot\frac{d}{2}$

Zur mathematischen Formulierung dieser Gesetzmässigkeiten wählen[26] wir als Variable x_n, d.h. den in der n-ten Zeitetappe zurückgelegten Weg, der im Graphen dem Abstand der $(n-1)$-ten Nullstelle zur nächsten entspricht. Dann gilt auf Grund der ersten Beobachtung für den Unterschied die Differenzengleichung

$$\Delta(x_n) = x_{n+1} - x_n = d$$

die sich auch als Rekursionsgleichung erster Ordnung

$$x_{n+1} = x_n + d$$

schreiben lässt. Die zweite Beobachtung liefert die Anfangsbedingung $x_1 = \frac{d}{2}$.

Weil es sich offensichtlich um eine arithmetische Folge handelt, findet man mit der seinerzeit gefundenen expliziten Formel für das allgemeine Glied einer arithmetischen Folge beim freien Fall auf Grund der beiden Beobachtungen für den in der n-ten Etappe zurückgelegten Weg die explizite Beschreibung

$$x_n = (2n-1)\cdot\frac{d}{2}$$

Galilei hat also beobachtet, dass sich die Differenzen zwischen zwei aufeinanderfolgender Nullstellen der Messkurve mit den ungeraden Zahlen multiplizieren und verkündete dieses Ergebnis stolz mit den Worten

> ... denn soviel ich weiss, hat noch niemand bewiesen, dass die von fallenden Körpern in gleichen Zeiten zurückgelegten Strecken sich zueinander verhalten wie die ungeraden Zahlen.

[26] Als Warung weisen wir darauf hin, dass das *nicht* die selbe Wahl ist, wie oben bei der gleichförmigen Bewegung! Unsere jetzige Wahl von x_n entspricht der damaligen Wahl der Differenzenfolge $\Delta(x_n)$, die konstant war und deshalb keine eigene Bezeichnung hatte. Der damaligen Folge x_n entspricht im aktuellen Beispiel des freien Falls eine Folge, die wir mit s_n bezeichnen werden.

Daraus lässt sich nun leicht der in der ersten n Etappen zurückgelegten Gesamtweg s_n berechnen, der im Graphen dem Abstand der n-ten Nullstelle vom Ursprung entspricht. Definitionsgemäss gilt für die Summenfolge

$$s_n = \sum_{j=1}^{n} x_j$$

Um diese Summenfolge, deren numerische Werte man auch in obiger Tabelle findet, durch eine Rekursionsgleichung zu beschreiben, beachten wir, dass auf Grund der Definition der Summenfolge für die Differenzen aufeinanderfolgender Glieder die Differenzengleichung erster Ordnung

$$\Delta(s_n) = s_{n+1} - s_n = x_{n+1}$$

gilt. Sie entspricht im kontinuierlichem Fall der Differentialgleichung des Integrals $F' = f(x)$. Um die bekannte Differenzengleichung von x_n ins Spiel bringen zu können, bilden wir die zweiten Differenzen der Summenfolge und erhalten die Differenzengleichung zweiter Ordnung

$$\Delta^2(s_n) = \Delta(s_{n+1}) - \Delta(s_n) = x_{n+1} - x_n = d$$

Weil sie nicht homogen ist, bilden wir nochmals die Differenzen und erhalten schliesslich die Differenzengleichung dritter Ordnung

$$\Delta^3(s_n) = 0$$

Sie lautet ausgeschrieben

$$\begin{aligned} \Delta^3(s_n) &= \Delta^2(s_{n+1}) - \Delta^2(s_n) \\ &= \big(s_{n+3} - 2s_{n+2} + s_{n+1}\big) - \big(s_{n+2} - 2s_{n+1} + s_n\big) \\ &= s_{n+3} - 3s_{n+2} + 3s_{n+1} - s_n = 0 \end{aligned}$$

Offenbar ist diese Summenfolge arithmetisch von der Ordnung 2 und wird durch die lineare, autonome homogene Rekursionssgleichung dritter Ordnung

$$s_{n+3} = 3s_{n+2} - 3s_{n+1} + s_n, \qquad s_0 = 0, s_1 = \frac{d}{2}, s_2 = 4 \cdot \frac{d}{2}$$

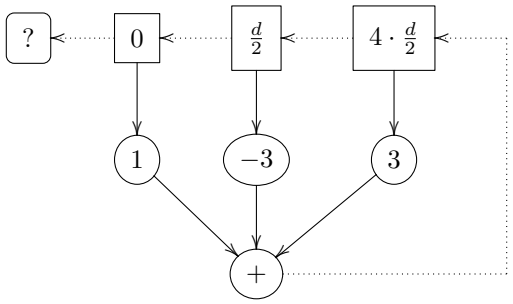
beschrieben. Ihr zugehöriges charakteristisches Polynom

$$\chi(\lambda) = 1 - 3\lambda + 3\lambda^2 - \lambda^3 = (1-\lambda)^3$$

hat den doppelt entarteten Eigenwert $\lambda = 1$ und die Begleitermatrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 3 \end{pmatrix}$$

Sie kann also durch das lineare Schieberegister



erzeugt werden und ist ebenfalls entartet. Die Matrix A hat den dreifachen Eigenwert $\lambda = 1$. Zur Bestimmung des Eigensystems gehen wir deshalb von der Matrix $A - E$ aus und suchen nun möglichst viele unabhängige, nichttriviale Lösungen für das lineare Gleichungssystem $(A - E) \cdot \vec{v} = \vec{0}$ bzw. für die Fixpunktgleichung $A \cdot \vec{v} = \vec{v}$. Seine erweiterte Matrix lautet

$$\left(\begin{array}{ccc|c} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 1 & -3 & 2 & 0 \end{array}\right)$$

Addition der ersten Zeile zur dritten liefert

$$\left(\begin{array}{ccc|c} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -2 & 2 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{13}(1) \end{array}\right]$$

Addition des (-2)-fachen der zweiten Zeile zur dritten liefert die Stufenform

$$\left(\begin{array}{ccc|c} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(-2) \end{array}\right]$$

die durch Addition der zweiten Zeile zur ersten reduziert wird.

$$\left(\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(1) \\ \\ \end{array}\right]$$

Aus ihr lesen wir die allgemeine Lösung

$$\vec{v} = t \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

ab. Offenbar ist es auch hier nicht möglich, drei unabhängige Eigenvektoren zu finden. Wegen dieser doppelten Entartung ist die Matrix A nicht diagonalisierbar. In diesem Beispiel ist es mit einfachen Mitteln möglich, eine explizite Formel für die Matrizenpotenz zu finden. Dazu beachtet man, dass ihre Einträge arithmetische Folgen zweiter Ordnung sein müssen und erhält

$$A^k = \frac{1}{2} \begin{pmatrix} (k-1)(k-2) & 2k(2-k) & k(k-1) \\ k(k-1) & 2(1+k)(1-k) & k(k+1) \\ k(k+1) & -2(2+k) & (k+1)(k+2) \end{pmatrix}$$

wie man durch Einsetzen in die Rekursionsgleichung der Potenz bestätigt. Auch sie kann man mit der Jordan'schen Normalform $J(A)$ leicht berechnen.

Auch diesmal ist es nicht all zu schwierig, für die Summenfolge einer arithmetischen Folge auch ohne Jordan'sche Normalform eine explizite Beschreibung zu erraten. Gemäss Definition ist

$$s_n = \sum_{j=1}^{n} x_j = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$$

Auf Grund der expliziten Beschreibung arithmetischer Folgen ist

$$x_j = x_1 + (j-1)d$$

Schreibt man s_n zweimal in umgekehrter Reihenfolge untereinander

$$\begin{array}{l} s_n = \quad x_1 \quad + \quad x_1 + d \quad + \cdots + x_1 + (n-2)d + x_1 + (n-1)d \\ s_n = x_1 + (n-1)d + x_1 + (n-2)d + \cdots + \quad x_1 + d \quad + \quad x_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{array}$$

und addiert vertikal, so liefert jede der n vertikalen Summen den selben Wert $2x_1 + (n-1)d$. Insgesamt ergibt sich also für die doppelte Summe

$$2s_n = n \cdot \big(2x_1 + (n-1)d\big)$$

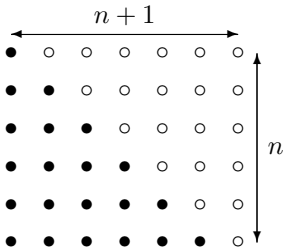
woraus man nach Division durch 2 die gesuchte explizite Summenformel

$$s_n = nx_1 + \frac{n(n-1)}{2}d$$

erhält, die man durch Einsetzen in die Rekursionsgleichung überprüfen kann. Im Spezialafall $x_1 = d = 1$ erhält man für die Summe der natürlichen Zahlen die Formel

$$\sum_{j=1}^{n} j = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

Sie lässt sich leicht geometrisch einsehen, wenn man einen Moment auf die folgende Figur für $n = 6$ starrt,



die bereits den alten Griechen bekannt war.

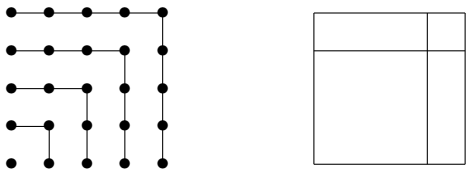
Beim freien Fall gilt $x_1 = \frac{d}{2}$ und damit erhält man für den in den ersten n Etappen zurückgelegten Weg die Beziehung

$$s_n = n^2 \cdot \frac{d}{2}$$

die heute statt den stolzen Worten Galilei's als Fallgesetz bezeichnet wird. Sie lässt sich durch Einsetzen in die Rekursiongleichung überprüfen. Auf Grund der Binomischen Formel ist nämlich

$$\Delta(s_n) = s_{n+1} - s_n = \Big((n+1)^2 - n^2\Big) \cdot \frac{d}{2} = (2n+1) \cdot \frac{d}{2} = x_{n+1}$$

Dass sich die ungeraden Zahlen zu den Quadratzahlen aufsummieren, lässt sich geometrisch an Hand der linken Figur



einsehen, in der die Folge der Quadratzahlen durch Zusammenzählen der durch geradlinig verbundene rechte Winkel zusammengefassten ungeraden Zahlen erhalten werden kann. Selbstverständlich hängt diese Faserung des Quadrates durch Gnomone mit der für den Beweis der Rekursionsgleichung benutzten Binomischen Formel zusammen, die in der rechten Figur geometrisch interpretiert ist.

In den üblichen Einheiten liefert die Folge x_n der pro Zeitintervall zurückgelegten Wege die Folge der mittleren Geschwindigkeiten $v_n = \frac{x_n}{T}$ $[\frac{\text{cm}}{\text{s}}]$. Die Folge der konstanten Geschwindigkeitszuwächse $\Delta(x_n) = d = 0.3924$ liefert die konstante[27] Beschleunigung $g = \frac{\Delta(v_n)}{T} = \frac{\Delta(x_n)}{T^2} = \frac{d}{T^2} = 981[\frac{\text{cm}}{\text{s}^2}]$.

In folgender Figur findet man die Graphen der beiden Folgen v_n und s_n.

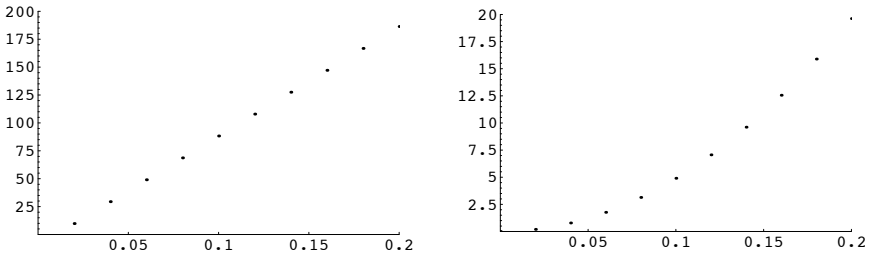


Abbildung 2.17: Geradliniger Verlauf der Geschwindigkeit v_n und quadratischer des zurückgelegten Weges s_n bei der gleichförmig beschleunigten Bewegung.

Offensichtlich wächst die Geschwindigkeit gleichförmig linear und der zurückgelegte Weg nimmt quadratisch zu.

Wiederholt man dieses Experiment mit anderen Zeitintervallen T, so ergeben sich numerisch andere Werte. Man stellt aber experimentell die charakteristische Eigenschaft einer *gleichförmig beschleunigten Bewegung* fest.

3. Der Wert von g ist unabhängig von der Wahl des Zeitintervalls T.

Deshalb ist es möglich und heute üblich, diese Resultate mit Hilfe von Analysis zu formulieren. Dazu fasst man die Zeit wieder kontinuierlich auf. Um die Werte

[27] Präzise Messungen der Fallbeschleunigung haben ergeben, dass sie von der geographischen Breite, von der Höhe über Meer, von der örtlichen Beschaffenheit der Erdrinde abhängt und zeitlich im gleichen Rhythmus wie Ebbe und Flut schwankt.

für einen festen Zeitpunkt $t \in \mathbb{R}_{\geq}$ zu erhalten, teilt man die Zeit so in n gleich lange Intervalle der Länge T ein, dass $t = n \cdot T$ bzw. $n = \frac{t}{T}$ gilt. Dann gilt für den konstanten Geschwindigkeitszuwachs auf Grund der dritten Beobachtung $d = g \cdot T^2$. Entsprechend wurden die Werte auf der horizontalen Achse in obige Graphen nicht einfach nummeriert, sondern in den üblichen Zeiteinheiten angegeben.

Damit erhält man für den mit Hilfe von t ausgedrückten Gesamtweg die von T unabhängige Beziehung

$$s_n(t) = n^2 \cdot \frac{d}{2} = n^2 \cdot \frac{g \cdot T^2}{2} = \frac{t^2}{T^2} \cdot \frac{g \cdot T^2}{2} = \frac{g}{2} \cdot t^2$$

Für die mittleren Geschwindigkeiten erhält man entsprechend

$$v_n(t) = x_n \cdot \frac{1}{T} = (2n-1) \cdot \frac{d}{2} \cdot \frac{1}{T} = (2n-1) \cdot \frac{g \cdot T^2}{2T} = (t - \frac{T}{2})g$$

Für die Momentangeschwindigkeit erhält man also

$$v(t) = \lim_{T \to 0} v_n(t) = \lim_{T \to 0} (t - \frac{T}{2})g = gt$$

Das Fallgesetz wird also heute durch die Formel $s(t) = \frac{g}{2}t^2$ und die Momentangeschwindigkeit durch ihre Ableitung $s'(t) = v(t) = gt$ beschrieben. Die zweite Ableitung beschreibt dann die konstante Beschleunigung $s''(t) = g$. In dieser Sprache handelt es sich bei unseren Daten um eine Messung der Beschleunigung.

Man erkennt an diesem Experiment deutlich, dass in der Praxis nicht kontinuierliche, sondern diskrete Information gemessen wird und dass die Messung von Ableitungen eine missliche Sache ist! Um so mehr wundert man sich, dass Physik im Jahrhundert der Information immer noch in der kontinuierlichen Sprache der Analysis und ihren — auf mystischen unendlich kleinen Grössen oder anspruchsvollen Grenzwerten beruhenden — Konzepten wie Momentangeschwindigkeit und Beschleunigung formuliert wird. Dies um so mehr, als zur numerischen Simulation der Lösung der betreffenden Differentialgleichung sowieso zu einer diskreten Differenzengleichung übergegangen werden muss und die Numerik von Ableitungen voller Ärger ist. Weil Computer in diskreten Zeitschritten arbeiten, können in Computersimulationsmodellen nämlich zeitkontinuierliche Systeme nur zeitdiskret dargestellt werden und müssen dazu zunächst diskretisiert werden. Es ist also schwer verständlich, warum heute von den Anwendern überhaupt noch zeitkontinuierliche Modelle benutzt werden. Eine der wenigen Stellen, wo in der Literatur mit Nachdruck auf diesen Missstand mit den Grundlagen der Physik aufmerksam gemacht wird, ist die amüsante Geschichte I/8-2 in den auch sonst empfehlenswerten "Lectures on Physics" von Feynman. In einem Artikel mit dem Titel "Simulating Physics with Computers" meinte er

> I want to talk about the possibility that there is to be an *exact* simulation, that the computer will do *exactly* the same as nature. If this is to be proved and the type of computer is as I've already explained, then it's going to be necessary that *everything* that happens in a finite volume of space and time would have to be exactly analyzable with a finite number of logical operations. The present theory of physics is not that way, apparently. It allows space to go

> down into infinitesimal distances, wavelength to get infinitely great, terms to be summed in infinite order, and so forth; and therefore, if this proposition is right, physical law is wrong.

Nicht nur ist Information physikalisch, sondern auch Physik informatisch. ◯

Aus diesem Beispiel dürfte klar geworden sein, wie die Verallgemeinerung auf arithmetische Folgen höheren Ordnung aussieht.

Satz. Eine arithmetische Folge k-ter Ordnung x_n erfüllt die lineare, autonome, homogen Rekursionsgleichung $\Delta^{k+1}(x_n) = 0$ der Ordnung $k+1$, die zum charakteristischen Polynom

$$\chi(\lambda) = (1-\lambda)^{k+1} = \sum_{j=0}^{k+1} (-1)^j \binom{k+1}{j} \lambda^j$$

gehört. Eine solche Folge kann explizit durch ein Polynom k-ten Grades

$$x_n = \sum_{j=0}^{k} a_j n^j$$

beschrieben werden.

En passant haben wir die diskrete Version des Hauptsatzes der Differential- und Integralrechnung angetroffen. Bildet man zu einer beliebigen Folge x_n ihre Summenfolge $s_n = \sum_{j=1}^{n} x_j$, so ist die Differenzenfolge dieser Summenfolge bis auf eine Zeitverschiebung die ursprüngliche Folge x_n. Daher gilt insbesondere folgendes Resultat, das im kontinuierlichen Fall der Tatsache entspricht, dass sich beim Integrieren der Grad eines Polynomes um 1 erhöht.

Satz. Die Summenfolge einer arithmetischen Folge k-ter Ordnung ist eine arithmetische Folge $(k+1)$-ter Ordnung.

Beispiel. Wir haben im letzten Beispiel gesehen, dass die Folge der Quadratzahlen $x_n = n^2$ eine arithmetische Folge zweiter Ordnung ist. Durch wiederholte Differenzenbildung erhalten wir die Werte folgender Tabelle:

n	0	1	2	3	4	5	6	7	8	9	10
x_n	0	1	4	9	16	25	36	49	64	81	100
Δx_n	1	3	5	7	9	11	13	15	17	19	
$\Delta^2 x_n$	2	2	2	2	2	2	2	2	2		
$\Delta^3 x_n$	0	0	0	0	0	0	0	0			
$\sum x_n$	0	1	5	14	30	55	81	130	194	275	375

Die Summenfolge der Quadratzahlen ist also eine arithmetische Folge dritter Ordnung und kann explizit durch ein kubisches Polynom

$$s_n = \sum_{j=0}^{n} j^2 = a_0 + a_1 n + a_2 n^2 + a_3 n^3$$

beschreiben werden. Um seine Koeffizienten zu bestimmen, verwenden wir die Anfangswerte der Folge und erhalten durch Einsetzen das lineare Gleichungssystem

$$\left\{\begin{array}{lcl} a_0 & = & 0 \\ a_0+a_1+a_2+a_3 & = & 1 \\ a_0+2a_1+4a_2+8a_3 & = & 5 \\ a_0+3a_1+9a_2+27a_3 & = & 14 \end{array}\right. \qquad \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 2 & 4 & 8 & 5 \\ 3 & 9 & 27 & 14 \end{array}\right)$$

Seine eindeutge Lösung liefert die gesuchten Koeffizienten $a_0 = 0$, $a_1 = \frac{1}{6}$, $a_2 = \frac{1}{2}$ und $a_3 = \frac{1}{3}$, woraus man die gesuchte Summenformel

$$s_n = \frac{1}{6}n + \frac{1}{2}n^2 + \frac{1}{3}n^3 = \frac{n(n+1)(2n+1)}{6}$$

für die Summe der Quadratzahlen erhält. ◯

Wie das letzte Beispiel zeigt, lässt sich eine arithmetische Folge höherer Ordnung umgekehrt leicht aus ihren Differenzenfolgen berechnen. Diese Idee wurde von Babage in seiner difference engine — dem ersten je entworfenen Computer — benutzt, um die Werte von Polynomen für natürliche Zahlen zu berechnen.

Beispiel. Zur automatischen Berechnung der Kubikzahlen beachten wir, dass es sich bei der gesuchten Folge $x_n = n^3$ um eine arithmetische Folge dritter Ordnung handelt. Für ihre sukkzessiven Differenzenfolgen erhalten wir die Werte der Tabelle

n	0	1	2	3	4	5	6	7	8	9	10
x_n	0	1	8	27	64	125	216	343	512	729	1000
Δx_n	1	7	19	37	61	91	127	169	217	271	
$\Delta^2 x_n$	6	12	18	24	30	36	42	48	54		
$\Delta^3 x_n$	6	6	6	6	6	6	6	6			
$\Delta^4 x_n$	0	0	0	0	0	0	0				

Damit lassen sich die Kubikzahlen berechnen, indem man die Tabelle von unten nach oben durch Summieren fortsetzt.

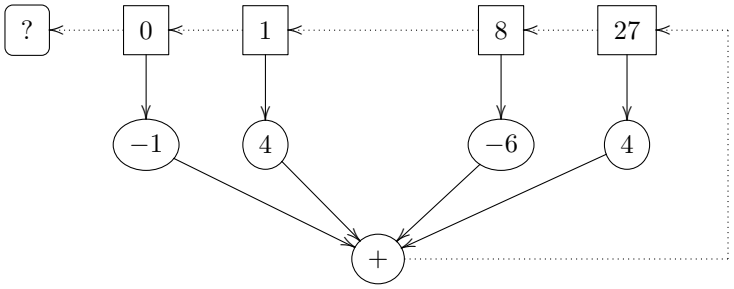
Als arithmetische Folge dritter Ordnung hat die Folge der Kubikzahlen das charakteristischen Polynom

$$\chi(\lambda) = (1-\lambda)^4 = 1 - 4\lambda + 6\lambda^2 - 4\lambda^3 + \lambda^4$$

und erfüllt daher die lineare, autonome, homogen Rekursionsgleichung

$$x_{n+4} = 4x_{n+3} - 6x_{n+2} + 4x_{n+1} - x_n$$

vierter Ordnung. Mit den Anfangswerten $x_0 = 0$, $x_1 = 1$, $x_2 = 8$, $x_3 = 27$ ergibt sich das zugehörige Schieberegister



Man findet es auch, indem man für die Rekursionsgleichung den Ansatz

$$x_{n+4} = \alpha x_{n+3} + \beta x_{n+2} + \gamma x_{n+1} + \delta x_n$$

macht und nun mit Hilfe der Anfangsbedingung die unbekannten Koeffizienten als Lösung eines linearen Gleichungssystems bestimmt. Weil seine Lösung eindeutig bestimmt ist und mit den Koeffizienten obiger Rekursiongleichung übereinstimmt, ist dieses Schieberegister minimal. ◯

Man beachte, dass längst nicht alle Folgen, die sich durch lineare Schieberegister beschreiben lassen, arithmetisch höherer Ordnung sind.

Beispiel. Als typisches Beispiel kehren wir zu den besprochenen Fibonacci-Zahlen zurück. Für sie liefert Differenzenbildung die Tabelle

n	0	1	2	3	4	5	6	7	8	9	10
f_n	0	1	1	2	3	5	8	13	21	34	55
Δf_n	1	0	1	1	2	3	5	8	13	21	34

Ein Blick auf diese Tabelle zeigt, dass die Differenzenfolge der Fibonacci-Folge die geschiftete Fibonacci-Folge liefert. Tatsächlich ist

$$\Delta f_n = f_{n+1} - f_n = (f_n + f_{n-1}) - f_n = f_{n-1}$$

Insbesondere wird daraus klar, dass die Fibonacci-Folge nicht arithmetisch sein kann. Ihr Wachstumsverhalten ist exponentiell und nicht polynomial. ◯

Das folgende Beispiel eines diskreten dynamischen Systems stammt aus der Ökologie und demonstriert eine Klasse wichtiger Anwendungen dynamischer Systeme auf sogn. zeitdiskrete stochastische Prozesse. Dabei geht es um die zeitliche Entwicklung von Wahrscheinlichkeitsverteilungen. Bei diesen stochastischen Prozessen kann das Ergebnis eines Zufallsexperimentes das Ergebnis des nächsten solchen Experimentes beeinflussen.

Beispiel. In einem Wald hat es Bäume der beiden Sorten 1 und 2, über die wir folgende ökologische Modellannahmen[28] treffen.

1. Innerhalb eines Jahres sterben die Bäume der ersten Sorte mit der Sterberate s_1 und jene der Sorte 2 mit der Sterberate s_2.

2. Wenn ein Baum stirbt, wächst an dieser Leerstelle im selben Jahr ein neuer Baum nach. Dabei wächst auf der Leerstelle ein Baum der Sorte 2 mit der Wachstumsrate w_2 und mit der Wachstumsrate w_1 wächst einer der Sorte 1.

Selbstverständlich interessieren wir uns für die Dynamik dieses Ökosystems. Insbesondere möchten wir ihr Langzeitverhalten kennen und wissen, wie sie vom Anfangszustand abhängt.

Den Zustand des System nach k Jahren beschreiben wir durch den Vektor

$$\vec{y}(k) = \begin{pmatrix} x_1(k) \\ x_2(k) \end{pmatrix} \in \mathbb{R}^2$$

[28] Sie sind, wie bei den meisten Anwendungen, vereinfacht. Der Leser möge sich überlegen, wo diese Vereinfachungen liegen und durch welche realistischeren Annahmen sie zu ersetzen sind.

im Zustandsraum $\mathbb{R}^2$. Seine j-te Komponente $x_j(k)$ gibt die Anzahl Bäume der j-ten Sorte nach k Jahren an.

In einem ersten Schritt müssen wir wissen, wie sich die Anzahl Bäume der beiden Sorten von einem Jahr zum nächsten ändert. Dazu suchen wir eine Differenzengleichung, die den Zustand des Waldes nach $k+1$ Jahren, d.h. den Vektor $\vec{y}(k+1)$ rekursiv mit Hilfe des Zustandes im Vorjahr $\vec{y}(k)$ ausdrückt. Im vorliegenden Beispiel wird diese Entwicklung durch die Modellannahmen beschrieben. Von den nach k Jahren $x_1(k)$ lebenden Bäumen der ersten Sorte und den $x_2(k)$ Bäumen der zweiten Sorte sterben im nächsten Jahr insgesamt

$$s(k) = s_1 x_1(k) + s_2 x_2(k)$$

ab und von jeder Sorte überleben

$$\begin{cases} \ddot{u}_1(k) = x_1(k) - s_1 x_1(k) = (1-s_1)x_1(k) \\ \ddot{u}_2(k) = x_2(k) - s_2 x_2(k) = (1-s_2)x_2(k) \end{cases}$$

An Stelle der $s(k)$ abgestorbenen Bäumen wachsen von jeder Sorte

$$\begin{cases} n_1(k) = w_1 s(k) \\ n_2(k) = w_2 s(k) \end{cases}$$

neue nach, so dass die Anzahl der Bäume im nächsten Jahr insgesamt durch das gekoppelte, autonome lineare System von Rekursiongleichungen

$$\begin{cases} x_1(k+1) &= w_1 s(k) + (1-s_1)x_1(k) = (w_1 s_1 + 1 - s_1)x_1(k) + w_1 s_2 x_2(k) \\ x_2(k+1) &= w_2 s(k) + (1-s_2)x_2(k) = w_2 s_1 x_1(k) + (w_2 s_2 + 1 - s_2)x_2(k) \end{cases}$$

beschrieben wird, das wir in matriziellen Form $\vec{y}(k+1) = A \cdot \vec{y}(k)$ schreiben können, falls wir die Systemmatrix

$$A = \begin{pmatrix} 1 - s_1 + s_1 w_1 & w_1 s_2 \\ w_2 s_1 & 1 - s_2 + s_2 w_2 \end{pmatrix}$$

definieren.

Um konkrete Zahlenwerte zu haben, nehmen wir an, innerhalb eines Jahres sterben $s_1 = 1\% = \frac{1}{100}$ der Bäume der Sorte 1 und $s_2 = 5\% = \frac{5}{100}$ der Sorte 2 und mit der Wahrscheinlichkeit $w_1 = 25\% = \frac{1}{4}$ wächst ein Baum der ersten Sorte und mit $w_2 = 75\% = \frac{3}{4}$ einer der zweiten Sorte nach. Dann erhalten wir für die numerische Systemmatrix

$$A = \begin{pmatrix} \frac{397}{400} & \frac{1}{80} \\ \frac{3}{400} & \frac{79}{80} \end{pmatrix}$$

Pflanzen wir ursprünglich im Wald eine Monokultur mit $x_1(0) = 10'000$ Bäumen der Sorte 1, so ist dann der Anfangszustand des Systems

$$\vec{y}(0) = \begin{pmatrix} 10'000 \\ 0 \end{pmatrix}$$

Nach einem Jahr ist der Wald im Zustand

$$\vec{y}(1) = A \cdot \vec{y}(0) = \begin{pmatrix} 9'925 \\ 75 \end{pmatrix}$$

und enthält also 75 Bäume der zweiten Sorte. Nach 57 Jahren ist der Wald im Zustand[29]

$$\vec{y}(57) = A^{57} \cdot \vec{y}(0) = \begin{pmatrix} 7'435.54 \\ 2'564.46 \end{pmatrix}$$

Ein Blick auf das Verhalten des Zustandes des Waldes in den ersten 200 Jahren in den beiden zugehörigen Graphen der linken Teilfigur lässt vermuten, dass sich der Zustand des Waldes nach sehr langer Zeit asymptotisch einem Grenzzustand

$$\vec{y}(\infty) = \lim_{k\to\infty} A^k \cdot \vec{y}(0) = \begin{pmatrix} 6'250 \\ 3'750 \end{pmatrix} = 10'000 \cdot \frac{1}{8} \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

nähert.

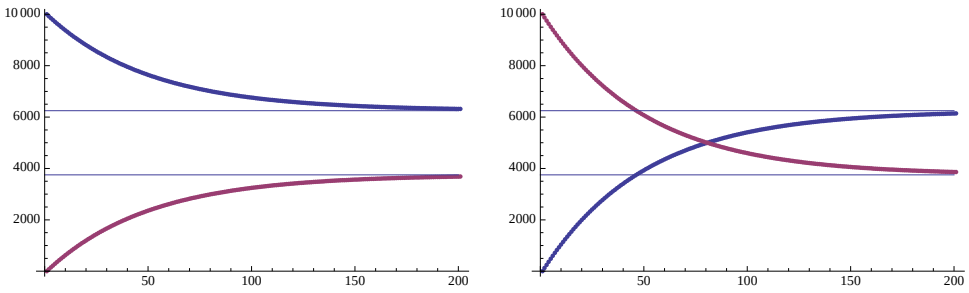


Abbildung 2.18: Entwicklung der Verteilung der beiden Baumsorten im Wald.

Nach sehr langer Zeit ist der Wald im Gleichgewichtszustand und enthält neben 6′250 Bäumen der Sorte 1 noch 3′750 Bäume der Sorte 2.

Interessieren könnte als nächstes, was mit dem Wald passiert, wenn man ihn in einen anderen Anfangszustand versetzt. Schöpfen wir ihn in der anderen Monokultur, indem wir also $x_2(0) = 10'000$ Bäume der zweiten Art pflanzen, erhalten wir mit dem Anfangszustand

$$\vec{y}(0) = \begin{pmatrix} 0 \\ 10'000 \end{pmatrix}$$

nach einem Jahr ein System im Zustand

$$\vec{y}(1) = A \cdot \vec{y}(0) = \begin{pmatrix} 125 \\ 9'875 \end{pmatrix}$$

sein. Ein Blick auf das Verhalten des Zustandes dieses Waldes in den ersten 200 Jahren in den beiden zugehörigen Graphen der rechten Teilfigur zeigt, dass sich das Ökosystem auch diesmal einem Grenzzustand nähert, der überraschenderweise mit jenem mit der anderen Anfangsbedingung übereinzustimmen scheint. Wir vermuten also, dass dieser Grenzzustand von der Anfangsverteilung unabhängig ist.

[29] Selbstverständlich wollen wir hier nicht Holz spalten. Falls der Leser sich durch die auftretenden Brüche gestört fühlt, kann er zur nächsten natürlichen Zahl runden.

Um das dynamische Verhalten des Systems genauer zu untersuchen und um das beobachtete Verhalten zu verstehen, benötigen wir die Potenzen A^k und berechnen dazu das Eigensystem der Systemmatrix A. Dazu gehen wir von der erweiterten Matrix

$$A-\lambda E=\left(\begin{array}{cc|c} \frac{397}{400}-\lambda & \frac{1}{80} & 0 \\ \frac{3}{400} & \frac{79}{80}-\lambda & 0 \end{array}\right), \qquad \left(\begin{array}{cc|c} 397-400\lambda & 5 & 0 \\ 3 & 395-400\lambda & 0 \end{array}\right)$$

des homogenen Systems $(A-\lambda E)\cdot\vec{x}=\vec{0}$ aus, die wir durch Multiplikation mit dem grössten gemeinsamen Nenner der Zeilen ganzzahlig gemacht haben, um die Rechnung zu vereinfachen. Vertauschen der beiden Zeilen liefert

$$\left(\begin{array}{cc|c} 3 & 395-400\lambda & 0 \\ 397-400\lambda & 5 & 0 \end{array}\right)$$

Addition des $-(397-400\lambda)$-fachen der ersten Zeile zum 3-fachen der zweiten liefert die Stufenform

$$B(\lambda)=\left(\begin{array}{cc|c} 3 & 395-400\lambda & 0 \\ 0 & \chi(\lambda) & 0 \end{array}\right)$$

wobei wir für das charakteristische Polynom

$$\chi(\lambda)=-156'800+316'800\lambda-160'000\lambda^2=3'200(-49+99\lambda-50\lambda^2)$$

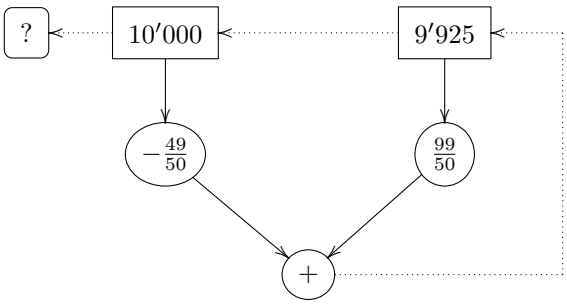
erhalten. Nachdem wir den grössten gemeinsamen Teiler seiner Koeffizienten d.h. $\mathrm{ggT}(156'800, 316'800, 160'000)=3'200$ ausgeklammert haben, ist also noch die quadratische Eigenwertgleichung

$$-49+99\lambda-50\lambda^2=0, \quad \text{bzw.} \quad \lambda^2-\frac{99}{50}\lambda+\frac{49}{50}=0$$

zu lösen und die zeitliche Entwicklung der ersten Baumsorte des Waldes erfüllt die Rekursionsgleichung

$$x_1(k+2)=\frac{99}{50}x_1(k+1)-\frac{49}{50}x_1(k)$$

und lässt sich durch das Schieberegister



beschreiben. Die zweite Baumsorte erfüllt die selbe Rekursionsgleichung aber eine andere Anfangsbedingung.

Die gesuchten Eigenwerte $\lambda_1=1$ und $\lambda_2=\frac{49}{50}$ erhält man als Lösungen der Eigenwertgleichung. Zur Berechnung der zugehörigen Eigenvektoren setzen wir nun die gefundenen Eigenwerte in die Stufenform ein.

1. Fall: $\lambda = 1$. Dann liefert die Stufenform

$$B(1) = \left(\begin{array}{cc|c} 3 & 5 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

den zugehörigen Eigenvektor

$$\vec{v}_1 = \frac{1}{8}\begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{5}{8} \\ \frac{3}{8} \end{pmatrix}$$

den wir so normiert haben, dass seine Spaltensumme 1 ergibt und der daher als Verteilung interpretiert werden kann. Man beachte, dass es sich bei $\vec{v}_1$ um einen Fixpunkt von A handelt, da nach Konstruktion $A \cdot \vec{v}_1 = \vec{v}_1$ gilt. Ferner scheint er sehr mit dem oben experimentell gefundenen Grenzzustand $\vec{y}(\infty)$ zusammenzufallen. Tatsächlich gilt $A \cdot \vec{y}(\infty) = \vec{y}(\infty)$ und ist die Grenzverteilung auch ein Fixpunkt des Systems. Startet das System in diesem Zustand, so bleibt es in diesem stationären Zustand.

2. Fall: $\lambda = \frac{49}{50}$. Dann liefert die zugehörige Stufenform

$$B\Big(\frac{49}{50}\Big) = \left(\begin{array}{cc|c} 3 & 3 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

mit dem zugehörigen Eigenvektoren

$$\vec{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Dieser Eigenvektor kann nicht als Verteilung interpretiert werden, da er Elemente mit unterschiedlichem Vorzeichen enthält. Versetzt man das System in diesen virtuellen Zustand, wird es nach sehr langer Zeit ausgestorben sein.

Bildet man mit Hilfe der beiden Eigenvektoren die Transformationsmatrix

$$X = \begin{pmatrix} 5 & 1 \\ 3 & -1 \end{pmatrix}, \qquad X^{-1} = -\frac{1}{8}\begin{pmatrix} -1 & -1 \\ -3 & 5 \end{pmatrix} = \frac{1}{8}\begin{pmatrix} 1 & 1 \\ 3 & -5 \end{pmatrix}$$

und ihre Inverse, erhält man die Diagonalmatrix

$$\Lambda = X^{-1} \cdot A \cdot X = \begin{pmatrix} 1 & 0 \\ 0 & \frac{49}{50} \end{pmatrix}$$

mit den Eigenwerten auf der Diagonalen. Die zugehörige Diagonalisierung

$$A = X \cdot \Lambda \cdot X^{-1}$$

liefert für die Matrizenpotenz die explizite Beschreibung

$$A^k = X \cdot \Lambda^k \cdot X^{-1} = \frac{1}{8}\begin{pmatrix} 5 + 3\Big(\frac{49}{50}\Big)^k & 5 - 5\Big(\frac{49}{50}\Big)^k \\ 3 - 3\Big(\frac{49}{50}\Big)^k & 3 + 5\Big(\frac{49}{50}\Big)^k \end{pmatrix}$$

Daraus ergeben sich für die Anzahl Bäume im Laufe der Zeit explizite Formeln. Ausgehend von einem beliebigen Anfangszustand

$$\vec{y}(0) = \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix}, \qquad x_1(0) + x_2(0) = 10'000$$

erhalten wir nach k Zeitschritten den Zustand $\vec{y}(k) = A^k \cdot \vec{y}(0)$, der in Komponenten die gesuchten Formeln liefert.

$$\left\{ \begin{array}{lcl} x_1(k) & = & \left(\frac{5}{8} + \frac{3}{8}\left(\frac{49}{50}\right)^k\right) \cdot x_1(0) + \left(\frac{5}{8} - \frac{5}{8}\left(\frac{49}{50}\right)^k\right) \cdot x_2(0) \\ x_2(k) & = & \left(\frac{3}{8} - \frac{3}{8}\left(\frac{49}{50}\right)^k\right) \cdot x_1(0) + \left(\frac{3}{8} + \frac{5}{8}\left(\frac{49}{50}\right)^k\right) \cdot x_2(0) \end{array} \right.$$

Man beachte, dass diese Formeln die allgemeine Gestalt

$$\left\{ \begin{array}{lcl} x_1(k) & = & c_{11}\lambda_1^k + c_{12}\lambda_2^k \\ x_2(k) & = & c_{21}\lambda_1^k + c_{22}\lambda_2^k \end{array} \right.$$

haben. Die vier Konstanten c_{ij} lassen sich also auch durch Einsetzen der Anfangswerte und der bekannten Eigenwerte λ_k bestimmen. Im vorliegenden Fall setzen wir $k = 0, 1$ ein und erhalten die beiden linearen Gleichungssysteme

$$\left\{ \begin{array}{lrcl} k=0: & c_{11} + c_{12} & = & x_1(0), \qquad c_{21} + c_{22} = x_2(0) \\ k=1: & \lambda_1 c_{11} + \lambda_2 c_{12} & = & x_1(1), \qquad \lambda_1 c_{21} + \lambda_2 c_{22} = x_2(1) \end{array} \right.$$

Man beachte, dass die beiden entstandenen Gleichungssysteme die selbe Koeffizientenmatrix

$$VdM(\lambda_1, \lambda_2) = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix}$$

mit der Determinante $\det(VdM) = (\lambda_2 - \lambda_1) \neq 0$ haben und deshalb tatsächlich die eindeutig bestimmte Lösung

$$\begin{array}{lcllcl} c_{11} & = & \frac{x_1(0)\lambda_2 - x_1(1)}{\lambda_2 - \lambda_1}, & c_{12} & = & \frac{x_1(1) - x_1(0)\lambda_1}{\lambda_2 - \lambda_1} \\ c_{21} & = & \frac{x_2(0)\lambda_2 - x_2(1)}{\lambda_2 - \lambda_1}, & c_{12} & = & \frac{x_2(1) - x_2(0)\lambda_1}{\lambda_2 - \lambda_1} \end{array}$$

liefern.

Einsetzen der numerischen Werte liefert im ersten Szenario der $x_1(0) = 10'000$ gepflanzten Bäume der ersten Sorte und der $x_2(0) = 0$ der anderen Sorte für die beiden Folgen die expliziten Darstellungen

$$x_1(k) = 6250 + 3750 \cdot \left(\frac{49}{50}\right)^k, \qquad x_2(k) = 3750 - 3750 \cdot \left(\frac{49}{50}\right)^k$$

Entsprechend findet man bei der anderen Monokultur mit den ursprünglich $x_2(0) = 10'000$ gepflänzten Bäumen der zweiten Sorte und der ursprünglich $x_1(0) = 0$ Bäumen der ersten Sorte die Folgen

$$\tilde{x}_1(k) = 6250 - 6250 \cdot \left(\frac{49}{50}\right)^k, \qquad \tilde{x}_2(k) = 3750 + 6250 \cdot \left(\frac{49}{50}\right)^k.$$

Diese expliziten Formeln für die Populationen lassen sich selbstverständlich auch direkt mit Hilfe der oben bestimmten Matrizenpotenz erhalten.

Mit Hilfe solcher expliziten Formeln lassen sich die aus der Schule bekannten Aufgaben lösen. Will man etwa wissen, nach welcher Zeit bei $10'000$ Bäumen der ersten Sorte noch $8'000$ solche Bäume vorhanden sind, führt das zur transzendenten Gleichung

$$x_1(k) = 8'000 = 6250 + 3750 \cdot \left(\frac{49}{50}\right)^k$$

mit der eindeutigen Lösung

$$k = \left\lceil \frac{\log(15) - \log(7)}{\log(50) - \log(49)} \right\rceil = 38$$

Um im zweiten Szenario zu bestimmen, nach wie langer Zeit sich die Anzahl Bäume der zweiten Sorte halbiert hat, ist die transzendente Gleichung

$$\tilde{x}_2(k) = \frac{\tilde{x}_2(0)}{2}, \qquad 3750 + 6250 \cdot \left(\frac{49}{50}\right)^k = 5000$$

zu lösen. Sie hat die eindeutig bestimmte Lösung

$$k = \left\lceil \frac{\log(5)}{\log(2) + 2\log(5) - \log(49)} \right\rceil = 80$$

Das ist auch gleichzeitig die Antwort auf die Frage, wie lange es dauert, bis im zweiten Szenario gleich viele Bäume beider Sorten vorhanden sind, weil dieses k auch die Gleichung $\tilde{x}_1(k) = \tilde{x}_2(k)$ löst.

Um nun das empirisch festgestellte asymptotische Verhalten zu bestätigen, beachten wir, dass wegen $|\lambda_2| < 1$ die Potenz λ_2^k für grosse k verschwinden wird und sich daher die Matrizenpotenz dem Grenzwert

$$A^\infty = \lim_{k\to\infty} A^k = \frac{1}{8}\begin{pmatrix} 5 & 5 \\ 3 & 3 \end{pmatrix} = \begin{pmatrix} \frac{5}{8} & \frac{5}{8} \\ \frac{3}{8} & \frac{3}{8} \end{pmatrix}$$

nähern wird. Diese Matrix hat den bereits gefundenen Eigenvektor $\vec{v}_1$ als gemeinsamen Spaltenvektor. Daher wird sie einen beliebigen Anfangszustand $\vec{y}(0)$ nach langer Zeit in den selben Gleichgewichtszustand

$$\begin{aligned} \vec{y}(\infty) &= A^\infty \cdot \vec{y}(0) = \frac{1}{8}\begin{pmatrix} 5x_1(0) + 5x_2(0) \\ 3x_1(0) + 3x_2(0) \end{pmatrix} = \frac{1}{8}\begin{pmatrix} 5(x_1(0) + x_2(0)) \\ 3(x_1(0) + x_2(0)) \end{pmatrix} \\ &= (x_1(0) + x_2(0)) \cdot \begin{pmatrix} \frac{5}{8} \\ \frac{3}{8} \end{pmatrix} = 10'000 \cdot \begin{pmatrix} \frac{5}{8} \\ \frac{3}{8} \end{pmatrix} = \begin{pmatrix} 6'250 \\ 3'750 \end{pmatrix} \end{aligned}$$

überführen. Der Grenzzustand ist also attraktiv und stabil.

Man beachte, dass in diesem numerischen Beispiel ein interessanter Erhaltungssatz gilt, indem der Wald zu allen Zeiten gleich viele Bäume enthält. Dieser Erhaltungssatz muss sich, wie alles Interessante dieses Systems, in einer Zusatzbedingung der Systemmatrix ausdrücken lassen. Um sie zu finden, gehen wir von einer beliebigen Systemmatrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Damit erfüllt das System die Rekursiongleichungen

$$\begin{cases} x_1(k+1) &= ax_1(k) + bx_2(k) \\ x_2(k+1) &= cx_1(k) + dx_2(k) \end{cases}$$

Damit die Summe $x_1(k) + x_2(k)$ für alle Zeiten und beliebige Zustände konstant ist, muss für alle Zustände der Erhaltungssatz

$$x_1(k+1) + x_2(k+1) = x_1(k) + x_2(k)$$

gelten. Dieser Erhaltungssatz kann matriziell so formuliert werden, dass man mit dem Spaltenvektor

$$\vec{d} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^2$$

der aus lauter 1 besteht, für jeden Zustand $\vec{y}(k)$ die Bedingung

$$\langle \vec{d}, \vec{y}(k+1) \rangle = \langle \vec{d}, \vec{y}(k) \rangle$$

verlangt. Auf Grund der Rekursion nimmt der Erhaltungssatz die Form

$$x_1(k+1) + x_2(k+1) = (a+c)x_1(k) + (b+d)x_2(k) \stackrel{!}{=} x_1(k) + x_2(k)$$

Weil der Erhaltungssatz insbesondere für die Zustände $\vec{y}(k)$ gelten muss, die durch die Standardbasisvektoren $\vec{e}_1$ und $\vec{e}_2$ beschrieben werden, muss die Systemmatrix A die beiden Bedingungen

$$t_1 = a + c = 1, \qquad t_2 = b + d = 1$$

erfüllen. Umgekehrt folgt aus diesen Bedingungen an die Systemmatrix der Erhaltungssatz. Der beobachtete Erhaltungssatz ist also gleichbedeutend damit, dass die Systemmatrix A lauter Spaltensummen 1 hat, wie man im numerischen Beispiel leicht bestätigt. Matriziell lässt sich diese Bedingung für eine stochastische Matrix durch die Matizengleichung

$$\vec{d}^{\,T} \cdot A = \vec{d}^{\,T}, \qquad A^T \cdot \vec{d} = \vec{d}$$

beschreiben. Ist nun $\vec{y}(k)$ eine Wahrscheinlichkeitsverteilung, so gilt definitionsgemäss $\langle \vec{d}, \vec{y}(k) \rangle = 1$. Daher gilt dann für eine stochastische Matrix A

$$\langle \vec{d}, \vec{y}(k+1) \rangle = \langle \vec{d}, A \cdot \vec{y}(k) \rangle = \langle A^T \cdot \vec{d}, \vec{y}(k) \rangle = \langle \vec{d}, \vec{y}(k) \rangle = 1$$

Damit ist der Erhaltungssatz gezeigt.

Im ursprünglichen ökologischen Beispiel ist die Bedingung des Erhaltungssatzes gleichbedeutend mit der Zusatzbedingung

$$w_1 + w_2 = 1$$

an die Ausgangsdaten, wie man durch eine ökologische Überlegung leicht einsieht.

Solche sogn. *stochastische Matrizen* P, deren sämtliche Elemente positiv und deren sämtliche Spaltensummen 1 sind, spielen in der Anwendung der Matrizenrechnung in der Wahrscheinlichkeitstheorie eine zentrale Rolle. Dort ist es üblich, eine etwas andere Formulierung zu benutzen. Als Zustand des diskreten stochastischen Systems verwendet in unserem Fall die *Verteilung* (stochastischer Vektor) der Bäume nach k Jahren

$$\vec{p}(k) = \begin{pmatrix} p_1(k) \\ p_2(k) \end{pmatrix}$$

Seine j-te Komponente $p_j(k) = \frac{x_j(k)}{x_1(k)+x_2(k)}$ bezeichnet definitionsgemäss den *Anteil* der Bäume der j-ten Sorte. Die Grösse $0 \leq p_j(k) \leq 1$ gibt also die Wahrscheinlichkeit an, nach k Jahren im Wald einen Baum der Sorte j anzutreffen.

Selbstverständlich erfüllt jeder solche Zustand die Eigenschaft

$$\sum_{j=1}^{2} p_j(k) = p_1(k) + p_2(k) = 1$$

einer diskreten Wahrscheinlichkeitsverteilung. Um nun die Dynamik dieser Wahrscheinlichkeitsverteilungen zu untersuchen, braucht man eine Differenzengleichung für die Anteile. Im vorliegenden Fall erhält man aus den Modellannahmen bzw. aus obigem System gekoppelter, autonomner linearer Differenzengleichunger für die Anzahlen $x_1(k)$ und $x_2(k)$ definitionsgemäss die Bedingungen

$$\left\{ \begin{array}{lcl} p_1(k+1) & = & \frac{x_1(k)}{x_1(k)+x_2(k)} = \frac{(w_1s_1+1-s_1)x_1(k)+w_1s_2x_2(k)}{x_1(k)+x_2(k)} \\ p_2(k+1) & = & \frac{x_2(k)}{x_1(k)+x_2(k)} = \frac{w_2s_1x_1(k)+(w_2s_2+1-s_2)x_2(k)}{x_1(k)+x_2(k)} \end{array} \right.$$

die mit der Definition der Anteile in das gekoppelte autonome, lineare System von Rekursionsgleichungen

$$\left\{ \begin{array}{lcl} p_1(k+1) & = & (w_1s_1 + 1 - s_1)p_1(k) + w_1s_2p_2(k) \\ p_2(k+1) & = & w_2s_1p_1(k) + (w_2s_2 + 1 - s_2)p_2(k) \end{array} \right.$$

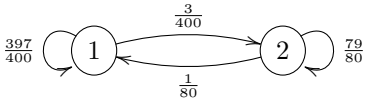
übergeht. Dieses System kann in matrizieller Form

$$\vec{p}(k+1) = P \cdot \vec{p}(k)$$

geschrieben werden, wobei P die bereits oben benutzte Übergangsmatrix

$$P = \begin{pmatrix} \frac{397}{400} & \frac{1}{80} \\ \frac{3}{400} & \frac{79}{80} \end{pmatrix}$$

bezeichnet. Alternativ wird ein solcher diskreter stochastischer Prozess durch den *Übergangsgraphen* beschrieben.



Die Modellannahmen lassen sich also mit folgenden Worten umformulieren:

1. Jedes Jahr werden $\frac{3}{400} = 0.75\%$ der Bäume der ersten Sorte durch Bäume der zweiten Sorte ersetzt und die restlichen $\frac{397}{400} = 99.25\%$ bleiben von der ersten Sorte.

2. Jedes Jahr werden $\frac{1}{80} = 1.25\%$ der Bäume der zweiten Sorte durch Bäume der ersten Sorte ersetzt und die restlichen $\frac{79}{80} = 98.75\%$ bleiben von der zweiten Sorte.

3. Ursprünglich beträgt der Anteil Bäume der ersten Sorte $p_1(0)$ und jener der zweiten Sorte $p_2(0)$.

Allgemein nehmen wir an, dass ein gewisses Experiment auf eine endliche Anzahl Arten $1, 2, \ldots, n$ ausgehen kann. Diese Ausgänge des Experimentes werden als *Zustände* des stochastischen Systems bezeichnet und in einem Graphen durch

seine Knoten markiert. Die Pfeile, die vom Zustand i ausgehen, geben die Übergangszustände an, die das System von i aus annehmen kann. Es gibt also genau dann einen Pfeil von i nach j, wenn das System im nächsten Zeitschritt vom Zustand i in den Zustand j übergehen kann. Das Matrizenelement p_{ji} in der Übergangsmatrix P gibt die Wahrscheinlichkeit an, mit der dieser Übergang stattfindet. Falls der Übergang $i \to j$ nicht möglich ist, setzen wir $p_{ji} = 0$. Deshalb heissen die Matrizenelemente auch *Übergangswahrscheinlichkeiten* des Prozesses und es muss $0 \leq p_{ji} \leq 1$ sein. Die Tatsache, dass die Spaltensumme einer stochastischen Matrix 1 ist, d.h. dass

$$\sum_{j=1}^{n} p_{ji} = 1$$

gilt, widerspiegelt sich im Umstand, dass in der Spalte $\vec{p}_i$ der Matrix P die Wahrscheinlichkeiten aller Übergänge stehen, die vom Zustand i aus möglich sind und dass sich das System irgendwohin entwickeln muss. Die Evolution eines solchen stochastischen Systems wird durch diese Matrix P beschrieben. Beschreiben wir den momentanen Zustand (Wahrscheinlichkeitsverteilung) des Systems zur Zeit k durch den Vektor $\vec{p}(k)$, so ist die Verteilung einen Moment später durch das Produkt

$$\vec{p}(k+1) = P \cdot \vec{p}(k), \qquad p_i(k+1) = \sum_{j=1}^{n} P_{ji} p_j(k)$$

gegeben. Weil die Matrix P zeitunabhängig ist, hängt die Wahrscheinlichkeit, vom Zustand $\vec{p}(k)$ in den Zustand $\vec{p}(k+1)$ überzugehen, nicht von der Geschichte des Prozesses ab. In diesem Sinne hat er kein Gedächtnis.

Man beachte, dass der neue Zustand $\vec{p}(k+1)$ genau dann die Eigenschaft

$$p_1(k+1) + p_2(k+1) = 1 = p_1(k) + p_2(k)$$

einer Wahrscheinlichkeitsverteilung erfüllt, falls der Erhaltungssatz gilt, von dem wir gesehen haben, dass er dazu äquivalent ist, dass die Systemmatrix stochastisch ist. Stochastische Matrizen sind also dadurch ausgezeichnet, dass sie Verteilungen in Verteilungen überführen.

Um die Verteilung der beiden Baumsorten im k-ten Jahr zu bestimmen, müssen wir also aus der Anfangsverteilung $\vec{p}(0)$ die Verteilung $\vec{p}(k) = P^k \cdot \vec{p}(0)$ nach k Zeitschritten bestimmen.

Im konkreten Beispiel der ersten Monokultur ist die Anfangsverteilung

$$\vec{p}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Nach einem Jahr haben die beiden Baumsorten im Wald die Verteilung

$$\vec{p}(1) = P \cdot \vec{p}(0) = \begin{pmatrix} 0.9925 \\ 0.0075 \end{pmatrix}$$

Der Wald enthält also 0.75% Bäume der zweiten Sorte. Nach 57 Jahren haben die Bäume im Wald die Verteilung

$$\vec{p}(57) = P^{57} \cdot \vec{p}(0) = \begin{pmatrix} 0.7435 \\ 0.2564 \end{pmatrix}$$

Man beachte, dass diese Zahlen wegen der Homogenität der Abbildung bis auf einen Faktor 10′000 mit den früher berechneten Anzahl Bäumen übereinstimmen. Insbesondere wird sich nach sehr langer Zeit die Verteilung der Bäume im Wald asymptotisch der Grenzverteilung

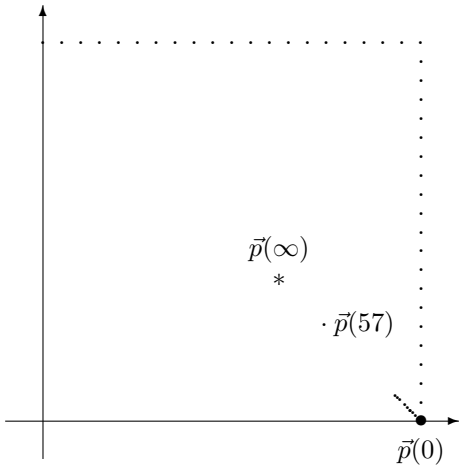
$$\vec{p}(\infty) = \lim_{k\to\infty} P^k \cdot \vec{p}(0) = \begin{pmatrix} 0.6250 \\ 0.3750 \end{pmatrix} = \frac{1}{8}\begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{5}{8} \\ \frac{3}{8} \end{pmatrix}$$

nähern. Dann sind also $\frac{5}{8} = 62.5\%$ der Bäume von der ersten und $\frac{3}{8} = 37.5\%$ von der zweiten Sorte.

Dieses Konvergenzverhalten kann man in der Figur der Verteilungsfolge

$$\vec{p}(0), \vec{p}(1), \ldots, \vec{p}(10), \ldots, \vec{p}(57), \ldots, \vec{p}(\infty)$$

auf dem Standardsimplex im Einheitsquadrat

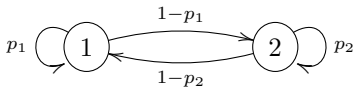


geometrisch beobachten. Die Existenz des Fixpunktes folgt also bereits aus dem Brouwer'schen Fixpunktsatz. Weil der zweite Eigenwert λ_2 betragsmässig recht nahe bei 1 liegt, ist die Konvergenz relativ langsam. ◯

Das am Beispiel beobachtete Verhalten ist für stochastische Matrizen typisch. Eine beliebige stochastische Matrix

$$P = \begin{pmatrix} p_1 & 1-p_2 \\ 1-p_1 & p_2 \end{pmatrix}, \qquad 0 \le p_1, p_2 \le 1$$

die zum Übergangsgraphen



gehört, hat die Determinante

$$\det(P) = p_1 + p_2 - 1, \qquad -1 \le \det(A) \le 1$$

und die Spur

$$\mathrm{tr}(P) = p_1 + p_2, \qquad 0 \le \mathrm{Tr}(A) \le 2$$

Damit gilt für ihr charakteristisches Polynom

$$\chi_P(\lambda) = p_1 + p_2 - 1 - (p_1 + p_2)\lambda + \lambda^2$$

Seine beiden Nullstellen sind aus dem Spektrum $\sigma_P = \{1, \det(P)\}$. Offenbar hat also jede stochastische Matrix P den Eigenwert $\lambda_1 = 1$. Da nämlich nach Voraussetzung $\vec{d}^{\,T} \cdot P = \vec{d}^{\,T}$ gilt, d.h. $\vec{d}^{\,T}$ ein sogn. Linkseigenvektor von P zum Eigenwert 1 ist, muss

$$\vec{d}^{\,T} \cdot (P - E) = (\vec{d}^{\,T} \cdot P) - (\vec{d}^{\,T} \cdot E) = \vec{d}^{\,T} - \vec{d}^{\,T} = \vec{0}$$

sein. Weil daher jede Spaltensumme der Matrix $P - E$ verschwindet, ist die Summe aller Zeilen von $P - E$ die Nullzeile und die Zeilen von $P - E$ sind linear abhängig, d.h. die Matrix $P - E$ ist singulär und ihr Kern verschwindet nicht. Es gibt also einen Vektor $\vec{v} \neq \vec{0}$ mit $(P - E) \cdot \vec{v} = \vec{0}$. Daraus folgt, dass $P \cdot \vec{v} = \vec{v}$ bzw. dass $\lambda = 1$ ein Eigenwert von P sein muss! Eigenwerte stochastischer Matrizen haben die Eigenschaft, dass für sie $|\lambda| \leq 1$ sein muss.

Im Normalfall ist also $\lambda_1 = 1$ betragsmässig echt grösser als alle anderen Eigenwerte und der zugehörige Eigenvektor ist die stationäre Verteilung. Für die zugehörigen Eigenvektoren erhält man

$$\vec{v}_1 = \frac{1}{p_1 + p_2 - 2} \begin{pmatrix} p_2 - 1 \\ p_1 - 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

und für die Matrizenpotenz ist

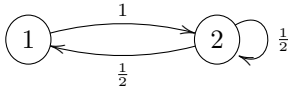
$$P^n = \frac{1}{p_1 + p_2 - 2} \begin{pmatrix} p_2 - 1 + \det^n(A)(p_1 - 1) & (p_2 - 1)(\det^n(A) - 1) \\ (1 - p_1)(\det^n(A) - 1) & p_1 - 1 + \det^n(A)(p_2 - 1) \end{pmatrix}$$

Sie strebt dem Grenzwert

$$P^\infty = \lim_{k \to \infty} P^k = \frac{1}{p_1 + p_2 - 2} \begin{pmatrix} p_2 - 1 & p_2 - 1 \\ p_1 - 1 & p_1 - 1 \end{pmatrix}$$

zu.

Beispiel. Führen wir auf dem Fibonacci-Graphen eine Zufallswanderung durch, indem wir jeden der möglichen Wege mit der selben Wahrscheinlichkeit einschlagen, erhalten wir den Übergangsgraphen



mit der Übergangsmatrix

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

und dem charakteristischen Polynom

$$\chi_P(\lambda) = -\frac{1}{2} - \frac{1}{2}\lambda + \lambda^2$$

und den beiden Eigenwerten aus dem Spektrum $\sigma_P = \{1, -\frac{1}{2}\}$. Starten wir die Wanderung in den beiden Destination mit der Anfangsverteilung $\vec{p}(0)$, so wird sie langfristig in die Grenzverteilung übergehen, die zum Fixpunkt

$$\vec{v}_1 = P^\infty \cdot \vec{p}(0) = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}$$

gehören. Wir werden uns also langfristig rund $\frac{1}{3}$ der Zeit im Knoten 1 aufhalten und die restliche Zeit im Knoten 2 sein.

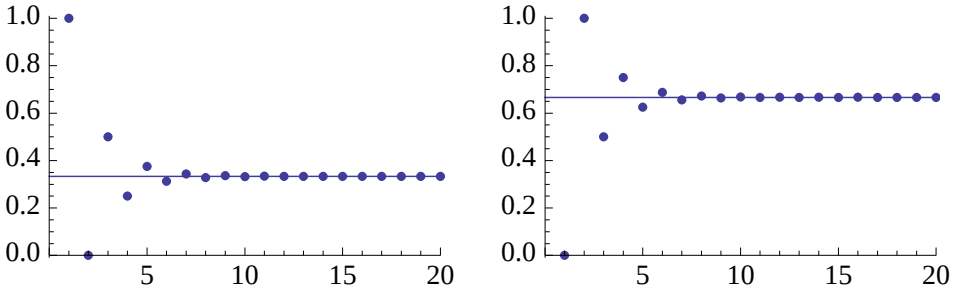


Abbildung 2.19: Aufenthaltswahrscheinlichkeiten bei der gleichverteilten Fibonacci-Wanderung.

Diese Zufallswanderungen entsprechen im bekannten Fibonacci-Baum einem Wahrscheinlichkeitsmass auf den Zweigen gemäss folgender Figur.
Aus der Matrizenpotenz

$$P^3 = \begin{pmatrix} \frac{1}{4} & \frac{3}{8} \\ \frac{3}{4} & \frac{5}{8} \end{pmatrix}$$

lesen wir ab, dass die Wahrscheinlichkeit, in drei Schritten vom Zustand 1 in den Zustand 1 zurückzukommen

$$p_{11}^3 = \frac{1}{4} = 1 \cdot \frac{1}{2} \cdot \frac{1}{2}$$

beträgt. Das heisst, dass bei einer sehr grossen Anzahl Anzahl Widerholungen von Irrfahrten der Länge 3 rund 25% entlang des einzigen Pfades von 1 nach 1 verlaufen werden. Entsprechend gilt

$$p_{12}^3 = \frac{3}{4} = \left(1 \cdot \frac{1}{2} \cdot 1\right) + \left(1 \cdot \frac{1}{2} \cdot \frac{1}{2}\right) = 1 - \frac{1}{4}$$

Daher werden rund 75% solcher Irrfahrten im Zustand 2 enden. Offensichtlich werden also die Wahrscheinlichkeiten entlang eines Pfades multipliziert und jene, die zu unterschiedlichen Pfaden gehören, die zum selben Zustand führen, werden addiert.

Entsprechend enthält die Matrizenpotenz

$$P^5 = \begin{pmatrix} \frac{5}{16} & \frac{11}{32} \\ \frac{11}{16} & \frac{21}{32} \end{pmatrix}$$

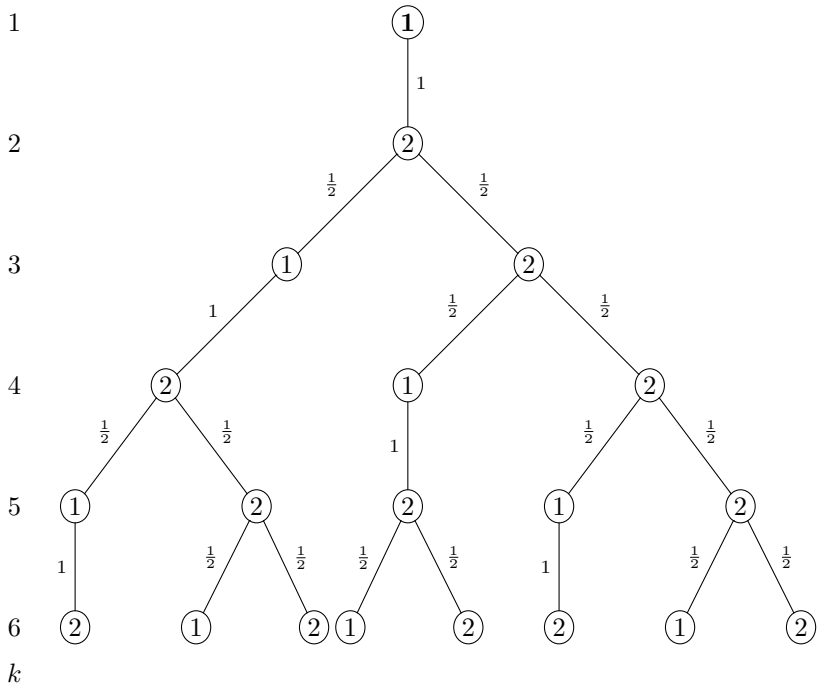


Abbildung 2.20: Das gleichverteilte Mass auf dem Fibonacci-Baum.

die Information, dass die Wahrscheinlichkeit in einem Wege der Länge 5 vom Zustand 1 zu sich selber zurückzukehren

$$p_{11}^5 = \frac{5}{16} = \left(1 \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{2}\right) + \left(1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{2}\right) + \left(1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$$

beträgt und das System bei einer solchen Irrfahrt entsprechend mit der Wahrscheinlichkeit

$$p_{12}^5 = \frac{11}{16} = \left(1{\cdot}\frac{1}{2}{\cdot}1{\cdot}\frac{1}{2}{\cdot}1\right){+}\left(1{\cdot}\frac{1}{2}{\cdot}1{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}\right){+}\left(1{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}{\cdot}1{\cdot}\frac{1}{2}\right){+}\left(1{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}{\cdot}1\right){+}\left(1{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}{\cdot}\frac{1}{2}\right)$$

in den Zustand 2 übergeht. Wegen mangelnder Alternativen ist es die Gegenwahrscheinlichkeit $p_{12}^5 = \frac{11}{16} = 1 - \frac{5}{16}$.

Andere Übergangswahrscheinlichkeiten auf dem selben Übergangsgraphen führen selbstverständlich zu einem anderen Baum-Mass. Die Probleme der Wahrscheinlichkeitsrechnung auf einem Mass-Baum werden aber auch dann zweckmässig mit den beiden Pfadregeln gelöst.

1. Die Wahrscheinlichkeit eines Pfades ist gleich dem Produkt aller Wahrscheinlichkeiten längs des Pfades.

2. Die Wahrscheinlichkeit, eine Zustandsmenge T zu treffen ist gleich der Summe der Wahrscheinlichkeiten aller Pfade, die nach T führen.

Dass sich die Verteilung in diesem Beispiel, in dem wir im Knoten 1 zur Wanderung gestartet sind, der Grenzveteilung alternierend nähert, hängt damit zusammen, dass der zweite Eigenwert λ_2 im Gegensatz zum früher betrachteten

Wald diesmal ein negatives Vorzeichen hat und daher die explizite Formel für die Komponentenfolgen

$$\begin{cases} p_1(k) &= c_{11}\lambda_1^k + c_{12}\lambda_2^k = \frac{1}{3} + c_{12}\Big(-\frac{1}{2}\Big)^k, \quad c_{12} = \frac{1}{6} \\ p_2(k) &= c_{21}\lambda_1^k + c_{22}\lambda_2^k = \frac{2}{3} + c_{22}\Big(-\frac{1}{2}\Big)^k, \quad c_{22} = -\frac{1}{6} \end{cases}$$

alternieren werden. Die Konvergenz ist schneller als im früheren Beispiel, weil dieser Eigenwert betragsmässig kleiner ist. ◯

Im Sonderfall, wo ein weiterer Eigenwert den Betrag $|\lambda| = 1$ hat, muss also $\det(P) = \pm 1$ sein.

Im ersten Sonderfall $p_1 = p_2 = 1$ ist die Übergangsmatrix

$$E_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

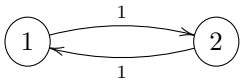
die Einheitsmatrix mit dem zugehörigen Übergangsgraphen, der in zwei Zusammmenhangskomponenten zerfällt.

Unter diesem Prozess bleibt jede Verteilung fix. Weil es nicht möglich ist, von jedem Zustand in jeden anderen überzugehen, ist dieses stochastische System nicht ergodisch.

Im anderen Sonderfall $p_1 = p_2 = 0$ ist die Übergangsmatrix

$$T_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

und der zugehörige Übergangsgraph ist



Dieses System oszilliert zwischen den beiden Zuständen hin und her und die einzige fixe Verteilung ist die Gleichverteilung. Dieser Prozess ist zwar ergodisch aber nicht regulär in dem Sinn, dass keine Potenz der Übergangsmatrix je strikt positive Elemente hat. Das typische Verhalten dieses Systems erkennt man, wenn man es im Anfangszustand $\vec{e}_1$ startet. Dann ist $T_2 \cdot \vec{e}_1 = \vec{e}_2$ und das System osziliert zwischen diesen beiden Zuständen hin und her, was nicht verwunderlich ist, wenn man beachtet, dass T_2 geometrisch als Spiegelung an der ersten Winkelhalbierenden beschreibt. Das periodische Verhalten dieses Systems lässt sich leicht numerisch beobachten.

k	0	1	2	3	4	5	...
x_k	0	1	0	1	0	1	...
y_k	1	0	1	0	1	0	...

Die Komponenten der zugehörigen Vektorfolge erfüllten die lineare Rekursion $x_{k+2} = x_k$ mit dem charakteristischen Polynom $\lambda^2 - 1$. Die Übergangsmatrix

T_2 kann als Begleiterpolynom dieses Polynoms aufgefasst werden und ist wegen $T_2^2 = E$ involutorisch. Sie hat die beiden Eigenwerte $\lambda_1 = 1$ und $\lambda_2 = -1$ mit den zugehörigen Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Sie bleiben bei der Geradenspiegelung T_2 fix, bzw. gehen in ihren Negativen über, weil sie auf dem Spiegel liegen bzw. senkrecht zu ihm stehen. Die beiden Eigenvektoren bilden die Spalten der Transformationsmatrix

$$H_2 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

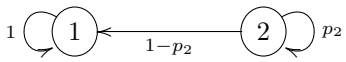
die als Hadamard-Transformation bekannt ist und in vielen Anwendungen eine zentrale Rolle spielt, weil sie als Drehung um den Winkel $\varphi = \frac{\pi}{2}$ die Ebene maximal stark verdreht. Wir werden später diesen Übergangsgraphen auf beliebige reguläre n-Ecke verallgemeinern und durch Untersuchung der zugehörigen Systemmatrix T_n, die den zyklischen Shift beschreibt, analog auf die diskrete Fourier-Transformation F_n stossen, die also eng mit der Hadamard-Transformation verwandt ist, aber im Gegensatz zu ihr transzendente Zahlen und transzendente Funktionen erfordert, die in einem Computer nicht exakt dargestellt werden können.

Ein weiterer Spezialfall spielt bei Studium stochastischer Systeme noch eine wichtige Rolle.

Falls $p_1 = 1$ und $p_2 \neq 1$ ist, hat die Übergangsmatrix

$$P = \begin{pmatrix} 1 & 1-p_2 \\ 0 & p_2 \end{pmatrix}, \qquad 0 \leq p_2 < 1$$

die zum Übergangsgraphen



zwar einen Eigenwert, der betragsmässig echt kleiner als 1 ist. Dieses stochastische System ist nicht ergodisch, hat aber einen absorbierenden Zustand, den es nie mehr verlassen kann und in dem es früher oder später landen wird.

Bisher haben wir autonome lineare Prozesse betrachtet, bei denen der Anfangszustand $\vec{a}$ nach k Schritten in den Zustand $\vec{y}(k) \in \mathbb{R}^n$ übergeht und die durch die lineare, homogenen Rekursionsgleichung $\vec{y}(k+1) = A \cdot \vec{y}(k)$ beschrieben werden. In den Anwendungen treten gelegentlich Prozesse auf, die von aussen durch eine zeitabhängige Störung $\vec{s}(k) \in \mathbb{R}^n$ gestört werden und die deshalb der inhomogen linearen Rekursionsgleichung

$$\vec{y}(k+1) = A \cdot \vec{y}(k) + \vec{s}(k), \qquad \vec{y}(0) = \vec{a}$$

genügen. Dabei nehmen wir an, der zeitliche Verlauf der Störung, d.h. die Vektorfolge $k \mapsto \vec{s}(k)$ sei bekannt. Um ein Gefühl für die Dynamik dieses allgemeineren Prozesses zu erhalten, berechnen wir rekursiv die ersten paar Zustände

und erhalten durch Einsetzen

$$\begin{aligned}
\vec{y}(0) &= \vec{a} \\
\vec{y}(1) &= A \cdot \vec{y}(0) + \vec{s}(0) = A \cdot \vec{a} + \vec{s}(0) \\
\vec{y}(2) &= A \cdot \vec{y}(1) + \vec{s}(1) = A^2 \cdot \vec{a} + A \cdot \vec{s}(0) + \vec{s}(1) \\
\vec{y}(3) &= A \cdot \vec{y}(2) + \vec{s}(2) = A^3 \cdot \vec{a} + A^2 \cdot \vec{s}(0) + A \cdot \vec{s}(1) + \vec{s}(2) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

Daraus erkennen wir bereits das gesuchte Muster und erhalten für die Lösung des inhomogenen Problems

$$\vec{y}(k) = A^k \cdot \vec{a} + \sum_{j=0}^{k-1} A^{k-j-1} \cdot \vec{s}(j) = A^k \cdot \vec{a} + \vec{p}(k), \qquad k \geq 1$$

Der erste Summand beschreibt das durch die Anfangsbedingung $\vec{a}$ verursachte freie (homogene) Systemverhalten, das durch Abschalten der äusseren Störung entsteht. Dazu kommt dann beim gestörten (inhomogenen) System als zweiter Summand noch die sogn. Partikulärlösung in Form einer Summe

$$\vec{p}(k) = \sum_{j=0}^{k-1} A^{k-j-1} \cdot \vec{s}(j) = A^k \cdot \sum_{j=0}^{k-1} A^{-(j+1)} \cdot \vec{s}(j) = A^k \cdot \vec{r}(k)$$

Die Bezeichnung für diesen Vektor ist gerechtfertigt, weil er die Eigenschaften

$$\vec{p}(k+1) = A \cdot \vec{p}(k) + \vec{s}(k), \qquad \vec{p}(0) = \vec{0}$$

hat, wie man leicht überprüft. Er erfüllt also die inhomogene Rekursionsgleichung und eine speziell einfache Anfangsbedingung.

Offenbar ist es im inhomogenen Fall nicht mehr so leicht, den gewünschten Zustand $\vec{y}(k)$ mit Hilfe der Matrizenpotenz A^k allein zu berechnen, weil man zur Lösung des zugehörigen homogenen Systems eben noch eine Partikulärlösung bestimmen und dazu eine Summe berechnen muss. Die explizite Berechnung dieser Summe kann, je nach der Gestalt der Störung, schwierig sein.

Im zeitkontinuierlichen Fall lautet das entsprechende lineare, inhomogene System von Differenzialgleichungen mit konstanten Koeffizienten

$$\vec{y}\,'(t) = A \cdot \vec{y}(t) + \vec{s}(t), \qquad \vec{y}(0) = \vec{a}$$

Auf Grund des diskreten Falls erwarten eine Lösung dieses Systems der Gestalt

$$\vec{y}(t) = e^{At} \cdot \vec{a} + \vec{p}(t).$$

Für die Partikulärlösung $\vec{p}(t)$ machen wir in Anlehnung an den diskreten Fall den Ansatz (sogn.Variation der Konstanten)

$$\vec{p}(t) = e^{tA} \cdot \vec{r}(t)$$

mit der Ableitung

$$\vec{p}\,'(t) = (e^{tA})' \cdot r(t) + e^{tA} \cdot \vec{r}\,'(t) = A \cdot e^{tA} \cdot \vec{r}(t) + e^{tA} \cdot \vec{r}\,'(t) = A \cdot \vec{p}(t) + e^{tA} \cdot \vec{r}\,'(t)$$

Damit es sich dabei um eine Lösung des inhomogenen Systems handelt, muss

$$e^{tA} \cdot \vec{r}\,'(t) = \vec{s}(t), \qquad \vec{r}\,'(t) = (e^{tA})^{-1} \cdot \vec{s}(t) = e^{-tA} \cdot \vec{s}(t)$$

gelten, was dazu führt, dass wir

$$\vec{r}(t) = \int_0^t e^{-A\tau} \cdot \vec{s}(\tau)\, d\tau$$

setzen müssen und sich damit die Partikulärlösung in der Form eines Integrals

$$\vec{p}(t) = e^{At} \cdot \int_0^t e^{-A\tau} \cdot \vec{s}(\tau)\, d\tau = \int_0^t e^{A\cdot(t-\tau)} \cdot \vec{s}(\tau)\, d\tau$$

ausdrücken lässt. Kennt man also den Propagator e^{At} von A, so ist zur Bestimmung der allgemeinen Lösung des Differentialgleichungssystems noch ein Integrationsproblem zu lösen, dass mehr oder weniger anspruchsvoll sein kann und im homogenen Fall $\vec{s}(t) = \vec{0}$ entfällt.

Wir nehmen nun an, die Störungfolge $\vec{s}(k)$ lasse sich ihrerseits durch eine autonome, homogene Rekursionsgleichung der Form

$$\vec{s}(k+1) = B \cdot \vec{s}(k), \qquad \vec{s}(k) = B^k \cdot \vec{s}(0)$$

beschreiben. Damit nimmt die Rekursionsgleichung die matrizielle Form

$$\vec{y}(k+1) = A \cdot \vec{y}(k) + E \cdot \vec{s}(k)$$

einer linearen, inhomogenen Rekursionsgleichung an. Sie wird homogen, indem wir mit dem Zustands- und Störungsvektor den erweiterten Zustand

$$\vec{x}(k) = \begin{pmatrix} \vec{y}(k) \\ \vec{s}(k) \end{pmatrix} \in \mathbb{R}^{2n}$$

der doppelten Dimension definieren. Er erfüllt das System von Rekursionsgleichungen

$$\begin{cases} \vec{y}(k+1) &= A \cdot \vec{y}(k) + E \cdot \vec{s}(k) \\ \vec{s}(k+1) &= B \cdot \vec{s}(k) \end{cases}$$

die matriziell in der homogenen Form

$$\vec{x}(k+1) = M \cdot \vec{x}(k)$$

geschrieben werden kann, falls man als Systemmatrix die Blockmatrix

$$M = \begin{pmatrix} A & E \\ 0 & B \end{pmatrix}$$

verwendet. Durch Übergang zur doppelten Dimension ist also das inhomogene System homogen geworden.

Beispiel. Die wichtigste Reihe überhaupt ist neben der arithmetischen die *geometrische Reihe*, weil sie sich ebenfalls explizit summieren lässt. Wir versuchen

nun, die Glieder einer geometrischen Folge zu summieren, d.h. wir suchen eine Formel für die Summe

$$y(k) = \sum_{j=0}^{k} q^j = 1 + q + q^2 + \cdots + q^n$$

Ideal wäre, wenn wir diese Summe, analog wie seinerzeit die Summe der arithmetischen Folge, durch eine explizite, elementare Formel beschreiben könnten.

Für $q = 1$ ist die Sache trivial, weil dann einfach $y(k) = k + 1$ gilt. In Zukunft werden wir also stillschweigend die Zusatzvoraussetzung $q \neq 1$ machen.

Als Summe erfüllt $y(k)$ die lineare, inhomogene Rekursionsgleichung einer Partialsummenfolge

$$y(k+1) = y(k) + q^{k+1}, \qquad y(0) = 1$$

Die zugehörige homogene Gleichung $y(k+1) = y(k)$ beschreibt einfach die konstanten Folgen. Die Störung ist hier $s(k) = q^{k+1}$ eine geometrische Folge und erfüllt daher die homogene Differenzengleichung

$$s(k+1) = q \cdot s(k), \qquad s(0) = q$$

der geometrischen Folge. Somit erfüllt der erweiterte Zustand

$$\vec{x}(k) = \begin{pmatrix} y(k) \\ s(k) \end{pmatrix}, \qquad \vec{x}(0) = \begin{pmatrix} 1 \\ q \end{pmatrix}$$

die homogene Differenzengleichung

$$\vec{x}(k+1) = M \cdot \vec{x}(k)$$

mit der Systemmatrix

$$M = \begin{pmatrix} 1 & 1 \\ 0 & q \end{pmatrix}$$

deren Potenzen nun zu bestimmen sind. Beispielsweise ist

$$M^2 = \begin{pmatrix} 1 & 1+q \\ 0 & q^2 \end{pmatrix}, \qquad M^3 = \begin{pmatrix} 1 & 1+q+q^2 \\ 0 & q^3 \end{pmatrix}$$

und damit gilt allgemein

$$M^k = \begin{pmatrix} 1 & \sum_{j=0}^{n-1} q^j \\ 0 & q^n \end{pmatrix}$$

Im wesentlichen kommt also die ursprüngliche geometrische Reihe darin vor. Die Matrix M hat das charakteristische Polynom $\chi_M(\lambda) = q - (1+q)\lambda + \lambda^2$ mit den beiden Eigenwerten $\lambda_1 = 1$ und $\lambda_2 = q$ mit den beiden Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ q-1 \end{pmatrix}$$

wie man leicht überprüft. Damit erhalten wir für die Transformationsmatrix und ihre Inverse

$$X = \begin{pmatrix} 1 & 1 \\ 0 & q-1 \end{pmatrix}, \qquad X^{-1} = \frac{1}{q-1} \begin{pmatrix} q-1 & -1 \\ 0 & 1 \end{pmatrix}$$

Die zugehörige Diagonalmatrix lässt sich leicht potenzieren und man erhält

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix}, \qquad \Lambda^k = \begin{pmatrix} 1 & 0 \\ 0 & q^k \end{pmatrix}$$

und damit die gesuchte Potenz

$$M^k = X \cdot \Lambda^k \cdot X^{-1} = \begin{pmatrix} 1 & \frac{q^k-1}{q-1} \\ 0 & q^k \end{pmatrix}$$

Damit lässt sich der Zustand $\vec{x}(k)$ explizit beschreiben. Man erhält

$$\vec{x}(k) = \begin{pmatrix} y(k) \\ s(k) \end{pmatrix} = M^k \cdot \vec{x}(0) = \begin{pmatrix} 1 & \frac{q^k-1}{q-1} \\ 0 & q^k \end{pmatrix} \cdot \begin{pmatrix} 1 \\ q \end{pmatrix} = \begin{pmatrix} \frac{q^{k+1}-1}{q-1} \\ q^{k+1} \end{pmatrix}$$

Ein Koeffizientenvergleich liefert die gesuchte Summenformel

$$y(k) = \sum_{j=0}^{k} q^j = \frac{q^{k+1}-1}{q-1}, \quad q \neq 1$$

Ersetzt man k durch $k-1$, nimmt sie die Form

$$y(k-1) = \sum_{j=0}^{k-1} q^j = \frac{q^k-1}{q-1}$$

an, mit der man die gefundene Matrizenpotenz bestätigt.

Wie bei der Summenformel der arithmetischen Folge geben wir nun für die gefundene Formel noch eine intuitiv einleuchtende Erklärungen, mit der sie sich einfach merken lässt. Dazu schreibt man die Summe $y(k)$ und darunter die mit q multiplizierte Summe $q \cdot y(k)$ aus:

$$\begin{array}{rcl} y(k) & = & 1 + q + q^2 + q^3 + \cdots + q^k \\ q \cdot y(k) & = & \quad q + q^2 + q^3 + \cdots + q^k + q^{k+1} \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \end{array}$$

Subtraktion der oberen Zeile von der unteren und berechnen der entstehenden Teleskopsumme liefert

$$q \cdot y(k) - y(k) = y(k) \cdot (q-1) = q^{k+1} - 1$$

und nach Division durch $q-1$ die oben berechnete Summenformel.

Für $|q| < 1$ konvergiert der Summenwert für wachsendes k, weil dann die Potenzen q^{k+1} schnell verschwinden. Wir erhalten also die geometrische Reihe

$$y(\infty) = 1 + q + q^2 + \cdots = \sum_{j=0}^{\infty} q^j = \frac{1}{1-q}, \qquad |q| < 1$$

Diese Formel lässt sich also algebraisch so merken, dass man die Reihe auf der linken Seite mit q multipliziert. Dann erhält man bis auf den ersten Summanden 1 die selbe Reihe, d.h. es gilt $q \cdot y(\infty) - y(\infty) = 1 = (q-1) \cdot y(\infty)$.

Um die Formel für die geometrische Reihe geometrisch plausibel zu machen, gehen wir vom rechtwinkligen Trapez $T_1 = P_0B_0B_1P_1$ aus, dessen Höhe B_0B_1 und länger Schenkel B_0P_0 die Länge 1 und dessen kürzerer Schenkel B_1P_1 die Länge q hat. Seine Seiten B_0B_1 und P_0P_1 schneiden sich für $0 \leq q < 1$ im Punkt B_∞. An den kürzeren Schenkel von T_1 wird nun ein ähnliches rechtwinkliges Trapez $T_2 = P_1B_1B_2P_2$ geklebt. Sein längerer Schenkel hat also die Länge q und sein kürzerer Schenkel hat aus Ähnlichkeitsgründen die Länge q^2. So fortfahrend erhalten wir eine Folge ähnlichee Trapeze T_n, deren kürzerer Schenkel die Länge $d(B_n, P_n) = q^n$ haben. Diese Längen bilden daher die geometrische Folge $1, q, q^2, \ldots, q^n$. Ferner ist die Gesamtlänge

$$s_n = d(B_0, B_n) = 1 + q + q^2 + \cdots + q^n$$

die gesuchte Summe und

$$s_\infty = d(B_0, B_\infty) = 1 + q + q^2 + \cdots$$

die gesuchte Reihe. Auf der Parallelen zur Geraden B_0B_∞ im Abstand 1 durch den Punkt P_0 befindet sich die Folge der zugehörigen Unterteilungspunkte $Q_1, Q_2, \ldots, Q_n$.

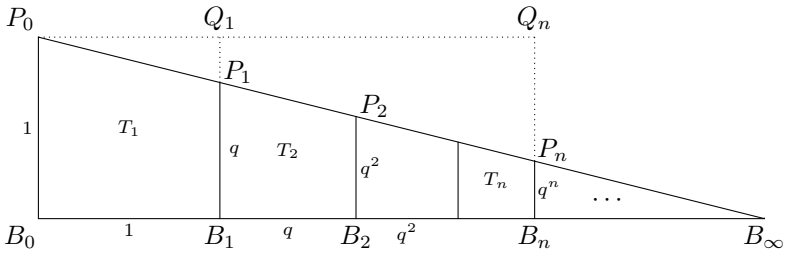


Abbildung 2.21: Begründung der Summenformel für die geometrische Folge.

Um die Summenformel für die geometrische Folge einzusehen, beachten wir, dass das rechtwinklige Dreiecke $D_1 = P_0P_1Q_1$, dessen Katheten die Längen $d(P_1, Q_1) = 1 - q$ und $d(Q_1, P_0) = 1$ haben, zum rechtwinkligen Dreieck $D_n = P_0P_nQ_n$ ähnlich ist, dessen Katheten nach Konstruktion die Längen $d(P_n, Q_n) = 1 - q^n$ und $d(P_0, Q_n) = s_n$ haben. Für die Verhältnisse gilt

$$\frac{s_n}{1 - q^n} = \frac{1}{1 - q}, \qquad s_n = \sum_{j=0}^{k} q^j = \frac{1 - q^n}{1 - q} = \frac{q^n - 1}{q - 1}$$

was der behaupteten Summenformel entspricht.

Um die Formel für die geometrische Reihe einzusehen, beachten wir, dass auch das rechtwinklige Dreieck $B_0B_\infty P_0$ mit den Kathetenlängen 1 und s_∞ zum Dreieck D_1 ähnlich ist. Für das Verhältnis gilt

$$\frac{s_\infty}{1} = \frac{1}{1 - q}, \qquad s_\infty = 1 + q + q^2 + q^3 + \cdots = \frac{1}{1 - q}$$

und damit die behauptete Formel für die geometrische Reihe.

Man findet bereits bei Euklid im Buch IX, Proposition 35 einen Beweis für die Summenformel der geometrischen Folge, der allerdings für einen modernen Leser in ziemlich altertümlicher und umständlicher Sprache formuliert ist. ◯

Beispiel. Die geometrische Reihe wurde in der Schule dazu verwendet, um periodische Dezimalzahlen in gewöhnliche Brüche zu verwandeln. Für $q = \frac{1}{10}$ liefert sie eine Begründung für die Beziehung

$$s_\infty = 1 + \frac{1}{10} + \frac{1}{100} + \cdots = 1.111\ldots = 1.\overline{1} = \frac{10}{9}$$

Nach Multiplikation mit $\frac{9}{10}$ erhält man die Beziehung

$$0.999\ldots = 0.\overline{9} = 1$$

die längst nicht alle Schulabgänger verstanden haben.

Um diese Beziehung zusätzlich etwas zu beleuchten, berechnen wir ihre n-te Partialsumme

$$\frac{9}{10} + \frac{9}{100} + \cdots + \frac{9}{10^n} = 0.\underbrace{99\ldots9}_{n} = \frac{9}{10} \cdot \frac{1 - \frac{1}{10^n}}{1 - \frac{1}{10}} = 1 - \frac{1}{10^n}$$

Die n-stellige Approximation $0.99\ldots9$ liefert tatsächlich nicht ganz 1. Der Unterschied wird aber sehr rasch beliebig klein und der Grenzwert der Folge von Partialsummen ist 1. Weil reelle Zahlen als Äquivalenzklassen konvergenter Folgen definiert sind, wobei zwei Folgen als äquivalent gelten, falls ihre Grenzwerte übereinstimmen, gilt die behauptete Beziehung tatsächlich und die Dezimaldarstellung einer rellen Zahl ist nicht eindeutig bestimmt. ◯

Beispiel. Lässt man eine Stahlkugel aus der Höhe h_0 frei auf eine glattgeschliffene horizontale Marmorplatte fallen, so springt sie wieder hoch, wird aber auf Grund von Reibungsverlusten nur noch den r-ten Teil der ursprünglichen Höhe erreichen. Die Frage stellt sich, wie lange sie so auf und ab hüpft, falls sie jedesmal, nachdem sie auf die Platte gefallen ist, nur noch den r-ten Teil der vorherigen Höhe erreicht.

Bezeichnen wir die erreichte Höhe nach dem n-ten Aufschlag mit h_n, so gilt auf Grund der Modellannahme die Rekursionsgleichung

$$h_{n+1} = r \cdot h_n$$

Die Fallhöhen bilden also eine geometrische Folge mit der sogn. Elastizitätskonstante $0 \leq r \leq 1$ als konstanten Quotienten und können daher explizit durch die Beziehung

$$h_n = h_0 \cdot r^n$$

beschrieben werden.

Um nun die Fallzeiten ins Spiel bringen zu können, erinnern wir uns an den von Galilei beim freien Fall beobachteten Zusammenhang zwischen dem zurückgelegten Weg und der benötigten Zeit

$$h = \frac{g}{2}t^2, \qquad t = \sqrt{\frac{2h}{g}}$$

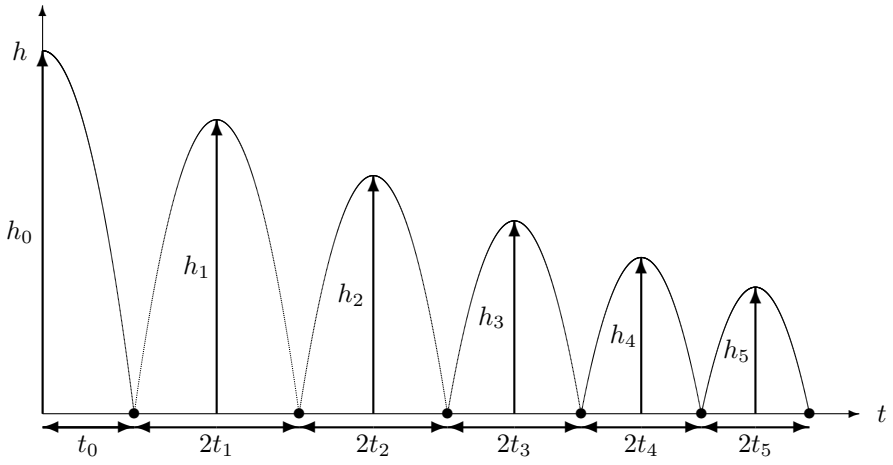


Abbildung 2.22: Weg-Zeit-Diagramm der hüpfenden Kugel für $r = \frac{9}{10}$.

Für die Folge der Fallzeiten erhält man eine weitere geometrische Folge

$$t_n = \sqrt{\frac{2h_n}{g}} = \sqrt{\frac{2h_0 r^n}{g}} = \sqrt{\frac{2h_0}{g}} \cdot \sqrt{r^n} = t_0 \cdot q^n$$

Dabei bezeichnet

$$t_0 = \sqrt{\frac{2h_0}{g}}$$

die Zeit zwischen dem Loslassen der Kugel bis zum ersten Aufschlag. Der konstante Restitutionskoeffizient

$$q = \sqrt{r}$$

beschreibt, wie gut die Kugel hüpft. Er kann experimentell bestimmt werden, indem man das Geräusch, das die Kugel beim Aufschlagen auf der Platte macht, mit einem Mikrophon aufnimmt und die Abstände zwischen den einzelnen diskreten Lautstärke-Maxima analysiert, die im Weg-Zeitdiagramm der Folge der markierten Punkte (•) entsprechen.

Die Gesamtzeit für n Sprünge, nachdem die Kugel das erste Mal auf den Tisch aufschlagen setzt sich aus den Fallzeiten der einzelnen Sprünge zusammen und beträgt, da die Steigzeit und die Fallzeit übereinstimmen

$$T_n = 2t_1 + \cdots + 2t_n = 2t_1(1 + q + q^2 + \cdots + q^n) = 2t_1 \sum_{j=0}^{n} q^n = 2t_1 \frac{q^{n+1} - 1}{q - 1}$$

da es sich dabei um eine gometrische Reihe handelt, die wir explizit aufsummieren können. Benutzen wir $t_1 = t_0 \cdot q$ und setzen fürt t_0 den bestimmten Wert ein, erhalten wir schliesslich die Beziehung

$$T_n = 2\sqrt{\frac{2h_0}{g}} \cdot q \cdot \frac{q^{n+1} - 1}{q - 1}$$

Interessant ist hier die Beobachtung, dass diese Zeit für wachsendes n nicht über alle Grenzen wächst, sondern sich für $q < 1$ rasch dem Grenzwert

$$T_\infty = 2\sqrt{\frac{2h_0}{g}} \cdot q \cdot \frac{1}{1-q}$$

nähert. Spätestens nach dieser endlichen Zeit wird die Kugel unendlich viele Hüpfer gemacht haben und auf der Platte ruhen.

In der Praxis wird man allerdings Hüpfer nicht mehr registrieren können, sobalt sie eine kritische Höhe ε unterschreiten. Es muss also

$$h_n \geq \varepsilon, \qquad h_0 r^n \geq \varepsilon, \qquad r^n \geq \frac{\varepsilon}{h_0}$$

gelten. Es können also höchstens

$$n_0 = \left\lfloor \frac{\log(\varepsilon) - \log(h_0)}{\log(r)} \right\rfloor$$

Hüpfer registriert werden.

Im numerischen Beispiel $h_0 = 80$ [cm] und $r = \frac{81}{100} = 81\%$ ist $q = \frac{9}{10} = 90\%$. Der erste Hüpfer ist $h_1 = 64.8$ [cm] hoch und dauert $2t_1 = 0.73$ [s]. Der Grenzwert ist $T_\infty = 7.27$ [s]. Für die Messgenauigkeit $\varepsilon = 10^{-2}$ [cm] ist $n_0 = 42$ und $T_{42} = 7.19$ [s]. Der letzte messbare Hüpfer dauert $2t_{42} = 9.7 \cdot 10^{-3}$ [s] und ist noch $h_{42} = 1.15 \cdot 10^{-2}$ [cm] hoch.

Bei jeder Kollision mit der Platte verliert die Kugel etwas an Energie und Impuls und die Frage stellt sich, wie sich diese physikalischen Grössen im Lauf der Zeit verhalten. Weil die Gesamtenergie zwischen zwei Hüpfern bei Abwesenheit von Luftreibung konstant bleibt, spielt für die Kugel der Masse m nur die potentielle Energie $E_{\text{pot}}(h) = mgh$ in der Höhe h und die kinetische Energie $E_{\text{kin}}(v) = \frac{1}{2}mv^2$ bei der Geschwindigkeit v eine Rolle und wegen der Energieerhaltung gilt $E_{\text{kin}}(v) = E_{\text{pot}}(h)$ d.h.

$$mgh = \frac{1}{2}mv^2, \qquad v(h) = \sqrt{2gh}$$

Weil sich die maximale Höhe beim n-ten Hüpfer von h_n zu $h_{n+1} = r \cdot h_n$ ändert und dann noch den r-Teil beträgt, ändert sich dabei die potentielle Energie von $E_{\text{pot}}(h_n)$ zu $E_{\text{pot}}(h_{n+1}) = r \cdot E_{\text{pot}}(h_n)$ und beträgt daher ebenfalls noch den r-ten Teil. Die Geschwindigkeit ändert sich dabei von $v(h_n)$ zu $v(h_{n+1}) = \sqrt{r} \cdot v(h_n)$ und beträgt noch den q-ten Teil. Damit ändert sich auch der Impuls $p = mv$ auf den q-ten Teil. Die Elastizitätskonstante r ist der feste Koeffizient in der geometrischen Folgen der Fallhöhen und der Energie bei aufeinanderfolgenden Hüpfern, während sich die Fallzeiten, Geschwindigkeiten und Impulse in Form einer geometrischen Folge mit dem Restitutionskoeffizienten $q = \sqrt{r}$ als konstanten Faktor ändern. ◯

Die für Skalare hergeleitete Summenformel für die geometrische Folge lässt sich auf Matrizen verallgemeinern. Für eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ gilt

$$(A-E)\cdot(E+A+A^2+\cdots+A^k) = A^{k+1}-E = (E+A+A^2+\cdots+A^k)\cdot(A-E)$$

wie man durch Ausmultiplizieren bestätigt. Falls also $A - E$ invertierbar ist, d.h. falls A keinen nicht trivialen Fixpunkt hat, gilt die Summenformel

$$E+A+A^2+\cdots+A^k = \sum_{j=0}^{k} A^j = (A-E)^{-1}\cdot(A^{k+1}-E) = (A^{k+1}-E)\cdot(A-E)^{-1}$$

Falls sämtliche Eigenwerte von A betragsmässig echt kleiner als 1 sind, konvergiert die geometrische Reihe gegen den Grenzwert

$$E + A + A^2 + \cdots = \sum_{j=0}^{\infty} A^j = (E - A)^{-1}$$

der geometrischen Reihe. Es handelt sich wohl neben der früher angetroffenen Exponentialreihe um die zweite fundamentale Potenzreihe, die nun also auch für Matrizen zur Verfügung steht.

2.6 Komplexe Zahlen

Wir haben darauf hingewiesen, dass Matrizen als verallgemeinerte Zahlen betrachtet werden können. Von Zahlen sind wir uns aber etwas mehr gewöhnt als nur die beiden Grundoperationen + und · zusammen mit den Rechenregeln einer Algebra. Wir erwarten, dass zu jedem Element, das von 0 verschieden ist, ein multiplikatives Inverses existiert. Eine Algebra mit dieser Eigenschaft nennt man einen *Schiefkörper*. In einem Schiefkörper gilt also zusätzlich zu den früher angegebenen Regeln einer Algebra:

20. Zu $A \neq 0$ gibt es ein multiplikatives Inverses.

Grob gesagt sind Schiefkörper algebraische Strukturen, in denen man so rechnen d.h. addieren und multiplizieren kann, wie man das in der Grundschule gelernt hat, solange man allerdings die Kommutativität der Multiplikation nicht benutzt. Wenn ein Schiefkörper sogar eine kommutative Multiplikation hat, reden wir von einem *Körper*.

In einem Körper $\mathbb{K}$ gibt es also zwei Operationen $+\colon \mathbb{K}\times\mathbb{K}\to\mathbb{K}$ (Addition) und $\cdot\colon \mathbb{K}\times\mathbb{K}\to\mathbb{K}$ (Multiplikation) für die gilt:

1. Die Addition eines Körpers ist assoziativ, kommutativ und hat ein Neutralelement $0 \in \mathbb{K}$ bezüglich der Addition. Ferner existiert zu jedem Element $a \in \mathbb{K}$ ein additives Inverses $-a$ mit $a + (-a) = 0$.

2. Die Multiplikation eines Körpers ist assoziativ, kommutativ und hat ein multipikatives Neutralelement $1 \in \mathbb{K}$. Ferner existiert zu jedem Element $0 \neq a \in \mathbb{K}$ ein multiplikatives Inverses a^{-1} mit $a \cdot a^{-1} = 1$.

3. Diese beiden Operationen sind distributiv $a\cdot(b+c) = a\cdot b+a\cdot c$ miteinander verknüpft.

Beispiel. Tatsächlich handelt die Sekundarschule vom Körper $(\mathbb{Q}, +, \cdot)$ der rationalen Zahlen, der dann in der Mittelschule mehr oder weniger explizit zum

Körper $(\mathbb{R}, +, \cdot)$ der reellen Zahlen erweitert wird. Die in der Primarschule behandelten natürlichen Zahlen $(\mathbb{N}, +, \cdot)$ und der ganzen Zahlen $(\mathbb{Z}, +, \cdot)$ mit den üblichen Grundoperationen sind noch keine Köper, weil in der Regel kein multiplikatives Inverses existiert.

Die reellen quadratischen Matrizen $\mathbb{R}^{n,n}$ mit der Matrizenaddition und dem Matrizenprodukt bilden für $n \geq 2$ keinen Schiefkörper, weil das multiplikative Inverse fehlt, was damit zusammenhängt, dass Nullteiler vorhanden sind. ◯

2.6.1 Konstruktion

Wir interessieren uns für Körper, die echt grösser sind als der Körper $\mathbb{R}$ der reellen Zahlen. Dieses Interesse begründet sich im Wunsch der italienischen Mathematiker des 14. Jahrhunderts, quadratische und kubische Gleichungen, wie sie in der Matrizenrechnung als charakteristische Gleichungen auftauchen, zu lösen. Solche Gleichungen haben oft keine reellen Lösungen, weil man gelegentlich Quadratwurzeln aus negativen Zahlen ziehen müsste, die bekanntlich unter den reellen Zahlen nicht existieren. Es stellte sich dann überraschenderweise heraus, dass wenn man solche — scheinbar sinnlosen — Objekte ohne mit der Wimper zu zucken hinschreibt und mit ihnen sehr sorgfältig rechnet, man sehr oft auf mirakulöse Weise die gesuchten reellen Lösungen erhält. Diese Objekte spielten also ursprünglich nur in gewissen Zwischenschritten, nicht aber im Endergebnis der Rechnungen, eine Rolle und wurden deshalb als *imaginäre* Zahlen bezeichnet, weil man nicht wusste, was sie sind, aber sich vorstellen konnte, dass sie irgendwie hilfreich sein sollten.

Der Durchbruch gelang dann 500 Jahre später Hamilton, als er erkannte, dass sich alle die benötigten, scheinbar sinnlosen, Objekte in der einheitlichen Normalform $x + yi$ beschreiben und damit einfach als Vektoren in der Ebene interpretieren lassen. Zu seiner sehr grossen Überraschung erkannt er ferner, dass sich auch die Operationen mit diesen Objekten geometrisch interpretieren lassen. Die Addition ist die Vektoraddition und damit nicht sonderlich überraschend. Die viel interessantere Multiplikation entpuppte sich als Drehstreckung: Beim Multiplizieren mit einer beliebigen komplexen Zahl werden die Punkte der Ebene um einen Faktor gestreckt und um einen Winkel gedreht. Es ist in erster Linie diese einfache geometrische Interpretation der komplexen Zahlen und ihrer Operationen, die sie für die Anwendungen unentbehrlich macht.

Auf Grund dieser engen Beziehung der komplexen Zahlen zur Geometrie der Ebene erwarten wir, dass wir die komplexen Zahlen irgendwie als Matrizen vom Typ $\mathbb{R}^{2,2}$ interpretieren können sollten. Viele Matrizen haben keine multiplikative Inversen, da Nullteiler existieren. Im allgemeinen sind also Matrizenringe keine Schiefkörper, denn wir können nicht dividieren. Wir stellen uns als nächstes die subtilere Frage, ob wir vielleicht eine *Teilmenge* $\mathbb{C} \subset \mathbb{R}^{2,2}$ von 2×2-Matrizen finden können, die einen Schiefkörper bildet, wobei wir natürlich die Rechenoperationen der Matrizenrechnung übernehmen wollen. Es soll also gelten:

1. Die Menge $\mathbb{C}$ ist ein Schiefkörper, d.h. jedes von 0 verschiedene Element aus $\mathbb{C}$ ist invertierbar.

2. Die Menge $\mathbb{C}$ ist abgeschlossen unter den Matrizenoperationen.

Die zweite Eigenschaft von $\mathbb{C}$ umfasst folgende Punkte:

- Die beiden Matrizen 0 und E gehören zu $\mathbb{C}$.
- Die Summe zweier Elemente aus $\mathbb{C}$ gehört wieder zu $\mathbb{C}$.
- Das skalare Vielfache eines Elementes aus $\mathbb{C}$ gehört wieder zu $\mathbb{C}$.
- Das Produkt zweier Elemente aus $\mathbb{C}$ gehört wieder zu $\mathbb{C}$.
- Die Inverse eines von 0 verschiedenen Elementes aus $\mathbb{C}$ gehört zu $\mathbb{C}$.

Um die erste Bedingung zu erfüllen, müssen wir die Elemente der Matrix

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

so wählen, dass die Determinante $ad-bc$ nur dann 0 ist, wenn die Elemente alle 0 sind. Dies erreichen wir zum Beispiel auf möglichst einfache Art dadurch, dass wir $d = a$ und $c = -b$ setzen. In diesem Fall lautet nämlich die Determinante $a^2 + b^2$. Dieser Ausdruck ist nur dann Null, wenn beide Zahlen a und b Null sind. Die einzige Matrix mit der Eigenschaft $d = a$ und $c = -b$, die also nicht invertierbar ist, ist die Nullmatrix. Wir werden also dazu geführt, die Matrizen der speziellen Form

$$Z_{a,b} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

zu untersuchen, die wir *komplexe Zahlen* nennen. Man kann eine solche Matrix als positiv orientiertes, orthogonales 2-Bein in der Ebene interpretieren, das aus den beiden Spaltenvektoren von $Z_{a,b}$ besteht, dessen beide Beine senkrecht zueinander stehen und die selbe Länge haben. Tatsächlich gilt für diese Matrizen die charakteristische Eigenschaft

$$Z_{a,b} \cdot Z_{a,b}^T = (a^2 + b^2)\ E_2.$$

Selbstverständlich müssen wir nun noch die zweite Bedingung überprüfen, d.h. untersuchen, ob diese Matrizen unter den Matrizenoperationen abgeschlossen sind. Es ist klar, dass die Nullmatrix 0 als komplexe Zahl aufgefasst werden kann. Sie wird die Rolle der Null spielen. Die Rolle der 1 übernimmt die Einheitsmatrix E, die in der Tat als komplexe Zahl aufgefasst werden kann. Für die Summe zweier komplexen Zahlen und für das skalare Vielfache einer komplexen Zahl ist sofort klar, dass es sich wieder um komplexe Zahlen handelt. Insbesondere kann die negative komplexe Zahl wieder als komplexe Zahl interpretiert werden. Also müssen wir noch untersuchen, ob das Produkt und die Inverse solcher Matrizen wiederum von dieser speziellen Gestalt sind. Das lässt sich aber einfach durch Nachrechnen bestätigen. Für das Produkt zweier komplexer Zahlen erhalten wir die Formel

$$\begin{aligned} Z_{a,b} \cdot Z_{c,d} &= \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} ac-bd & -(ad+bc) \\ ad+bc & ac-bd \end{pmatrix} \\ &= Z_{ac-bd,ad+bc} \end{aligned}$$

Als Produkt ergibt sich in der Tat eine komplexe Zahl. Für die reziproke komplexe Zahl der komplexen Zahl $Z_{a,b}$ müssen wir die inverse Matrix verwenden,

die nach unserer Überlegung immer dann existiert, wenn die komplexe Zahl $Z_{a,b}$ nicht Null ist. Wir erhalten:

$$Z_{a,b}^{-1} = \frac{1}{a^2+b^2} \begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \frac{1}{a^2+b^2} \cdot Z_{a,-b} \qquad (Z_{a,b} \neq 0)$$

Wiederum handelt es sich in der Tat um eine komplexe Zahl. Man beachte, dass auch für komplexen Zahlen Division durch 0 keinen Sinn hat. Unsere Idee, mit Hilfe der Matrizen neue Zahlen zu erklären, scheint also zu funktionieren und führt zur folgenden Definition.

Definition. Unter der Menge der *komplexen Zahlen* verstehen wir die Menge der Matrizen

$$\mathbb{C} = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mid a, b \in \mathbb{R} \right\} \subset \mathbb{R}^{2,2}$$

Man nennt a den *Realteil* und b den *Imaginärteil* der komplexen Zahl $Z_{a,b}$. Beides sind also reelle Zahlen.

Die Grundoperationen d.h. die Addition, die Subtraktion, die Multiplikation und die Division komplexer Zahlen ist, wie oben erläutert, mit Hilfe der entsprechenden Matrizenoperationen erklärt. Die Division ist einzig durch 0 nicht definiert.

Es scheint auf den ersten Blick, als ob wir beim Multiplizieren komplexer Zahlen aufpassen müssten, wenn wir die beiden komplexe Zahlen vertauschen. Das ist erfreulicherweise nicht so, wie wir nachträglich überrascht feststellen.

Satz. Die Multiplikation komplexer Zahlen ist kommutativ.

Beweis. Zum Beweis müssen wir einfach die beiden komplexen Zahlen in anderer Reihenfolge multiplizieren und mit obigem Produkt vergleichen. Es gilt:

$$Z_{c,d} \cdot Z_{a,b} = \begin{pmatrix} c & -d \\ d & c \end{pmatrix} \cdot \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \begin{pmatrix} ac-bd & -(ad+bc) \\ ad+bc & ac-bd \end{pmatrix} = Z_{a,b} \cdot Z_{c,d}$$

Weil das selbe Resulat herauskommt, ist der Beweis erbracht. □

Unsere Überlegungen haben gezeigt, wie man mit komplexen Zahlen rechnet. Ferner zeigen die Rechenregeln der Matrizenalgebra, dass für das Rechnen mit komplexen Zahlen die selben Rechenregeln wie für das Rechnen mit reellen Zahlen gelten. Wir stellen also fest, dass die komplexen Zahlen einen Körper bilden.

Einer der Gründe, warum die komplexen Zahlen gegenüber den reellen Zahlen Vorteile haben liegt darin, dass die komplexen Zahlen mehr innere Symmetrie haben, die wir von den Matrizen her kennen. Wir stellen nämlich fest:

Satz. Zu jeder komplexen Zahl $Z_{a,b}$ ist die Transponierte $Z_{a,b}^T$ wieder eine komplexe Zahl.

Beweis. Für die komplexe Zahl $Z_{a,b} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ erhalten wir

$$Z_{a,b}^T = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} = Z_{a,-b}$$

Diese Matrix liegt erneut in der Menge $\mathbb{C}$ und beschreibt damit eine gewisse komplexe Zahl. □

Man nennt die Transponierte der komplexen Zahl $Z_{a,b}$ die *konjugiert* komplexe Zahl und schreibt $\overline{Z_{a,b}}$ dafür. Aus den Regeln für den Umgang mit der Transposition und der Kommutativität der Multiplikation komplexer Zahlen folgt, dass sich die Konjugation komplexer Zahl mit den Grundoperationen verträgt. Es gilt nämlich:

$$\overline{Z_{a,b} + Z_{c,d}} = \overline{Z_{a,b}} + \overline{Z_{c,d}}, \qquad \overline{Z_{a,b} \cdot Z_{c,d}} = \overline{Z_{a,b}} \cdot \overline{Z_{c,d}}, \qquad \overline{\overline{Z_{a,b}}} = Z_{a,b}$$

Eine solche innere Symmetrie einer Struktur nennt man einen *Automorphismus*. Selbstverständlich gibt es in jedem Körper immer mindestens einen Automorphismus: die Identität. Im Gegensatz zu den reellen Zahlen, die nur den trivialen Automorphismus besitzen, hat der Körper $\mathbb{C}$ der komplexen Zahlen weitere, nicht triviale Automorphismen. Neben der stetigen Konjugation, die den Unterkörper $\mathbb{R} \subset \mathbb{C}$ fest lässt, gibt es sogar überabzählbar viel sogn. wilde Automorphismen, die allerdings nicht durch irgend eine Formel beschrieben werden können.

Man kann die reellen Zahlen als Spezialfälle der komplexen Zahlen auffassen, indem man die folgende Zuordnung betrachtet:

$$\mathbb{R} \to \mathbb{C}, \qquad r \mapsto Z_{r,0} = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}$$

Sie ordnet also jeder reellen Zahl eine komplexe Zahl mit dem Imaginärteil 0 zu. Man stellt fest, dass die Addition und die Multiplikation komplexer Zahlen gerade so erklärt sind, dass sie sich mit den entsprechenden Operationen reeller Zahlen vertragen. Die reellen Zahlen 0 und 1 gehen unter dieser Zuordnung in die Nullmatrix bzw. in die Einheitsmatrix über. Man kann also die reellen Zahlen als Teil der komplexen Zahlen auffassen. Deshalb werden wir in Zukunft die komplexe Zahl $Z_{r,0}$ weiterhin kurz mit r bezeichnen. Die reellen Zahlen sind unter den komplexen Zahlen dadurch charakterisiert, dass für sie $\overline{A} = A$ gilt, d.h. dass ihr Imaginärteil verschwindet.

In einer solchen Situation sagt man, die Zuordnung $\mathbb{R} \to \mathbb{C}$ sei eine *Körpererweiterung*. Eine weitere Körpererweiterung ist bereits aus der Schule geläufig: Die Einbettung $\mathbb{Q} \to \mathbb{R}$, die einem Bruch eine Dezimalzahl zuordnet. Beispielsweise kennen wir von dort die Zuordnung

$$\frac{10}{81} \mapsto 0.12345679\overline{012345679}$$

und die periodischen Dezimalzahlen entsprechen genau den Brüchen.

Mit Hilfe der Konjugation kann jeder komplexen Zahl eine positive reelle Zahl zugeordnet werden.

Definition. Es sei $Z_{a,b}$ eine komplexe Zahl. Unter ihrer *Norm* verstehen wir die reelle Zahl

$$N\big(Z_{a,b}\big) = Z_{a,b} \cdot \overline{Z_{a,b}} \in \mathbb{R}$$

Streng genommen handelt es sich beim Produkt

$$Z_{a,b} \cdot \overline{Z}_{a,b} = Z_{a,b} \cdot Z_{a,b}^T = (a^2 + b^2)E_2$$

um eine gewisse Diagonalmatrix, die wir aber als reelle Zahl $a^2 + b^2 = N(Z_{a,b})$ interpretieren können. Diese reelle Zahl stimmt mit der Determinante der Matrix $Z_{a,b}$ überein und wir erkennen, dass die Norm einer komplexen Zahl als Summe von zwei Quadraten sogar eine positive reelle Zahl ist.

Die Norm ist mit gewissen Grundoperationen verträglich. Zunächst sind Norm und Konjugation verträglich. Es ist nämlich

$$N\big(\overline{Z_{a,b}}\big) = N\big(Z_{a,b}\big)$$

Überraschenderweise ist die Norm mit dem Produkt komplexer Zahlen veträglich. Wie bei den reellen Zahlen, wo bekanntlich

$$a^2 \cdot b^2 = (a \cdot b)^2$$

ist, gilt auch für komplexe Zahlen die fundamentale *Normproduktregel*.

$$N\big(Z_{a,b} \cdot Z_{c,d}\big) = N\big(Z_{a,b}\big) \cdot N\big(Z_{c,d}\big)$$

In Komponenten ausgedrückt besagt die Normproduktregel

$$(a^2 + b^2) \cdot (c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2$$

Aus ihr folgt eine bemerkenswerte Eigenschaft natürlicher Zahlen, die schon vor 2000 Jahren bekannt war[30]. Es gilt nämlich, dass ein Produkt zweier Summen von 2 perfekten Quadraten als Summe von von 2 perfekten Quadraten ausgedrückt werden kann, wie das numerische Beispiel

$$(4 + 16) \cdot (9 + 25) = (6 - 20)^2 + (10 + 12)^2, \quad \text{bzw.} \quad 20 \cdot 34 = 196 + 484$$

illustriert. Die Normproduktregel ist ein nicht trivialer Sachverhalt, der allerdings einfach nachzurechnen ist, wenn man ihn einmal vermutet hat. Matriziell folgt er aus der Tatsache, dass die Multiplikation komplexer Zahlen kommutativ ist bzw. aus der Tatsache, dass die Determinante multiplikativ ist. Man beachte, dass die Normproduktregel äquivalent zum Multiplikationsgesetz der komplexen Zahlen ist.

Weil die Norm $N\big(Z_{a,b}\big) \geq 0$ einer komplexen Zahl $Z_{a,b}$ reell und positiv ist, können wir die Quadratwurzel ziehen.

Definition. Die Quadratwurzel der Norm d.h. der Ausdruck

$$|Z_{a,b}| = \sqrt{N\big(Z_{a,b}\big)} = \sqrt{a^2 + b^2}$$

heisst *Betrag* der komplexen Zahl $Z_{a,b}$.

Im Spezialfall einer reellen Zahl

$$Z_{r,0} = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}$$

[30] Man bezeichnet sie auch als Brahmaguptas Zweiquadrate Identität.

ist $|Z_{r,0}| = |r|$ der aus der Schule bekannte Betrag und daher sind die Notationen verträglich.

Mit Hilfe der Norm lässt sich die Reziproke $Z_{a,b}^{-1}$ einer komplexen Zahl $Z_{a,b}$ ausdrücken. Definitionsgemäss gilt $Z_{a,b}^{-1} \cdot Z_{a,b} = E$. Multiplizieren wir diese Gleichung von rechts mit der Konjugierten $\overline{Z_{a,b}}$, erhalten wir die Beziehung $Z_{a,b}^{-1} \cdot Z_{a,b} \cdot \overline{Z_{a,b}} = \overline{Z_{a,b}}$ bzw. mit der Definition der Norm $Z_{a,b}^{-1} \cdot N\big(Z_{a,b}\big) = \overline{Z_{a,b}}$. Division durch die reelle Zahl $N\big(Z_{a,b}\big)$ liefert die gesuchte Formel

$$Z_{a,b}^{-1} = \frac{1}{N\big(Z_{a,b}\big)} \cdot \overline{Z_{a,b}}, \qquad (Z_{a,b} \neq 0)$$

die wir bereits kennen.

Wegen der Kommutativität der Multiplikation komplexer Zahlen und weil jede komplexe Zahl $Z_{c,d} \neq 0$ eine Inverse hat, können wir für komplexe Zahlen in gewohnter Manier den Quotienten definieren und weiterhin Bruchrechnen wie es sich gehört.

Definition. Es seien $Z_{a,b}$ und $Z_{c,d} \neq 0$ zwei komplexe Zahlen. Unter ihrem *Quotienten* verstehen wir den Bruch

$$\frac{Z_{a,b}}{Z_{c,d}} = Z_{c,d}^{-1} \cdot Z_{a,b} = Z_{a,b} \cdot Z_{c,d}^{-1}, \qquad (Z_{c,d} \neq 0)$$

Neben diesen vertrauten Eigenschaften eines Körpers haben die komplexen Zahlen zusätzliche, völlig neue und überraschende algebraische Eigenschaften. Um zu sehen, worum es geht, nehmen wir den Prototyp einer komplexen Zahl unter die Lupe, die nicht von einer reellen Zahl herkommt.

Satz. Es gibt eine komplexe Zahl I mit der Eigenschaft $I^2 = -E$.

Beweis. Man wähle die komplexe Zahl

$$I = Z_{0,1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

und rechne nach. □

Man nennt I die *imaginäre Einheit* von $\mathbb{C}$. Sie löst die Gleichung $z^2 = -1$. Tatsächlich ist auch $-I$ Lösung dieser Gleichung, wie man leicht verifiziert. Die Existenz einer Zahl I mit der Eigenschaft $I^2 = -1$ galt lange Zeit als fraglich oder gar schockierend. In unserem Zusammenhang ist ihre Existenz selbstverständlich und der eigentliche Schock liegt in der willkürlichen Wahl, mit der wir I vor $-I$ bevorzugen! Treffen nämlich nicht alle Mitmenschen die selbe Wahl, sind Missverständnisse früher oder später unvermeidlich. Immerhin gibt es im Zusammenhang mit der Wahl der imaginären Einheit nur eine Möglichkeit für ein Missverständnis.

Beispiel. Für die imaginäre Einheit I erhalten wie die konjugierte komplexe Zahl

$$\overline{I} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -I$$

und damit $I \cdot \overline{I} = 1 = N(I)$. Daher ist $|I| = 1$ und $|-I| = 1$. ◯

Beispiel. Die weiteren Potenzen der Zahl I lassen sich leicht bestimmen. Es gilt: $I^3 = -I$ und $I^4 = E$ und sie wiederholen sich dann periodisch.
Für den reziproken Wert der imaginären Einheit gilt:

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

Daher sind also die Potenzen der imaginären Einheit periodisch mit der Periode 4. Es gilt

$$I^n = \begin{cases} E & \text{falls } n \equiv 0 \bmod 4 \\ I & \text{falls } n \equiv 1 \bmod 4 \\ -E & \text{falls } n \equiv 2 \bmod 4 \\ -I & \text{falls } n \equiv 3 \bmod 4 \end{cases}$$

Der Leser ist gut beraten, sich diese Formeln zu merken. Sie besagen, dass die vier Matrizen der Menge $\{E, I, -E, -I\}$ mit der Matrizenmultiplikation eine zyklische Gruppe der Ordnung 4 bilden. ◯

Wir stellen also fest, dass in den komplexen Zahlen gewisse Gleichungen Lösungen haben, die in den reellen Zahlen unlösbar sind. Das macht den Körper der komplexen Zahlen besonders angenehm. Zum Beispiel werden wir zeigen, dass jede quadratische Gleichung in den komplexen Zahlen lösbar ist. Dies ist ein Spezialfall des berühmten *Fundamentalsatzes der Algebra*.

Satz. Im Körper $\mathbb{C}$ der komplexen Zahlen besitzt jede Polynomgleichung

$$c_n z^n + \cdots + c_2 z^2 + c_1 z + c_0 = 0, \quad c_n \neq 0$$

mit reellen oder komplexen Koeffizienten $c_k \in \mathbb{C}$ genau n Lösungen, wenn man deren Vielfachheit berücksichtigt.

Der Beweis des Fundamentalsatzes erfordert Überlegungen, die nicht in den momentanen Rahmen passen und ist etwas anspruchsvoll. Deshalb soll er hier nicht geliefert werden. Für $n \leq 4$ gibt es sogar Formeln, um diese Lösungen mit Hilfe von Radikalen, d.h. geschachtelten Wurzelausdrücken zu bestimmen, die aber für die Handrechnung ab $n = 3$ nicht sehr praktisch sind.

Beispiel. Die Gleichung $x^5 - 5x^4 + 5x^3 - 25x^2 + 4x - 20 = 0$ besitzt die einzige reelle Lösung 5, die als komplexe Zahl der Matrix

$$\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} = 5E$$

entspricht. Daneben hat diese Gleichung die vier komplexen Lösungen

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = I, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -I, \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix} = 2I, \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix} = -2I$$

die wegen den reellen Koeffizienten des Polynoms als konjugiert komplexe Paare auftreten, wie der Leser durch Einsetzen leicht nachrechnen kann. ◯

Der Fundamentalsatz der Algebra hat eine Formulierung, die in den Anwendungen oft benutzt wird. Sie besagt, dass die einzigen irreduziblen komplexen Polynome linear sind.

Satz. Jedes Polynom

$$c_n z^n + \cdots + c_2 z^2 + c_1 z + c_0, \quad c_n \neq 0$$

vom Grad n lässt sich in Linearfaktoren zerlegen. Es existieren also (nicht notwendigerweise verschiedene) komplexe Zahlen $z_1, \ldots, z_n \in \mathbb{C}$ so dass

$$c_n z^n + \cdots + c_2 z^2 + c_1 z + c_0 = c_n \cdot (z - z_1) \cdot (z - z_2) \cdots (z - z_n)$$

gilt. Diese Faktorisierung ist bis auf Reihenfolge der Faktoren eindeutig bestimmt. Die Linearfaktoren entsprechen eindeutig den Nullstellen des Polynoms.

In vielen elementaren Lehrbüchern werden die komplexen Zahlen nicht in Matrizenform, d.h. als Matrizen $Z_{a,b}$, sondern entweder in Normalform, d.h. als Linearkombinationen $a + bi$, oder wie in vielen Taschenrechnern in Vektorform, d.h. als Zeilenvektoren $(a, b) \in \mathbb{R}^2$ reeller Zahlen, $a, b \in \mathbb{R}$ erklärt. Das dortige Vorgehen hat den Vorteil, dass in jenen Notation weniger unnötige Information enthalten ist, die geschrieben, bzw. gespeichert werden muss. Es ist aber dann überhaupt nicht klar, wo die mystische Zahl i mit der Eigenschaft $i^2 = -1$ herkommt, bzw. wie man solche Paare multiplizieren soll. Die Matrizendarstellung der komplexen Zahlen durch gewisse 2×2-Matrizen umgeht diese Probleme und lässt alle arithmetischen Operationen mit komplexen Zahlen als selbstverständlich erscheinen, weil dann klar ist, wie das Produkt zweier komplexer Zahlen zu erklären ist.

Die scheinbar verschiedenen Vorgehensweisen, komplexe Zahlen via Matrizen bzw. als Linearkombinationen oder als Vektoren zu definieren, laufen aber auf das Selbe heraus, da sich nämlich jede komplexe Zahl in Matrizenform als Linearkombination bzw. durch die Angabe eines Paares von zwei reellen Zahlen eindeutig festlegen lässt, wie folgende Rechnung zeigt:

$$Z_{a,b} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = a \cdot Z_{1,0} + b \cdot Z_{0,1} = a \cdot E + b \cdot I$$

Die Mathematiker schreiben seit Euler[31] die komplexen Zahlen in Normalform

$$a + bi = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = Z_{a,b} \qquad a, b \in \mathbb{R}$$

Statt E schreibt man dann also kurz $1 \in \mathbb{C}$ und für die imaginäre Einheit I verwendet man das Symbol $i \in \mathbb{C}$ für das dann also

$$i^2 = -1$$

gilt. Wir hätten die Menge der komplexen Zahlen also auch mit Hilfe dieser Normalform als

$$\mathbb{R}[i] = \{a + bi \mid a, b \in \mathbb{R}\}$$

definieren können, wobei zusätzlich die Relation $i^2 = -1$ gilt. Informatiker implementieren den Datentyp der komplexen Zahlen meistens in Vektorform.

[31] 1707 – 1783.

Der Leser gewöhne sich also an, zwischen diesen verschiedenen Darstellungsarten einer komplexe Zahl umzurechnen und sich von Fall zu Fall zu überlegen, in welcher Form die Sache möglichst durchsichtig erscheint. Es sei schon jetzt darauf hingewiesen, dass bald noch zwei weitere solche Darstellungsarten dazukommen werden. Eine komplexe Zahl kann nämlich statt in Matrizenform, in Normalform oder in Vektorform auch in Polarform oder in Exponentialform dargestellt werden. Diese Dartellungsarten haben wiederum für gewisse Zwecke beträchtliche Vorteile. So gesehen besteht die Kunst im Umgang mit komplexen Zahlen vor allem darin, jeweils die passende Darstellungsform zu wählen und zwischen ihnen umrechnen zu können. Gefragt ist also in der Mathematik eine gewisse geistige Beweglichkeit — stures oder pedantisches Verharren steht dem Verständnis nur entgegen.

Die Grundrechnungsarten für komplexe Zahlen lassen sich leicht von der Matrizendarstellung auf die Normalformen übertragen und lauten dann wie folgt.

Für die *Summe* und die *Differenz* von zwei komplexen Zahlen $z_1 = a + bi \in \mathbb{C}$ und $z_2 = c + di \in \mathbb{C}$ gilt

$$z_1 \pm z_2 = (a + bi) \pm (c + di) = (a \pm c) + (b \pm d)i \in \mathbb{C}$$

Das *Produkt* von zwei komplexen Zahlen $z_1 = a + bi \in \mathbb{C}$ und $z_2 = c + di \in \mathbb{C}$ berechnet man mit Hilfe der Produktformel

$$z_1 \cdot z_2 = (a + bi) \cdot (c + di) = (ac - bd) + (ad + bc)i \in \mathbb{C}$$

Zur Berechnung eines Produktes von zwei komplexen Zahlen nach dieser Formel müssen also 4 Produkte und 2 Additionen reeller Zahlen berechnet werden. Dass es auch mit nur 3 reellen Multiplikationen geht, zeigt die magische Formel

$$(a + bi) \cdot (c + di) = (a - b)d + (c - d)a + \Big((a - b)d + (c + d)b\Big)i \in \mathbb{C}$$

die man durch Nachrechnen leicht bestätigt. Allerdings sind dann 5 Additionen erforderlich.

Die *Konjugierte* der komplexen Zahl $z = a + bi \in \mathbb{C}$ ist

$$\overline{z} = a - bi \in \mathbb{C}$$

Der *Realteil* der komplexen Zahl $z = a + bi$ ist $\Re(z) = a$ und der *Imaginärteil* $\Im(z) = b$. Sowohl Real- als auch Imaginärteil einer komplexen Zahl sind reell. Die komplexe Zahl z ist genau dann reell, wenn $\Im(z) = 0$ ist. Eine komplexe Zahl z mit der Eigenschaft $\Re(z) = 0$ nennt man manchmal auch *imaginär*.

Beispiel. Für jede komplexe Zahl $z = a + bi \in \mathbb{C}$ definieren wir die reelle *Norm* $z \cdot \overline{z} = a^2 + b^2 = |z|^2 \geq 0$. Es ist das Quadrat des Betrages $|z| \in \mathbb{C}$. ◯

Man beachte, dass diese Grundoperationen formal wie im Reellen ausgeführt werden. Beim Rechnen mit komplexen Zahlen in Normalform muss man zusätzlich noch die Beziehung $i^2 = -1$ berücksichtigen. Am Schluss der Rechnung wird das Ergebnis in die Normalform $a + bi$ gebracht. Insbesondere gelten die vom Rechnen mit reellen Zahlen gewohnten Rechengesetze (Körper) auch für das Rechnen mit komplexen Zahlen. Aufpassen muss man einzig mit Ungleichungen. Für zwei komplexe Zahlen kann man nicht vernünftig sagen, welche

von ihnen die grössere ist und es macht insbesondere keinen Sinn, eine komplexe Zahl als positiv oder negativ zu bezeichnen.

Beispiel. Es ist $(2+2i)+(4+i) = 6+3i$ und $(3+\frac{1}{2}i)(2-4i)+(i-1) = 6+i-12i-2i^2+i-1 = 7-10i$. Auch das Konjugieren ist einfach. Es gilt etwa $\overline{4+2i} = 4-2i$.

Für die Norm der komplexen Zahl $2+3i$ erhalten wir die reelle Zahl $N(2+3i) = 4+9 = 13$. Damit gilt für ihren Betrag $|2+3i| = \sqrt{4+9} = \sqrt{13}$. ◯

CAS. Selbstverständlich kann man in Sage auch problemlos mit komplexen Zahlen umgehen. Obige Beispiele lassen sich etwa durch den Kode erhalten, der für sich selber spricht. ◇

Weil man mit komplexen Zahlen genau so rechnet wie mit reellen Zahlen, gilt die *Binomische Formel* auch für komplexe Zahlen.

Satz. Für beliebige komplexe Zahlen $x, y \in \mathbb{C}$ und jede natürliche Zahl $n \geq 0$ ist

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

Für eine von Null verschiedene komplexe Zahl $z = a+bi$ schreiben wir für die *reziproke* komplexe Zahl wie üblich $\frac{1}{z}$. Unsere Matrixdarstellung liefert die Normalform

$$\frac{1}{z} = \frac{1}{a+bi} = \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i, \qquad (z \neq 0)$$

Allgemeiner ergibt sich der *Quotient* von zwei komplexen Zahlen $z_1 = x+yi$ und $z_2 = a+bi \neq 0$ in Normalform durch die Formel

$$z_1 \cdot \frac{1}{z_2} = \frac{z_1}{z_2} = \frac{x+yi}{a+bi} = \frac{xa+yb}{a^2+b^2} + \frac{ya-xb}{a^2+b^2}i, \qquad (z_2 \neq 0)$$

Man kann sich diese Formel dadurch merken, dass man beachtet, dass der Nenner bei Multiplikation mit der konjugiert komplexen Zahl $a-bi$ reell wird. Erweitern mit $a-bi$ liefert nämlich

$$\frac{z_1}{z_2} = \frac{x+iy}{a+bi} = \frac{(x+yi)(a-bi)}{(a+bi)(a-bi)}$$

und durch Ausmultiplizieren die behauptete Formel.

Beispiel. Es ist

$$\frac{1}{4-5i} = \frac{4+5i}{(4-5i)(4+5i)} = \frac{4}{41} + \frac{5}{41}i$$

und

$$\frac{8-i}{7-i} = \frac{(8-i)(7+i)}{(7-i)(7+i)} = \frac{57+i}{49+1} = \frac{57}{50} + \frac{1}{50}i$$

Selbstverständlich gelten die üblichen Regeln für das Bruchrechnen. ◯

CAS. Auch die Division funktioniert in Sage wir erwartet. ◇

Der Leser fange nun um Himmels Willen nicht damit an, sich zu fragen, was der i-te Teil eines Kuchens ist. Die komplexen Zahlen dienen anderen Zwecken als zum fairen Teilen von Kuchen unter Geschwistern. Wer Kuchen teilen möchte, kommt mit den positiven rationalen Zahlen aus $\mathbb{Q}$ durchs Leben. Reelle Zahlen wie $\frac{\sqrt{2}}{2}$ oder $-\frac{1}{57}$ beschreiben keine Kuchenstücke. Irrationale Zahlen der Form $\sqrt{2}$ oder π braucht der Banker nicht und auch der Physiker würde besser darüber nachdenken, ob er sie wirklich messen kann und in seinen Modellen haben will. Für einen modernen Naturwissenschaftler existiert nur, was messbar ist; nur das kann der Ingenieur verkaufen und alle anderen Geschichten haben höchstens literarischen oder theologischen Wert. Der Bauer, der seine Schafe zählen will, wird kaum an den negativen Zahlen interessiert sein und der Ackerbauer nicht an der 0. Für ihre Zwecke reichen die strikt positiven natürlichen Zahlen völlig aus. Den heutigen Ingenieur empfehlen die Mathematiker die Verwendung der komplexen Zahlen in erster Linie zur prägnanten Beschreibung von Schwingungsphänomenen oder von Drehungen in der ebenen Geometrie — etwa in der Computer-Graphik. Der morgige Ingenieur wird sie zur Beschreibung von Quantensystemen benötigen. Wer mit den komplexen Zahlen glaubt Mühe zu haben, bedenke die Worte von Titchmarsh:

> There are certainly many people who regard $\sqrt{2}$ as something perfectly obvious, but jib at i. This is because they think they can visualize the former as something in physical space, but not the latter. Actually i is a much simpler concept.

Wir werden bald auch i und $\frac{1}{i} = -i$ und die restlichen komplexen Zahlen geometrisch veranschaulichen.

Es sei allerdings nicht verschwiegen, dass die Verwendung der komplexen Zahlen in der Ingenieur-Mathematik — ausser in der Quantenmechanik — nicht wirklich nötig ist. Kompetente Praktiker können die Grundideen ihres Faches auch reell erläutern und müssen sich nicht hinter dem Gebrauch von komplexem Formelsalat verstecken. Statt allerdings in seriösen Anwendungen mit trigonometrischen Formeln zu rechnen wie wild, empfiehlt es sich es dann, die komplexen Zahlen zu verwenden.

Beispiel. Um den Real- und den Imaginärteil der komplexen Zahl

$$z = \Big(\frac{8-i}{5+i}\Big)^4$$

zu bestimmen, berechnen wir zunächst den Quotienten in Klammern

$$\frac{8-i}{5+i} = \frac{(8-i)(5-i)}{(5+i)(5-i)} = \frac{40-1+i(-5-8)}{25+1} = \frac{1}{2}(3-i)$$

und wenden dann die Binomische Formel an. Es ergibt sich

$$\begin{aligned} z &= \frac{1}{16}(3-i)^4 = \frac{1}{16}\big(3^4 - 4\cdot 3^3 i + 6\cdot 3^2\cdot i^2 - 4\cdot 3 i^3 + i^4\big) \\ &= \frac{1}{16}\big(81 - 54 + 1 + i(-108+12)\big) = \frac{1}{4}(7-24i) = \frac{7}{4} - 6i \end{aligned}$$

Alles wie im Reellen — nur dass halt mehr Geduld erforderlich ist. ◯

CAS. Auch solche Rechnungen lassen sich leicht automatisieren. ◇

Obwohl wir bald eine andere Art kennen lernen werden, wie sich Potenzen einer komplexen Zahl $z = a + ib$ berechnen lassen, wollen wir uns überlegen, was in ihrer Normalform involviert ist. Mit Hilfe der Binomischen Formel erhalten wir

$$z^n = (a+ib)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}(ib)^k = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} \cdot i^k \cdot b^k$$

Wenn wir nun beachten, dass sich für irgend ein Folge a_k die zugehörige Reihe dadurch berechnen lässt, dass man zunächst über die geraden und dann über die ungeraden Glieder summiert, d.h. dass

$$\sum_{k=0}^{n} a_k = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} a_{2k} + \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} a_{2k+1}$$

gilt, so nimmt obige Beziehung die Form

$$z^n = (a+ib)^n = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} i^{2k} \cdot \binom{n}{2k} \cdot a^{n-2k} \cdot b^{2k} + \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} i^{2k+1} \cdot \binom{n}{2k+1} \cdot a^{n-(2k+1)} \cdot b^{2k+1}$$

an. Aus den früher berechneten Potenzen der Zahl i ergeben sich die Werte

$$i^l = \begin{cases} (-1)^l & \text{für } l = 2k \\ (-1)^l \cdot i & \text{für } l = 2k+1 \end{cases}$$

Setzen wir diese Werte in obiger Summe ein, erhalten wir für die gesuchte Normalform der Potenz $z^n = (a+ib)^n$ den Ausdruck

$$z^n = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \cdot \binom{n}{2k} a^{n-2k} \cdot b^{2k} + i \Big(\sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} (-1)^k \cdot \binom{n}{2k+1} a^{n-(2k+1)} \cdot b^{2k+1} \Big)$$

Die Grundoperationen für das Rechnen mit komplexen Zahlen erfüllen folgende Verträglichkeitsbedingungen mit der Konjugation.

Korollar. Für beliebige komplexe Zahlen z_1, z_2, z gilt:

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}, \qquad \overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}, \qquad \overline{\overline{z}} = z$$

$$z \cdot \overline{z} = |z|^2, \qquad |\overline{z}| = |z|$$

Aus $z = \Re(z) + \Im(z)i$ und $\overline{z} = \Re(z) - \Im(z)i$ ergeben sich durch Addition und Subtraktion für den Real- und den Imaginärteil einer komplexen Zahl z die Darstellungen mit Hilfe der Konjugation.

$$\Re(z) = \frac{1}{2}\Big(z + \overline{z}\Big), \qquad \Im(z) = \frac{1}{2i}\Big(z - \overline{z}\Big)$$

Offensichtlich ist $z \in \mathbb{C}$ genau dann reell, falls $z = \overline{z}$ gilt.

Der Betrag hat folgende charakteristischen Eigenschaften eines Betrages.

Korollar. Für alle komplexe Zahlen $z, z_1, z_2 \in \mathbb{C}$ und jede reelle Zahl $r \in \mathbb{R}$ gilt:

1. $|rz| = |r| \cdot |z|$.
2. $|z_1 + z_2| \leq |z_1| + |z_2|$.
3. $|z| = 0$ genau dann wenn $z = 0$.

Mit Hilfe von komplexen Zahlen lassen sich quadratische Gleichungen

$$az^2 + bz + c = 0; \quad a, b, c \in \mathbb{C}, a \neq 0$$

uneingeschränkt lösen. Die Gleichung $z^2 + 1 = 0$ hat bekanntlich in den reellen Zahlen keine Lösung. Solche Lösungen gibt es aber sehr wohl in den komplexen Zahlen, nämlich $z_1 = i$ und $z_2 = -i$. Allgemein besitzt obige quadratische Gleichung immer komplexe Lösungen. Um sie zu bestimmen geht man in gewohnter Manier durch quadratisches Ergänzen vor. Nach Division durch $a \neq 0$ geht die quadratische Gleichung in die normierte Form

$$z^2 + pz + q = 0$$

über, wobei $p = \frac{b}{a}$ und $q = \frac{c}{a}$ ist. Quadratisches Ergänzen liefert nun die Scheitelform

$$\left(z + \frac{p}{2}\right)^2 = \frac{p^2}{4} - q = \frac{p^2 - 4q}{4}$$

Wurzelziehen liefert die beiden (nicht notwendig verschiedenen) komplexen Lösungen

$$\begin{aligned} z_1 &= -\frac{p}{2} + \frac{1}{2}\sqrt{p^2 - 4q} = -\frac{b}{2a} + \frac{1}{2a}\sqrt{b^2 - 4ac} \\ z_2 &= -\frac{p}{2} - \frac{1}{2}\sqrt{p^2 - 4q} = -\frac{b}{2a} - \frac{1}{2a}\sqrt{b^2 - 4ac} \end{aligned}$$

Gedankenloses Wurzelziehen, wie wir es soeben für diese Allerweltsformel praktiziert haben, ist unter Anfängern zu Unrecht beliebt. Sie benützen beispielsweise für die imaginäre Einheit i die Bezeichnung $\sqrt{-1}$ und erwarten dann selbstverständlich, dass die aus der Schule überlieferten Wurzelgesetze weiterhin gelten. Dass dies nicht der Fall ist, zeigt die Paradoxie

$$-1 = i \cdot i = \sqrt{-1} \cdot \sqrt{-1} = \sqrt{(-1) \cdot (-1)} = \sqrt{1} = 1$$

Wenn man nicht genau weiss, was man macht, wird das Resultat im besten Fall sinnlos und im schlechtesten Fall falsch. Wer noch nie die Theorie der Riemannschen Flächen studiert hat, sollte im eigenen Interesse die Finger vom Wurzelzeichen lassen. Wer unbedingt mit dem Wurzelsymbol angeben will, muss sich vorher vergewissern, dass der Radikand eine positive reelle Zahl ist. Die meisten Programmierer von Taschenrechner gehören diesbezüglich in die Klasse der Ignoranten und deshalb ist der Ausgabe der Wurzeltaste nur bei positivem, reellem Radikanden zu trauen.

In unserem Beispiel müssten wir zunächst definieren, was mit *der* Quadratwurzel einer komplexen Zahl $w = u + iv \in \mathbb{C}$ gemeint ist. Das machen wir aber eben nicht, weil man ohne die Theorie der Riemannschen Flächen nicht vernünftig sagen kann, was man unter der Wurzelfunktion versteht. Für die Anwendungen

benötigen man aber zum Glück keine solche Funktion, sondern höchstens die Lösungsmenge der speziellen quadratischen Gleichung

$$z^2 = w$$

Diese Gleichung können wir aber auch lösen, ohne vorher eine komplexe Wurzelfunktion definiert zu haben. Der Anfänger halte also zwei verschiedene Aufgaben sorgfältig auseinander.

- Einerseits kann man sich um die Definition einer *Wurzelfunktion* bzw. um die Definition des Symbols $\sqrt{w}$ bemühen. Darum foutieren wir uns bzw. schicken den neugierigen Leser in eine Bibliothek, um sich dort über Riemann'schen Flächen kundig zu machen.
- Bei der anderen — einfacheren — Aufgabe geht es darum, bei der Vorgabe der komplexen Zahl $w = u + iv \in \mathbb{C}$ die spezielle *quadratische Gleichung* $z^2 = w$ zu lösen.

Zur Bestimmung sämtlicher komplexen Lösungen der Gleichung $z^2 = w$ machen wir für die gesuchte Lösung den Ansatz $z = x + iy \in \mathbb{C}$ und erhalten durch Einsetzen in die spezielle quadratische Gleichung nach dem Ausmultiplizieren

$$z^2 = x^2 - y^2 + i(2xy) = u + iv.$$

Durch Vergleichen von Real- und Imaginärteil erhalten wir daraus das reelle Gleichungssystem

$$\left\{ \begin{array}{rcl} x^2 - y^2 & = & u \\ 2xy & = & v \end{array} \right.$$

Das Problem, die quadratische Gleichung $z^2 = w$ komplex zu lösen, ist damit auf das gleichwertige Problem reduziert, dieses reelle Gleichungssystem zu lösen. Weil es nicht linear ist, müssen wir uns auf einiges gefasst machen. Weil beide Gleichungen quadratisch sind, haben wir überhaupt eine Chance eine symbolische Lösung zu erhalten. Im typischen Fall $u = v = 1$ lauten die beiden Gleichungen

$$x^2 - y^2 = 1, \qquad 2xy = 1$$

Ein Blick auf die zugehörigen Graphen suggeriert, dass es sich dabei um zwei um $\frac{\pi}{4}$ gedrehten Hyperbeln handelt.

Ausgehend von der gestrichelten Hyperbel mit der Gleichung $2xy = 1$, liefert die Drehmatrix für die Drehung mit dem Drehwinkel $\varphi = -\frac{\pi}{4}$

$$D_{-\frac{\pi}{4}} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

die Formeln für die zugehörige Koordinatentransformation

$$\left\{ \begin{array}{rcl} x & = & \frac{\sqrt{2}}{2}\tilde{x} - \frac{\sqrt{2}}{2}\tilde{y} \\ y & = & \frac{\sqrt{2}}{2}\tilde{x} + \frac{\sqrt{2}}{2}\tilde{y} \end{array} \right.$$

Damit wird die gedrehte Hyperbel durch die Gleichung

$$2\Big(\frac{\sqrt{2}}{2}\tilde{x} - \frac{\sqrt{2}}{2}\tilde{y}\Big) \cdot \Big(\frac{\sqrt{2}}{2}\tilde{x} + \frac{\sqrt{2}}{2}\tilde{y}\Big) = 1$$

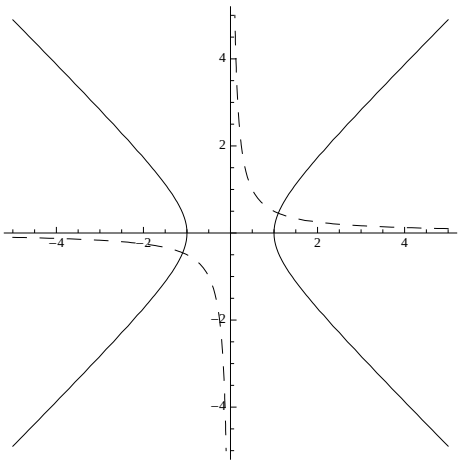


Abbildung 2.23: Graphen der beiden Hyperbeln $x^2 - y^2 = 1$ und $2xy = 1$.

beschrieben, die nach Vereinfachen mit Hilfe der dritten binomischen Formel tatsächlich in die Gleichung

$$\tilde{x}^2 - \tilde{y}^2 = 1$$

der anderen, fett gezeichneten, Kurve der Einheitshyperbel übergeht. Aus der Figur entnimmt man, dass im Normalfall das Problem zwei Lösungen haben wird, die den beiden Schnittpunkten der beiden Hyperbeln entsprechen.

1. Ist w reell, d.h. $v = 0$, so ist auf Grund der zweiten Gleichung $x = 0$ oder $y = 0$.

   (a) Für $u = 0$ und $v = 0$ muss $x^2 = y^2$ sein. Daher kommt dann nur $x = 0$ und $y = 0$ und damit $z = 0$ als Lösung in Frage.

   (b) Für $u > 0$ und $v = 0$ muss $x^2 = u$ und $y = 0$ sein. Die quadratische Gleichung $x^2 = u$ hat dann zwei reelle Lösungen, für die wir dem $\sqrt{\ }$-Süchtigen zuliebe $x = \pm\sqrt{u}$ schreiben. Damit ist $z = \pm\sqrt{u}$. Er hätte wohl nichts anderes erwartet.

   (c) Ist aber $v = 0$ und $u < 0$, so muss $x = 0$ und $y^2 = -u$ sein. In diesem Fall ist $z = \pm\sqrt{-u}i$.

2. Es sei nun w nicht reell, d.h. $v \neq 0$. Durch Quadrieren der beiden Gleichungen und Addieren erhalten wir $x^4 - 2x^2y^2 + y^4 + 4x^2y^2 = u^2 + v^2$ bzw.

   $$(x^2 + y^2)^2 = u^2 + v^2$$

   Weil die rechte Seite positiv ist, können wir auch ihre Lösungsmenge mit Hilfe von Wurzeln beschreiben und erhalten für die Lösung

   $$x^2 + y^2 = \sqrt{u^2 + v^2}$$

   Sie ist eindeutig bestimmt, weil für $x^2 + y^2$ nur positive Zahlen in Frage kommen. Zusammen mit der ersten Gleichung des Gleichungssystems

erhalten wir daraus die beiden positiven (!) Werte

$$x^2 = \frac{1}{2}(\sqrt{u^2+v^2}+u), \qquad y^2 = \frac{1}{2}(\sqrt{u^2+v^2}-u)$$

und daraus schliesslich

$$x = \frac{x}{|x|}\cdot\sqrt{\frac{1}{2}(\sqrt{u^2+v^2}+u)}, \qquad y = \frac{y}{|y|}\cdot\sqrt{\frac{1}{2}(\sqrt{u^2+v^2}-u)}$$

Aus der zweiten Gleichung des Systems folgt die Vorzeichenbedingung

$$\frac{x}{|x|}\cdot\frac{y}{|y|} = \frac{v}{|v|}$$

und daraus die beiden gesuchten Lösungen der Gleichung $z^2 = w$

$$\begin{aligned} z_1 &= \sqrt{\frac{1}{2}(\sqrt{u^2+v^2}+u)} + i\frac{v}{|v|}\cdot\sqrt{\frac{1}{2}(\sqrt{u^2+v^2}-u)} \\ z_2 &= -\Big(\sqrt{\frac{1}{2}(\sqrt{u^2+v^2}+u)} + i\frac{v}{|v|}\cdot\sqrt{\frac{1}{2}(\sqrt{u^2+v^2}-u)}\Big) \end{aligned}$$

Zwei scheinbar eindrückliche Formeln, die aber auch für den $\sqrt{\ }$-Addikten völlig unnötig sind, weil man die Gleichung $z^2 = w$ besser mit Hilfe der komplexen Exponentialfunktion löst, wie wir bald besprechen werden. Ein letzter Blick auf diese $\sqrt{\ }$-Orgie zeigt, dass man Polynomgleichungen gar nicht mit Hilfe von Wurzeln lösen will, auch wenn man es kann!

Beispiel. Die quadratische Gleichung $z^2 = 57$ hat die beiden reellen Lösungen $z = \pm\sqrt{57}$. Das Symbol $\pm$ deutet hier bloss an, dass die beiden Werte $\sqrt{57}$ und $-\sqrt{57}$ als Lösung der quadratischen Gleichung in Frage kommen. Der Wert der reellen Zahl $\sqrt{57}$ ist für den positiven reellen Radikanden 57 nämlich positiv, weil Quadratwurzeln für positive reelle Radikanden so definiert worden sind!

Im Gegensatz dazu hat die quadratische Gleichung $z^2 = -32$ die beiden imaginären Lösungen $z = \pm 4\sqrt{2}i$. Wiederum bezeichnet das Symbol $\sqrt{2}$ eine eindeutig bestimmte reelle Zahl, da der Radikand 2 auch hier positiv ist. Das Symbol $\pm$ beschreibt wiederum eine Auswahlsendung, die als Lösung einer quadratischen Gleichung in Frage kommt und ist nicht etwa als Funktionssymbol zu interpretieren, weil die Werte von Funktionen eindeutig bestimmt sein müssen und wir in diesem Zusammenhang keine Auswahlsendungen dulden können. ◯

Beispiel. Die quadratische Gleichung $z^2 + \sqrt{2}z + 1 = 0$ nimmt nach dem quadratischen Ergänzen die Scheitelform

$$\Big(z + \frac{\sqrt{2}}{2}\Big)^2 = -\frac{1}{2}$$

an. Daraus erhalten wir die beiden linearen Gleichungen

$$z_1 + \frac{\sqrt{2}}{2} = \frac{\sqrt{2}}{2}i, \qquad z_2 + \frac{\sqrt{2}}{2} = -\frac{\sqrt{2}}{2}i$$

und die beiden gesuchten konjugiert komplexen Lösungen $z_1 = \frac{\sqrt{2}}{2}(-1+i)$ und $z_2 = \frac{\sqrt{2}}{2}(-1-i)$. ◯

Beispiel. Um die quadratische Gleichung

$$z^2 - 2(1+i)z + (3-2i) = 0$$

mit komplexen Koefizienten zu lösen, ergänzt man quadratisch und erhält die Scheitelform

$$\big(z-(1+i)\big)^2 = 2i - (3-2i) = -3+4i$$

Aus den soeben hergeleiteten Formeln für die Lösungen der speziellen quadratischen Gleichung

$$z^2 = w$$

mit $w = -3+4i$, d.h. für $u = -3$ und $v = 4$ ergeben sich die Werte

$$z_1 = 1+2i, \qquad z_2 = -1-2i$$

für die Lösungen der speziellen Gleichung. Für die beiden Lösungen der ursprünglichen quadratischen Gleichung erhalten wir durch Einsetzen schliesslich die beiden linearen Gleichungen

$$z_1 - (1+i) = 1+2i, \quad z_2 - (1+i) = -1-2i$$

mit den Lösungen

$$z_1 = 2+3i, \qquad z_2 = -i$$

wie man durch Einsetzen bestätigt. Offenbar hat also jede quadratische Gleichung über den komplexen Zahlen zwei Lösungen. Wer, wie ein Zahnarzt, auf der Extraktion schlechter Wurzeln besteht, kehre nun zur Allerweltsformel zurück und lerne an Hand der Beispiele mindestens in diesem Fall die Quadratwurzeln richtig zu interpretieren. ◯

CAS. Auch für das Lösen quadratischer Gleichungen ist Sage hilfreich. ◇

Nachdem wir mit Hilfe komplexer Zahlen quadratische Gleichungen uneingeschränkt lösen können, gehen wir noch auf die Frage ein, wie man kubische Gleichungen löst.

Zunächst beachten wir, dass sich die allgemeine kubische Gleichung

$$x^3 + ax^2 + bx = c = 0$$

auf einen Spezialfall reduzieren lässt, indem man $x = z - \frac{a}{3}$ substituiert. Dann erhält man nämlich

$$\begin{aligned} x^3 + ax^2 + bx + c &= \Big(z-\frac{a}{3}\Big)^3 + a\Big(z-\frac{a}{3}\Big)^2 + b\Big(z-\frac{a}{3}\Big) + c \\ &= z^3 + \Big(b - \frac{a^2}{3}\Big)z + \Big(\frac{2a^3}{27} - \frac{ab}{3} + c\Big) \end{aligned}$$

Die kubische Gleichung hat dann die reduzierte Form

$$z^3 + pz + q = 0, \quad p = b - \frac{a^2}{3}, q = \frac{2a^3}{27} - \frac{ab}{3} + c.$$

in der der Koeffizient des quadratischen Terms fehlt.

und Eine reduzierte kubische Gleichung lässt sich immer mit Hilfe von Radikalen lösen. Für die reduzierte kubische Gleichung

$$z^3 = pz + q$$

lässt sich nämlich mit der magischen Formel von Cardano

$$z = \sqrt[3]{\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}} + \sqrt[3]{\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}}$$

die Lösungsmenge bestimmen.

Beispiel. Beispielsweise liefert diese Formel für die reelle Gleichung $z^3 = 15z+4$ den Ausdruck

$$z = \sqrt[3]{2 + \sqrt{2^2 - 5^3}} + \sqrt[3]{2 - \sqrt{2^2 - 5^3}} = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i}$$

Mit den Beziehungen $(2 + i)^3 = 2 + 11i$ und $(2 - i)^3 = 2 - 11i$ ist

$$z_1 = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i} = (2 + i) + (2 - i) = 4.$$

Dabei handelt es sich tatsächlich um eine reelle Lösung der gegebenen Gleichung, wie man durch Einsetzen leicht kontrolliert. Die hier benötigte komplexe Lösung $x_1 = 2+i$ der speziellen kubischen Gleichung $x^3 = w$ (man redet auch von einer dritten Wurzel von $w = 2 + 11i$) bestimmt man, wie im Fall der speziellen quadratischen Gleichung $x^2 = w$, mit Hilfe der Polardarstellung.

Für die beiden anderen dritten Wurzeln von $w = 2 + 11i$ erhält man durch Multiplikation mit Hilfe der beiden primitiven dritten Einheitswurzeln

$$x_2 = \zeta_3 \cdot x_1 = \Big(-\frac{1}{2} + \frac{\sqrt{3}}{2} i \Big) \cdot (2 + i) = -\Big(\frac{\sqrt{3}}{2} + 1 \Big) + \Big(\sqrt{3} - \frac{1}{2} \Big) i$$

$$x_3 = \zeta_3^2 \cdot x_1 = \Big(-\frac{1}{2} - \frac{\sqrt{3}}{2} i \Big) \cdot (2 + i) = \Big(\frac{\sqrt{3}}{2} - 1 \Big) - \Big(\sqrt{3} + \frac{1}{2} \Big) i$$

Die kubische Gleichung hat also noch die beiden reellen Lösungen

$$z_2 = x_2 - \overline{x_2} = -\sqrt{3} - 2, \qquad z_3 = x_3 - \overline{x_3} = \sqrt{3} - 2.$$

Die komplexen Zahlen sind also am Schluss der Rechnung scheinbar verschwunden! ◯

2.6.2 Geometrische Interpretation

Die Vektordarstellung einer komplexe Zahl $z = a + bi \in \mathbb{C}$ suggeriert, dass sich diese komplexe Zahl geometrisch als Punkt in der Ebene auffassen lässt, indem man den Realteil a von z in Richtung der ersten Achse und den Imaginärteil b von z in Richtung der zweiten Achse abträgt. In diesem Zusammenhang redet man deshalb von der *reellen Achse* und von der *imaginären Achse*. Auf der reellen Achse liegen die reellen Zahlen der Form $a = a + 0i$, während auf der imaginären Achse die *imaginären Zahlen* der Form $bi = 0 + bi$ liegen.

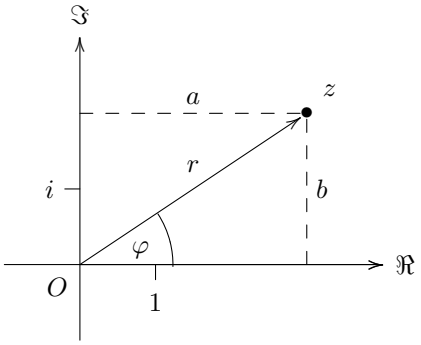


Abbildung 2.24: Geometrische Interpretation komplexer Zahlen.

Es gibt also eine eineindeutige Beziehung zwischen den komplexen Zahlen und den Punkten der *komplexen Ebene* $\mathbb{C}$ bzw. den Vektoren in $\mathbb{R}^2$. Elektriker reden deshalb statt von komplexen Zahlen manchmal auch von Zeigern und meinen damit diese (Orts-)Vektoren. Der Betrag der komplexen Zahl $z = a + bi$ ist

$$|z| = |a + bi| = \sqrt{a^2 + b^2} = r$$

Er kann geometrisch als Abstand des Punktes vom Ursprung interpretiert werden. Unter dem *Polarwinkel* φ der komplexen Zahl $z = a + bi \neq 0$ verstehen wir jenen Winkel, um den man die positive reelle Achse im Gegenuhrzeigersinn drehen muss, bis sie mit dem Vektor z übereinstimmt. Da die reellen Zahlen den Punkten auf der reellen Achse entsprechen, ist der Polarwinkel der positiven reellen Zahlen 0. Jener der negativen reellen Zahlen ist π. Die konjugiert komplexe Zahl $\overline{z} = a - bi$ ist das Spiegelbild von z an der reellen Achse. Insbesondere ist $\overline{i} = -i$. Der Polarwinkel der konjugiert komplexen Zahl ist $-\varphi$.

Entscheidend für die Anwendungen der komplexen Zahlen ist nun die Tatsache, dass sich die oben erklärten Grundoperationen der komplexen Zahlen einfach geometrisch deuten lassen.

Für die Addition ist die Sache einfach. Die Summe zweier komplexer Zahlen z_1 und z_2 geschieht komponentenweise und entspricht daher der Addition von Vektoren. Wenn die komplexen Zahlen z_1 und z_2 zwei Punkten der Ebene entsprechen, so entspricht $z_1 + z_2$ jenem Punkt der Ebene, der als vierter Eckpunkt des durch $0, z_1, z_2$ aufgespannten Parallelogrammes entsteht.

Beispiel. Die Addition $(2 + 2i) + (4 + i) = 6 + 3i$ lässt sich geometrisch mit Hilfe der Parallelogrammkonstruktion interpretieren.

Die Addition komplexer Zahlen entspricht also der Vektoraddition und deshalb ist die Darstellung in Normalform bzw. in Vektorform anschaulich. ◯

Aus dieser Figur kann man sofort ablesen, wie sich der Abstand der beiden Punkte z_1 und z_2 mit Hilfe der Addition berechnen lässt. Es gilt folgende Version des Satzes von Pythagoras:

Definition. Für den Abstand zweier komplexen Zahlen z_1 und z_2 gilt:

$$d(z_1, z_2) = |z_2 - z_1|$$

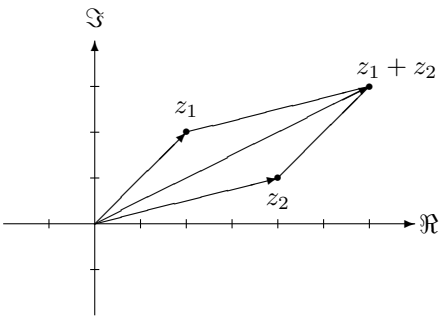


Abbildung 2.25: Geometrische Interpretation der Addition komplexer Zahlen.

Aus den Betragseigenschaften ergeben sich die Eigenschaften einer *Metrik*.

Korollar. Für beliebige komplexe Zahlen $z_1, z_2, z_3 \in \mathbb{C}$ gilt:

1. $d(z_1, z_2) = d(z_2, z_1)$.
2. $d(z_1, z_3) \leq d(z_1, z_2) + d(z_2, z_3)$.
3. $d(z_1, z_2) = 0$ genau dann wenn $z_1 = z_2$.

Neben der additiven Struktur hat die Euklid'sche Ebene erstaunlicherweise auch eine verträgliche multiplikative Struktur, die mit Drehungen und Streckungen in Beziehung steht, wie wir nun sehen werden. Etwas weniger auf der Hand als die geometrische Interpretation der Addition liegt nämlich die geometrische Interpretation der Multiplikation komplexer Zahlen. Um zu verstehen, was beim Multiplizieren zweier komplexer Zahlen z_1 und z_2 geometrisch passiert, beachten wir zunächst, dass die Normproduktregel

$$|z_1 \cdot z_2|^2 = (z_1 \cdot z_2)\overline{(z_1 \cdot z_2)} = z_1 \cdot z_2 \cdot \overline{z_1} \cdot \overline{z_2} = z_1 \cdot \overline{z_1} \cdot z_2 \cdot \overline{z_2} = |z_1|^2 \cdot |z_2|^2$$

gilt. Weil Beträge immer positive reelle Zahlen liefern, gilt also für den Betrag des Produktes die Produktregel.

Korollar. Für den Betrag des Produktes zweier komplexen Zahlen z_1 und z_2 gilt:

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|$$

Der Betrag eines Produktes ist also das Produkt der Beträge und aus diese Produktregel besagt, dass der Betrag Produkttreu ist. Entsprechendes gilt für die Summe nicht.

Es bleibt die Frage zu klären, was mit den Polarwinkeln zweier komplexen Zahlen beim Multiplizieren passiert. Weil wir über den Betrag des Produktes bereits Bescheid wissen, können wir annehmen, dass die beiden komplexen Zahlen z_1 und z_2 beide den Betrag 1 haben und damit die entsprechenden Punkte auf dem Einheitskreis liegen. Wegen des letzten Korollars liegt dann auch das Produkt

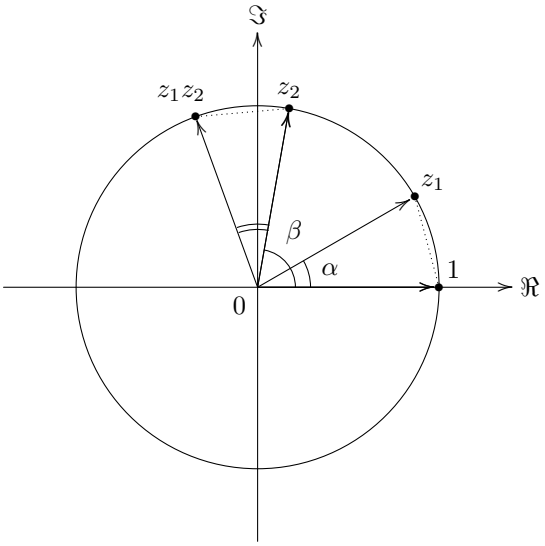


Abbildung 2.26: Zur geometrische Interpretation der Multiplikation komplexer Zahlen.

$z_1 \cdot z_2$ auf dem Einheitskreis. Wir nehmen an, der Polarwinkel von z_1 sei α und jener von z_2 sei β. Wir möchten den Polarwinkel φ des Produktes $z_1 \cdot z_2$ berechnen.
Das letzte Korollar liefert für den Abstand

$$d(z_2, z_1z_2) = |z_1z_2 - z_2| = |z_2 \cdot (z_1 - 1)| = |z_2| \cdot |z_1 - 1| = |z_1 - 1| = d(z_1, 1)$$

wobei wir die Annahme $|z_2| = 1$ benutzt haben. Dehr sind die beiden punktierten Sehnen gleich lang und daher sind die beiden Dreiecke $(1, 0, z_1)$ und $(z_2, 0, z_1z_2)$ kongruent und es gilt $\angle(z_2, 0, z_1z_2) = \alpha$. Damit ist entweder $\varphi = \alpha + \beta$ oder $\varphi = \alpha - \beta$. Um uns zwischen diesen beiden Alternativen entscheiden zu können, vertauschen wir die Rollen von z_1 und z_2. Eine analoge Überlegung zeigt, dass dann entweder $\varphi = \alpha + \beta$ oder $\varphi = \beta - \alpha$ gelten muss. Aus Symmetriegründen kommt also nur die erste Alternative $\varphi = \alpha + \beta$ in Frage. Unsere Überlegung zeigt den folgenden fundamentalen Satz über das Produkt komplexer Zahlen:

Satz. Für den Polarwinkel φ des Produktes zweier komplexen Zahlen z_1 und z_2 mit den Polarwinkeln α und β gilt:

$$\varphi = \alpha + \beta$$

Der Polarwinkel eines Produktes ist also die Summe der Polarwinkel.

Wir fassen diese, auch für die ebene Computer-Graphik zentrale Interpretation der Multiplikation komplexer Zahlen in etwas andere Worte und benutzen dabei die geometrische Sprache.

Beispiel. Falls $z = x+yi \in \mathbb{C}$ eine beliebige komplexe Zahl und $r \in \mathbb{R}$ eine reelle Zahl bezeichnet, so ist $r \cdot z = (rx) + (ry)i$. Für den Betrag gilt $|r \cdot z| = |r| \cdot |z|$.

Multiplikation mit einer reellen Zahl entspricht geometrisch einer *Streckung*. Falls $r > 0$ ist, so haben z und rz die selbe Richtung. Im Fall $r < 0$ dagegen sind die Richtungen von z und $r \cdot z$ entgegengesetzt.

Ferner ist $i \cdot z = -y + xi$. Daher liefert Multiplikation mit i eine *Drehung* um den Winkel $\varphi = \frac{\pi}{2}$ im Gegenuhrzeigersinn.

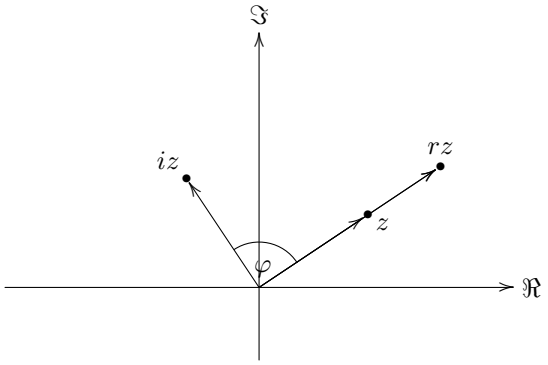


Abbildung 2.27: Multiplikation mit $r \in \mathbb{R}$ und i.

Multiplizieren wir z mit der beliebigen komplexen Zahl $w = a + bi$, erhalten wir das Produkt $w \cdot z = (a + bi) \cdot z = az + bzi$. Daher wird z zunächst um den reellen Faktor a gestreckt. Dann wird z um den reellen Faktor b gestreckt und das Ergebnis um den Winkel $\frac{\pi}{2}$ im Gegenuhrzeigersinn gedreht. Schliesslich werden die beiden Ergebnisse addiert.

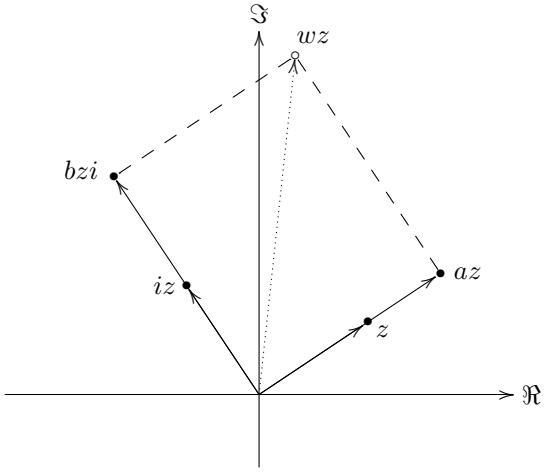


Abbildung 2.28: Multiplikation mit $w = a + bi \in \mathbb{C}$.

Zusammengefasst können wir also sagen, dass Multiplikation mit einer beliebigen komplexen Zahl $w \in \mathbb{C}$ einer *Drehstreckung* entspricht. Ein beliebiger Vektor $z \in \mathbb{C}$ wird beim Bilden des Produktes $w \cdot z$ um den Faktor $|w|$ gestreckt und um den Polarwinkel von w gedreht. Insbesondere geht dabei die komplexe Zahl $z = 1$ in die komplexe Zahl w über.

Jede komplexe Zahl $w \in \mathbb{C}$ kann geometrisch als Beschreibung einer Drehstreckung in der Ebene mit dem Drehzentrum O, dem Polarwinkel von w als Drehwinkel und dem Betrag $|w|$ als Streckfaktor interpretiert werden. Bei der Multiplikation von komplexen Zahlen multiplizieren sich die Beträge und addieren sich die Polarwinkel. Beim Dividieren durch eine komplexe Zahl $w \neq 0$ werden die Streckung und die Drehung rückgängig gemacht. ◯

Nachdem wir auf elementare Art die geometrisch Bedeutung der Multiplikation mit einer komplexen Zahl verstanden haben, lässt es sich nun auch leicht in der Matrizendarstellung einsehen, dass eine komplexe Zahl der Form

$$Z_{a,b} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

tatsächlich einer Drehstreckung entspricht. Dazu beachten wir, dass die Drehung D_φ um den Winkel φ bzw. die Streckung S_r mit dem Streckfaktor $r \geq 0$ bekanntlich durch die kommutierenden Matrizen

$$D_\varphi = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}, \qquad S_r = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}, \qquad [D_\varphi, S_r] = 0$$

dargestellt werden. Daher wird die Drehstreckung um den Winkel φ und dem Streckfaktor r durch die komplexe Zahl

$$\begin{aligned} S_r \cdot D_\varphi &= \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix} \cdot \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix} = \begin{pmatrix} r\cos(\varphi) & -r\sin(\varphi) \\ r\sin(\varphi) & r\cos(\varphi) \end{pmatrix} \\ &= Z_{r\cos(\varphi), r\sin(\varphi)} \end{aligned}$$

dargestellt. Um eine beliebige komplexe Zahl $z = a + ib \in \mathbb{C}$ auf diese Art als Drehstreckung auffassen zu können, sind wir gezwungen, den Drehwinkel φ und r so zu wählen, dass die *Umrechnungsformeln*

$$a = r \cdot \cos(\varphi), \qquad b = r \cdot \sin(\varphi)$$

gelten. Wir haben nun also für eine komplexe Zahl eine weitere Darstellungsart gefunden, nämlich ihre *Polarform*

$$Z_{a,b} = S_r \cdot D_\varphi.$$

Sie ist offenbar dann speziell geeignet, wenn Ebenendrehungen involviert sind.

Die Umrechnungsformeln lassen sich bekanntlich geometrisch interpretieren. Aus der Figur lesen wir ebenfalls die *inversen Umrechnungsformeln* zwischen der Normalform und der Polarform einer komplexen Zahl ab. In ältern Büchern sagt man, man rechne von kartesischen in Polarkoordinaten um und umgekehrt.

Korollar. Es sei $z = a+bi \neq 0$ eine komplexe Zahl. Für ihren Betrag $|z| = r \geq 0$ und ihren Polarwinkel $\varphi \in [0, 2\pi)$ gilt:

$$a = r \cdot \cos(\varphi), \qquad b = r \cdot \sin(\varphi)$$

und nach "Auflösen" dieser nichtlinearen Gleichungen nach r bzw. φ:

$$r = \sqrt{a^2 + b^2}, \qquad \tan(\varphi) = \frac{b}{a}$$

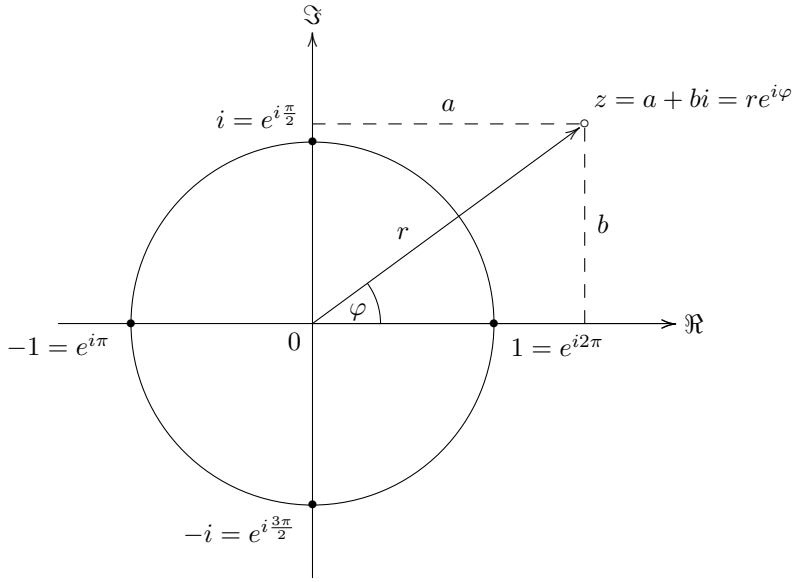


Abbildung 2.29: Bedeutung der Polardarstellung komplexer Zahlen.

Den sogn. Hauptwert des Winkels $\varphi \in (-\pi, \pi]$, den man häufig auch Hauptargument von $z \neq 0$ nennt und dann mit $\mathrm{Arg}(z)$ bezeichnet, bestimmt man durch Quadrantenüberlegungen, die hinter folgender Fallunterscheidung stecken, das zum numerischen Rechnen in diversen Programmiersprachen verwendet wird.

$$\mathrm{Arg}(z) = \varphi = \begin{cases} \arctan\left(\frac{b}{a}\right) & a > 0 \\ \arctan\left(\frac{b}{a}\right) + \pi & a < 0 \text{ und } b \geq 0 \\ \arctan\left(\frac{b}{a}\right) - \pi & a < 0 \text{ und } b < 0 \\ \frac{\pi}{2} & a = 0 \text{ und } b > 0 \\ -\frac{\pi}{2} & a = 0 \text{ und } b < 0 \\ \text{unbestimmt} & a = 0 \text{ und } b = 0 \end{cases}$$

Will man sich nicht (willkürlich) auf einen Winkel festlegen, benutzt man sie alle und redet dann vom Argument

$$\arg(z) = \{\mathrm{Arg}(z) + k2\pi, k \in \mathbb{Z}\}, \quad z \neq 0$$

Um den Polarwinkel $\varphi \in [0, 2\pi)$ zu erhalten, muss man zum Hauptwert noch 2π addieren, falls er negativ ist. Der Leser hüte sich also vor blindem Benutzen der arctan-Taste!

Beispiel. Die komplexe Zahl $1 + i$ liegt im ersten Quadranten und hat den Betrag $r = \sqrt{2}$ und den Polarwinkel $\varphi = \frac{\pi}{4}$. Die komplexe Zahl $z = 1 + \sqrt{3}i$ liegt ebenfalls im ersten Quadranten, hat den Betrag $|z| = 1$ und das Hauptargument $\mathrm{Arg}(z) = \frac{\pi}{3}$. Analog hat die imaginäre Einheit den Polarwinkel $\frac{\pi}{2}$, der mit ihrem Hauptargument übereinstimmt. Analog erhalten wir die Hauptargumente

$\mathrm{Arg}(1) = 0$, $\mathrm{Arg}(-1) = \pi$ und $\mathrm{Arg}(1-i) = -\frac{\pi}{4}$. Man beachte, dass zu $-i$ der Polarwinkel $\frac{3\pi}{2}$ und das Hauptargument $\mathrm{Arg}(-i) = -\frac{\pi}{2}$ gehören. ◯

2.6.3 Die Euler'sche Formel

Wir haben festgestellt, dass die Multiplikation komplexer Zahlen eng mit Drehungen zusammenhängt. Daher spielen jene komplexen Zahlen eine wichtige Rolle, die durch Drehung von $1 \in \mathbb{C}$ um den Ursprung entstehen, d.h. der Einheitskreis S^1.

Die Drehmatrix D_φ stellt die folgende wichtige komplexe Zahl dar.

$$e^{i\varphi} = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix} \in \mathbb{C}$$

Ihre Normalform liefert die berühmte Euler'sche Formel.

Satz. Für jeden Winkel $\varphi \in \mathbb{R}$ gilt $e^{i\varphi} = \cos(\varphi) + i\sin(\varphi)$.

Die komplexen Zahlen der Form $e^{i\varphi}$ gehören zu den Punkten auf dem Einheitskreis in der komplexen Ebene. Ihre Norm ist nach dem Satz von Pythagoras tatsächlich 1. Es ist also

$$N(e^{i\varphi}) = |e^{i\varphi}| = 1$$

Vorläufig ist $e^{i\varphi}$ nur eine Abkürzung für eine gewisse komplexe Zahl auf dem Einheitskreis. Bald werden sich aber Zusammenhänge zur Exponentialfunktion zeigen und praktische Rechenregeln ergeben, die diese Abkürzung rechtfertigen.

Aus der Beschreibung der komplexen Zahlen mit Hilfe von Drehstreckungen bzw. aus der Darstellung in Polarform folgt, dass sich jede beliebige komplexe Zahl $z = a + bi \neq 0$ auch in der *Exponentialform*

$$z = r \cdot e^{i\varphi}$$

darstellen lässt, wobei die positive reelle Zahl r den *Betrag* und φ den *Polarwinkel* der komplexen Zahl z bezeichnet. Weil die Exponentialform einfach die verkappte Polarform ist, lässt sich mit Hilfe der angegebenen Umrechnungsformeln zum Umrechnen von der Normalform in Polarform auch einfach zwischen den Normalform und der Exponentialform umrechnen.

Beispiel. Es ist

$$e^{i\frac{\pi}{2}} = i, \qquad e^{i\pi} = -1, \qquad e^{2\pi i} = 1.$$

Insbesondere gilt also $e^{i\pi} + 1 = 0$ — eine Formel, die bei einigen Autoren mystische Schwärmereien produziert. Sie verknüpft die fünf wichtigsten Zahlen der klassischen Mathematik: $0, 1, e, i, \pi$. Aus der $2\pi-$Periodizität der Kreisfunktionen folgt nicht nur die Beziehung $e^{2\pi i} = 1$, sondern allgemeiner, dass die komplexe Exponantialfunktion 2π-periodisch ist!

Die im letzten Abschnitt bei der Umrechnung von kartesischer in Polarform besprochenen Beispiele übersetzen sich in die Beziehungen $1 + i = \sqrt{2}\, e^{i\frac{\pi}{4}}$ und $1 + \sqrt{3}i = 2\, e^{i\frac{\pi}{3}}$. ◯

CAS. Zur Umrechnung zwischen kartesischer und Polarform bzw. des Betrages und des Hauptargumentes einer komplexen Zahl greift man bei Bedarf aus Sage mittels folgendes Kodes auf das eingebaute Programm Maxima zu. ◇

Der vom Reellen her unvertraute Umstand, dass die komplexe Exponentialfunktion 2π-periodisch ist, hat einige ungewohnte Konsequenzen. Beispielsweise hat die Exponentialgleichung

$$e^{i\varphi} = 1$$

nicht nur, wie im Reellen, die einzige Lösung $\varphi_0 = 0$, sondern unendlich viele Lösungen der Form

$$\varphi_k = 2k\pi, \quad k \in \mathbb{Z}$$

Weil die Exponentialfunktion in der reellen Analysis eine zentrale Rolle spielt, wird man für die komplexe Analysis einige Konzepte gründlich überdenken müssen. Das lohnt sich sich um so mehr, als die komplexe Analysis dann einfacher wird und viele Artefakte der reellen Analysis befriedigend erklärt.

Beispielsweise hat man im Reellen den Logarithmus kurzerhand als Umkehrung der Exponentialfunktion erklärt. Weil die komplexe Exponentialfunktion periodisch ist, kann sie unmöglich umkehrbar sein, was zur Konsequenz hat, dass so etwas wie eine komplexe Logarithmusfunktion gar nicht existieren kann! Man muss also beim Versuch, alle komplexen Lösungen z der Exponentialgleichung

$$e^z = w$$

zu bestimmen, etwas vorsichtig sein. Machen wir für die gesuchte Lösung den Ansatz $z = x+iy \in \mathbb{C}$, erhalten wir durch Einsetzen in die Exponentialgleichung und benutzen des Additionstheorems

$$e^{x+iy} = w = e^x \cdot e^{iy}$$

Falls wir nun die rechte Seite der Exponentialgleichung w in Exponentialform $w = r \cdot e^{i\varphi}$ darstellen, wird daraus die Gleichung

$$e^x \cdot e^{iy} = r \cdot e^{i\varphi}$$

aus der wir ablesen, dass die beiden Gleichungen

$$e^x = r, \qquad e^{iy} = e^{i\varphi}$$

erfüllt sein müssen. Diese beiden Gleichungen für die reellen Zahlen x und y lassen sich aber leicht lösen. Für $r > 0$ erhalten wir aus der ersten Gleichung die Lösung $x = \log(r)$, wobei hier der reelle Logarithmus der strikt positiven reellen Zahl r zu verwenden ist. Die zweite Gleichung, die wir dank des Additionstheorems auch in der Form

$$e^{i(y-\varphi)} = 1$$

schreiben können, haben wir oben gelöst und festgestellt, dass ihre Lösungsmenge durch die reellen Zahlen

$$y_k = \varphi + k2\pi, \qquad k \in \mathbb{Z}$$

gegeben ist. Die Exponentialgleichung hat also für $w = 0$ (wie im Reellen) gar keine Lösung. Für $w = r \cdot e^{i\varphi} \neq 0$ hat sie aber, anders als im Reellen, unendlich viele Lösungen den Form

$$z_k = \log(r) + (\varphi + 2\pi k)i = \log(|w|) + (\varphi + 2\pi k)i, \quad k \in \mathbb{Z}$$

wie man durch Einsetzen leicht überprüft. Alle diese abzählbar vielen, äquidistanten komplexen Zahlen, die auf einer Vertikalen mit dem Realteil $\log(r)$ liegen, haben also das Recht, als Logarithmus von w bezeichnet zu werden. Wenn schon, bezeichnet das Symbol $\log(w)$ also eine ganze Menge von komplexen Zahlen und kann daher nicht als Funktion auf $\mathbb{C} \setminus \{0\}$ aufgefasst werden. Wie bei Quadratwurzeln wird erst dann restlos klar, was man unter dem Logarithmus $\log(w)$ verstehen soll, wenn man die Riemannschen Flächen zur Verfügung hat. Auch hier braucht der Anwender aber in der Regel eigentlich keine Logarithmusfunktion, sondern höchstens die Lösungsmenge der Exponentialgleichung.

Beispiel. Die Exponentialgleichung $e^z = 1$ hat also die komplexen Lösungen $z_k = k2\pi i, k \in \mathbb{Z}$, die alle auf der imaginären Achse liegen und uns bereits bekannt sind. ◯

Durch Konjugieren erhalten wir aus der Euler'schen Formel

$$\overline{e^{i\varphi}} = \overline{\cos(\varphi) + i\sin(\varphi)} = \cos(\varphi) - i\sin(\varphi) = \cos(-\varphi) + i\sin(-\varphi) = e^{-i\varphi}$$

Korollar. Für jeden Winkel $\varphi \in \mathbb{R}$ gilt $\overline{e^{i\varphi}} = e^{-i\varphi}$.

Berechnen wir den Realteil von $e^{i\varphi}$, erhalten wir

$$\Re(e^{i\varphi}) = \cos(\varphi) = \frac{1}{2}\Big(e^{i\varphi} + \overline{e^{i\varphi}}\Big) = \frac{1}{2}\Big(e^{i\varphi} + e^{-i\varphi}\Big)$$

Für den Imaginärteil erhalten wir entsprechend

$$\Im(e^{i\varphi}) = \sin(\varphi) = \frac{1}{2i}\Big(e^{i\varphi} - \overline{e^{i\varphi}}\Big) = \frac{1}{2i}\Big(e^{i\varphi} - e^{-i\varphi}\Big)$$

Wir stellen also fest, dass sich die Kreisfunktionen mit Hilfe der komplexen Exponentialfunktion bzw. den Hyperbelfunktionen ausdrücken.

Korollar. Für jeden Winkel $\varphi \in \mathbb{R}$ gilt:

$$\begin{aligned}\cos(\varphi) &= \frac{1}{2}\Big(e^{i\varphi} + e^{-i\varphi}\Big) = \cosh(i\varphi)\\ \sin(\varphi) &= \frac{1}{2i}\Big(e^{i\varphi} - e^{-i\varphi}\Big) = -i\sinh(i\varphi)\end{aligned}$$

Durch Quotientenbildung erhält man daraus

$$\tan(\varphi) = -i \cdot \frac{e^{i\varphi} - e^{-i\varphi}}{e^{i\varphi} + e^{-i\varphi}} = -i \cdot \tanh(i\varphi), \qquad \varphi \neq k \cdot \frac{\pi}{2}, k \in \mathbb{Z}$$

Mit Hilfe dieser Formeln lassen sich die Kreisfunktionen durch die Exponentialfunktion bzw. durch die Hyperbelfunktionen ausdrücken. Das hat den Vorteil, dass sich jedes Problem im Zusammenhang mit Kreisfunktionen in ein äquivalentes Problem mit (komplexen) Exponentialfunktionen umformulieren lässt. Dieses umformulierte Problem kann dann meistens einfacher behandelt werden, da nun nur noch die (komplexe) Exponentialfunktion $z = e^{ix}$ involviert ist, die sich im Vergleich zu den äquivalenten Formeln mit Kreisfunktionen leichter

manipulieren lässt. In den physikalischen Anwendungen werden beispielsweise Schwingungsphänomene vorteilhaft komplex formuliert und man verzichtet vorteilhaft auf die ganze Trigonometrie.

Durch Auflösen obiger Formeln lassen sich umgekehrt die Hyperbelfunktionen eines reellen Argumenten mit Hilfe der Kreisfunktionen eines imaginären Argumentes ausdrücken.

$$\begin{array}{rcl} \cosh(x) & = & \cos(ix) \\ \sinh(x) & = & -i \cdot \sin(ix) \\ \tanh(x) & = & -i \cdot \tan(ix) \end{array}$$

Offenbar entsprechen sich die Beziehungen zwischen den Kreisfunktionen und den Hyperbelfunktionen eineindeutig.

Beispiel. Bei der Beschreibung periodischer Prozesse spielen die Kreisfunktionen eine zentrale Rolle. Damit die Schwingung beliebig stark oszillieren kann, wir die Zeitmessung zu einem beliebigen Zeitpunkt beginnen können und um beliebig schnelle Schwingungsprozesse mitzuberücksichtigen, strecken wir die Auslenkung die Zeitskala und verschieben den Nullpunkt. So erhalten wir die *allgemeine Kosinusfunktion*[32].

$$h(t) = A \cdot \cos(\omega t + \alpha) = \Re\Big(A \cdot e^{i(\omega t + \alpha)}\Big)$$

Die Physiker reden auch von einer *harmonische Schwingung* und nennen $h(t)$ die *Auslenkung* oder Elongation zur Zeit t. Der Faktor A heisst *Amplitude* oder maximale Auslenkung der harmonischen Schwingung. Mit ω bezeichnen sie die *Kreisfrequenz* und mit

$$T = \frac{1}{f} = \frac{2\pi}{\omega}$$

die *Periodendauer*. Der Reziprokwert der Periodendauer $f = \frac{\omega}{2\pi}$ heisst bei ihnen *Frequenz* und den Winkel α nennen sie *Nullphasenwinkel* und der Ausdruck $\varphi(t) = \omega t + \alpha$ heisst *Phasenwinkel*.

Die harmonische Schwingung, die gegeben ist durch die Formel

$$h(t) = 3\cos\left(2t + \frac{2\pi}{3}\right) = \Re\Big(3 \cdot e^{i(2t + \frac{2\pi}{3})}\Big)$$

hat die Amplitude 3, die Kreisfrequenz 2, den Nullphasenwinkel $\frac{2\pi}{3}$ und den Graphen

Auf Grund des Graphen wird klar, welche Einflüsse die drei Parameter A, ω, α in der harmonischen Schwingung auf die Kosinusfunktion haben. Die Änderung der Amplitude A bewirkt eine Vergrösserung bzw. Verkleinerung der Auslenkung auf das $|A|$-fache , je nach dem, ob $|A| \geq 1$ bzw. $|A| \leq 1$ ist und eine Spiegelung an der horizontalen Achse, wenn A negativ ist.

[32] Andere Autoren beschreiben harmonische Schwingungen durch einen sinusförmigen, statt durch einen kosinusförmigen Verlauf und müssen dann halt statt des Realteils $\Re$ den Imaginärteil $\Im$ benutzen. Die allgemeine Sinusfunktion $A\sin(\omega t + \beta)$ lässt sich dank des Additionstheorems als allgemeine Kosinusfunktion darstellen, indem man α so wählt, dass gilt:

$$\cos(\beta) = -\sin(\alpha), \qquad \sin(\beta) = \cos(\alpha).$$

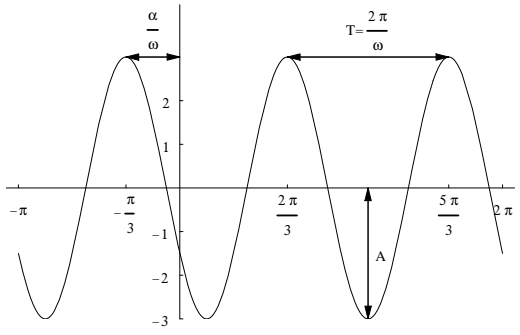


Abbildung 2.30: Graph der harmonischen Schwingung $3\cos\left(2t+\frac{2\pi}{3}\right)$.

Die Änderung der Kreisfrequenz ω bewirkt auf den Graphen der Kosinusfunktion eine Streckung bzw. eine Stauchung in horizontaler Richtung auf das $\frac{1}{|\omega|}$-fache, je nach dem, ob $|\omega| \geq 1$ bzw. $|\omega| \leq 1$ ist und eine Spiegelung an der horizontalen Achse, falls ω negativ ist.

Die Periodendauer T kann als Länge einer Schwingungsperiode interpretiert werden. Für alle Argumente t gilt nämlich die *Periodizitätsbedingung*

$$h(t+T) = h(t)$$

Diese wichtige Beziehung, die ausdrückt, dass die verallgemeinerte Kosinusfunktion die *Periode* T besitzt, folgt aus der 2π-Periodizität der Kosinusfunktion durch eine kleine Rechnung. Es gilt in der Tat $h(t+T) = A\cos\left(\omega(t+T)+\alpha\right) = A\cos\left(\omega(t+\frac{2\pi}{\omega})+\alpha\right) = A\cos(\omega t+\alpha+2\pi) = A\cos(\omega t+\alpha) = h(t)$. Geometrisch bedeutet diese Beziehung, dass der Graph von h mit sich selber zur Deckung gelangt, falls er um die Strecke T parallel zur t-Achse verschoben wird.

Die Änderung des Nullphasenwinkels α bewirkt auf die Kosinusfunktion eine Verschiebung der Kurve in Richtung der t-Achse und zwar um $\frac{\alpha}{\omega}$ nach rechts, falls $\frac{\alpha}{\omega} \leq 0$ und um $\frac{\alpha}{\omega}$ nach links, falls $\frac{\alpha}{\omega} \geq 0$. Dass die Verschiebung um $\frac{\alpha}{\omega}$ und *nicht* etwa um α stattfindet, entnimmt man der Form $h(t) = A\cos\left(\omega(t+\frac{\alpha}{\omega})\right)$. Wegen der Periodizität der Kosinusfunktion können wir uns auf Nullphasenwinkel im Grundinterval $0 \leq \alpha < 2\pi$ beschränken. Notfalls reduziert man ihn modulo 2π.

Viele Überlegungen im Zusammenhang mit harmonischen Schwingungen werden speziell anschaulich, wenn man sie mit einer geeigneten gleichförmigen Kreisbewegung in Verbindung bringt. Dazu beachtet man, dass für $A \geq 0$ die in der harmonischen Schwingung vorkommende komplexe Funktion

$$\vec{r}(t) = A\cdot e^{i(\omega t+\alpha)} = \underbrace{A\cdot e^{i\alpha}}_{\substack{\text{komplexe}\\ \text{Amplitude}}} \cdot \underbrace{e^{i\omega t}}_{\substack{\text{Zeit-}\\ \text{funktion}}} = A\cdot\begin{pmatrix}\cos(\omega t+\alpha)\\ \sin(\omega t+\alpha)\end{pmatrix}$$

als Parametrisierung eines Kreises vom Radius A betrachtet werden kann. Den rechten komplexen Fakor nennt man in der Physik *Zeitfunktion* und der linke Faktor heisst dort *komplexe Amplitude*. Sie entsteht also als Produkt aus der (reellen) Amplitude A mit der *Phase* $e^{i\alpha}$, die zum *Nullphasenwinkel* α gehört. Der Phasenfaktor ist also eine komplexe Zahl vom Betrag 1. ◯

Beispiel. Im Zusammenhang mit der Berechnung des Matrizenexponentials e^{At} einer Matrix A mit komplexen Eigenwerten stösst man auf die Funktion

$$g(t) = e^{(a+ib)t} = e^{at} \cdot e^{ibt} = e^{at} \cdot (\cos(bt) + i \sin(bt)), \quad a, b \in \mathbb{R}$$

deren Real- bzw. Imaginärteil

$$g_1(t) = \Re(g(t)) = e^{at} \cos(bt), \qquad g_2(t) = \Im(g(t)) = e^{at} \sin(bt)$$

gedämpfte harmonische Schwingungen beschreiben.

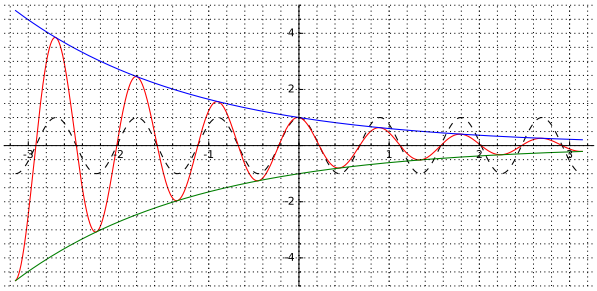


Abbildung 2.31: Graphen der harmonischen Schwingung $h(t) = \cos(7t)$, der gedämpften harmonischen Schwingung $g_1(t) = e^{\frac{t}{2}} \cos(7t)$ und der Einhüllenden.

Dabei handelt es sich also um harmonische Schwingungen, deren Amplitude mit Hilfe einer Exponentialfunktion moduliert wird. ◯

Mit Hilfe der komplexen Exponentialfunktion wird der ganze Formelkram der Goniometrie überflüssig, weil im umformulierten Problem nur noch (komplexe) rationale Funktionen der Exponentialfunktion $z = e^{ix}$ involviert sind, die sich im Vergleich zu den äquivalenten Formeln mit Kreisfunktionen leichter manipulieren lassen.

Beispiel. Um die Halbwinkelformel

$$\cos^2\left(\frac{x}{2}\right) = \frac{1 + \cos(x)}{2}$$

einzusehen, ersetzen wir die Kreisfunktionen durch die Exponentialfunktion. Für die linke Seite der fraglichen Gleichung erhalten wir

$$\cos^2\left(\frac{x}{2}\right) = \left(\frac{e^{i\frac{x}{2}} + e^{-i\frac{x}{2}}}{2}\right)^2 = \frac{e^{ix} + 2e^{i0} + e^{-ix}}{4} = \frac{e^{ix} + 2 + e^{-ix}}{4}$$

wobei wir natürlich die binomische Formel und das Exponentialgesetze benutzt haben. Für die rechte Seite ergibt sich

$$\frac{1 + \cos(x)}{2} = \frac{1 + \frac{e^{ix} + e^{-ix}}{2}}{2} = \frac{2 + e^{ix} + e^{-ix}}{4}$$

Weil die beiden Ausdrücke übereinstimmen, ist die Behauptung gezeigt. Man beachte, dass dieser Ausdruck für alle Winkel x positiv ist und daher die Quadratwurzel gezogen werden kann. Wir erhalten die Halbwinkelformel

$$\left|\cos\left(\frac{x}{2}\right)\right| = \sqrt{\frac{1+\cos(x)}{2}} = \frac{1}{2}\sqrt{2+2\cos(x)}$$

wie sie üblicherweise in der Literatur vorkommt. ◯

Beispiel. Dem trigonometrischen Satz von Pythagoras

$$\cos^2(\varphi) + \sin^2(\varphi) = 1$$

entspricht der hyperbolische Satz von Pythagoras

$$\cosh^2(i\varphi) - \sinh^2(i\varphi) = \cosh^2(x) - \sinh^2(x) = 1$$

den man selbstverständlich auch mit Hilfe der Funktionalgleichung der Exponentialfunktion nachrechnen kann.

Dieser Umstand motiviert die Bezeichnungsweise der hyperbolischen Funktionen. Sie stehen zur Einheitshyperbel mit der Gleichung

$$x^2 - y^2 = 1$$

in der selben Beziehung wie die Kreisfunktionen zum Einheitskreis mit der Gleichung

$$x^2 + y^2 = 1.$$

Allgemeiner entsprechen sich Euklid'sche und Minkowski Geometrie bis auf einen Faktor $\pm i$, der nach dem Quadrieren zu einem Vorzeichenwechsel führt. Beispielsweise wird die Einheitshyperbel durch die hyperbolische Drehung

$$H_\varphi = \begin{pmatrix} \cosh(\varphi) & \sinh(\varphi) \\ \sinh(\varphi) & \cosh(\varphi) \end{pmatrix}$$

in sich übergeführt und das hyberbolische Additionstheorem

$$L_{\varphi+\psi} = L_\varphi \cdot L_\psi$$

nimmt die Form

$$\begin{aligned} \sinh(\varphi+\psi) &= \sinh(\varphi)\cosh(\psi) + \cosh(\varphi)\sinh(\psi) \\ \cosh(\varphi+\psi) &= \cosh(\varphi)\cosh(\psi) + \sinh(\varphi)\sinh(\psi) \end{aligned}$$

an. Durch Quotientenbildung und Erweitern mit $\cosh(\varphi)\cdot\cosh(\psi)$ erhalten wir daraus das Additionstheorem für den hyperbolische Tangens

$$\tanh(\varphi+\psi) = \frac{\tanh(\varphi)+\tanh(\psi)}{1+\tanh(\varphi)\cdot\tanh(\psi)}$$

Die Minkowski-Geometrie kann durch die spezielle Relativitätstheorie physikalisch interpretiert werden. Falls sich ein Beobachter B_1 gegenüber einem anderen

Beobachter B_0 mit konstanter Geschwindigkeit v in x-Richtung gleichförmig bewegt und sie zur Zeit $t = 0$ im Ursprung sind, so hängen die Raumzeitkoodinaten[33] (t, x) des Beobachters B_0 eines Ereignisses mit den Raumzeitkoordinaten (t', x') des Beobachters B_1 für das selbe Ereignis durch Multiplikation mit der hyperbolischen Drehmatrix (Lorentz-Boost, bzw. Lorentz-Transformation)

$$L_\varphi = \begin{pmatrix} \cosh(\varphi) & \sinh(\varphi) \\ \sinh(\varphi) & \cosh(\varphi) \end{pmatrix}, \qquad \text{d.h. durch} \quad \begin{cases} t = \cosh(\varphi)t' + \sinh(\varphi) \cdot x' \\ x = \sinh(\varphi)t' + \cosh(\varphi) \cdot x' \end{cases}$$

zusammen. Dabei gilt für den "Boostwinkel" φ die Beziehung

$$\tanh(\varphi) = v, \qquad \cosh(\varphi) = \frac{1}{\sqrt{1-v^2}}, \qquad \sinh(\varphi) = \frac{v}{\sqrt{1-v^2}}$$

Das hyberbolische Additionstheorem $L_{\varphi+\psi} = L_\varphi \cdot L_\psi$ beschreibt die Raumzeitkoordinaten (t'', x'') eines Beobachters B_2, der sich bezüglich B_1 mit der Geschwindigkeit $\tanh(\psi)$ bewegt[34].

Das Additionstheorem für den hyperbolische Tangens

$$w = \frac{v+w}{1+v \cdot u}$$

spielt die Rolle des relativistischen Additionstheorem für Geschwindigkeiten.

Der Faktor $\pm i$ zwischen Euklidscher und Minkowski-Geometrie ist auch der Grund, warum in der speziellen Relativitätstheorie die Zeit t manchmal mit i oder mit $-i$ multipliziert wird. Durch diesen mathematischen Trick (Physiker reden von einer Wick-Rotation) geht die Minkowski-Geometrie in die bekanntere Euklidsche Geometrie über. Eine Minkowski-Transformation kann so als Rotation aufgefasst werden und die Schrödinger-Gleichung wird zur Wärmeleitungsgleichung. Man darf diese Analogie allerdings nicht zu weit treiben. Während die Grundfigur der Euklidschen Geometrie — der Einheitskreis — und Drehgruppe kompakt sind, gilt das für die Grundfigur der Minkowski-Geometrie — die Einheitshyperbel — und für die Lorentz-Gruppe nicht. ◯

Besonders eingängig sind die Additionstheoreme, wenn man sie mit Hilfe von komplexen Zahlen ausdrückt. Hinter den Additiontheoremen steckt aus geometrischer Sicht bekanntlich die Funktionalgleichung der Drehung:

$$D_{\varphi+\psi} = D_\varphi \cdot D_\psi$$

Drehen wir in der komplexen Ebene 1 um den Winkel φ, erhalten wir die komplexe Zahl $e^{i\varphi}$ auf dem Einheitskreis. Multiplizieren wir diese komplexe Zahl mit $e^{i\psi}$, so erhalten wir die komplexe Zahl $e^{i\varphi} \cdot e^{i\psi}$. Nach unserer Einsicht über das Verhalten von Polarwinkeln beim Multiplizieren bzw. auf Grund der Funktionalgleichung der Drehung erhalten wir die selbe komplexe Zahl auch als $e^{i(\varphi+\psi)}$.

[33] Wobei heutzutage Zeiten t und Distanzen x mit der selben Einheit gemessen werden! Wählt man als Zeiteinheit die Sekunde, so benutzt man zur Messung der Länge einer Strecke diejenige Zeit, die das Licht braucht, um diese Strecke im Vakuum zurückzulegen. Zum Umrechnen beachtet man, dass das Licht im Vakuum pro Sekunde abmachungsgemäss genau $c = 299'792'458$ Meter (1 Lichtsekunde) weit kommt. In diesem natürlichen Masssystem hat die Lichtgeschwindigkeit den exakten Wert 1 und andere Geschwindigkeiten sind Bruchteile!

[34] Wer SI-Einheiten vorzieht, muss jedes vorkommende t mit c multiplizieren!

Die Funktionalgleichung der Drehung wird also mit der Euler'schen Formel zur Funktionalgleichung der Exponentialfunktion.

Korollar. Für beliebige Winkel φ, ψ gilt die Funktionalgleichung der Exponentialfunktion

$$e^{i(\varphi+\psi)} = e^{i\varphi} \cdot e^{i\psi}$$

Mit Hilfe der Euler'schen Formel lässt sich diese Funktionalgleichung als speziell suggestive Form der Additionstheorme interpretieren.

Korollar. Für beliebige Winkel $\varphi, \psi \in \mathbb{R}$ gelten die *Additionstheoreme*:

$$\begin{aligned} \sin(\varphi+\psi) &= \sin(\varphi)\cos(\psi) + \cos(\varphi)\sin(\psi) \\ \cos(\varphi+\psi) &= \cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi) \end{aligned}$$

Durch Quotientenbildung und Erweitern mit $\cos(\varphi)\cdot\cos(\psi)$ erhalten wir daraus

$$\tan(\varphi+\psi) = \frac{\tan(\varphi)+\tan(\psi)}{1-\tan(\varphi)\cdot\tan(\psi)}$$

Sie gehen für gleiche Winkel $\varphi = \psi$ in die *Doppelwinkelformeln*

$$\begin{aligned} \sin(2\varphi) &= 2\sin(\varphi)\cos(\varphi) \\ \cos(2\varphi) &= \cos^2(\varphi) - \sin^2(\varphi) \\ \tan(2\varphi) &= \frac{2\tan(\varphi)}{1-\tan^2(\varphi)} \end{aligned}$$

über, aus denen sofort hervorgeht, dass die Kreisfunktionen nicht linear sind.

Beweis. Für die linke Seite der Funktionalgleichung der Exponentialfunktion liefert die Euler'sche Formel

$$e^{i(\varphi+\psi)} = \cos(\varphi+\psi) + \sin(\varphi+\psi)i$$

Für die rechte Seite erhalten wir ebenfalls mit Hilfe der Euler'schen Formel

$$\begin{aligned} e^{i\varphi}\cdot e^{i\psi} &= \big(\cos(\varphi) + i\sin(\varphi)\big)\cdot\big(\cos(\psi)+i\sin(\psi)\big) = \\ &\big(\cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi)\big) + \big(\cos(\varphi)\sin(\psi) + \sin(\varphi)\cos(\psi)\big)i \end{aligned}$$

Dabei haben wir beim Übergang von der ersten zur zweiten Zeile die Multiplikation komplexer Zahlen verwendet. Ein Vergleich von Real- und Imaginärteilen dieser beiden komplexen Zahlen liefert die beiden Additionstheoreme. □

Weil die Additionstheoreme auch für komplexe Argumente gelten, können wir die trigonometrischen Funktionen leicht für ein beliebiges komplexes Argument $z = a + ib$ für $a, b \in \mathbb{R}$ berechnen. Zunächst ist dank des Additionstheorems

$$\begin{aligned} \cos(z) &= \cos(a+ib) = \cos(a)\cdot\cos(ib) - \sin(a)\cdot\sin(ib) \\ \sin(z) &= \sin(a+ib) = \sin(a)\cdot\cos(ib) + \cos(a)\cdot\sin(ib) \end{aligned}$$

Wie wir gesehen haben, lassen sich die trigonometrischen Funktionen eines imaginären Argumentes durch die hyperbolischen Funktionen ausdrücken. Es ist

$$\begin{aligned} \cos(ib) &= \cosh(b) \\ \sin(ib) &= i\sinh(b) \end{aligned}$$

Duch Einsetzen erhalten wir die gesuchten Formeln

$$\begin{aligned} \cos(z) &= \cos(a)\cdot\cosh(b) - i\sin(a)\cdot\sinh(b) \\ \sin(z) &= \sin(a)\cdot\cosh(b) + i\cos(a)\cdot\sinh(b) \end{aligned}$$

Nicht nur Identitäten für die trigonometrische Funktionen lassen sich durch Übergang zu den komplexen Zahlen behandeln. Auch Bestimmungsgleichungen mit Kreisfunktionen (sogn. goniometrische Gleichungen) führen nach dem Ersetzen der Kreisfunktionen zu rationalen Bestimmungsgleichungen in der Unbestimmten $z = e^{ix}$.

Beispiel. Sollen für $\alpha = \frac{3\pi}{2}$ im Intervall $[0, 2\pi)$ sämtliche reellen Lösungen der goniometrischen Gleichung

$$\cos(x) + \sin(2x)\sin(\alpha) = 2\cos^2(x)\cos(\alpha)$$

bestimmt werden, so beachtet man zunächst, dass sich diese Gleichung wegen $\sin(\alpha) = -1$ und $\cos(\alpha) = 0$ sofort zur Gleichung

$$\cos(x) - \sin(2x) = 0$$

vereinfacht.

Diese vereinfachte Gleichung kann man entweder — wie in der Schule — reell lösen, indem man die Doppelwinkelformel $\sin(2x) = 2\sin(x)\cos(x)$ benutzt. Dann lautet die vereinfachte Gleichung

$$\cos(x) - 2\sin(x)\cos(x) = 0 = \cos(x)\Big(1 - 2\sin(x)\Big)$$

Daher sind noch die beiden Gleichungen

$$\cos(x) = 0, \quad 1 - 2\sin(x) = 0, \quad \text{bzw.} \quad \cos(x) = 0, \quad \sin(x) = \frac{1}{2}$$

zu lösen und man erhält im Grundintervall $[0, 2\pi)$ die vier Lösungen $x_1 = \frac{\pi}{2}$, $x_2 = \frac{3\pi}{2}$, $x_3 = \frac{\pi}{6}$, $x_4 = \frac{5\pi}{6}$.

Alternativ ersetzt man die beiden trigonometrischen Funktionen durch die komplexe Exponentialfunktion. Mit der Substitution

$$z = e^{ix} \neq 0, \qquad \overline{z} = e^{-ix} = z^{-1}$$

ist

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2} = \frac{z + z^{-1}}{2}, \qquad \sin(x) = \frac{e^{ix} - e^{-ix}}{2i} = \frac{z - z^{-1}}{2i}$$

Damit gilt für den doppelten Winkel

$$\sin(2x) = \frac{e^{i2x} - e^{-i2x}}{2i} = \frac{z^2 - z^{-2}}{2i}, \qquad \cos(2x) = \frac{e^{i2x} + e^{-i2x}}{2} = \frac{z^2 + z^{-2}}{2}$$

und die vereinfachte goniometrische Gleichung nimmt die Form

$$\frac{z + z^{-1}}{2} - \frac{z^2 - z^{-2}}{2i} = \frac{1 + iz + iz^3 - z^4}{2iz^2} = 0$$

an. Dieser rationale Ausdruck verschwindet genau dann, wenn sein Zähler

$$1 + iz + iz^3 - z^4 = 0$$

verschwindet. Damit haben wir die goniometrische Gleichung auf eine äquivalente algebraische Gleichung vierten Grades übergeführt. Mehr oder weniger vom Schiff[35] sieht man, dass diese Gleichung die beiden komplexen Lösungen $z_{1/2} = \pm i$ und damit die komplexe Faktorisierung

$$(z + i)(z - i)(1 + iz - z^2) = 0$$

hat. Daher erhalten wir die vier Lösungen $z_1 = i$, $z_2 = -i$, $z_3 = \frac{1}{2}\big(\sqrt{3} + i\big)$ und $z_4 = \frac{1}{2}\big(-\sqrt{3} + i\big)$. Daraus erkennt man, dass die Lösungen einer algebraischen Gleichung nicht konjugiert komplex zu sein brauchen, falls die Koeffizienten des Polynoms nicht reell sind. Diese vier komplexen Zahlen haben den Betrag 1 und gehören zu den vier gesuchten Polarwinkeln $x_1 = \frac{\pi}{2}$, $x_2 = \frac{3\pi}{2}$, $x_3 = \frac{\pi}{6}$ und $x_4 = \frac{5\pi}{6}$ im Grundintervall $[0, 2\pi)$. ◯

Zum Üben geben wir eine weiteres Beispiel dieses in der Schule so beliebten Aufgabentyps.

Beispiel. Sollen alle reellen Lösungen im Grundintervall $[0, 2\pi)$ der goniometrischen Gleichung

$$\sin(x) + \sin(2x) + \sin(3x) = 0$$

bestimmt werden, so erkennt man vom Schiff aus, dass diese Gleichung aus Symmetriegründen die trivialen Lösungen $x_1 = 0$, $x_2 = \pi$, $x_3 = \frac{\pi}{2}$ und $x_4 = \frac{3\pi}{2}$ hat. Es geht also darum, die restlichen Lösungen zu finden.

Ersetzt man wie in der letzten Aufgabe die trigonometrischen Funktionen durch die komplexe Exponentialfunktion mit Hilfe von $z = e^{ix} \neq 0$ und

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2} = \frac{z + z^{-1}}{2}, \qquad \sin(x) = \frac{e^{ix} - e^{-ix}}{2i} = \frac{z - z^{-1}}{2i}$$

so geht die goniometrische Gleichung in die algebraische Gleichung

$$\frac{z - z^{-1}}{2i} + \frac{z^2 - z^{-2}}{2i} + \frac{z^3 - z^{-3}}{2i} = \frac{z - z^{-1} + z^2 - z^{-2} + z^3 - z^{-3}}{2i} = 0$$

[35] Wem das nicht gelingt beachtet, dass sich auch die Nullstellen einer komplexen Gleichung $f(z) = 0$ für differenzierbares f mit Hilfe des Tangentenverfahrens

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}$$

numerisch approximativ bestimmen lassen.

über, deren Zähler nach Multiplikation mit z^3 zur algebraischen Gleichung sechsten Grades

$$z^6 + z^5 + z^4 - z^2 - z - 1 = 0$$

wird. Man beachte, dass ihre Koeffizienten reell und daher ihre Nullstellen paarweise konjugiert komplex sind. Die bereits vom Schiff aus festgestelten Lösungen entsprechen den Werten $z_1 = 1$, $z_2 = -1$, $z_3 = i$ und $z_4 = -i$ und bestätigen dieses Muster. Wir erwarten also noch zwei weitere Lösungen und faktorisieren dazu das gefundene Polynom

$$(z-1)\cdot(z+1)\cdot(z-i)\cdot(z+i)\cdot(z^2+z+1) = (z^2-1)\cdot(z^2+1)\cdot(z^2+z+1)$$

Offensichtlich müssen wir also noch die beiden konjugiert komplexen Lösungen der quadratischen Gleichung

$$z^2 + z + 1 = 0 = \left(z + \frac{1}{2}\right)^2 - \frac{1}{4} + 1 = 0, \qquad \left(z + \frac{1}{2}\right)^2 = -\frac{3}{4}$$

bestimmen. Sie haben die Normalformen $z_5 = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ und $z_6 = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$ und gehören zu den beiden Polarwinkeln $x_5 = \frac{2\pi}{3}$ und $x_6 = \frac{4\pi}{3}$. Dabei handelt es sich um die beiden restlichen der gesuchten reellen Lösungen.

Selbstverständlich hätte man diese Lösungen auch gefunden, wenn man im Reellen geblieben wäre und die mittlere der Dreifachwinkelformeln

$$\begin{aligned}
\cos(3\varphi) &= \cos^3(\varphi) - 3\cos(\varphi)\sin^2(\varphi) \\
\sin(3\varphi) &= 3\cos^2(\varphi)\sin(\varphi) - \sin^3(\varphi) \\
\tan(3\varphi) &= \frac{3\tan(\varphi) - \tan^3(\varphi)}{1 - 3\tan^2(\varphi)}
\end{aligned}$$

verwendet hätte. Der Weg über das Komplexe zeigt allerdings die Stuktur des Problems besser und hat insbesondere den Vorteil, dass sofort klar wird, wie viele Lösungen die Gleichung hat, weil das dank des Fundamentalsatazes der Algebra für das entstehende Polynom klar ist. ◯

Mit vollständiger Induktion folgt aus der Funktionalgleichung der Exponentialfunktion das Exponentialgesetz.

Korollar. Für jeden Winkel $\varphi \in \mathbb{R}$ und jede natürlich Zahl $n \in \mathbb{N}$ gilt

$$e^{in\varphi} = (e^{i\varphi})^n$$

Diese Beziehung ist in den Schulbüchern als Formel von De Moivre bekannt. Aus ihr folgt, dass die Funktionen

$$S^1 \to \mathbb{C}, \qquad \varphi \mapsto e^{in\varphi}$$

genau den Restriktionen der homogenen harmonischen Polynomen vom Grad n in den zwei Variablen $x = \cos(\varphi)$ und $y = \sin(\varphi)$ entsprechen. In der Tat ist für $n = 0$

$$e^{i0\varphi} = 1$$

eine Konstante, d.h. ein Polynom vom Grad 0.

Für $n = 1$ ist nach der Eulerschen Formel

$$e^{i\varphi} = \cos(\varphi) + i\sin(\varphi) = x + iy$$

Für $n = 2$ ist nach der Formel von De Moivre

$$e^{i2\varphi} = (e^{i\varphi})^2 = (\cos(\varphi) + i\sin(\varphi))^2 = (x + iy)^2 = x^2 + 2ixy - y^2$$

tatsächlich ein homogenes Polynom vom Grad 2 in x und y.
Entsprechend ist nach der Binomischen Formel für $n = 3$

$$e^{i3\varphi} = (e^{i\varphi})^3 = (\cos(\varphi) + i\sin(\varphi))^3 = (x + iy)^3 = x^3 + 3ix^2y - 3xy^2 - iy^3$$

ein homogenes Polynom vom Grad $n = 3$ in x und y.
Allgemein ist

$$e^{in\varphi} = (e^{i\varphi})^n = (\cos(\varphi) + i\sin(\varphi))^n = (x + iy)^n = \sum_{k=0}^{n} i^k \cdot \binom{n}{k} \cdot x^{n-k} \cdot y^k$$

ein homogenes harmonisches Polynomen vom Grad n in x und y, dessen Real- und Imaginärteil wir früher bei der Besprechnung der Normalform einez komplexen Zahl $z = x + iy$ angetroffen haben.

In höheren Dimensionen sind die entsprechenden Funktionen als *Kugelflächenfunktionen* bekannt und nur mühsamer zu erhalten. Kugelflächenfunktionen haben in der Physik eine Bedeutung als Lösung gewisser partieller Differentialgleichungen. Sie treten beispielsweise bei der Berechnung des Wasserstoffspektrums und von Atomorbitalen auf, da die beschreibende zeitunabhängige Schrödingergleichung den Laplace-Operator enthält man das Problem am besten in Kugelkoordinaten löst. Auch die in der Elektrostatik auftretenden Randwertprobleme (Dirichlet- bzw. Poisson-Problem) können durch die Entwicklung nach Kugelflächenfunktionen gelöst werden. In der Geophysik und Gedäsie werden die Kugelflächenfunktionen bei der Approximation des Geoids und des Magnetfeldes verwendet.

Setzen wir die Euler'sche Formel in die De Moivre'sche Formel ein, ergibt sich

$$\cos(n\varphi) + i\sin(n\varphi) = \big(\cos(\varphi) + i\sin(\varphi)\big)^n$$

Um die Normalform der rechten Seite dieser Gleichung zu erhalten, benutzen wir die früher berechnete Normalform der Potenz einer komplexen Zahl und erhalten für die gesuchte Normalform der Formel von de Moivre den Ausdruck

$$\begin{aligned}\big(\cos(\varphi) + i\sin(\varphi)\big)^n &= \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \binom{n}{2k} \cos^{n-2k}(\varphi)\sin^{2k}(\varphi) + \\ &\quad i\Big(\sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} (-1)^k \binom{n}{2k+1} \cos^{n-(2k+1)}(\varphi)\sin^{2k+1}(\varphi)\Big)\end{aligned}$$

Ein Vergleich der Real- und der Imaginärteile dieses Ausdruckes mit der linken Seite der ursprünglichen Gleichung liefert folgende *n-fach Winkelformeln*. Sie spielen bei der Untersuchung regulärer n-Ecke eine zentrale Rolle.

Satz. Für beliebige Winkel $\varphi \in \mathbb{R}$ und eine natürliche Zahl $n \in \mathbb{N}$ gilt:

$$\begin{aligned}
\cos(n\varphi) &= \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \binom{n}{2k} \cos^{n-2k}(\varphi) \sin^{2k}(\varphi) \\
\sin(n\varphi) &= \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} (-1)^k \binom{n}{2k+1} \cos^{n-(2k+1)}(\varphi) \sin^{2k+1}(\varphi) \\
\tan(n\varphi) &= \frac{\displaystyle\sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} (-1)^k \binom{n}{2k+1} \tan^{2k+1}(\varphi)}{\displaystyle\sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \binom{n}{2k} \tan^{2k}(\varphi)}
\end{aligned}$$

Die Koeffizienten in den n-fach Winkelformeln für die Kreisfunktionen sind gewisse Binomialkoeffizienten: Jede zweite Zahl aus den Diagonalen des Pascal'schen Dreiecks, wobei die Vorzeichen abwechseln.

Beweis. Für den Beweis der dritten Behauptung beachte man, dass aus den ersten beiden Beziehungen die Gleichungen

$$\begin{aligned}
\frac{\cos(n\varphi)}{\cos^n(\varphi)} &= \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^k \binom{n}{2k} \tan^{2k}(\varphi), \\
\frac{\sin(n\varphi)}{\cos^n(\varphi)} &= \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} (-1)^k \binom{n}{2k+1} \tan^{2k+1}(\varphi)
\end{aligned}$$

folgen, aus denen sich die n-fach Winkelformel für den Tangens durch Quotientenbildung ergibt. □

Beispiel. Es gelten folgende 4-fach Winkelformeln:

$$\begin{aligned}
\cos(4\varphi) &= \cos^4(\varphi) - 6\cos^2(\varphi)\sin^2(\varphi) + \sin^4(\varphi) \\
\sin(4\varphi) &= 4\cos^3(\varphi)\sin(\varphi) - 4\cos(\varphi)\sin^3(\varphi) \\
\tan(4\varphi) &= \frac{4\tan(\varphi) - 4\tan^3(\varphi)}{1 - 6\tan^2(\varphi) + \tan^4(\varphi)}
\end{aligned}$$

Die Koeffizienten in der n-fach Winkelformel für den Tangens sind also die Binomialkoeffizienten, die zwischen Nenner und Zähler in einem Zickzack-Muster abwechseln und im Nenner mit 1 beginnen. Die Vorzeichen alternieren in 2-er Schritten. ◯

Beispiel. Der belgische Mathematiker van Roomen forderte im Jahr 1593, wie es damals in der Forschung üblich war, die Mathematiker der ganzen Welt heraus, indem er von ihnen verlangte, die Monstergleichung

$$\begin{aligned}
&x^{45} - 45x^{43} + 945x^{41} - 12'300x^{39} + 111'150x^{37} - 740'259x^{35} + 3'764'565x^{33} \\
&-14'945'040x^{31} + 46'955'700x^{29} - 117'679'100x^{27} + 236'030'652x^{25} \\
&-378'658'800x^{23} + 483'841'800x^{21} - 488'494'125x^{19} + 384'942'375x^{17} \\
&-232'676'280x^{15} + 105'306'075x^{13} - 34'512'075x^{11} + 7'811'375x^9 \\
&-1'138'500x^7 + 95'634x^5 - 3'795x^3 + 45x = r
\end{aligned}$$

für eine gegebene feste Zahl $r \in \left[0, \frac{1}{2}\right]$ zu lösen. Anlässlich eines Besuches beim französischen König Henri *IV* sprach der holländische Botschafter dem König ironisch sein Beileid dafür aus, dass Frankreich keine Mathematiker habe, die dieser Herausforderung gewachsen seien. Durch so viel Nationalismus angestachelt, liess der König seinen Untertanen Vietà kommen, der sich das Problem ansah und innerhalb weniger Minuten die positiven Nullstellen dieser Gleichung fand, indem er auf den Zusammenhang zwischen Algebra und Geometrie hinwies. Er bemerkte nämlich, dass wenn man $\sin(45\varphi)$ nach der Formel von de Moivre entwickelt, mit Hilfe des Satzes von Pythagoras die geraden Potenzen von $\cos(\varphi)$ durch $\sin(\varphi)$ ersetzt und dann $x = 2\sin(\varphi)$ setzt, die Hälfte der linken Seite der Monstergleichung entsteht. Die gesuchte Lösung erhielt er also, indem er ausgehend von

$$2\sin(45\varphi) = r, \qquad \varphi = \frac{\arcsin(\frac{r}{2})}{45}$$

die Werte von $x = 2\sin(\varphi)$ berechnete. Ob die beiden Politiker allerdings dieser Überlegung folgen konnten, ist nicht überliefert. ◯

Die Formel von De Moivre lässt sich zum *Potenzgesetz* verallgemeinern.

Korollar. Es sei $z = r \cdot e^{i\varphi}$ eine komplexe Zahl. Dann gilt $z^n = r^n \cdot e^{in\varphi}$.

Diese Beziehung kann zur Berechnung von hohen Potenzen einer komplexen Zahl z benutzt werden. Dazu bringt man zunächst die komplexe Zahl auf Exponentialform $z = r \cdot e^{i\varphi}$ und kann dann deren n-te Potenz z^n in Exponentialfform leicht berechnen. Von dieser komplexen Zahl wiederum lassen sich dann Real- und Imaginärteil leicht bestimmen, indem man in die Normalform umrechnet.

Beispiel. Zur Berechnung der zehnten Potenz der komplexe Zahl $z = 1 + i$ bestimmen wir zunächst die Exponentialform von z. Es gilt $z = \sqrt{2} \cdot e^{i\frac{\pi}{4}}$. Damit erhalten wir $z^{10} = (\sqrt{2})^{10} \cdot e^{i\frac{5\pi}{2}} = 32 \cdot e^{i\frac{\pi}{2}} = 32i$. ◯

Man beachte, dass aus der Formel von De Moivre folgt, dass der Polarwinkel von z^n der n-fache Polarwinkel von z ist. Das hat zur Folge, dass die Funktion $z \mapsto z^n$ die komplexe Ebene $\mathbb{C}$ n-fach um den Ursprung windet. Aus dieser geometrischen Beobachtung folgt der Fundamentalsatz der Algebra.

Beispiel. Für eine natürliche Zahl $n \geq 2$ und eine beliebige komplexe Zahl $w \neq 0$ hat die Gleichung

$$z^n = w$$

wegen des Fundamentalsatzes der Algebra genau n Lösungen. Insbesondere sind n-te Wurzeln einer komplexen Zahl $w \neq 0$ nicht eindeutig bestimmt. So etwas wie *die* n-te Wurzel eine komplexen Zahl w gibt es also nicht[36]. Deshalb führen wir für n-te Wurzeln gar nicht erst eine eigene Bezeichnung ein.

[36] In den Anwendungen wird häufig eine dieser Lösungen willkürlich ausgezeichnet. Wer in der Mathematik aber willkürliche Wahlen trifft, darf nicht erstaunt sein, wenn die entstehende verstümmelte Theorie unnatürlich wird! In unserem Fall würde eine solche Wahl die Symmetrie des regulären n-Ecks zerstören, was offensichtlich einem barbarischen, mutwilligen Zerstörungsakt gleichkommt. Viele Resultate — etwa im Umkreis der Fourier-Theorie zur Untersuchung periodischer Phänomene — folgen gerade aus der *Gesamtheit* dieser Symmetrien. Wenn man also in einem mathematische Problem keine *kanonische* Wahl treffen kann, ist es empfehlenswert, überhaupt keine Wahl zu treffen und sich nach einer natürlicheren Theorie umzusehen!

Um die Menge *aller* Lösungen der Gleichung $z^n = w$ zu finden, stellt man zunächst w in Exponentialform $w = r \cdot e^{i\varphi}$ dar und macht für die gesuchte Lösung den Ansatz $z = \rho \cdot e^{i\lambda}$. Setzen wir diesen Ansatz in die definierende Gleichung ein, geht sie über in

$$\rho^n \cdot e^{in\lambda} = r \cdot e^{i\varphi}$$

woraus die beiden Gleichungen

$$\rho^n = r, \qquad e^{in\lambda} = e^{i\varphi}$$

folgen. Aus der ersten Gleichung erhalten wir für den Betrag

$$\rho = \sqrt[n]{r} = \sqrt[n]{|w|}$$

Aus der zweiten Gleichung folgt, dass der Polarwinkel λ der gesuchten komplexen Zahlen ferner die Bedingung

$$e^{in\lambda} = e^{i\varphi}, \quad \text{bzw.} \quad e^{i(n\lambda-\varphi)} = 1$$

erfüllen. Diese Bedingung ist *nicht* äquivalent zu $n\lambda - \varphi = 0$ sondern, wegen der Periodizität der Kreisfunktionen, zur Bedingung

$$n\lambda - \varphi = k2\pi \quad \text{für } k \in \mathbb{Z}$$

Daraus folgt, dass der gesuchte Polarwinkel λ einen der Werte

$$\lambda_k = \frac{\varphi}{n} + k\frac{2\pi}{n} \quad \text{für } k \in \mathbb{Z}$$

annehmen muss. Zwei Werte von k, die sich um ein Vielfaches von n unterscheiden, liefern für λ zwei Winkel, die sich um 2π unterscheiden. Daher können die zugehörigen komplexen Zahlen z nicht unterschieden werden. Es bleiben also genau n verschiedene Winkel für λ übrig, die zu den Werten $0 \le k \le n-1$ gehören. Die gesuchten n Lösungen der Gleichung

$$z^n = w$$

haben also die Exponentialformen

$$z_k = \sqrt[n]{r} \cdot e^{i(\frac{\varphi}{n} + k\frac{2\pi}{n})}, \quad \text{für } 0 \le k \le n-1$$

Zur Kontrolle berechne man die n-te Potenz von z_k. Diese n komplexen Zahlen heissen n-te *Wurzeln* der komplexen Zahl $w = r \cdot e^{i\varphi}$. Sie liegen auf einem Kreis mit dem Zentrum in 0, dessen Radius $\sqrt[n]{r}$ ist und gehören zu den Winkeln $\frac{\varphi}{n} + k\frac{2\pi}{n}$, die den Kreis in n gleiche Teile teilen.
Für $w = 1$ redet man deshalb auch von der *Kreisteilungsgleichung* $z^n = 1$ mit den sogn. n-ten komplexen *Einheitswurzeln*

$$\zeta_n^k = e^{i(k\cdot\frac{2\pi}{n})} = \cos\left(k \cdot \frac{2\pi}{n}\right) + i\sin\left(k \cdot \frac{2\pi}{n}\right), \quad \text{für } 0 \le k \le n-1$$

als Lösungen. Sie liegen auf dem Einheitskreis und bilden ein reguläres n-Eck mit einer Ecke in $\zeta_n^0 = 1$. Die Zahl $\zeta_n = e^{i(\frac{2\pi}{n})}$ heisst *primitive* n-te Einheitswurzel.

Die restlichen Einheitwurzeln sind Potenzen der primitiven Einheitswurzel, da für $0 \leq k \leq n-1$ die Gleichung

$$\zeta_n^k = (\zeta_n)^k$$

gilt, was die gewählte Bezeichnungsweise rechtfertigt. Allgemeiner heisst $\zeta \in \mathbb{C}$ *primitive* n-te Einheitswurzel, falls $\zeta^n = 1$ ist, aber für alle $1 \leq k \leq n-1$ die Bedingung $\zeta^k \neq 1$ gilt. ◯

Beispiel. Die Kreisteilungsgleichung

$$z^n = 1$$

und ihre Lösungen, d.h. die komplexen n-ten Einheitswurzeln

$$\zeta_n^k = e^{i(k \cdot \frac{2\pi}{n})} = \cos\left(k \cdot \frac{2\pi}{n}\right) + i \sin\left(k \cdot \frac{2\pi}{n}\right), \qquad 0 \leq k \leq n-1$$

spielen im Zusammenhang mit der sogn. Fourier-Transformation eine zentrale Rolle, weil sie als Eigenwerte und Eigenvektoren einer fundamentalen linearen Abbildung

$$T_n \colon \mathbb{R}^n \to \mathbb{R}^n, \qquad \vec{x} \mapsto T_n \cdot \vec{x}$$

auftreten. Der *zyklische Shift* T_n ordnet einem Vektor $\vec{x} \in \mathbb{R}^n$ den Vektor $T_n \cdot \vec{x}$ zu, dessen Komponenten durch

$$(T_n \cdot \vec{x})_j = \vec{x}_{j+1 \bmod n}, \qquad 1 \leq j \leq n$$

erklärt sind. Weil T_n gerade durch die orthogonale Begleitermatrix

$$T_n := B(z^n - 1) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n,n}$$

des Kreisteilungspolynoms dargestellt wird, in der jede Zeile aus der vorangehenden durch einen zyklischen Rechtsshift hervorgeht[37], hat sie das Kreisteilungspolynom als Minimalpolynom und daher die n verschiedenen komplexen n-ten Einheitswurzeln als Eigenwerte.

Wie alle zirkulären Matrizen, wird die Matrix T_n durch die Vandermonde-Matrix der komplexen Einheitswurzeln

$$V(1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-2}, \zeta_n^{n-1})$$

diagonalisiert, für die also

$$\mathrm{F}_n = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \zeta_n^1 & \zeta_n^2 & \cdots & \zeta_n^{n-2} & \zeta_n^{n-1} \\ 1 & \zeta_n^2 & \zeta_n^{2\cdot 2} & \cdots & \zeta_n^{2\cdot(n-2)} & \zeta_n^{2\cdot(n-1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \zeta_n^{n-2} & \zeta_n^{2\cdot(n-2)} & \cdots & \zeta_n^{(n-2)\cdot(n-2)} & \zeta_n^{(n-2)\cdot(n-1)} \\ 1 & \zeta_n^{n-1} & \zeta_n^{2\cdot(n-1)} & \cdots & \zeta_n^{(n-1)\cdot(n-2)} & \zeta_n^{(n-1)\cdot(n-1)} \end{pmatrix} \in \mathbb{C}^{n,n}$$

[37] Solche Matrizen bezeichnet man als zyklisch.

gilt. Das typische Element der symmetrischen Fouriermatrix ist also[38]

$$(\mathrm{F}_n)_{j,k} = \zeta_n^{(j-1)\cdot(k-1)} = e^{i\cdot\frac{2\pi}{n}(j-1)\cdot(k-1)}, \qquad 1 \le j,k \le n$$

Geometrisch erhält man die Spaltenvektoren der Matrix F_n, d.h. Eigenvektoren von T_n, indem man die Ecken des regelmässigen n-Ecks, das von den komplexen n-ten Einheitswurzeln gebildet wird, im Gegenuhrzeigersinn in Schritten der wachsenden Weiten $0 \le k \le n-1$ durchläuft.

Weil die komplexen n-ten Einheitswurzeln alle voneinander verschieden sind, ist die Matrix F_n invertierbar. Es stellt sich heraus, dass diese Inverse F_n^{-1} dank der Orthogonalität von T_n eine besonders einfache Gestalt hat. Sie ist bis auf einen Faktor n unitär.

Satz. Die Fouriermatrix $F_n \in \mathbb{C}^{n,n}$ erfüllt die Orthogonalitätsbeziehung

$$\mathrm{F}_n \cdot \overline{\mathrm{F}}_n^T = nE_n = \overline{\mathrm{F}}_n^T \cdot \mathrm{F}_n.$$

Dabei bezeichnet $\overline{F}_n \in \mathbb{C}^{n,n}$ die Matrix, die aus F_n durch Konjugieren sämtlicher Elemente hervorgeht.

Beweis. Um die Orthogonalitätsbeziehung einzusehen, müssen wir einfach das Matrizenprodukt $\mathrm{F}_n \cdot \overline{\mathrm{F}}_n^T$ berechnen. Nun gilt definitionsgemäss

$$(\overline{\mathrm{F}}_n^T)_{k,l} = \zeta_n^{-(l-1)\cdot(k-1)} = e^{-i\frac{2\pi}{n}(l-1)\cdot(k-1)}, \qquad 1 \le k,l \le n$$

und nach Definition des Matrizenproduktes gilt für seine Elemente

$$\begin{aligned}
(\mathrm{F}_n \cdot \overline{\mathrm{F}}_n^T)_{j,l} &= \sum_{k=1}^{n} (F_n)_{j,k} \cdot (\overline{\mathrm{F}}_n^T)_{k,l} = \sum_{k=1}^{n} \zeta_n^{(j-1)(k-1)} \cdot \zeta_n^{-(l-1)(k-1)} \\
&= \sum_{k=1}^{n} \zeta_n^{(j-1)(k-1)-(l-1)(k-1)} = \sum_{k=1}^{n} \zeta_n^{(k-1)\cdot(j-l)} \\
&= \sum_{k=1}^{n} (\zeta_n^{(j-l)})^{k-1} = \begin{cases} n & \text{falls } j = n \\ \frac{\left(\zeta_n^{(j-l)}\right)^n - 1}{\zeta_n^{(j-l)} - 1} & \text{falls } j \ne n \end{cases}
\end{aligned}$$

Dabei ist das letzte Gleichheitszeichen auf Grund der Summenformel für die geometrische Reihe gerechtfertigt. Falls wir nun berücksichtigen, dass

$$\zeta_n^{(j-l)} = e^{i\frac{2\pi}{n}\cdot(j-l)}, \qquad 1 \le j,l \le n$$

eine komplexe n-te Einheitswurzel ist und daher definitionsgemäss

$$\left(\zeta_n^{(j-l)}\right)^n = 1, \qquad \text{bzw.} \qquad \left(\zeta_n^{(j-l)}\right)^n - 1 = 0$$

[38] Insbesondere in diesem Zusammenhang ist es üblich, die Indizes von 0 bis $n-1$ statt von 1 bis n laufen zu lassen. Dann hat das typische Element der Fourier-Matrix die Form

$$(\mathrm{F}_n)_{j,k} = \zeta_n^{j\cdot k} = e^{i\frac{2\pi}{n}j\cdot k}, \qquad 0 \le j,k \le n-1$$

Selbstverständlich müssen dann alle Beziehungen entsprechend modifiziert werden.

gilt, erhalten für die Elemente des gesuchten Matrizenproduktes

$$(\mathrm{F}_n \cdot \overline{\mathrm{F}}_n^T)_{j,l} \quad = \quad \begin{cases} n & \text{falls } j = n \\ 0 & \text{falls } j \neq n \end{cases}$$

was genau der behaupteten Orthogonalitätsbeziehung entspricht. □

Auf Grund der soeben bewiesenen Beziehung gilt also für die Inverse von F_n

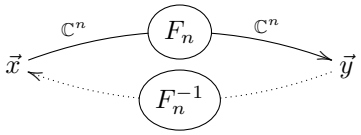
$$F_n^{-1} = \frac{1}{n} \cdot \overline{F}_n^T \in \mathbb{C}^{n,n}$$

und das typische Element der inversen Fouriermatrix hat die Form

$$(\mathrm{F}_n^{-1})_{j,k} = \frac{1}{n} \cdot \zeta_n^{-(j-1)\cdot(k-1)}, \qquad 1 \leq j,k \leq n$$

Geometrisch erhält man die Spaltenvektoren der Matrix F_n^{-1}, indem man das von den komplexen n-ten Einheitswurzeln gebildete regelmässige n-Eck im Uhrzeigersinn in Schritten der wachsenden Weiten $0 \leq k \leq n-1$ durchläuft.

Multiplikation eines Vektors $\vec{x} \in \mathbb{C}^n$ mit der Fouriermatrix F_n wird von den Anwendern[39] als *diskrete Fouriertransformation* (DFT_n) oder als *Fouriersynthese* bezeichnet. Entsprechend wird die Multiplikation eines Vektors $\vec{y} \in \mathbb{C}^n$ mit der inversen Fouriermatrix F_n^{-1} als *inverse diskrete Fouriertransformation* (IDFT_n) oder als *Fourieranalyse* bezeichnet. Die Fouriersynthese $\vec{x} \mapsto F_n \cdot \vec{x}$ und ihre inverse Fourieranalyse $\vec{y} \mapsto F_n^{-1} \cdot \vec{y}$ liefern also ein Paar inverser linearer Prozesse.



CAS. Die Folge der Fourier-Matrizen F_n kann in Sage mit Hilfe des folgenden Kodes definiert werden:

```
def DFT(n):
    return matrix(n, n, lambda j, k: e^(i*2*pi/n*j*k))
```

Analog definiert man die Folge der inversen Fouriermatrix F_n^{-1} durch den Kode

```
def IDFT(n):
    return matrix(n, n, lambda j, k: 1/n*e^(-i*2*pi/n*j*k))
```

Die diskrete Forier-Transformation eines Vektors $\vec{x} \in \mathbb{C}^n$ berechnet man dann, indem man ihn mit der Matrix `DFT(n)` multipliziert. ◇

Das Produkt der Matrix F_n mit einem Vektor $\vec{x} \in \mathbb{C}^n$ ist definitionsgemäss

$$y_j = (F_n \cdot \vec{x})_j = \sum_{k=1}^{n} \zeta_n^{(j-1)\cdot(k-1)} \cdot x_k, \quad 1 \leq j \leq n, \qquad \text{(Synthese, DFT)}$$

Entsprechend gilt für das Produkt der Matrix F_n^{-1} mit dem Vektor $\vec{y} \in \mathbb{C}^n$

$$x_k = (F_n^{-1} \cdot \vec{y})_k = \frac{1}{n} \cdot \sum_{j=1}^{n} \zeta_n^{-(k-1)\cdot(j-1)} \cdot y_j, \quad 1 \leq k \leq n, \qquad \text{(Analyse, IDFT)}$$

[39] In der Physik, Datenanalyse, Signalverarbeitung etc. sind auch leicht andere Konventionen im Gebrauch, was die Verteilung des skalaren Faktors betrifft.

In Formeln dieser Art wird stillschweigend abgemacht, dass die Indizes ausserhalb von $1, \ldots, n$ zyklisch wieder in diesen Indexbereich abgebildet werden, d.h. zyklisch modulo n genommen werden, wie man kurz sagt.

Beispiel. Weil die diskrete Fourriertransformation mit der Faltung verträglich ist, lassen sich mit ihr Polynome

$$f(x) = \sum_{i=0}^{m-1} a_i x^i, \qquad g(x) = \sum_{j=0}^{n-1} b_j x^j \in \mathbb{C}[x]$$

der Grade $\deg(f) = m-1$ bzw. $\deg(g) = n-1$, d.h. mit m bzw. n Koeffizienten, multiplizieren. Dazu beachten wir, dass sich komplexe Polynome nicht nur durch die Angabe ihrer Koeffizientenvektoren, sondern alternativ auch durch ihre Werte an gewissen Stützstellen vollständig beschreiben lassen. In der Stützstellendarstellung lassen sich dann die beiden Polynome sehr effizient elementweise multiplizieren. Am Schluss kehren wir zur Koeffizientendarstellung zurück. Es ergibt sich folgendes Vorgehen zur Berechung des Polynomproduktes

$$p(x) = \sum_{k=0}^{l-1} c_k x^k = \sum_{k=0}^{m+n-2} c_k x^k = f(x) \cdot g(x), \qquad c_k = \sum_{i+j=k} a_i \cdot b_j$$

vom Grad $\deg(p) = l-1 = m+n-2$ mit insgesamt $l = m+n-1$ Koeffizienten:

1. Bestimme die Werte von f und g an den l Stützstellen $x_0, \ldots, x_{l-1}$, wobei $l-1 = m+n-2$ die Summe der Grade der beiden Polynome ist.
2. Berechne die Produkte $p(x_k) = f(x_k) \cdot g(x_k)$ an diesen l Stützstellen.
3. Bestimme aus den Werten $p(x_k)$ an den Stützstellen die Koeffizienten c_k.

Ein Blick auf die Formel für die Fourier-Koeffizienten zeigt,

$$y_k = (F_l \cdot \vec{x})_k = \sum_{j=0}^{l-1} x_j \cdot \zeta_l^{k \cdot j} = \sum_{j=0}^{l-1} x_j \cdot (\zeta_l^k)^j, \qquad 0 \le k \le l-1$$

dass wir als Stützstellen zweckmässigerweise die l-ten komplexen Einheitswurzeln, d.h.

$$x_k = \zeta_l^k = e^{i\frac{2\pi}{l} \cdot k}, \qquad 0 \le k \le l-1$$

wählen, weil wir dann im ersten Schritt mit den Koeffizientenvektoren $\vec{a}$ und $\vec{b}$ der beiden gegebenen Polynome einfach eine DFT_l, d.h. eine Multiplikation mit der Matrix F_l, durchführen müssen und dann im dritten Schritt mit Hilfe der entsprechenden IDFT_l die Koeffizienten des Produktes erhalten.

Um beispielsweise die Polynome $f(x) = 2x^2 + 3x - 4$ und $g(x) = x - 1$ zu multiplizieren, für die $m = 3$, $n = 2$ und damit $l = 3 + 2 - 1 = 4$ gilt, bestimmen wir zunächst die diskrete Fourietransformation ihrer beiden Koeffizientenvektoren

$$\vec{a} = \begin{pmatrix} -4 \\ 3 \\ 2 \\ 0 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{C}^4$$

und erhalten die beiden Vektoren

$$\mathrm{DFT}_4(\vec{a}) = F_4 \cdot \vec{a} = \begin{pmatrix} 1 \\ -6+3i \\ -5 \\ -6-3i \end{pmatrix}, \quad \mathrm{DFT}_4(\vec{b}) = F_4 \cdot \vec{b} = \begin{pmatrix} 0 \\ -1+i \\ -2 \\ -1-i \end{pmatrix} \in \mathbb{C}^4.$$

Ihr kompontentenweises Produkt liefert

$$\mathrm{DFT}_4(\vec{c}) = \mathrm{DFT}_4(\vec{a}) * \mathrm{DFT}_4(\vec{b}) = \begin{pmatrix} 0 \\ 3-9i \\ 10 \\ 3+9i \end{pmatrix} \in \mathbb{C}^4$$

Daher gilt für den gesuchten Koeffizientenvektor $\vec{c}$ des Produktes

$$\vec{c} = \mathrm{IDFT}_4(\mathrm{DFT}_4(\vec{a}) * \mathrm{DFT}_4(\vec{b})) = \frac{1}{4}\begin{pmatrix} 16 \\ -28 \\ 4 \\ 8 \end{pmatrix} = \begin{pmatrix} 4 \\ -7 \\ 1 \\ 2 \end{pmatrix} \in \mathbb{C}^4$$

wie man mit dem Sage-Kode bestätigt. Damit gilt für das Produkt der beiden Polynome

$$p(x) = f(x) \cdot g(x) = 4 - 7x + x^2 + 2x^3$$

wie man selbstverständlich durch Ausmultiplizieren bestätigt. ◯

Wer glaubt, dass das im letzten Beispiel skizzierte Verfahren zum Multiplizieren von Polynomen mit Hilfe der Fourier-Transformation aufwändiger sei als naives Ausmultiplizieren, täuscht sich gewaltig! Weil das Matrix-Vektor-Produkt $F_n \cdot \vec{x}$ und auch $F_n^{-1} \cdot \vec{y}$ die Form einer Faltung haben, können sie sehr schnell, d.h. mit $\mathcal{O}(n \cdot \log_2(n))$ Multiplikationen, durchgeführt werden, obwohl der beschriebene Standardalgorithmus zur Berechnung des Matrizenproduktes

$$F_n \cdot \vec{x} = \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \in \mathbb{C}^n$$

die Komplexität $\mathcal{O}(n^2)$ hat. Die Idee der sogn. FFT, die auf Gauss zurückgeht, besteht darin, dass man für eine Zweierpotenz[40] $n = 2^q$ den Datenvektor $\vec{x} \in \mathbb{C}^n$ in zwei Vektoren

$$\vec{x}_{\text{ungerade}} = \begin{pmatrix} x_1 \\ x_3 \\ \cdots \\ x_{n-1} \end{pmatrix} \in \mathbb{C}^{\frac{n}{2}}, \qquad \vec{x}_{\text{gerade}} = \begin{pmatrix} x_2 \\ x_4 \\ \cdots \\ x_n \end{pmatrix} \in \mathbb{C}^{\frac{n}{2}}$$

der halben Dimension aufspaltet und dann mit Hilfe der beiden Vektoren

$$\vec{y}_{\text{ungerade}} := F_{\frac{n}{2}} \cdot \vec{x}_{\text{ungerade}}, \qquad \vec{y}_{\text{gerade}} := F_{\frac{n}{2}} \cdot \vec{x}_{\text{gerade}}$$

[40] Dies kann mit geeigneter Auffüllung mit Koeffizienten 0 immer erreicht werden.

den gesuchten Vektor $\vec{y}$ der unbekannten Fourier-Koeffizienten konstruiert. Dieses Verfahren kann rekursiv angewandt werden: Ein Problem der Dimension $n = 2^q$ wird zu 2 Problemen der Dimension $\frac{n}{2} = 2^{q-1}$ reduziert etc. bis es schliesslich zu $n = 2^q$ trivialen Problemen der Dimension 1 wird. Die Kosten pro Schritt sind linear, d.h. von der Ordnung $\mathcal{O}(n)$ und es sind insgesamt $q = \log_2(n)$ Reduktionsschritte nötig, was die behauptete Komplexität erklärt.

Die für viele Anwendungen wichtige Idee der schnellen Fourier-Transformation kann man auch an Hand der folgenden Umformung erkennen, in der die Bestimmung des Fourier-Koeffizienten y_k für $0 \leq k \leq n-1$ dank der Beziehung

$$\zeta_n^{k\cdot 2j} = \zeta_{\frac{n}{2}}^{k\cdot j}, \qquad (n \text{ gerade})$$

auf zwei kleinere Summen mit geraden bzw. ungeraden Werten von k zurückgeführt wird.

$$\begin{aligned} y_k &= (F_n \cdot \vec{x})_k = \sum_{j=0}^{n-1} x_j \cdot \zeta_n^{k\cdot j} = \sum_{j=0}^{\frac{n}{2}-1} x_{2j}\zeta_n^{k\cdot 2j} + \sum_{j=0}^{\frac{n}{2}-1} x_{2j+1}\zeta_n^{k\cdot(2j+1)} \\ &= \sum_{j=0}^{\frac{n}{2}-1} x_{2j}\zeta_{\frac{n}{2}}^{k\cdot j} + \zeta_n^k \cdot \sum_{j=0}^{\frac{n}{2}-1} x_{2j+1}\zeta_{\frac{n}{2}}^{k\cdot j} \end{aligned}$$

Damit kann die Berechnung der Fourier-Koeffizienten rekursiv auf zwei Probleme halber Grösse (und linearem Organisationsaufwand) zurückgeführt werden, so dass der erforderliche Aufwand nur $\mathcal{O}(n \cdot \log_2(n))$ Multiplikationen beträgt. Kombiniert man die schnelle Fouriertransformation mit der beschriebenen Methode zur Multiplikation von Polynomen $f, g \in \mathbb{C}[x]$, lässt sich ein solches Produkt mit höchstens $\mathcal{O}(n \cdot \log_2(n))$ (komplexen) Multiplikationen berechnen.

Bekanntlich lässt sich die Multiplikation ganzer Zahlen auf die Multiplikation von Polynomen zurückführen. Um beispielsweise die ganzen Zahlen $a = 2357$ und $b = 2468$ zu multiplizieren, betrachten wir die beiden zugehörigen Polynome

$$f(x) = 7 + 5x + 3x^2 + 2x^3, \qquad g(x) = 8 + 6x + 4x^2 + 2x^3$$

und berechnen ihr Produkt

$$p(x) = f(x) \cdot g(x) = 56 + 82x + 82x^2 + 68x^3 + 34x^4 + 14x^5 + 4x^6.$$

Das gesuchte Produkt ergibt sich dann durch Substitution

$$a \cdot b = f(10) \cdot g(10) = p(10) = 5'817'076.$$

Die schnelle Fouriertransformation liefert also insbesondere einen sehr effizienten Algorithmus zur Multiplikation (grosser) ganzer Zahlen. ◯

Den Spezialfall $n = 2$ der Quadratwurzeln haben wir seinerzeit bereits in Normalform behandelt und betrachten ihn nun im Licht der Exponentialform. Um die spezielle quadratische Gleichung $z^2 = w$ zu lösen, stellen wir w zuerst in Exponentialform $w = r \cdot e^{i\varphi}$ dar. Dann findet man leicht die beiden Lösungen in Exponentialform

$$z_k = \sqrt{r} \cdot e^{i(\frac{\varphi}{2} + k\pi)}, \quad k = 0, 1$$

die man schliesslich in Normalform umrechnet. Dabei treten folgende Fälle auf:

- Falls w reell und $w \geq 0$ ist, so ist $r = w$ und $\varphi = 0$. Die beiden komplexen Lösungen haben dann wegen $e^{i0} = 1$ und $e^{i\pi} = -1$ die Form $z_1 = \sqrt{w}$ und $z_2 = -\sqrt{w}$.

- Wenn w reell und $w < 0$ ist, so ist $r = -w$ und $\varphi = \pi$. Die beiden komplexen Lösungen sind dann wegen $e^{i\frac{\pi}{2}} = i$ und $e^{i\frac{3\pi}{2}} = -i$ rein imaginär und lauten $z_1 = \sqrt{-w}i$ und $z_2 = -\sqrt{-w}i$.

- Im allgemeinen Fall erhält man zwei komplexe Zahlen z_1 und z_2, die ein reguläres 2-Eck bilden, d.h. punktsymmetrisch bezüglich des Ursprungs sind, so dass als $z_2 = -z_1$ sein muss.

Beispiel. Um die Gleichung $z^2 = -3 + 4i$ zu lösen, berechnen wir zunächst die Exponentialform von $w = -3 + 4i = 5 \cdot e^{i\varphi}$. Es gilt

$$5\cos(\varphi) = -3, \qquad 5\sin(\varphi) = 4, \qquad \tan(\varphi) = -\frac{4}{3}.$$

Der Polarwinkel von w beträgt also $\varphi \approx 2.2143$ bzw. im Gradmass $\varphi \approx 126.86°$. Daher liegt w im zweiten Quadranten und hat den Betrag $r = 5$.

Um die Normalform der beiden komplexen Lösungen in Exponentialform

$$z_k = \sqrt{5} \cdot e^{i\frac{\varphi}{2} + k\cdot\pi}, \qquad k = 0, 1$$

mit rationalen Mitteln zu bestimmen, setzen wir nicht einfach einen Näherungswert für φ ein, sondern ziehen die sogn. *Halbwinkelformeln* der Kreisfunktionen

$$\begin{aligned}
\cos\left(\frac{\varphi}{2}\right) &= \frac{1}{2}\sqrt{2 + 2\cos(\varphi)} = \frac{\sqrt{5}}{5}, \\
\sin\left(\frac{\varphi}{2}\right) &= \frac{1}{2}\sqrt{2 - 2\cos(\varphi)} = \frac{2\sqrt{5}}{5} \\
\tan\left(\frac{\varphi}{2}\right) &= \frac{1 - \cos(\varphi)}{\sin(\varphi)} = 2
\end{aligned}$$

heran, die man aus den Doppelwinkelformeln durch Auflösen erhält.

Mit Hilfe der Quadrantenrelationen

$$\begin{aligned}
\sin(\varphi + \pi) &= -\sin(\varphi) \\
\cos(\varphi + \pi) &= -\cos(\varphi) \\
\tan(\varphi + \pi) &= \tan(\varphi)
\end{aligned}$$

erhalten wir daraus die benötigten exakten Werte

$$\cos\left(\frac{\varphi}{2} + \pi\right) = -\frac{\sqrt{5}}{5}, \qquad \sin\left(\frac{\varphi}{2} + \pi\right) = -\frac{2\sqrt{5}}{5}, \qquad \tan\left(\frac{\varphi}{2} + \pi\right) = 2$$

Für die Normalform $z_k = x_k + iy_k$ der beiden Lösungen z_k im ersten und dritten Quadranten, die wir bereits in Exponentialform kennen, gilt demnach

$$x_0 = \sqrt{5} \cdot \cos\left(\frac{\varphi}{2}\right) = 1, \qquad y_0 = \sqrt{5} \cdot \sin\left(\frac{\varphi}{2}\right) = 2$$

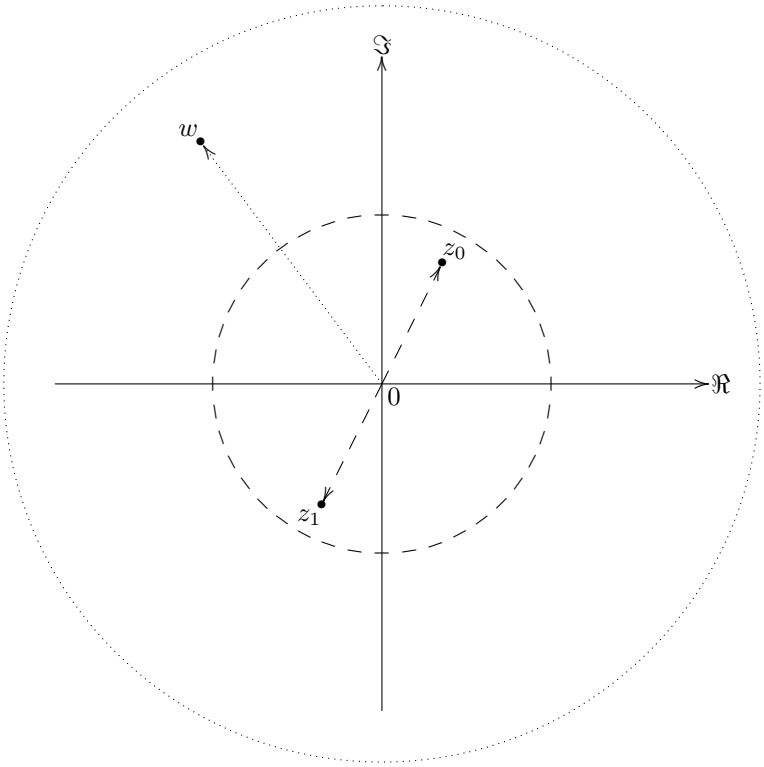


Abbildung 2.32: Die komplexen Lösungen der Gleichung $z^2 = -3 + 4i$.

bzw.

$$x_1 = \sqrt{5} \cdot \cos\left(\frac{\varphi}{2} + \pi\right) = -1, \qquad y_1 = \sqrt{5} \cdot \sin\left(\frac{\varphi}{2} + \pi\right) = -2$$

Insgesamt erhalten wir so die bereits früher bestimmten Werte $z_0 = 1 + 2i$ und $z_1 = -1 - 2i$, für die in der Tat $z_k^2 = -3 + 4i$ ist.

Man beachte, dass die beiden komplexen Zahlen, die als Lösung der speziellen Gleichung $z^2 = w$ vorkommen, punktsymmetrisch zum Ursprung liegen. ◯

Beispiel. Die vierten Wurzeln von -4 sind definitionsgemäss die Lösungen der Gleichung $z^4 = -4$. Wir erhalten zunächst die Exponentialform $-4 = 4 \cdot e^{i\pi}$ und damit die gesuchten Lösungen $z_k = \sqrt{2} \cdot e^{i(\frac{\pi}{4} + k\frac{\pi}{2})}$ für $0 \leq k \leq 3$. In Normalform lauten diese Lösungen $z_0 = 1 + i$, $z_1 = -1 + i$, $z_2 = -1 - i$, $z_3 = 1 - i$, was man leicht durch Nachrechnen bestätigt.

Diese vier komplexen Zahlen bilden die Ecken eines achsenparallelen Quadrates mit Zentrum im Ursprung und der Seitenlänge 2.

Die vier komplexen Lösungen der Gleichung $z^4 = -1$ liegen auf dem Einheitsheitskreis und bilden die Ecken eines konzentrischen Quadrates. Ihre Normalformen erhält man aus den z_k durch Division durch $\sqrt{2}$. ◯

Beispiel. Die vierten komplexen Einheitswurzeln sind Lösungen der Gleichung

$$z^4 = 1.$$

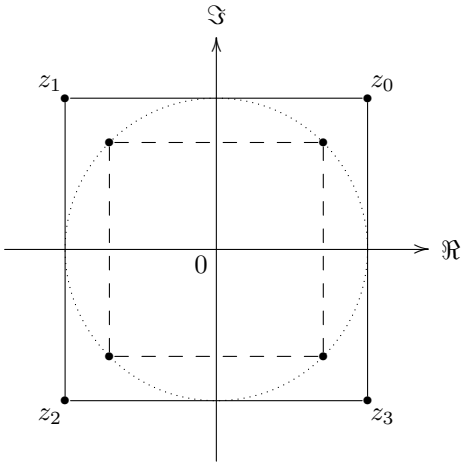


Abbildung 2.33: Die komplexen Lösungen der Gleichung $z^4 = -4$.

Wir erhalten für die vier Lösungen $\zeta_4^0 = 1$, $\zeta_4 = i$, $\zeta_4^2 = -1$, $\zeta_4^3 = -i$. Diese vier Punkte der komplexen Ebene liegen auf den Koordinatenachsen und bilden ein Quadrat im Einheitskreis gemäss folgender Figur. Eine der Ecken dieses Quadrates ist selbstverständich 1, da $1^4 = 1$ ist. Für die Normalformen der 4-ten Einheitswurzeln gilt also

k	ζ_4^k	$x_k + iy_k$	$N(x_k + iy_k)$
0	$e^{i0\cdot\frac{2\pi}{4}}$	$\cos(0) + i\sin(0)$	1
1	$e^{i1\cdot\frac{2\pi}{4}}$	$\cos\left(1\cdot\frac{2\pi}{4}\right) + i\sin\left(1\cdot\frac{2\pi}{4}\right)$	i
2	$e^{i2\cdot\frac{2\pi}{4}}$	$\cos\left(2\cdot\frac{2\pi}{4}\right) + i\sin\left(2\cdot\frac{2\pi}{4}\right)$	-1
3	$e^{i3\cdot\frac{2\pi}{4}}$	$\cos\left(3\cdot\frac{2\pi}{4}\right) + i\sin\left(3\cdot\frac{2\pi}{4}\right)$	$-i$

Diese Koordinaten sind dann zyklisch modulo 4 zu nehmen.
Die zugehörige Fourier-Matrix lautet also

$$F_4 = \begin{pmatrix} \zeta_4^0 & \zeta_4^0 & \zeta_4^0 & \zeta_4^0 \\ \zeta_4^0 & \zeta_4^1 & \zeta_4^2 & \zeta_4^3 \\ \zeta_4^0 & \zeta_4^2 & \zeta_4^0 & \zeta_4^2 \\ \zeta_4^0 & \zeta_4^3 & \zeta_4^2 & \zeta_4^1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}$$

Sie ist umkehrbar und ihre Inverse lautet

$$F_4^{-1} = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}$$

Man beachte, dass die vierten komplexen Einheitswurzeln mit der reellen Faktorisierung

$$z^4 - 1 = (z^2 - 1) \cdot (z^2 + 1)$$

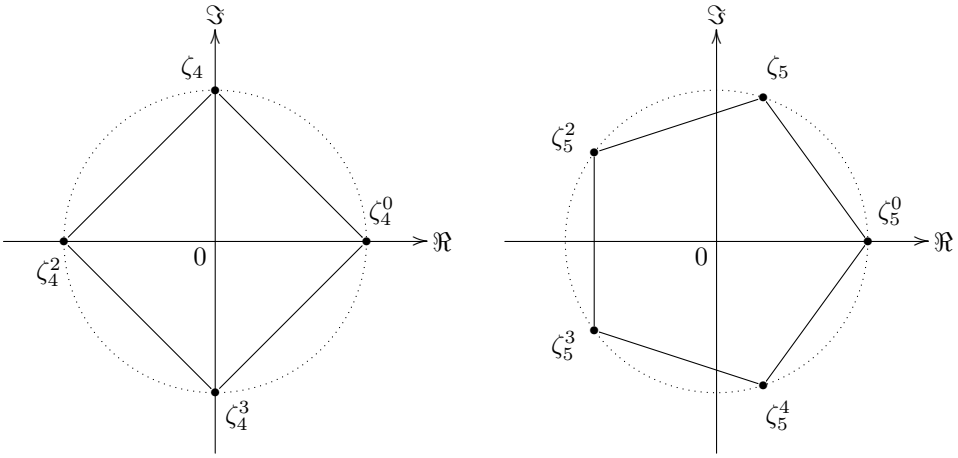


Abbildung 2.34: Die vierten und fünften komplexen Einheitswurzeln.

zusammenhängen, das nach dem Fundamentalsatz deer Algebra die komplexe Faktorisierung

$$z^4 - 1 = (z-1)\cdot(z+1)\cdot(z-i)\cdot(z+i)$$

in vier Linearfaktoren besitzt. ◯

Beispiel. Entsprechend liegen die fünften Einheitswurzeln auf den Ecken eines regulären 5-Eckes mit einer Ecke in $\zeta_5^0 = 1$ und lösen die Kreisteilungsgleichung

$$z^5 = 1.$$

Die gesuchten Lösungen dieser Gleichung können in Exponentialform durch die komplexen Zahlen

$$z_k = \zeta_5^k = e^{i(k\cdot\frac{2\pi}{5})} = \cos\left(k\cdot\frac{2\pi}{5}\right) + i\sin\left(k\cdot\frac{2\pi}{5}\right), \qquad 0 \le k \le 4$$

beschrieben werden. Sie liegen auf dem Einheitskreis und ihr Zentriwinkel ist $\frac{2\pi}{5} = 72°$. Eine der Ecken dieses regulären 5-Ecks ist selbstverständich 1, da $1^5 = 1$ ist. Für die Normalformen der 5-ten Einheitswurzeln gilt

k	z_k	$x_k + iy_k$	$N(x_k + iy_k)$
0	$e^{i0\cdot\frac{2\pi}{5}}$	$\cos(0) + i\sin(0)$	1
1	$e^{i1\cdot\frac{2\pi}{5}}$	$\cos\left(\frac{2\pi}{5}\right) + i\sin\left(\frac{2\pi}{5}\right)$	$0.309017 + 0.951057i$
2	$e^{i2\cdot\frac{2\pi}{5}}$	$\cos\left(2\cdot\frac{2\pi}{5}\right) + i\sin\left(2\cdot\frac{2\pi}{5}\right)$	$-0.809017 + 0.587785i$
3	$e^{i3\cdot\frac{2\pi}{5}}$	$\cos\left(3\cdot\frac{2\pi}{5}\right) + i\sin\left(3\cdot\frac{2\pi}{5}\right)$	$-0.809017 - 0.587785i$
4	$e^{i4\cdot\frac{2\pi}{5}}$	$\cos\left(4\cdot\frac{2\pi}{5}\right) + i\sin\left(4\cdot\frac{2\pi}{5}\right)$	$0.309017 - 0.951057i$

Diese Koordinaten sind zyklisch modulo 5 zu nehmen.

Statt nur durch die Werte transzendenter Formeln oder gar einfach nur durch numerische Näherungwerte zu beschreiben, möchten wir, wie üblich, nach Möglichkeit die kartesischen Koordinaten dieser Punkte auf dem Einheitskreis mit Hilfe

von Radikalen beschreiben. Das solllte hier möglich sein, weil wir von der Kreisteilungsgleichung fünften Grades $z^5 = 1$ eine ganzzahlige Lösung $z_0 = 1$ kennen, die wir abspalten können und die Faktorisierung

$$z^5 - 1 = (z-1) \cdot (1 + z + z^2 + z^3 + z^4)$$

erhalten, so dass wir also nur noch die komplexen Nullstellen des fünften, ganzzahligen irreduziblen Kreisteilungspolynom[41]

$$\Phi_5(z) = 1 + z + z^2 + z^3 + z^4 = \left(z - \zeta_5\right) \cdot \left(z - \zeta_5^2\right) \cdot \left(z - \zeta_5^3\right) \cdot \left(z - \zeta_5^4\right)$$

benötigen. Weil es sich bei den gesuchten Real- und Imaginärteilen der 5-ten Einheitswurzeln um Werte von Kreisfunktionen handelt, liegt es nahe, in diesem Polynom auf Grund der Euler'schen Formel die Substitution

$$w = z + z^{-1} = z + \frac{1}{z} = \frac{z^2+1}{z} = 2\cos\left(\frac{2\pi}{5}\right)$$

zu machen. Sie lässt sich Umkehren, weil die zugehörige quadratische Gleichung

$$z^2 - wz + 1 = 0 = \left(z - \frac{w}{2}\right)^2 - \frac{w^2}{4} + 1$$

die beiden Lösungen

$$z_1 = \frac{w + \sqrt{w^2-4}}{2}, \qquad z_2 = \frac{w - \sqrt{w^2-4}}{2}$$

hat. Statt einfach diese Ausdrücke zu substituieren, ist es bequemer, das Kreisteilungspolynom durch z^2 zu dividieren und den rationalen Ausdruck

$$f(z) = \frac{\Phi_5(z)}{z^2} = \frac{1}{z^2} + \frac{1}{z} + 1 + z + z^2$$

zu untersuchen. Man beachte, dass dieses Ausdruck genau dann verschwindet, wenn sein Zähler verschwindet, d.h. an einer der gesuchten Nullstellen von $\Phi_5(z)$. Es geht also nun darum, die Nullstellen von $f(z)$ zu bestimmen. In $f(z)$ ist nun die Substitution einfacher, wenn wir nämlich beachten, dass

$$w^2 = z^2 + 2 + z^{-2}$$

gilt. Unter der gewünschten Substitution geht offenbar $f(z)$ in das Polynom

$$\tilde{f}(w) = w^2 + w - 1 = \left(w + \frac{1}{2}\right)^2 - \frac{1}{4} - 1 = \left(w + \frac{1}{2}\right)^2 - \frac{5}{4}$$

zweiten Grades über. Seine Nullstellen sind

$$w_1 = \frac{-1 + \sqrt{5}}{2}, \qquad w_2 = \frac{-1 - \sqrt{5}}{2}$$

[41] Unter dem n-ten Kreisteilungspolynom $\Phi_n(z)$ versteht man das ganzzahlige, normierte Polynom grössten Grades, das $z^n - 1$ teilt und zu allen $x^d - 1$ mit $d < n$ teilerfremd ist.

Seine komplexen Nullstellen sind die primitiven n-ten Einheitswurzeln $\zeta_n^k = e^{i(k \cdot \frac{2\pi}{n})}$, wobei k die zu n teilerfremden Zahlen zwischen 1 und n durchläuft. Daher gelten die Faktorisierungen

$$\Phi_n(z) = \prod_{\substack{1 \le k \le n \\ \mathrm{ggT}(k,n)=1}} (z - \zeta_n^k) = \prod_{\substack{1 \le k \le n \\ \mathrm{ggT}(k,n)=1}} (z - e^{i(k \cdot \frac{2\pi}{n})}), \qquad z^n - 1 = \prod_{d|n} \Phi_d(z).$$

Diese beiden Nullstellen von $\tilde{f}$ entsprechen nach der Rücksubstitution den vier komplexen Zahlen

$$z_4 = \frac{-1+\sqrt{5}}{4} + \frac{\sqrt{2(5+\sqrt{5})}}{4}i, \qquad z_1 = \frac{-1+\sqrt{5}}{4} - \frac{\sqrt{2(5+\sqrt{5})}}{4}i$$

und

$$z_2 = \frac{-1-\sqrt{5}}{4} + \frac{\sqrt{2(5-\sqrt{5})}}{4}i, \qquad z_3 = \frac{-1-\sqrt{5}}{4} - \frac{\sqrt{2(5-\sqrt{5})}}{4}i$$

Das fünfte Kreisteilungspolynom hat also die reelle Faktorisierung

$$\Phi_5(z) = \Big(z^2 + \Big(\frac{1}{2}+\frac{1}{2}\sqrt{5}\Big)z + 1\Big) \cdot \Big(z^2 + \Big(\frac{1}{2}+\frac{1}{2}\sqrt{5}\Big)z + 1\Big)$$

Weil sich die Koordinaten der gefundenen vier Punkte auf dem Einheitskreis dank diesen Formeln durch geschachtelte Quadratwurzeln ausdrücken lassen, können diese fünf Punkte in der Ebene mit Hilfe von Zirkel und Lineal konstruiert werden. Die konstruktive Bewältugung der regelmässigen 5-Ecks mit Zirkel und Lineal gehört zu den Höhepunkten der griechischen Geometrie.

Dementsprechend lassen sich auch die zugehörige Fouriermatrix

$$F_5 = \begin{pmatrix} \zeta_5^0 & \zeta_5^0 & \zeta_5^0 & \zeta_5^0 & \zeta_5^0 \\ \zeta_5^0 & \zeta_5^1 & \zeta_5^2 & \zeta_5^3 & \zeta_5^4 \\ \zeta_5^0 & \zeta_5^2 & \zeta_5^4 & \zeta_5^1 & \zeta_5^3 \\ \zeta_5^0 & \zeta_5^3 & \zeta_5^1 & \zeta_5^4 & \zeta_5^2 \\ \zeta_5^0 & \zeta_5^4 & \zeta_5^3 & \zeta_5^2 & \zeta_5^1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & z_1 & z_2 & z_3 & z_4 \\ 1 & z_2 & z_4 & z_1 & z_3 \\ 1 & z_3 & z_1 & z_4 & z_2 \\ 1 & z_4 & z_3 & z_2 & z_1 \end{pmatrix}$$

und ihre Inverse F_5^{-1} vollständig mit geschachtelten Quadratwurzeln beschreiben. ◯

Die Konstruktion des regulären n-Ecks mit Zirkel und Lineal gelingt genau dann, wenn $n = 2^k \cdot p_1 \cdots p_m$ ist, wobei die p_j paarweise verschiedene Primzahlen der speziellen Form $2^{2^r}+1$ sind. Genau in diesen Fällen lassen sich also die n-ten komplexen Einheitswurzeln durch Radikale ausdrücken.

Beispiel. Aus der Schule erinnern wir uns, dass sich das reguläre 6-Eck leicht mit Zirkel und Lineal konstruieren lässt und sich daher die komplexen 6-Einheitswurzeln, d.h. die Lösungen der Kreisteilungsgleichung

$$z^6 = 1$$

durch Radikale ausdrücken lassen sollten.

Die Lösungen dieser Gleichung können in Exponentialform durch

$$z_k = \zeta_6^k = e^{i(k\cdot\frac{2\pi}{6})} = e^{i(k\cdot\frac{\pi}{3})} = \cos\Big(k\cdot\frac{\pi}{3}\Big) + i\sin\Big(k\cdot\frac{\pi}{3}\Big), \qquad 0 \le k \le 5$$

beschrieben werden. Sie liegen auf dem Einheitskreis und ihr Zentriwinkel beträgt $\frac{\pi}{3} = 60°$. Eine der Ecken dieses regulären 6-Ecks ist selbstverständich 1,

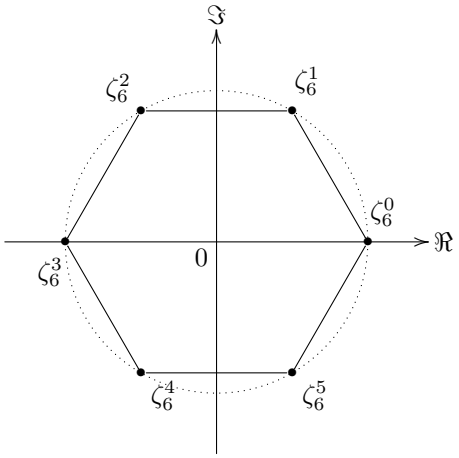


Abbildung 2.35: Die sechsten komplexen Einheitswurzeln.

da $1^6 = 1$ ist. Weil diesmal die Eckenzahl gerade ist, wird auch die gegenüberliegende Ecke $\zeta_6^3 = -1$ auf dem regulären 6-Eck liegen, da ja auch $(-1)^6 = 1$ gilt. Für die Normalformen der 6-ten Einheitswurzeln gilt

k	z_k	$x_k + iy_k$		$N(x_k + iy_k)$
0	$e^{i0\cdot\frac{\pi}{3}}$	$\cos(0) + i\sin(0)$	1	1
1	$e^{i1\cdot\frac{\pi}{3}}$	$\cos\left(\frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{3}\right)$	$0.5 + \frac{\sqrt{3}}{2}i$	$0.5 + 0.866025$
2	$e^{i2\cdot\frac{\pi}{3}}$	$\cos\left(2\cdot\frac{\pi}{3}\right) + i\sin\left(2\cdot\frac{\pi}{3}\right)$	$-0.5 + \frac{\sqrt{3}}{2}i$	$-0.5 + 0.866025$
3	$e^{i3\cdot\frac{\pi}{3}}$	$\cos\left(3\cdot\frac{\pi}{3}\right) + i\sin\left(3\cdot\frac{\pi}{3}\right)$	-1	-1
4	$e^{i4\cdot\frac{\pi}{3}}$	$\cos\left(4\cdot\frac{\pi}{3}\right) + i\sin\left(4\cdot\frac{\pi}{3}\right)$	$-0.5 - \frac{\sqrt{3}}{2}i$	$-0.5 - 0.866025$
5	$e^{i5\cdot\frac{\pi}{3}}$	$\cos\left(5\cdot\frac{\pi}{3}\right) + i\sin\left(5\cdot\frac{\pi}{3}\right)$	$0.5 - \frac{\sqrt{3}}{2}i$	$0.5 - 0.866025$

Diese Koordinaten sind zyklisch modulo 6 zu nehmen.

Weil wir bereits zwei verschiedene ganzzahlige Nullstellen der Kreisteilungsgleichung kennen, erhalten wir sofort die Faktorisierung

$$z^6 - 1 = (z-1)\cdot(z+1)\cdot(1+z^2+z^4)$$

und weil sich der biquadratische Faktor weiter faktorisieren lässt, sogar sofort die reelle Faktorisierung

$$z^6 - 1 = (z-1)\cdot(z+1)\cdot(1+z+z^2)\cdot(1-z+z^2)$$

Es geht also noch darum, die komplexen Nullstellen des dritten, ganzzahligen irreduziblen Kreisteilungspolynoms

$$\Phi_3(z) = 1 + z + z^2 = \left(z-\zeta_6^2\right)\cdot\left(z-\zeta_6^4\right) = (z-\zeta_3)\cdot(z-\zeta_3^2)$$

und des sechsten, ganzzahligen irreduziblen Kreisteilungspolynoms

$$\Phi_6(z) = 1 - z + z^2 = \left(z-\zeta_6\right)\cdot\left(z-\zeta_6^5\right)$$

exakt zu bestimmen.
Für die eine komplexe Nullstslle von $\Phi_3(z)$ erhalten wir

$$\zeta_6^2 = \zeta_3 = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$$

und damit aus Symmetriegründen für die andere

$$\zeta_6^4 = \zeta_3^2 = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$$

Entsprechend erhalten wir für die eine komplexe Nullstelle von $\Phi_6(z)$ den Ausdruck

$$\zeta_6 = \frac{1}{2} + \frac{\sqrt{3}}{2}i$$

und damit aus Symmetriegründen für die andere

$$\zeta_6^5 = \frac{1}{2} - \frac{\sqrt{3}}{2}i$$

Auch die zugehörige Fouriermatrix

$$F_6 = \begin{pmatrix} \zeta_6^0 & \zeta_6^0 & \zeta_6^0 & \zeta_6^0 & \zeta_6^0 & \zeta_6^0 \\ \zeta_6^0 & \zeta_6^1 & \zeta_6^2 & \zeta_6^3 & \zeta_6^4 & \zeta_6^5 \\ \zeta_6^0 & \zeta_6^2 & \zeta_6^4 & \zeta_6^0 & \zeta_6^2 & \zeta_6^4 \\ \zeta_6^0 & \zeta_6^3 & \zeta_6^0 & \zeta_6^3 & \zeta_6^0 & \zeta_6^3 \\ \zeta_6^0 & \zeta_6^4 & \zeta_6^2 & \zeta_6^0 & \zeta_6^4 & \zeta_6^2 \\ \zeta_6^0 & \zeta_6^5 & \zeta_6^4 & \zeta_6^3 & \zeta_6^2 & \zeta_6^1 \end{pmatrix}$$

und ihre Inverse F_6^{-1} lassen sich offenbar vollständig mit Quadratwurzeln beschreiben. ◯

Auch alle anderen aus der Schule bekannten Beziehungen über Kreisfunktionen und sämtliche in der Schule behandelten planimetrischen Resultate kann man sehr bequem und einfach mit Hilfe der komplexen Zahlen unter der Verwendung der Eulerschen Formel herleiten, indem man die entsprechenden Real- und Imaginärteile vergleicht. So gesehen versteht der Mathematiker nicht so recht, warum man heutzutage in der Schule statt der schwerfälligen und zu Formelsalat neigenden Trigonometrie nicht systematisch die viel bequemeren komplexen Zahlen verwendet, die erst noch für die beste aller zeitgenössischen physikalischen Theorien — die Quantenmechanik — fundamental sind, weil in der zeitabhängigen Schrödingergleichung

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t} = \Big(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x^2} + V(x)\Big)\Psi(x,t)$$

auf der linken Seite die komplexe Einheit explizit vorkommt. Diese lineare partielle Differentialgleichung beschreibt die zeitliche Entwicklung der sogn. Wellenfunktion $\Psi(x,t)$ eines Teilchens der Masse m, das sich entlang der x-Achse im Potential $V(x) \in \mathbb{R}$ bewegt. Diese Wellenfunktion muss also notgedrungen komplex sein. Ihre Norm

$$\rho(x,t) := |\Psi(x,t)|^2$$

ist reell und kann als Wahrscheinlichkeitsdichte interpretiert werden, d.h.

$$P_{a,b}(t) := \int_a^b |\Psi(x,t)|^2 \, dx = \int_a^b \rho(x,t) \, dx$$

liefert die Wahrscheinlichkeit, das Teilchen zur Zeit t im Intervall $[a,b]$ zu finden.
Um dieses Wahrscheinlichkeitsgesetz differentiell zu formulieren, definiert man mit Hilfe der Wellenfunktion $\Psi(x,t)$ den sogn. *Wahrscheinlichkeitsstrom*

$$J(x,t) := \frac{\hbar}{m} \Im\Big(\overline{\Psi}(x,t) \cdot \frac{\partial \Psi(x,t)}{\partial x}\Big)$$

Interpretert man die Wahrscheinlichkeitsdichte $|\Psi(x,t)|^2$ als Anzahl Teilchen pro Längeneinheit, kann der Wahrscheinlichkeitsstrom als zugehörige Flussrate interpretiert werden. In der Tat überprüft man mit Hilfe der Schrödingergleichung die Erhaltungsgleichung[42]

$$\frac{\partial J}{\partial x} + \frac{\partial \rho}{\partial t} = 0$$

Die gesuchte differentielle Version des Wahrscheinlichkeitsgesetzes lautet damit

$$\frac{dP_{a,b}(t)}{dt} = J(a,t) - J(b,t)$$

wobei die rechte Seite die Rate angibt, mit der die Wahrscheinlichkeit am linken bzw. am rechten Rand des Intervalls $[a,b]$ abfliesst.

Komplexe Zahlen sind auch für einfache physikalische Anwendungen bequemer.

Beispiel. In vielen physikalischen Anwendungen müssen harmonische Schwingungen gleicher Frequenz überlagert werden. Wie alle Resultate über Kreisfunktionen kann auch dieses Problem einfach mit Hilfe komplexer Zahlen gelöst werden. Dazu stellen wir zunächst die beiden gegebenen harmonischen Schwingungen als Realteile einer geeigneten komplexen Exponentialform dar. Es ist

$$\begin{aligned} h_1(x) &= \Re\big(A_1 \cdot e^{i(\omega x + \alpha_1)}\big) = A_1 \cdot \cos(\omega x + \alpha_1) \\ h_2(x) &= \Re\big(A_2 \cdot e^{i(\omega x + \alpha_2)}\big) = A_2 \cdot \cos(\omega x + \alpha_2) \end{aligned}$$

Nun werden diese beiden komplexen Zahlen addiert. Weil die beiden Kreisfrequenzen nach Voraussetzung übereinstimmen, erhalten wir

$$A_1 e^{i(\omega x + \alpha_1)} + A_2 e^{i(\omega x + \alpha_2)} = e^{i\omega x}\big(A_1 e^{i\alpha_1} + A_2 e^{i\alpha_2}\big)$$

Den linken Faktor nennt man in der Physik *Zeitfunktion* und die Summe im rechten Faktor heisst dort manchmal *komplexe Amplitude*. Sie wird nun in Exponentialform dargestellt. Wir bestimmen also A und α so, dass die folgende Gleichung gilt.

$$A_1 e^{i\alpha_1} + A_2 e^{i\alpha_2} = A e^{i\alpha}$$

[42] Die an die Differentialgleichung

$$\mathrm{div}(\vec{J}) + \frac{\partial \rho}{\partial t} = 0$$

der Ladungserhaltung aus der Elektrotechnik zwischen dem Stromstärkevektor $\vec{J}$ und der Ladungsdichte ρ erinnert und besagt, dass die Änderung der Ladung $Q_V(t)$ in einem festen Volumen V nur auf Grund des Flusses von $\vec{J}$ durch die Oberfläche ∂V des Volumens möglich ist, d.h. in der Integralform, dass $\frac{dQ_V(t)}{dt} = -\int_{\partial V} \langle \vec{J}, d\vec{n} \rangle$ gilt.

Weil der Realteil einer Summe die Summe der Realteile ist, erhalten wir die gesuchte Superposition dann nämlich aus der Beziehung

$$\begin{aligned} h(x) &= h_1(x) + h_2(x) = \Re\big(A_1 \cdot e^{i(\omega x+\alpha_1)}\big) + \Re\big(A_2 \cdot e^{i(\omega x+\alpha_2)}\big) \\ &= \Re\big(A_1 \cdot e^{i(\omega x+\alpha_1)} + A_2 \cdot e^{i(\omega x+\alpha_2)}\big) = \Re\big(e^{i\omega x} \cdot Ae^{i\alpha}\big) \\ &= \Re(Ae^{i(\omega x+\alpha)}) = A\cos(\omega x+\alpha) \end{aligned}$$

Offenbar spielt die Zeitfunktion hier keine Rolle und es genügt, A und α zu bestimmen. Mit Hilfe der Euler'schen Formel ist

$$\begin{aligned} A_1e^{i\alpha_1} + A_2e^{i\alpha_2} &= A_1\big(\cos(\alpha_1) + i\sin(\alpha_1)\big) + A_2\big(\cos(\alpha_2) + i\sin(\alpha_2)\big) \\ &= \big(A_1\cos(\alpha_1) + A_2\cos(\alpha_2)\big) + i\big(A_1\sin(\alpha_1) + A_2\sin(\alpha_2)\big) \end{aligned}$$

Daher gilt für den Betrag A der komplexen Amplitude

$$\begin{aligned} A^2 &= \big(A_1\cos(\alpha_1) + A_2\cos(\alpha_2)\big)^2 + \big(A_1\sin(\alpha_1) + A_2\sin(\alpha_2)\big)^2 \\ &= A_1^2 + A_2^2 + 2A_1A_2\big(\cos(\alpha_1)\cos(\alpha_2) + \sin(\alpha_1)\sin(\alpha_2)\big) \\ &= A_1^2 + A_2^2 + 2A_1A_2\cos(\alpha_1 - \alpha_2) \end{aligned}$$

Für den Polarwinkel dieser komplexen Zahl gilt

$$\tan(\alpha) = \frac{A_1\sin(\alpha_1) + A_2\sin(\alpha_2)}{A_1\cos(\alpha_1) + A_2\cos(\alpha_2)}$$

und wir erhalten den Fundamentalsatz über harmonische Schwingungen, wonach sich gleichfrequente harmonische Schwingungen zu einer harmonischen Schwingung überlagern. ◯

Das folgende numerische Beispiel soll diese Überlegung illustrieren.

Beispiel. Zur Superposition der beiden harmonischen Schwingungen der selben Frequenz $h_1(x) = \sin(\frac{\pi}{2} - x)$ und $h_2(x) = \frac{1}{2}\cos(x + \frac{\pi}{3})$. stellen wir in einem ersten Schritt beide harmonischen Schwingungen mit Hilfe der Komplementärwinkelformel $\sin(x) = \cos(\frac{\pi}{2} - x)$ als Kosinusschwingungen dar. In unserem Fall gilt für die erste harmonische Schwingung $h_1(x) = \sin(\frac{\pi}{2} - x) = \cos(x)$. Auf Grund des Hauptsatzes ist diese Superposition von der Form $h(x) = A\cos(x + \alpha)$. Zur Berechnung der Amplitude A und des Nullphasenwinkels α benötigen wir zunächst die komplexen Formen. Es ist $h_1(x) = \Re(e^{ix})$ und $h_2(x) = \Re(\frac{1}{2}e^{i(x+\frac{\pi}{3})})$. Für die komplexe Amplitude erhalten wir

$$1 + \frac{1}{2}e^{i\frac{\pi}{3}} = 1 + \frac{1}{2}\Big(\cos(\frac{\pi}{3}) + i\sin(\frac{\pi}{3})\Big) = 1 + \frac{1}{2}\Big(\frac{1}{2} + i\frac{\sqrt{3}}{2}\Big) = \frac{5}{4} + i\frac{\sqrt{3}}{4}$$

im ersten Quadranten. Für ihren Polarwinkel gilt $\tan(\alpha) = \frac{\sqrt{3}}{5}$. Von den beiden Winkeln im Grundintervall

$$\alpha_1 = \arctan\Big(\frac{\sqrt{3}}{5}\Big), \qquad \alpha_2 = \arctan\Big(\frac{\sqrt{3}}{5}\Big) + \pi$$

kommt aus Quandrantengründen nur der Winkel $\alpha = \alpha_1 \approx 0.334\ldots$ als Nullphasenwinkel in Frage. Den Betrag der Superposition erhalten wir aus der Beziehung $A^2 = \frac{25}{16} + \frac{3}{16} = \frac{7}{4}$. ◯

Beispiel. Eines der überraschenden — aber auch typischen — Phänomene im Zusammenhang mit harmonischen Schwingungen besteht in der Auslöschung, d.h. im Umstand, dass die Superposition harmonischer Schwingungen verschwinden kann, ohne, dass das für eine von ihnen der Fall ist.

Um reelle Zahlen $a, b \in \mathbb{R}$ so zu bestimmen, dass für alle $t \in \mathbb{R}$ die reelle Gleichung

$$\sin(t) + a \sin\left(t + \frac{2\pi}{3}\right) + b \sin\left(t - \frac{\pi}{6}\right) = 0$$

gilt, komplexifizieren wir die gegebene reelle Gleichung und erhalten

$$\Im\left(e^{it}\right) + a\Im\left(e^{i(t+\frac{2\pi}{3})}\right) + b\Im\left(e^{i(t-\frac{\pi}{6})}\right) = 0$$

Weil a, b reell sein sollen, ist also die Gleichung

$$\Im\left(e^{it} + ae^{i(t+\frac{2\pi}{3})} + be^{i(t-\frac{\pi}{6})}\right) = 0$$

zu lösen, die durch Verwenden des Additionstheorems die Gestalt

$$\Im\left(e^{it} + ae^{it} \cdot e^{i\frac{2\pi}{3}} + be^{it} \cdot e^{-i\frac{\pi}{6}}\right) = 0$$

annimmt. Ausklammern des gemeinsamen Zeitfaktors e^{it} macht daraus die Gleichung

$$\Im\left(e^{it} \cdot \left(1 + a \cdot e^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}}\right)\right) = 0$$

Weil diese Gleichung für alle $t \in \mathbb{R}$ erfüllt sein soll, dürfen wir den gemeinsamen Faktor e^{it} weglassen und müssen noch die Gleichung

$$1 + ae^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}} = 0$$

lösen. Diese Gleichung ergibt sich auch, wenn man die gegebene Gleichung in der üblichen Form komplexifiziert. Dann lautet sie nämlich

$$\frac{e^{it} - e^{-it}}{2i} + a\frac{e^{i(t+\frac{2\pi}{3})} - e^{-i(t+\frac{2\pi}{3})}}{2i} + b\frac{e^{i(t-\frac{\pi}{6})} - e^{-i(t+\frac{\pi}{6})}}{2i} = 0$$

oder nach Multiplikation mit $2i$

$$e^{it} - e^{-it} + a(e^{i(t+\frac{2\pi}{3})} - e^{-i(t+\frac{2\pi}{3})}) + b(e^{i(t-\frac{\pi}{6})} - e^{-i(t+\frac{\pi}{6})}) = 0$$

Durch Verwenden des Additionstheorems und Ausklammern der gemeinsamen zeitabhängigen Faktoren e^{it} und e^{-it} wird daraus

$$e^{it} \cdot \left(1 + ae^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}}\right) - e^{-it} \cdot \left(1 + ae^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}}\right) = 0$$

bzw. nach Ausklammern des gemeinsamen Faktors die Gleichung

$$(e^{it} - e^{-it}) \cdot \left(1 + ae^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}}\right) = 0$$

Sie kann nur dann für alle $t \in \mathbb{R}$ erfüllt sein, wenn die Gleichung

$$1 + ae^{i\frac{2\pi}{3}} + be^{-i\frac{\pi}{6}} = 0$$

gilt. In kartesischer Darstellung entspricht sie der Gleichung

$$1 + a\Big(-\frac{1}{2} + \frac{\sqrt{3}}{2}i\Big) + b\Big(\frac{\sqrt{3}}{2} - \frac{1}{2}i\Big) = 0$$

oder nach Multiplikation mit 2 der Gleichung

$$2 + a\big(-1 + \sqrt{3}i\big) + b\big(\sqrt{3} - i\big) = 0$$

bzw. nach dem Zusammenfassen von Real- und Imaginärteil der Gleichung

$$\big(2 - a + \sqrt{3}b\big) + i\big(\sqrt{3}a - b\big) = 0$$

Sie entspricht dem linearen Gleichungungssystem

$$\left\{ \begin{array}{rcl} -a + \sqrt{3}b & = & -2 \\ \sqrt{3}a - \quad b & = & 0 \end{array} \right.$$

mit der eindeutigen Lösung $a = -1$ und $b = -\sqrt{3}$.
Wer komplexe Zahlen lieber ganz vermeidet, kann reell argumentieren und mit Hilfe des Additionstheorems die gegebene reelle Gleichung in die Form

$$\sin(t) + a\big(\sin(t)\cos\big(\frac{2\pi}{3}\big) + \cos(t)\sin\big(\frac{2\pi}{3}\big)\big) + b\big(\sin(t)\cos\big(\frac{\pi}{6}\big) - \cos(t)\sin\big(\frac{\pi}{6}\big)\big) = 0$$

bringen und dann die numerischen Werte einsetzen. Dabei erhält man

$$\sin(t) - \frac{a}{2}\sin(t) + \frac{\sqrt{3}a}{2}\cos(t) + \frac{\sqrt{3}b}{2}\sin(t) - \frac{b}{2}\cos(t) = 0$$

bzw. nach Multiplikation mit 2 und Ausklammern der gemeinsamen Faktoren $\sin(t)$ und $\cos(t)$

$$\sin(t) \cdot \big(2 - a + \sqrt{3}a\big) + \cos(t) \cdot \big(\sqrt{3}a - b\big) = 0$$

Weil $\sin(t)$ und $\cos(t)$ linear unabhängig sind, kann diese Gleichung nur dann für alle $t \in \mathbb{R}$ gelten, wenn die beiden Faktoren einzeln verschwinden, was obigem linearen Gleichungssystem entspricht und die selbe Lösung liefert. ◯

Beispiel. Oft benötigt man die Amplitude und den Nullphasenwinkel der Überlagerung von $n+1$ harmonischen Schwingungen mit gleicher Amplitude A, gleicher Kreisfrequenz ω und den Nullphasenwinkeln $0, \delta, 2\delta, \ldots, n\delta$. Dazu berechnen wir zunächst die Exponentialform der Summe

$$\begin{array}{rcl} s_n(x) & = & Ae^{i\omega x} + Ae^{i(\omega x+\delta)} + \cdots + Ae^{i(\omega x+k\delta)} + \cdots + Ae^{i(\omega x+n\delta)} \\ & = & \displaystyle\sum_{k=0}^{n} Ae^{i(\omega x+k\delta)} = Ae^{i\omega x} \cdot \sum_{k=0}^{n} e^{ik\delta} \end{array}$$

Wegen der Formel von De Moivre gilt

$$e^{ik\delta} = (e^{i\delta})^k$$

und unsere Summe hat die Gestalt

$$s_n(x) = Ae^{i\omega x} \cdot \sum_{k=0}^{n} (e^{i\delta})^k$$

Offensichtlich handelt es sich hier um eine geometrische Reihe mit dem konstanten Quotienten $e^{i\delta}$. Falls δ kein Vielfaches von 2π ist, ist $e^{i\delta} \neq 1$ und wir können die Summenformel für die geometrische Reihe benutzen. Sie liefert für unsere Summe

$$s_n(x) \quad = \quad Ae^{i\omega x} \cdot \frac{1-(e^{i\delta})^{n+1}}{1-e^{i\delta}} = Ae^{i\omega x} \cdot \frac{e^{i\frac{n+1}{2}\delta}(e^{-i\frac{n+1}{2}\delta} - e^{i\frac{n+1}{2}\delta})}{e^{i\frac{\delta}{2}}(e^{-i\frac{\delta}{2}} - e^{i\frac{\delta}{2}})}$$

Wir erinnern uns nun an die Tatsache, dass sich die Kreisfunktionen mit Hilfe der Exponentialfunktion ausdrücken lassen. Wir haben seinerzeit die Beziehung $\sin(\varphi) = \frac{1}{2i}(e^{i\varphi} - e^{-i\varphi})$ gefunden, die nach Multiplikation mit $2i$ die Form $e^{i\varphi} - e^{-i\varphi} = 2i\sin(\varphi)$ annimmt. Damit geht die Summe über in

$$\begin{aligned} s_n(x) &= Ae^{i\omega x} \cdot \frac{e^{i\frac{n+1}{2}\delta}(-2i\sin(\frac{n+1}{2}\delta))}{e^{i\frac{\delta}{2}}(-2i\sin(\frac{\delta}{2}))} = Ae^{i\omega x} \cdot \frac{\sin(\frac{n+1}{2}\delta)}{\sin(\frac{\delta}{2})} e^{i\frac{n}{2}\delta} \\ &= \Big(A\frac{\sin(\frac{n+1}{2}\delta)}{\sin(\frac{\delta}{2})}\Big) e^{i(\omega x + \frac{n}{2}\delta)} \end{aligned}$$

womit wir für $\delta \neq m \cdot 2\pi$ die gesuchte Polarform gefunden haben. Falls δ ein Vielfaches von 2π ist, gilt natürlich für die Summe

$$s_n(x) = A(n+1)e^{i\omega x}$$

Real- und Imaginärteil des gefundenen Ausdruckes für $s_n(x)$ liefern für $A = 1$ und $\omega = 0$ die folgenden *Summenformeln* für Kreisfunktionen:

Korollar. Für den Winkel δ und jede natürliche Zahl $n \geq 1$ gilt:

$$\sum_{k=0}^{n} \cos(k\delta) \quad = \quad \begin{cases} n+1 & \text{falls } \delta = m\ 2\pi \text{ für } m \in \mathbb{Z} \\ \frac{\sin(\frac{n+1}{2}\delta)\cos(\frac{n}{2}\delta)}{\sin(\frac{\delta}{2})} & \text{falls } \delta \neq m\ 2\pi \text{ für } m \in \mathbb{Z} \end{cases}$$

$$\sum_{k=0}^{n} \sin(k\delta) \quad = \quad \begin{cases} 0 & \text{falls } \delta = m\ 2\pi \text{ für } m \in \mathbb{Z} \\ \frac{\sin(\frac{n+1}{2}\delta)\sin(\frac{n}{2}\delta)}{\sin(\frac{\delta}{2})} & \text{falls } \delta \neq m\ 2\pi \text{ für } m \in \mathbb{Z} \end{cases}$$

Diese Summenformeln spielen beim Integrieren der Kreisfunktionen eine zentrale Rolle. ◯

Weil die Drehmatrizen die charakteristische Funktionalgleichung der Exponentialgleichung erfüllen, d.h. wegen

$$D_{\varphi+\psi} = D_\varphi \cdot D_\psi$$

eine Summe in ein Produkt überführt, erfährt die Euler'sche Formel, die wir seinerzeit ohne viel Federlesen aus dem Hut gezaubert haben, nachträglich eine

gewisse Rechtfertigung. Die Exponentialfunktion gehört aber eigentlich in die Analysis. Aus dem Analysis-Unterricht erinnern wir uns nämlich daran, dass die Exponentialfunktion durch eine gewisse Differentialgleichung definiert ist. Für die Ableitung der Exponentialfunktion $e^{\lambda\varphi}$ gilt:

$$(e^{\lambda\varphi})' = \lambda \cdot e^{\lambda\varphi}$$

Die Exponentialfunktion $e^{\lambda\varphi}$ erfüllt also das charakteristische Anfangswertproblem

$$f' = \lambda f, \qquad f(0) = 1$$

Um die Bezeichnung $D_\varphi = e^{i\varphi}$ besser rechtfertigen zu können, müssen wir also einsehen, dass die Drehmatrix D_φ diese Differentialgleichung der Exponentialfunktion für $\lambda = i$ d.h. die Gleichung $(e^{i\varphi})' = i \cdot e^{i\varphi}$ erfüllt. Durch Ableiten der Drehmatrix erhalten wir einerseits

$$D_\varphi = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}, \qquad D'_\varphi = \begin{pmatrix} -\sin(\varphi) & -\cos(\varphi) \\ \cos(\varphi) & -\sin(\varphi) \end{pmatrix}$$

Andererseits gilt:

$$I \cdot D_\varphi = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix} = \begin{pmatrix} -\sin(\varphi) & -\cos(\varphi) \\ \cos(\varphi) & -\sin(\varphi) \end{pmatrix}$$

Ein Vergleich zeigt, dass die Drehmatrix D_φ die Bedingungen

$$D'_\varphi = I \cdot D_\varphi, \qquad D_0 = E$$

erfüllt. Diese Beziehung geht mit unseren Abmachungen tatsächlich in die charakteristische Differentialgleichung der Exponentialfunktion

$$(e^{i\varphi})' = i \cdot e^{i\varphi}$$

über. Die Euler'sche Formel

$$e^{i\varphi} = \cos(\varphi) + i\sin(\varphi)$$

haben wir oben eingesehen, indem wir die rechte Seite abgeleitet haben und dabei

$$\big(\cos(\varphi) + i\sin(\varphi)\big)' = -\sin(\varphi) + i\cos(\varphi) = i\big(\cos(\varphi) + i\sin(\varphi)\big) = ie^{i\varphi}$$

erhalten. Das ist aber gerade die charakteristische Eigenschaft der Exponentialfunktion $e^{i\varphi}$. Nachträglich ist also die Verwendung der Exponentialfunktion auch aus dem analytischen Blickwinkel gerechtfertigt. Für die zweite Ableitung gilt mit $D''_\varphi = -D_\varphi$ die Differentialgleichung der harmonischen Schwingung, die in komplexer Schreibweise die Form

$$(e^{i\varphi})'' = -e^{i\varphi}$$

annimmt.

Die soeben hergeleiteten Formeln lassen sich geometrisch interpretieren. Dazu denken wir uns im Physiker-Lingo ein Teilchen, das sich entlang der Kurve mit der Parameterdarstellung

$$t \mapsto e^{it}$$

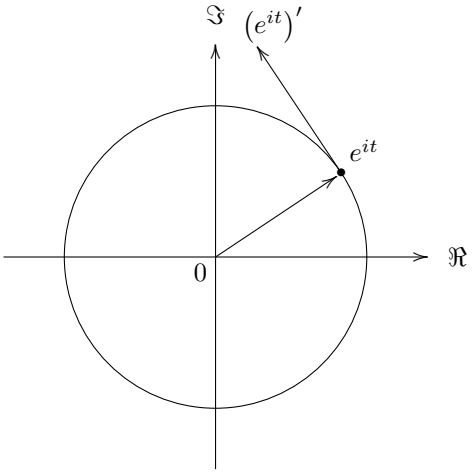


Abbildung 2.36: Geschwindigkeit bei der gleichförmigen Kreisbewegung.

bewegt. Wie wir wissen, handelt es sich bei dieser Kurve um den Einheitskreis. Die Parametergleichung beschreibt also eine gleichförmige Kreisbewegung auf dem Einheitskreis. Die Ableitung dieser Kurvengleichung $(e^{it})'$ kann aber bekanntlich als *Geschwindigkeit* des Teilchens interpretiert werden.

Weil das Teilchen für $t = 2\pi$ genau einmal um den Einheitskreis vom Umfang 2π gelaufen ist, muss der Betrag der Geschwindigkeit dieser gleichförmigen Kreisbewegung 1 sein. Der Geschwindigkeitsvektor steht ferner bei einer gleichförmigen Kreisbewegung orthogonal zu Radiusvektor. Da Drehung um $\frac{\pi}{2}$ durch Multiplikation mit i erreicht werden kann, hat der Geschwindigkeitsvektor die Darstellung ie^{it}. Daher gilt also tatsächlich $(e^{it})' = ie^{it}$. Entsprechend liefert die zweite Ableitung einer Kurvengleichung die *Beschleunigung*. In unserem Fall hat diese Beschleunigung den Betrag 1 und ist radial nach innen gerichtet. Der Beschleunigungsvektor kann daher durch $-e^{it}$ beschrieben werden. Die Formel $(e^{it})'' = -e^{it}$ beschreibt also die Zentripetalbeschleunigung.

Komplexe Zahlen werden schon seit langem intensiv in der Physik und in der Elektrotechnik benutzt. Elektrotechniker schreiben für eine typische komplexe Zahl übrigens manchmal $a + bj$, da sie den Buchstaben i aus unerfindlichen Gründen für die Bezeichnung von elektrischen Strömen zur Verfügung haben möchten. Entsprechend beschreiben sie eine in x-Richtung fortschreitende harmonische Welle statt in der unter Physikern üblichen Form

$$f(x,t) = e^{i(kx-\omega t)}$$

die Lösung der üblichen Wellengleichung

$$\frac{\partial^2 f(x,t)}{\partial x^2} = \frac{k^2}{\omega^2}\frac{\partial^2 f(x,t)}{\partial t^2}$$

ist, meist durch den Ausdruck $e^{j(\omega t-kx)}$. Zwischen der Kreisfrequenz ω, der Frequenz ν, der Periodenlänge T, der Wellenzahl k, der Wellenlänge λ und der Phasengeschwindigkeit v gelten die Beziehungen

$$\omega = 2\pi\nu = \frac{2\pi}{T}, \quad k = \frac{2\pi}{\lambda}, \quad v = \frac{\omega}{k} = \frac{\lambda}{T} = \lambda\nu$$

Die zeitabhängige Schrödingergleichung für das (nichtrelativistische) freie Teilchen der kinetischen Energie $E = \frac{m}{2}v^2$ im Potential $V(x) = 0$ lautet im Physiker-Lingo

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t} = \Big(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x^2}\Big)\Psi(x,t)$$

und hat die Lösung

$$\Psi(x,t) = e^{i(kx-\frac{E}{\hbar}t)}, \qquad E = \hbar\omega, \quad k = \sqrt{\frac{2mE}{\hbar^2}}$$

die Schrödinger, zusammen mit den Gleichungen von Einstein und de Broglie

$$E = \hbar\omega, \qquad \lambda = \frac{h}{p} \quad \text{bzw.} \quad p = \hbar k$$

die den Welle-Teilchen-Dualismus ausdrücken, zur Wahl seiner Gleichung gedient hat.

Für mathematische Zwecke ist die Beschränkung auf feste Standardbezeichnungen und die pedantische Verwendung willkürlicher Normen unflexibel und fatal; es gibt einfach mehr wichtige mathematische Ideen als Buchstaben im lateinischen Alphabet und Euler hat seine Wahl, die durch den Anfangsbuchstaben des Wortes imaginär (in der lateinischen Version) gut motiviert war, in seinem dreibändigen Lehrbuch der Integralrechnung mit dem Titel „Institutionum Calculi Integralis“ im Jahr 1770, also etwa 150 Jahre vor der Entdeckung der Elektrizität, getroffen und dabei das paradoxe Symbol $\sqrt{-1}$ vermieden. Um den lächerlichen Sprachenkrieg zwischen den beiden Anwendergruppen zu vermeiden, schlägt ihnen der Autor vor, in Zukunft zur Übersetzung die banale Formel

$$j := -i$$

zu benutzen, die dem einzigen nicht trivialen Automorphismus von $\mathbb{C}$

$$\mathbb{C} \to \mathbb{C}, \qquad a + bi \mapsto a - bi = a + bj$$

zu Grunde liegt.

Der mathematische Grund für die intensive Verwendung der komplexen Zahlen in den meisten Anwendungen liegt darin, dass man mit Hilfe von komplexen Zahlen Drehungen in der Ebene und ihre Verknüpfung auf besonders einfache Art beschreiben kann. Insbesondere lassen sich die ganze ebene Schulgeometrie und die Trigonometrie durch leichtes Rechnen mit komplexen Zahlen herleiten. Komplexe Zahlen sind dafür logisch — vorläufig nicht aber pädagogisch — entbehrlich. Als Folge davon lassen sich dann mit Hilfe der komplexen Zahlen alle Schwingungs- und Wellenphänomene und allgemeiner durch Verwendung der Fourier-Theorie alle periodischen Phänomene übersichtlich beschreiben. Obwohl sich mit Hilfe der komplexen Zahlen viele Rechnungen in diesen Bereichen stark vereinfachen, werden die komplexen Zahlen in den Anfängervorlesungen nicht systematisch benutzt, weil für den Anfänger durch den Gebrauch der komplexen Zahlen oft die geometrischen Hintergründe verschleiert werden. In der fortgeschrittenen linearen Algebra, wo es darum geht, Matrizen mit Hilfe der Eigenwerte auf gewisse speziell einfache Normalformen zu bringen und in der Quantenmechanik, wo es darum geht, Aussagen über physikalische Prozesse zu machen, kommt man aber nicht um die komplexen Zahlen herum.

Angenehm an den komplexen Zahlen ist neben dem Fundamentalsatz der Algebra ihre enge Beziehung zu den Drehungen. Aus ihr folgt der Fundamentalsatz sogar im Wesentlichen. Wirklich überraschend an den komplexen Zahlen ist also nicht die Zahl i mit ihren neuen algebraischen Eigenschaften, sondern die Produktformel und die Tatsache, dass sie mit der Euklid'schen Norm d.h. mit der Geometrie verträglich ist. Durch Identifikation von $\mathbb{C}$ mit der Ebene $\mathbb{R}^2$ liefert die Produktformel der komplexen Zahlen nämlich eine Abbildung

$$\mu_2\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \qquad \big((a,b),(c,d)\big) \mapsto (ac-bd, ad+bc)$$

die das Element $(1,0) \in \mathbb{R}^2$, das auch als Abbildung

$$\eta_2\colon \{*\} \to \mathbb{R}^2, \qquad * \mapsto (1,0)$$

aufgefasst werden kann, als beidseitige Einheit hat. Dass diese Multiplikation assoziativ ist und das besagte Neutralelement besitzt, lässt sich elementefrei so ausdrücken, dass man sagt, dass die folgenden beiden Diagramme kommutieren:

$$\begin{array}{ccc}
\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 & \xrightarrow{\mathrm{Id}\times\mu_2} & \mathbb{R}^2 \times \mathbb{R}^2 \\
\downarrow{\scriptstyle \mu_2\times\mathrm{Id}} & & \downarrow{\scriptstyle \mu_2} \\
\mathbb{R}^2 \times \mathbb{R}^2 & \xrightarrow{\mu_2} & \mathbb{R}^2
\end{array}
\qquad
\begin{array}{ccccc}
\{*\} \times \mathbb{R}^2 & \xrightarrow{\eta_2\times\mathrm{Id}} & \mathbb{R}^2 \times \mathbb{R}^2 & \xrightarrow{\eta_2\times\mathrm{Id}} & \mathbb{R}^2 \times \{*\} \\
\downarrow{\scriptstyle \sim} & & \downarrow{\scriptstyle \mu_2} & & \downarrow{\scriptstyle \sim} \\
\mathbb{R}^2 & = & \mathbb{R}^2 & = & \mathbb{R}^2
\end{array}$$

Aus Dimensionsgründen ist der Graph der Abbildung μ_2, deren Bauart an die Additionstheoreme erinnert, nicht gut anschaulich zu verstehen. Schränken wir sie auf die Einheitsvektoren aus dem Einheitskreis $S^1 \subseteq \mathbb{R}^2$ ein, folgt aus der Normproduktregel, dass auch ihr Bild wiederum auf dem Einheitskreis liegt. Daher erhalten wir durch Einschränken die assoziative Multiplikationsabbildung

$$\mu_2\colon \mathrm{S}^1 \times \mathrm{S}^1 \to \mathrm{S}^1, \qquad \big((a,b),(c,d)\big) \mapsto (ac-bd, ad+bc)$$

und die Einheitsabbildung

$$\eta_2\colon \{*\} \to S^1, \qquad * \mapsto (1,0)$$

Dass auch sie assoziativ sind und $(1,0)$ als beidseitiges Neutralelement haben, drückt sich dadurch aus, dass die beiden entsprechenden Diagramme kommutieren.

$$\begin{array}{ccc}
S^1 \times S^1 \times S^1 & \xrightarrow{\mathrm{Id}\times\mu_2} & S^1 \times S^1 \\
\downarrow{\scriptstyle \mu_2\times\mathrm{Id}} & & \downarrow{\scriptstyle \mu_2} \\
S^1 \times S^1 & \xrightarrow{\mu_2} & S^1
\end{array}
\qquad
\begin{array}{ccccc}
\{*\} \times S^1 & \xrightarrow{\eta_2\times\mathrm{Id}} & S^1 \times S^1 & \xrightarrow{\eta_2\times\mathrm{Id}} & S^1 \times \{*\} \\
\downarrow{\scriptstyle \sim} & & \downarrow{\scriptstyle \mu_2} & & \downarrow{\scriptstyle \sim} \\
S^1 & = & S^1 & = & S^1
\end{array}$$

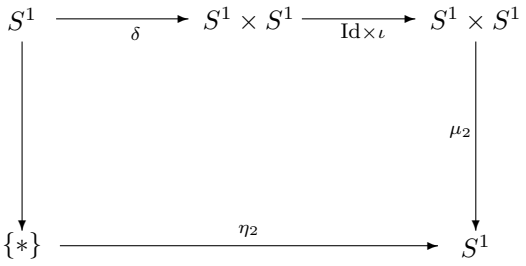
In Exponentialform hat diese Multiplikationsabbildung die Gestalt

$$\mu_2\colon \mathrm{S}^1 \times \mathrm{S}^1 \to \mathrm{S}^1, \qquad \big(e^{i\alpha}, e^{i\beta}\big) \mapsto e^{i(\alpha+\beta)}$$

Sie hat eine Inverse

$$\iota\colon S^1 \to S^1, \qquad e^{i\varphi} \mapsto e^{-i\varphi}$$

die zusätzlich folgendes Diagramm kommutativ macht:



Dabei bezeichnet

$$\delta\colon S^1 \to S^1 \times S^1, \qquad x \mapsto (x, x)$$

die Diagonalabbildung und die unmarkierte Abbildung $S^1 \to \{*\}$ bezeichnet die (eindeutig bestimmte) Projektion auf den Punkt.

Weil sich das kartesische Produkt

$$S^1 \times S^1 = \{\big((a,b),(c,d)\big) \mid a^2 + b^2 = 1, c^2 + d^2 = 1\} \subseteq \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4$$

als Torus im 3-dimensionalen Raum einbetten lässt, indem man gegenüberliegende Seiten des Quadrates $[0, 2\pi] \times [0, 2\pi]$ etwa mit Hilfe der Parametrisierung

$$(\alpha, \beta) \mapsto \begin{pmatrix} (3 + \cos(\alpha)) \cdot \cos(\beta) \\ (3 + \cos(\alpha)) \cdot \sin(\beta) \\ \cos(\alpha) \end{pmatrix}$$

verklebt, liefert die Multiplikationsabbildung eine Abbildung vom Torus auf den Einheitskreis. Die Fasern (Urbilder) der einzelnen Kreispunkte sind auf dem Parameterbereich $[0, 2\pi] \times [0, 2\pi]$, dessen Punkte durch die beiden Polarwinkel α und β beschrieben werden, die in der folgenden Figur eingezeichneten Strecken, auf denen die Summe modulo 2π jeweils den selben Wert liefert. Deshalb bestehen mit Ausnahme der durch die Diagonale parametrisierten Faser die Parameterbereiche aller anderen Fasern aus zwei Streckenstücken, die an gegenüberliegenden Rändern verkebt werden müssen und dann auf dem Torus eine geschlossene Kurve ergeben.

Auf dem veklebten Torus sind die entstehenden Fasern also parallele, geschlossene Kurven, die sich einmal gleichmässig entlang eines Breitenkreises und einmal entlang eines Längenkreises um den Torus winden. Diese parallelen Fasern fasern den ganzen Torus. In folgender Figur sind 8 verschiedene Fasern in unterschiedlichen Farbwerten dargestellt.

In dieser Figur findet man neben der räumlichen Torusfaserung auch die zugehörigen Bilder in den entsprechenden Farbwerten.

Diese fundamentale Abbildung und ihre Verträglichkeit mit der Norm enthält also den wesentlichen Gehalt der komplexen Zahlen und damit jenen der Trigonometrie und der Ebenengeometrie mit ihren Drehungen. Wir sind auf sie gestossen, weil wir die Multiplikationsabbildung μ_2 der komplexen Zahlen genauer unter die Luppe genommen haben. Analog kann man nun auch die bekanntere Multiplikationsabbildung

$$\mu_1\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad (x, y) \mapsto x \cdot y$$

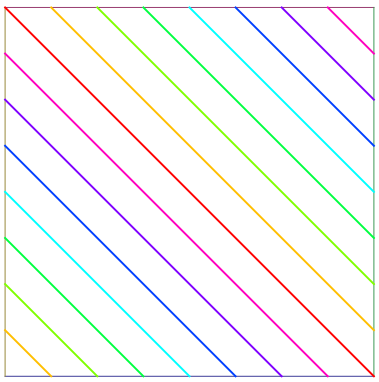


Abbildung 2.37: Der Torus als Quadrat mit identifizierten Kanten.

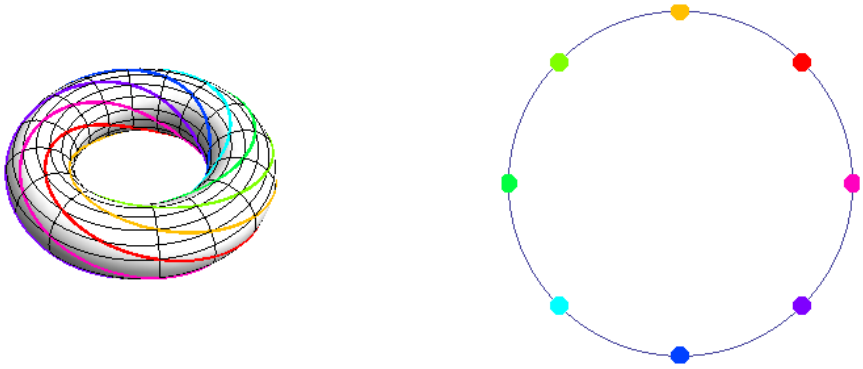


Abbildung 2.38: Multiplikation komplexer Zahlen als Torusfaserung.

unter die Lupe nehmen. Sie hat bekanntlich $1 \in \mathbb{R}$ als Einheit und ihr Graph lässt sich aus Dimensionsgründen als einschaliges Hyperboloid veranschaulichen. Die aus der Schule wohlbekannte Normproduktregel für $\mathbb{R}$

$$(x \cdot y)^2 = x^2 \cdot y^2$$

liefert durch Einschränkung auf den zugehörigen „Einheitskreis“, der in dieser Dimension aus zwei Punkten besteht,

$$S^0 = \{x \in \mathbb{R} \mid |x| = 1\} = \{\pm 1\}$$

die Multiplikationsabbildung

$$\mu_1 \colon S^0 \times S^0 \to S^0 \qquad (x, y) \mapsto x \cdot y$$

mit der symmetrischen Multiplikationstabelle

$\cdot$	1	-1
1	1	-1
-1	-1	1

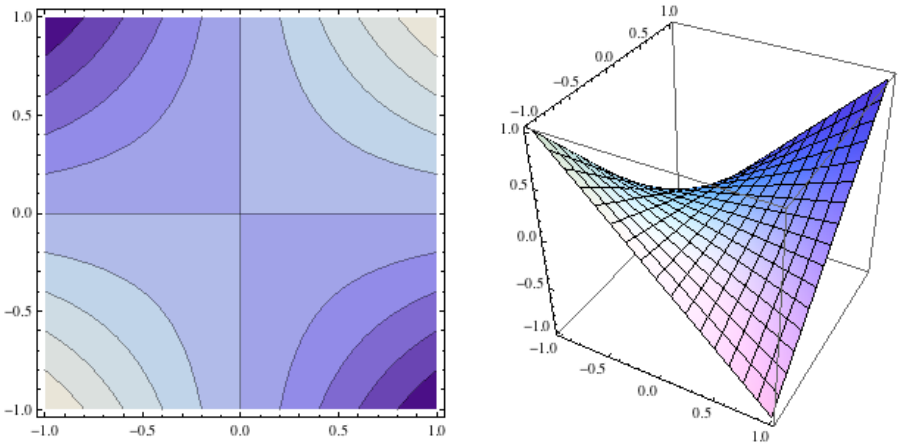


Abbildung 2.39: Die Multiplikationsabbildung $(x, y) \mapsto x \cdot y$ reeller Zahlen mit Hilfe von Niveaulinien und als Graph geometrisch dargestellt.

Daraus entnehmen wir insbesondere als interessante Information die Beziehung

$$(-1) \cdot (-1) = 1,$$

die viele von uns seinerzeit als Schüler etwas überrascht hat. Das zugehörige Neutralelement ist durch die Abbildung

$$\eta_1 \colon \{*\} \to S^0, \qquad * \mapsto 1$$

gegeben wie man an Hand der Tabelle bestätigt.

2.7 Quaternionen

Wenn die reellen und die komplexen Zahlen schon so nützlich sind, fragt man sich selbstverständlich, ob sich solche verallgemeinerten Zahlen auch in höheren Dimensionen finden lassen, die das Leben auch dort vereinfachen. Wir interessieren uns also für Produktabbildungen auf $\mathbb{R}^m$ und für zugehörige Multiplikationsabbildungen

$$\mu_m \colon S^{m-1} \times S^{m-1} \to S^{m-1}$$

In höheren Dimensionen können wir keine weitere Körper mehr erwarten. Aus dem Fundamentalsatz der Algebra folgt folgendes Resultat von Frobenius.

Satz. Für $m \geq 3$ ist es unmöglich, auf $\mathbb{R}^m$ eine Produktabbildung zu definieren, so dass $\mathbb{R}^m$ mit der üblichen Addition eine Körpererweiterung von $\mathbb{R}$ wird.

Nach diesem Satz haben wir die einzigen endlichdimensionalen Körpererweiterung von $\mathbb{R}$, nämlich die beiden Körper $\mathbb{R}$ und $\mathbb{C}$ bereits gefunden. Ein weiterer Grund, warum wir in diesem Grundkurs die komplexen Zahlen so lang wie vernünftig möglich vermeiden besteht darin, dass in höheren Dimensionen keine

entsprechenden verallgemeinerten Zahlen mehr zur Verfügung stehen und wir möglichst viele Überlegungen dimensionsunabhängig durchführen wollen.

Weil die Raumdrehungen in den Anwendungen eine fundamentale Rolle spielen, wünschen wir uns natürlich trotz dieses negativen Bescheides dringend einen Zahlenbereich, der die Untersuchung der Drehungen in $\mathbb{R}^3$ erleichtert. Auf Grund des soeben formulierten Satzes ist es allerdings hoffnungslos, in der Dimension 3 einen Körper zu erwarten. Es bleibt uns also nicht anderes übrig, als entweder die Hoffnung ganz zu begraben oder aber die Anforderungen zu lockern. Weil wir an Hand einfacher Beispiele von Raumdrehungen erkennen, dass dort das Kommutativgesetz nicht erfüllt sein kann, verzichten wir in Zukunft auf das Kommutativgesetz der Multiplikation. Damit die Verträglichkeit mit der Euklid'schen Geometrie gewährleistet ist, erwarten wir allerdings weiterhin eine Normproduktregel

$$N(\vec{x} \cdot \vec{y}) = N(\vec{x}) \cdot N(\vec{y})$$

wobei $N(\vec{x}) = \langle \vec{x}, \vec{x} \rangle^2$ die Euklid'schen Norm bezeichnet. Es ist nun nicht schwierig zu sehen, dass auch ein solches Produkt in der Dimension $m = 3$ nicht existieren kann. Wählen wir nämlich die beiden Vektoren aus $\mathbb{R}^3$

$$\vec{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{y} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$$

so ist $N(\vec{x}) = 3$ und $N(\vec{y}) = 21$. Weil aber

$$(1^2 + 1^2 + 1^2) \cdot (1^2 + 2^2 + 4^2) = 3 \cdot 21 = 63$$

nicht Summe von drei Quadratzahlen ist, wie man durch Aufzählen aller Möglichkeiten erkennt, kann die Normproduktregel in drei Dimensionen nicht gelten!

Nach langem Suchen fand Hamilton am 16. Oktober 1843 einen weiteren Schiefkörper in Form der sogn. Quaternionen zu seiner Überraschung erst in der Dimension 4. Weil auch die Quaternionen als Matrizen-Algebra dargestellt werden können, geben wir eine Beschreibung der Quaternionen nach dem selben Muster, mit dem wir im letzten Abschnitt die komplexen Zahlen beschrieben haben.

Quaternionen spielen in viele Anwendungen im Umkreis von Raumdrehungen eine zentrale Rolle, weil das Rechnen mit ihnen bequemer, effizienter und stabiler ist, als mit den früher benutzten Drehmatrizen und Euler-Winkeln mit ihrer involvierten Trigonometrie. Mit Quaternionen lässt sich das Problem einer unerwünschten Blockade der kardanischen Aufhängung vermeiden. Deshalb sollten sie für Probleme in der Computer-Graphik, Robotik, Navigation und beim Berechnen von Planetenbahen verwendet werden. Ihre Effizienz erkennt man, wenn man die Komplexität des Quaternionenproduktes mit jener beim Multiplizieren der zugehörigen Drehmatrizen vergleicht.

Methode	Additionen	Multiplikationen	Speicher
Quaternionen	12	16	4
Drehmatrizen	18	27	9

Wir definieren Quaternionen, wie komplexe Zahlen, durch gewisse Matrizen, die wir im Moment aus dem Zylinder zaubern.

Definition. Unter der Menge der *Quaternionen* verstehen wir die Menge der Matrizen

$$\mathbb{H} = \left\{ \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} \mid a, b, c, d \in \mathbb{R} \right\},$$

Die Spaltenvektoren der Matrix $A \in \mathbb{H}$ sind paarweise orthogonal und haben alle die selbe Norm $a^2 + b^2 + c^2 + d^2$. Daher gilt $A \cdot A^T = (a^2 + b^2 + c^2 + d^2)E_4$. Wir identifizieren die reellen Quaternionen aE, für die $b = c = d = 0$ gilt, mit der reellen Zahl $a \in \mathbb{R}$.

Die schiefsymmetrischen Basisquaternionen

$$I = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad J = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

erfüllen die fundamentalen Beziehungen

$$I^2 = J^2 = -E, \qquad I \cdot J = -J \cdot I$$

Definiert man zusätzlich die schiefsymmetrische Matrix

$$K = I \cdot J = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

so erfüllen diese drei Matrizen die äquivalenten Beziehungen

$$I^2 = J^2 = K^2 = IJK = -E$$

die seinerzeit Hamilton angegeben hat. Mit diesen Basisquaternionen lässt sich jedes Quaternion auf eindeutige Art als Linearkombination in der Normalform

$$A_{a,b,c,d} = aE + bI + cJ + dK$$

ausdrücken, wobei a, b, c, d reelle Zahlen sind. Dafür benutzen wir, wie bei den komplexen Zahlen, manchmal auch knapper die Vektorform d.h. das Quadrupel $q = (a, b, c, d)$, wenn keine Verwechslungen zu befürchten sind. Die assoziative Quaternionenalgebra lässt sich nun nach dem selben Muster entwickeln, das wir oben für die komplexen Zahlen benutzt haben. Insbesondere rechnet man leicht nach, dass die angegebene Matrizenmenge $\mathbb{H}$ abgeschlossen unter den Matrizenoperationen ist. Beim Rechnen mit Quaternionen muss man einzig auf die Tatsache aufpassen, dass die Multiplikation von Quaternionen nicht mehr kommutativ sein muss, wie die Produkte

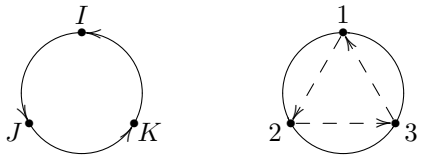
$$IJ = K = -JI$$

zeigen. Für die Multiplikationstabelle der Quaternionen erhalten wir

$\cdot$	E	I	J	K
E	E	I	J	K
I	I	$-E$	K	$-J$
J	J	$-K$	$-E$	I
K	K	J	$-I$	$-E$

$\cdot$	e_0	e_1	e_2	e_3
e_0	e_0	e_1	e_2	e_3
e_1	e_1	$-e_0$	e_3	$-e_2$
e_2	e_2	$-e_3$	$-e_0$	e_1
e_3	e_3	e_2	$-e_1$	$-e_0$

Diese Multiplikationstabelle lässt sich leicht an Hand des folgenden Quaternionenzirkels merken.



Zunächst erinnert man sich, dass E das multiplikative Neutralelement ist und dass die Quadrate von I, J, K die Eigenschaft

$$I^2 = J^2 = K^2 = -E$$

haben. Multipliziert man nun zwei aufeinanderfolgende Basisquaternionen im Gegenuhrzeigersinn, erhält man das nächste. Multipliziert man sie entgegengesetzt zur Pfeilrichtung, ergibt sich das negative des nächsten. Beispielsweise ist $I \cdot J = K$, aber $J \cdot I = -K$. Die Linearkombination der Normalform eines Quaternions $q = (a, b, c, d)$ wird gelegentlich auch in der vektoriellen Form

$$q = a1 + be_1 + ce_2 + de_3$$

geschrieben. Dabei identifiziert man die Einheitsmatrix E mit dem multiplikativen Neutralelement $E = e_0 = 1$ und schreibt für die Basisquaternionen $I = e_1$, $J = e_2$ und $K = e_3 = e_1 \cdot e_2$. Diese Elemente können als Standardbasisvektoren von $\mathbb{R}^4$ aufgefasst werden, wobei man mit 0 statt mit 1 zu zählen beginnt und daher das Vektorpfeilchen weglässt. In dieser Bezeichnungsweise nimmt die Multiplikationstabelle die oben angegebene äquivalente Form an. Man beachte, dass sie bis auf die Diagonale die Multiplikationstabelle des Vektorprodukten in $\mathbb{R}^3$ enthält.

In der Literatur findet man auch andere, aber gleichwertige Matrizendarstellungen der Quaternionen. Beispielsweise kann man $q = (a, b, c, d) \in \mathbb{H}$ als Linearkombination $q = aE + bI + cJ + dK$ durch die komplexen Matrizen

$$\begin{pmatrix} z_1 & -z_2 \\ \overline{z}_2 & \overline{z}_1 \end{pmatrix} = \begin{pmatrix} a+id & -b-ic \\ b-ic & a-id \end{pmatrix} \in \mathbb{C}^{2,2}$$

darstellen. In dieser komplexen Darstellung sind die Basisquaternionen also

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad J = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} \quad K = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

Sie erfüllen die erwähnten Relationen, wie man leicht überprüft. So gesehen verallgemeinern die Quaternionen die komplexen Zahlen, für die $c = d = 0$ gilt. Vertauscht man diese Basisquaternionen antizyklisch $I \to K \to J \to I$, erfüllen sie die selben Relationen und man erhält die äquivalente komplexe Darstellung

$$\begin{pmatrix} \overline{z}_1 & -i\overline{z}_2 \\ iz_2 & z_1 \end{pmatrix} = \begin{pmatrix} a-id & -c-ib \\ c-ib & a+id \end{pmatrix} \in \mathbb{C}^{2,2}$$

Diese Matrizen lassen sich als Linearkombination der Spinmatrizen

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

in der Form $q = a\sigma_0 - ib\sigma_1 - ic\sigma_2 - id\sigma_3$ zerlegen lassen, die von den Physikern in der Quantenmechanik gerne benutzt wird. Offenbar sind im Umgang mit Quaternionen nur die Relationen zwischen den Basisquaternionen der Multiplikationstabelle wirklich relevant und wir benutzen wieder unsere reelle Darstellung.

Das Produkt von zwei beliebigen Quaternionen $A_1, A_2 \in \mathbb{H}$ liefert ein weiteres Quaternion

$$(a_1, b_1, c_1, d_1) \cdot (a_2, b_2, c_2, d_2) = (a_3, b_3, c_3, d_3) \in \mathbb{H}$$

das in Komponenten durch die Produktformeln

$$\begin{array}{rcl} a_3 & = & a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2 \\ b_3 & = & a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2 \\ c_3 & = & c_1a_2 + d_1b_2 + a_1c_2 - b_1d_2 \\ d_3 & = & d_1a_2 - c_1b_2 + b_1c_2 + a_1d_2 \end{array}$$

gegeben ist. Sie lässt sich leicht aus obiger Multiplikationstabelle gewinnen und zeigt, dass das Quaternionenprodukt die Multiplikation komplexer Zahlen von zwei auf vier Dimensionen verallgemeinert, indem man $c_1 = c_2 = 0$ und $d_1 = d_2 = 0$ setzt. Der Verlust der Kommutativität bei der Multiplikation von Quaternionen deutet darauf hin, dass wir damit endlich auch Dinge beschreiben können, die nicht vertauschbar sind. Dazu gehören Rotationen im Raum oder Messungen in der Physik.

Obwohl Quaternionen also im Gegensatz zu den komplexen Zahlen nicht kommutativ sind, erfüllen sie alle anderen algebraischen Eigenschaften der reellen und der komplexen Zahlen, wenn man sie geeignet interpretiert. Das *konjugierte Quaternion* $\overline{A} = aE - bI - cJ - dK$ beschreibt eine Abbildung

$$\alpha\colon \mathbb{H} \to \mathbb{H}, \qquad A \mapsto \overline{A}$$

der Quaternionen, die wie erwartet durch Transposition der darstellenden Matrix A d.h. durch $\alpha(A) = \overline{A} = A^T$ erklärt ist. Die Transposition beschreibt allerdings für die Quaternionen keinen Automorphismus, da wegen der fehlenden Kommutativität der Quaternionen einerseits $(I \cdot J)^T = K^T = -K$ gilt. Anderseits ist aber $I^T \cdot J^T = (-I) \cdot (-J) = I \cdot J = K \neq -K$. Die Abbildung α liefert wegen der Verträglichkeit der Transposition mit der Multiplikation einen sogn. *Antiautomorphismus*, für den also $\alpha(A_1 \cdot A_2) = \alpha(A_2) \cdot \alpha(A_1)$ gilt.

Das Produkt $A \cdot \overline{A}$ hat auch für Quaternionen eine fundamentale Beziehung zum Betragsquadrat, da die Spaltenvektoren von A ein Orthogonalsystem bilden, dessen Betragsquadrate übereinstimmen. Es gilt die Beziehung

$$A \cdot \overline{A} = a^2 + b^2 + c^2 + d^2 = N(A) \in \mathbb{R}$$

Sie kann auch in der Form $\det(A) = N(A)$ formuliert werden und beschreibt anschaulich, wie weit $(a, b, c, d) \in \mathbb{H}$ vom Ursprung entfernt ist.

Wegen der Multiplikativität der Determinante erfüllt diese Quaternionennorm die bemerkenswerte Normproduktregel

$$N(A) \cdot N(B) = N(A \cdot B)$$

Sie nimmt in Komponenten die Form des Vier-Quadrate-Satzes

$$(a_1^2 + b_1^2 + c_1^2 + d_1^2) \cdot (a_2^2 + b_2^2 + c_2^2 + d_2^2) = (a_3^2 + b_3^2 + c_3^2 + d_3^2)$$

an, wobei wie oben

$$\begin{aligned} a_3 &= a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2 \\ b_3 &= a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2 \\ c_3 &= c_1a_2 + d_1b_2 + a_1c_2 - b_1d_2 \\ d_3 &= d_1a_2 - c_1b_2 + b_1c_2 + a_1d_2 \end{aligned}$$

gilt. Er lässt sich, wenn man ihn einmal vermutet, leicht nachrechnen und ist äquivalent zum Multiplikationsgesetz der Quaternionen. Sie bildet die Ursache für die bemerkenswerte Tatsache, dass jede natürliche Zahl die Summe von vier ganzzahligen Quadraten ist. Diese Formel liefert uns analog zur Situation bei den komplexen Zahlen eine Multiplikationsabbildung

$$\mu_4 \colon \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}^4, \qquad \big((a_1, b_1, c_3, d_4), (a_2, b_2, c_2, d_1)\big) \mapsto (a_3, b_3, c_3, d_3)$$

und besitzt die Einheit $(1, 0, 0, 0)$. Sie lässt sich dank des Vierquadratesatzes auf eine Multiplikationsabbildung der 3-Sphäre

$$S^3 = \{(a, b, c, d) \mid a^2 + b^2 + c^2 + d^2 = 1\} \subset \mathbb{R}^4$$

einschränken, wir mit den Einheitsquatrernionen identifizieren können. und liefert dort die Multiplikationsabbildung

$$\mu_4 \colon S^3 \times S^3 \to S^3, \qquad \big((a_1, b_1, c_3, d_4), (a_2, b_2, c_2, d_1)\big) \mapsto (a_3, b_3, c_3, d_3)$$

Diese Gruppenstruktur der S^3 spielt, wie seinerzeit die Multiplikationsstruktur von S^0 der reellen Zahlen und die Multiplikationsstruktur von S^1 der komplexen Zahlen in der klassischen Mathematik, in der heutigen Mathematik und ihren Anwendungen eine zentrale Rolle.

Wie früher kann man damit zeigen, dass alle Quaternionen mit Ausnahme der 0 invertierbar sind. Die Inversen von A lassen sich mit Hilfe der Norm

$$A^{-1} = \frac{1}{N(A)}\overline{A}, \qquad N(A) \neq 0$$

berechnen. Insbesondere erhalten wir das gegenüber den komplexen Zahlen leicht unterschiedliche folgendes Resultat.

Satz. Die Quaternionen bilden einen Schiefkörper, der nicht kommutativ ist.

Das folgende numerische Beispiel dient zur llustration dieser Rechenregeln.

Beispiel. Für das Quaternion $x = (2, 1, 3, 1) = 2E + I + 3J + K$ ist das konjugierte Quaternion $\overline{x} = (2, -1, -3, -3) = 2E - I - 3J - K$ und damit $|x|^2 = 4 + 1 + 9 + 1 = 15$. Mit $y = (0, 1, 0, -2) = I - 2K$ gilt für das Produkt $x \cdot y = (3, -4, 4, -5) = 3E - 4I + 4J - 5K$. ◯

Zwei Teilmengen von Quaternionen spielen für die geometrischen Anwendungen eine zentrale Rolle.

Definition. Quaternionen $A = aE + xI + yJ + zK$, deren Realteil $a = 0$ ist, heissen *reine oder vektorielle Quaternionen* und solche mit $N(A) = 1$ heissen *Einheitsquaternionen*.

Die reinen Quaternionen haben also die Form $v = xI + yJ + zK$ und können mit dem Punkt (x, y, z) des dreidimensionalen Raumes bzw. mit dem Vektor

$$\vec{v} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3$$

identifiziert werden. Obwohl die reinen Quaternionen unter Summen abgeschlossen sind, gilt das Entsprechende für Produkte nicht. Falls der Vektor $\vec{w} \in \mathbb{R}^3$ mit dem zugehörigen vektoriellen Quaternion $w = \tilde{x}I + \tilde{y}J + \tilde{z}K$ identifiziert wird, so gilt gilt für das Quaternionenprodukt

$$v \cdot w = -(x\tilde{x} + y\tilde{y} + z\tilde{z})E + (y\tilde{z} - \tilde{y}z)I - (x\tilde{z} - \tilde{x}z)J + (x\tilde{y} - \tilde{y}x)K$$

Das Produkt vektorieller Quaternionen hat also einen Realteil, der bis auf das Vorzeichen das Euklid'sche Skalarprodukt $\langle \vec{v}, \vec{w} \rangle$ ist. Sein vektorieller Anteil ist das Vektorprodukt $\vec{v} \times \vec{w}$ dieser beiden Vektoren. Deshalb ist $v \cdot w$ genau dann ein vektorielles Quaternion, falls die beiden Vektoren orthogonal sind. Es ist genau dann ein reelles Quaternion, falls $\vec{v} \times \vec{w} = \vec{0}$ ist, d.h. falls $\vec{v}$ und $\vec{w}$ kollinear sind.

Dank dieser Gleichung kann man umgekehrt die Operationen der Vektorgeometrie mit Hilfe des Quaternionenproduktes ausdrücken. Für die beiden Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^3$ mit den beiden zugehörigen vektoriellen Quaternionen v, w gilt

$$\langle \vec{v}, \vec{w} \rangle = \frac{1}{2}\Big(\overline{v} \cdot w + w \cdot \overline{v}\Big), \quad \vec{v} \times \vec{w} = \frac{1}{2}\Big(v \cdot w - w \cdot v\Big)$$

Mit ihnen lassen sich die Eigenschaften dieser beiden Operationen, die historisch jünger als das Quaternionenprodukt sind, leicht bestätigen. Tatsächlich kann man Quaternionen auch in der vektoriellen Form

$$q = aE + xI + yJ + zK = (a, x, y, z) = (a, \vec{v}) \in \mathbb{H}^4, \quad a \in \mathbb{R}, \vec{v} \in \mathbb{R}^3$$

darstellen, wobei a den Realteil und $\vec{v} \in \mathbb{R}^3$ den vektoriellen Teil von q bezeichnet. In dieser Bezeichnungsweise lässt sich das Quaternionenprodukt der beiden Quaternionen $q_1 = (a_1, \vec{v})$ und $q_2 = (a_2, \vec{w})$ durch die Beziehung

$$q_1 \cdot q_2 = \big(a_1 a_2 - \langle \vec{v}, \vec{w} \rangle, a_1\vec{w} + a_2\vec{v} + \vec{v} \times \vec{w}\big) \in \mathbb{H}$$

mit Hilfe der Operationen der Vektoralgebra ausdrücken.

Wie bei den komplexen Zahlen haben die Einheitsquaternionen $A \in S^3$ eine eindeutige Polardarstellung bzw. eine zugehörige Exponentialform

$$A = \cos(\varphi)E + \sin(\varphi)u = e^{\varphi u}, \qquad N(u) = 1$$

wobei das vektorielle Quaternion u einem Einheitsvektor $\vec{u} \in \mathbb{R}^3$ entspricht.

Allgemeiner hat ein beliebiges Quaternion $A = (a_1, a_2, a_3, a_4) \neq 0$ mit dem zugehörigen vektoriellen Quaternion $a = (0, a_2, a_3, a_4)$ die Exponentialdarstellung

$$A = |A| \cdot \Big(\cos(\varphi)E + \sin(\varphi)u\Big) = |A| \cdot e^{\varphi u}, \qquad N(u) = 1$$

Dabei ist

$$\tan(\varphi) = \frac{|A|}{a_1}, \qquad u = \frac{a}{|A|}$$

Beispiel. Für $A = (3, 2, 2, 1)$ ist $|A| = 3\sqrt{2}$ und $a = (0, 2, 2, 1)$ mit $|a| = 3$. Daraus erhalten wir $\varphi = \frac{\pi}{4}$ und die Polardarstellung

$$A = 3\sqrt{2} \cdot \Big(\cos\big(\frac{\pi}{4}\big) + \frac{2I + 2J + K}{3} \sin\big(\frac{\pi}{4}\big) \Big)$$

Man kann sie durch ausmultiplizieren bestätigen. ◯

Die quaternionale Exponentialfunktion liefert also eine surjektive Abbildung

$$\mathbb{R}^3 \to S^3, \qquad \varphi u \mapsto e^{\varphi u}, \quad |\vec{u}| = 1$$

Die Inverse solcher $A \in S^3$ existiert und es gilt wie bei den komplexen Zahlen

$$A^{-1} = \frac{1}{N(A)} \overline{A} = \cos(\varphi) E - \sin(\varphi) u = e^{-\varphi u}$$

Bisher verläuft die Theorie der Quaternionen also ganz analog zu jener der komplexen Zahlen. Insbesondere kann man zeigen, dass die vektoriellen Quaternionen den Tangentialraum in E an S^3 bilden. Die Situation mit S^3 ist aber im Vergleich mit jener in S^1 deutlich komplexer, weil die Multiplikation von Quaternionen im allgemeinen nicht kommutativ ist. Das hat beispielsweise zur Folge, dass für die quaternionale Exponentialfunktion das Additionstheorem nicht in der aus der Schule bekannten Form zu gelten braucht!

Die reinen Einheitsquaternionen entsprechen den Punkten auf der 2-Sphäre S^2 im dreidimensionalen Raum. Weil sie nicht unter Multiplikation abgeschlossen sind, erhalten wir so auf S^2 keine Multiplikation. Weil sich alle Punkte auf der Einheitssphäre mit Hilfe von Kugelkoordinaten beschreiben lassen, können sie auch durch die reinen Einheitsquaternionen

$$\cos(\vartheta)\cos(\varphi) I + \cos(\vartheta)\sin(\varphi) J + \sin(\vartheta) K$$

für beliebige Winkel ϑ und φ beschreiben werden. Man rechnet leicht nach, dass alle diese Quaternionen ein Quadrat haben, das -1 ist.

Die Identifikation der Punkte des Raumes $\mathbb{R}^3$ mit den reinen Quaternionen ist der Schlüssel für ihre Anwendungen in der Raumgeometrie. Ferner stellt es sich heraus, dass die Transformation $A \mapsto Q \cdot A \cdot Q^{-1}$ reine Quaternionen $A \in \mathbb{R}^3$ wieder in reine Quaternionen überführt, falls Q invertierbar ist, d.h. insbesondere, falls $Q \in S^3$ ein Einheitsquaternion bezeichnet. Interpretiert man also die reinen Quaternionen als Vektoren in $\mathbb{R}^3$, wird durch diese Transformation eine gewissen Abbildung des Raumes induziert, die eine einfache geometrische Interpretation hat.

Satz. Falls die Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^3$ den beiden vektoriellen Quaternionen v, w entsprechen, so beschreibt die Abbildung

$$\rho_w \colon \mathbb{R}^3 \to \mathbb{R}^3, \qquad v \mapsto -w \cdot v \cdot w^{-1}$$

eine Spiegelung an der Ebene durch den Ursprung mit dem Normalenvektor $\vec{w}$.

Diese Abbildung ist nämlich linear und es gilt $\rho_w(w) = -w$. Für jeden Vektor $\vec{x} \perp \vec{w}$ ist ferner $\rho_w(x) = -w \cdot x \cdot w^{-1} = -x \cdot w \cdot w^{-1} = -x$.

Weil die zweifachen Ebenenspiegelungen genau den Drehungen entsprechen erhalten wir sofort das folgende wichtige Resultat.

Satz. Die Drehungen von $\mathbb{R}^3$ entprechen genau den Abbildungen $-\rho_w$ für einen gewissen Vektor $\vec{w} \neq \vec{0}$.

Genauer lassen sich diese Transformationen geometrisch als Drehung des dreidimensionalen Raumes um eine Achse interpretieren. Es gilt also folgenes, für die Anwendungen der Vektorgeometrie hilfreiches Resultat.

Satz. Falls das Einheitsquaternion $Q \in S^3$ die Polardarstellung

$$Q = \cos(\varphi)E + \sin(\varphi)u = e^{\varphi u}$$

mit dem Einheitsvektor $\vec{u} \in \mathbb{R}^3$ hat, so liefert für jeden Vektor $\vec{v} \in \mathbb{R}^3$ mit dem zugehörigen vektoriellen Quaternion v die Konjugation $v \mapsto Q \cdot v \cdot Q^{-1}$ wiederum ein vektorielles Quaternion, desssen zugehöriger Vektor aus $\vec{v}$ durch Drehung um die Drehachse $\vec{u}$ mit dem Drehwinkel 2φ entsteht.

Beispiel. Das Einheitsquaternion $Q = \cos(\theta) + \sin(\theta)K$ hat das multiplikative Inverse $Q^{-1} = \cos(\theta) - \sin(\theta)K$. Daher gilt $Q \cdot K \cdot Q^{-1} = K$. Ferner gilt für das vektorielle Quaternion

$$v = r\big(\cos(\phi)I + \cos(\phi)J\big)$$

die Beziehung

$$Q^{-1} \cdot v \cdot Q = r\big(\cos(2\theta + \phi)I + \sin(2\theta + \phi)J\big)$$

woraus folgt, dass die Abbildung $v \mapsto Q \cdot v \cdot Q^{-1}$ in diesem Fall tatsächlich eine Drehung um die z-Achse mit dem Drehwinkel 2θ beschreibt. ◯

Als Folge dieses Resultates kann jede Drehung von $\mathbb{R}^3$ mit dem Einheitsvektor $\vec{u}$ als Achse und dem Drehwinkel φ durch Konjugation mit einem gewissen Einheitsquaternion $Q = \cos(\varphi)E + \sin(\varphi)u$ beschrieben werden. Weil $(-Q) \cdot A \cdot (-Q)^{-1} = Q \cdot A \cdot Q^{-1}$ gilt, bewirkt das Einheitsquaternion $-Q$ die selbe Drehung wie Q. Man stellt weiter fest, dass die beiden antipodischen Einheitsquaternionen $\pm Q$ die einzigen Einheitsquaternionen sind, die diese Drehung bewirken. Daher entspricht jede Drehung des Raumes $\mathbb{R}^3$ genau einem Antipodenpaar von Einheitsquaternionen.

Unter dieser Beziehung zwischen Raumdrehungen und Einheitsquaternionen entspricht die Komposition von zwei Raumdrehungen mit der Dreachsen $\vec{u}_1$ bzw. $\vec{u}_2$ und den Drehwinkeln φ_1 bzw. φ_2 dem Produkt $Q_1 \cdot Q_2$ der zugehörigen Einheitsquaternionen.

In der Computer-Graphik werden also die Quaternionen zweckmässig an Stelle der mühsamen Euler-Winkel zur Beschreibungen von Raumdrehungen benutzt. Steuerungen von anspruchsvollen Fluggeräten wie Helikoptern oder des space shuttle und Computer-Simulationen basieren auf Quaternionen, wenn sie zuverlässig funktionieren sollen. Weil sich Skalar- und Vektorprodukt mit Hilfe des Quaternionenproduktes einheitlich behandeln lassen, können mit ihnen die Kreiseltheorie in der Mechanik oder die Maxwell'schen Gleichungen in der Elektrodynamik in besonders übersichtlicher Form geschrieben werden. Quaternionen spielen auch in der Quantenmechanik beim Studium des Elektronen-Spins und

damit bei der Erklärung des Periodensystems in der Chemie eine fundamentale Rolle, weil die Einheitsquaternionen isomorph zu SU_2, der Symmetriegruppe des allereinfachsten quantenmechanischen Systems mit Spin $\frac{1}{2}$, ist.

Nachdem Hamilton in der Dimension vier so erfolgreich war, stellte sich die Frage, ob sich sein Erfolg in noch höheren Dimensionen wiederholen lässt. Eine erste — negative — Antwort gibt ein weiterer Satz von Frobenius.

Satz. Der Vektorraum $\mathbb{R}^m$ hat höchstens in den Dimensionen $1, 2, 4$ die Struktur eines Schiefkörpers.

Will man also einen der restlichen Euklid'schen Räume mit einem Produkt ausrüsten, muss man die Wünsche weiter lockern.

Gibt man zusätzlich zu den Inversen noch die Assoziativität des Produktes auf und fragt, in welchen Dimensionen m es möglich ist, eine bilineare Multiplikation

$$\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m, \qquad (\vec{x}, \vec{y}) \mapsto \vec{x} \cdot \vec{y}$$

zu erklären, die $\vec{e}_1$ als Neutralelement besitzt und die mit der Norm verträglich ist, d.h. dass die *Normproduktregel*

$$|\vec{x} \cdot \vec{y}|^2 = |\vec{x}|^2 \cdot |\vec{y}|^2$$

erfüllt ist, so fand Cayley in der Dimension 8 eine weitere sogn. *normierte Algebra* $\mathbb{O}$, deren Produkt allerdings nicht mehr assoziativ ist und die deshalb nicht durch Matrizen dargestellt werden kann. Ihre Elemente sind die Linearkombinationen

$$a_0 1 + a_1 e_1 + a_2 e_2 + a_3 e_3 + a_4 e_4 + a_5 e_5 + a_6 e_6 + a_7 e_7, \qquad a_j \in \mathbb{R}$$

Das Element 1 spielt die Rolle des multiplikativen Neutralelementes und die restlichen Basisoktaven erfüllen neben den Relationen

$$e_j \cdot e_j = -1, \qquad e_j \cdot e_k = -e_k \cdot e_j, \quad (j \neq k)$$

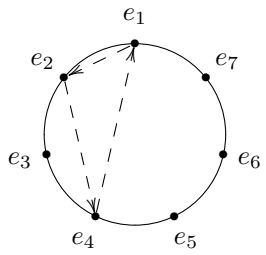
Beziehungen, die in der kompakten Form

$$e_j \cdot e_{j+1} = e_{j+3}$$

zusammengefasst werden können. Dabei werden die Indizes zyklisch permutiert und modulo 7 verschoben. Genauer gilt für die Basisoktaven die Multiplikationstabelle

$\cdot$	1	e_1	e_2	e_3	e_4	e_5	e_6	e_7
1	1	e_1	e_2	e_3	e_4	e_5	e_6	e_7
e_1	e_1	-1	e_4	e_7	$-e_2$	e_6	$-e_5$	$-e_3$
e_2	e_2	$-e_4$	-1	e_5	e_1	$-e_3$	e_7	$-e_6$
e_3	e_3	$-e_7$	$-e_5$	-1	e_6	e_2	$-e_4$	e_1
e_4	e_4	e_2	$-e_1$	$-e_6$	-1	e_7	e_3	$-e_5$
e_5	e_5	$-e_6$	e_3	$-e_2$	$-e_7$	-1	e_1	e_4
e_6	e_6	e_5	$-e_7$	e_4	$-e_3$	$-e_1$	-1	e_2
e_7	e_7	e_3	e_6	$-e_1$	e_5	$-e_4$	$-e_2$	-1

Diese Multiplikationstabelle lässt sich leicht an Hand des folgenden Oktavenzirkels merken.



Dabei multipliziert man zwei Elemente im gestrichelten Dreieck im Gegenuhrzeigersinn um das dritte Element als Produkt zu erhalten. Multipliziert man sie entgegengesetzt zur Pfeilrichtung, ergibt sich das negative Produkt. Anschliessend wird die Figur um den Winkel $\frac{2\pi}{7}$ gedreht und analog vorgegangen. Aus der angegebenen Lage erhält man beispielsweise durch Multiplizieren im Gegenuhrzeigersinn die erste Zeile der folgenden Tabelle.

$$\begin{array}{lll}
e_1 \cdot e_2 = e_4 & e_2 \cdot e_4 = e_1 & e_4 \cdot e_1 = e_2 \\
e_2 \cdot e_3 = e_5 & e_3 \cdot e_5 = e_2 & e_5 \cdot e_2 = e_3 \\
e_3 \cdot e_4 = e_6 & e_4 \cdot e_6 = e_3 & e_6 \cdot e_3 = e_4 \\
e_4 \cdot e_5 = e_7 & e_5 \cdot e_7 = e_4 & e_7 \cdot e_4 = e_5 \\
e_5 \cdot e_6 = e_1 & e_6 \cdot e_1 = e_5 & e_1 \cdot e_5 = e_6 \\
e_6 \cdot e_7 = e_2 & e_7 \cdot e_2 = e_6 & e_2 \cdot e_6 = e_7 \\
e_7 \cdot e_1 = e_3 & e_1 \cdot e_3 = e_7 & e_3 \cdot e_7 = e_1
\end{array}$$

Die restlichen Einträge der Multiplikationstabelle ergeben sich entsprechend durch Multiplizieren im Uhrzeigersinn. Dass dieses Mutiplikationsgesetz nicht assoziativ sein kann, zeigt schon das Gegenbeispiel

$$-e_6 = (e_1 \cdot e_2) \cdot e_3 \neq e_1 \cdot (e_2 \cdot e_3) = e_6$$

Das nichtassoziative Produkt der Oktaven produziert einige beachtenswerte geometrische Anomalien. Weil nur die projektive oktavische Ebene $\mathbb{OP}^2$ zur Verfügung steht, sind mit den Oktaven relativ wenige Symmetriegruppen, die sogn. Ausnahmegruppen, verknüpft, die aber in der modernen Physik eine zentrale Rolle spielen. Die Automorphismengruppe von $\mathbb{O}$ ist nicht die ganze 21-dimensionale Drehgruppe SO_7, sondern bloss eine Untergruppe, die 14-dimensionale Ausnahmegruppe G_2. In allen Dimensionen $n \geq 3$ besteht die Drehgruppe SO_n aus lauter inneren Automorphismus der Form $A \mapsto Q^{-1} \cdot A \cdot Q$ für $Q \in \mathrm{O}_n$. Auch für ihre universelle Überlagerung Spin_n sind für $n \geq 8$ alle Automorphismen von der Form $a \mapsto s^{-1} \cdot a \cdot s$, wobei $s \in \mathrm{Pin}_n$ ist. Einzig die Gruppe Spin_8 besitzt weitere Ausnahmeautomorphismen. Ein solcher exotischer Automorphismus von Spin_8 wird als Trialität bezeichnet.

In einer berühmten Arbeit zeigte dann Hurwitz, dass normierte (nicht notwendigerweise assoziative) Algebren höchstens in den Dimensionen $m = 1, 2, 4, 8$ existieren können. Die reellen Zahlen $\mathbb{R}$, die komplexen Zahlen $\mathbb{C}$, die Quaternionen $\mathbb{H}$ und die Oktaven $\mathbb{O}$ liefern tatsächlich normierte Algebren in den Dimenionen $m = 1, 2, 4, 8$, wobei die ersten drei assoziativ sind. Nur für diese m kann ein Produkt zweier Summen von m perfekten Quadraten als Summe von von m perfekten Quadraten ausgedrückt werden.

Als weitere, scheinbar weitreichende, Lockerung der Wünsche kann man die Bilinearität des Produktes aufgeben und nur noch Stetigkeit der Multiplikationsabbildung verlangen. Die beschriebenen Multiplikationen liefern durch Einschränkung auf die Einheitsvektoren stetige Abbildungen

$$\mu\colon S^{m-1} \times S^{m-1} \to S^{m-1}$$

die e_0 als Neutralelement haben. Weil es scheinbar viel mehr stetige Abbildungen gibt, als solche, die von einem bilinearen Produkt herrühren, stellte sich die Frage, für welche m eine solche stetige Multiplikation (H-Struktur) auf der m-Sphäre existiert. Mit Hilfe der sogn. Hopf-Konstruktion liefert jede solche Multiplikationsabbildung eine stetige Abbildung

$$H(\mu)\colon S^{2m-1} \to S^m$$

mit einer sehr einschränkenden Eigenschaft. Aus einer sehr anspruchsvollen Arbeit von Adams folgt schliesslich, dass eine solche Multiplikation nur für $m = 1, 2, 4, 8$ existieren kann. In diesen niedrigen Dimensionen herrscht also tatsächlich ein sehr spezieller Ausnahmezustand, der für viele Phänomene in höheren Dimensionen verantwortlich und noch lange nicht vollständig verstanden ist.

Will man das Assoziativgesetz der Multiplikation beibehalten, um eine Chance zu haben, die Algebra durch Matrizen darzustelllen, bietet es sich an, in höheren Dimensionen Nullteiler zuzulassen und sich an den Relationen der Erzeugenden in den niedrigdimensionalen Beispielen zu orientieren. In der zweidimensionalen reellen Algebra der komplexen Zahlen $\mathbb{C}$ gibt es ein erzeugendes Element I, dessen Quadrat $-E$ ergibt. Analog findet man in der vierdimensionalen Quaternionenalgebra $\mathbb{H}$ drei solche erzeugende Elemente I, J, K, deren Quadrate $-E$ ergeben und die antikommutieren. Sie erfüllen also die Relationen

$$I^2 = J^2 = K^2 = -E, \qquad I \cdot J = -J \cdot I, I \cdot K = -K \cdot I, J \cdot K = -K \cdot J$$

Es liegt also nahe, in höheren Dimensionen eine m-dimensionale assoziative Algebra hyperkomplexer Zahlen zu erklären, in der $m-1$ solche antikommutierende Elemente existieren, deren Quadrate -1 ergibt. Wir werden gleich sehen, warum dann ihre Dimension $m = 2^n$ eine Zweierpotenz ist.

Definition. Für jede natürliche Zahl n definiert man die *Clifford-Algebra* Cl_n als reelle, assoziative Algebra mit $1 \in \mathrm{Cl}_n$, die durch n antikommutierernde Basiselemente $e_1, e_2, \ldots, e_n \in \mathrm{Cl}_n$ erzeugt wird, deren Quadrate -1 sind.

Diese Algebra-Erzeugenden von Cl_n erfüllen definitionsgemäss die Relationen

$$e_i^2 = -1, \qquad e_i \cdot e_j = -e_j \cdot e_i \quad (i \neq j)$$

und Cl_n wird aus allen reellen Linearkombinationen der formalen Produkten

$$e_{i_1} e_{i_2} \cdots e_{i_r}, \quad 1 \leq i_1 < i_2 < \cdots < i_r \leq n,\ 0 \leq r \leq n$$

dieser Erzeugenden gebildet. Das leere Produkt bezeichnen wir mit 1. Insbesondere gibt es also eine natürliche Einbettung

$$\mathbb{R}^n \subset \mathrm{Cl}_n, \qquad \vec{x} \mapsto x_1 e_1 + \cdots + x_n e_n$$

Zählt man die formalen Produkte der Basiselemente, erkennt man für die Dimension von Cl_n die Beziehung

$$\dim(\mathrm{Cl}_n) = 1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$$

Beispiel. Für $n = 0$ benötigen wir nur das leere Produkt 1. Die Clifford-Algebra $\mathrm{Cl}_0 \cong \mathbb{R}$ ist im wesentlichen die Algebra der reellen Zahlen. ◯

Beispiel. Für $n = 1$ gehen wir von den 2 Basiselementen $1, e_1$ aus und es ist $e_1^2 = -1$. Die Zuordnung

$$\mathrm{Cl}_1 \to \mathbb{C}, \quad x + ye_1 \mapsto x + yi, \quad x, y \in \mathbb{R}$$

und die Multiplikationstabelle

$\cdot$	1	e_1
1	1	e_1
e_1	e_1	-1

zeigen, dass die 2-dimensionale Algebra $\mathrm{Cl}_1 \cong \mathbb{C}$ im wesentlichen die Algebra der komplexen Zahlen ist. ◯

Beispiel. Für $n = 2$ benutzen wir die 4 Basiselemente $1, e_1, e_2, e_1e_2$. Sie erfüllen die Relationen $e_1^2 = e_2^2 = -1$ und $e_1e_2 = -e_2e_1$. Die Zuordnung

$$\mathrm{Cl}_2 \to \mathbb{H}, \quad a1 + be_1 + ce_2 + de_1e_2 \mapsto aE + bI + cJ + dK, \quad a, b, c, d \in \mathbb{R}$$

und die Multiplikationstabelle

$\cdot$	1	e_1	e_2	e_1e_2
1	1	e_1	e_2	e_1e_2
e_1	e_1	-1	e_1e_2	$-e_2$
e_2	e_2	$-e_1e_2$	-1	e_1
e_1e_2	e_1e_2	e_2	$-e_1$	-1

zeigen, dass $\mathrm{Cl}_2 \cong \mathbb{H}$ im wesentlichen die Algebra der Quaternionen ist. ◯

Beispiel. Für $n = 3$ benutzen wir die 8 Basiselemente

$$1, e_1, e_2, e_1e_2, e_3, e_1e_3, e_2e_3, e_1e_2e_3$$

Sie erfüllen die Relationen $e_1^2 = e_2^2 = e_3^2 = -1$ und $e_1e_2 = -e_2e_1$, $e_1e_3 = -e_3e_1$ und $e_2e_3 = -e_3e_2$. Damit erhalten wir die Multiplikationstabelle

$\cdot$	1	e_1	e_2	e_1e_2	e_3	e_1e_3	e_2e_3	$e_1e_2e_3$
1	1	e_1	e_2	e_1e_2	e_3	e_1e_3	e_2e_3	$e_1e_2e_3$
e_1	e_1	-1	e_1e_2	$-e_2$	e_1e_3	$-e_3$	$e_1e_2e_3$	$-e_2e_3$
e_2	e_2	$-e_1e_2$	-1	e_1	e_2e_3	$-e_1e_2e_3$	$-e_3$	e_1e_3
e_1e_2	e_1e_2	e_2	$-e_1$	-1	$e_1e_2e_3$	$-e_2e_3$	$-e_1e_3$	$-e_3$
e_3	e_3	$-e_1e_3$	$-e_2e_3$	$e_1e_2e_3$	-1	e_1	e_2	$-e_1e_2$
e_1e_3	e_1e_3	e_3	$-e_1e_2e_3$	$-e_2e_3$	$-e_1$	-1	e_1e_2	e_2
e_2e_3	e_2e_3	$e_1e_2e_3$	e_3	e_1e_3	$-e_2$	$-e_1e_2$	-1	$-e_1$
$e_1e_2e_3$	$e_1e_2e_3$	$-e_2e_3$	e_1e_3	e_3	$-e_1e_2$	e_2	$-e_1$	-1

Ein beliebiges Element von Cl_3 hat die vektorielle Form

$$a_1 1 + a_2 e_1 + a_3 e_2 + a_4 e_1 e_2 + a_5 e_3 + a_6 e_1 e_3 + a_7 e_2 e_3 + a_8 e_1 e_2 e_3, \quad a_k \in \mathbb{R}$$

Um eine matrizielle Darstellung $\rho\colon \mathrm{Cl}_3 \to M_{2^3}(\mathbb{R})$ dieser Algebra zu erhalten, ◯

To do! Weitere Hinweise auf Clifford-Algebra. Beschreibung der Matrizendarstellung von Cl_3 der Dimension 2^3 und Nachweis, dass dann Nullteiler vorhanden sind. Periodizität, Beziehung zu SO_n.

2.8 Endliche Körper

Für praktische Zwecke in der Informatik hat der Umgang mit ungenauen, unendlichen Körpern — wie $\mathbb{R} \subseteq \mathbb{C} \subseteq \mathbb{H}$ — den nicht unbeträchtlichen Nachteil, dass dann einerseits beim Potenzieren sehr grosse Zahlen entstehen können, was zu Überlauf führen kann und sich andererseits die Rundungsfehler zu unbrauchbaren Ergebnissen anhäufen können. Beiden Problemen und damit dem Sumpf der Numerik kann man bei der Verwendung endlicher Körper

$$\mathbb{K}_q = \{r_1, r_2, \ldots, r_q\}$$

aus dem Weg gehen. In einem solchen Körper gibt es bloss q viele Elemente d.h. insbesondere nur *endlich viele* Zahlen, mit denen sich aber analog wie mit den reellen oder komplexen Zahlen rechnen lässt. Wir werden die Theorie der endlichen Körper so aufbauen, dass im Umgang mit ihnen einzig die Matrizenrechnung und das Rechnen mit ganzen Zahlen beschränkter Grösse notwendig ist und daher weder Rundungsfehler noch Überlauf auftreten.

Galois $(1811-1832)$ hat die Theorie der endlichen Körper entwickelt, als er sich mit der Frage beschäftigt hat, wann sich Polynomgleichungen durch Radikale lösen lassen und damit so nebenbei einige der klassischen schwierigen Probleme der Mathematik auf überraschende Weise erledigte. Die Dreiteilung eines beliebigen Winkels, die Verdoppelung eines Würfels, die Quadratur des Kreises sind mit Zirkel und Lineal nicht möglich und eine allgemeine algebraische Gleichung vom Grad 5 kann nicht durch Radikale gelöst werden. Weil die endlichen Köper lange Zeit rein-mathematischen Zwecken dienten und keine Anwendungen ausserhalb der Mathematik hatten, gilt die Theorie der endlichen Körpern unter vielen mathematischen Ignoranten immer noch als abstrakt und schwierig. Wir werden hier einen elementaren Weg einschlagen und dabei insbesondere Quotienten- und Polynomringe und die Erweiterungstheorie der Körper vermeiden, sondern die endlichen Körper analog zur Körpererweiterung $\mathbb{R} \subseteq \mathbb{C}$, an die erfahrungsgemäss die meisten Praktiker dank der Differential- und Integralrechnung fest glauben, mit Hilfe von Matrizen konstruieren[43]. Die endlichen Körper dienen in der Informatik als diskrete Struktur, die als Verallgemeinerung der Primkörper in der linearen Kodierungstheorie und in der Kryptologie eine fundamentale Rolle spielen. Der Autor kennt keinen Weg, wie sich die endlichen Körper auf natürliche Art über $\mathbb{C}$ realisieren und damit geometrisch interpretieren bzw. wie sich damit physikalische Theorien diskret approximieren lassen.

[43] Der Connaisseur erinnert sich, dass der Zerfällungskörper $\mathbb{E}$ eines irreduziblen, separablen Polynoms $p(x) \in \mathbb{K}[x]$ über dem Körper $\mathbb{K}$ durch den Isomorphismus $\mathbb{E} \cong \mathbb{K}[B(r(x))]$ matriziell beschrieben werden kann, wobei $r(x)$ die Resolvente von $p(x)$ und $B(r(x))$ seine Begleitermatrix ist.

Beispiel. Der einfachste endliche Körper, für den man wohl im Zeitalter der Informatik keine Propaganda mehr zu machen braucht, hat die Charakteristik 2 und besteht aus den beiden Binärzahlen $\mathbb{Z}_2 = \{0, 1\}$ mit den beiden Operationen

$+_2$	0	1
0	0	1
1	1	0

$\cdot_2$	0	1
0	0	0
1	0	1

x	$-x$
0	0
1	1

x	x^{-1}
0	–
1	1

(kleines Einsundeins bzw. kleines Einmaleins). In Zukunft werden wir die lästige Bezeichnung der beiden Rechenoperationen mit unteren Indizes vermeiden. Aus dem Kontex sollte klar werden, in welchen Körper gerechnet wird. Man kann sich diese Tabellen dadurch merken, dass man beachtet, dass es sich um die 2-er Reste bei der üblichen Addition und Multiplikation der ganzen Zahlen handelt. Die Projektion

$$\pi_2 \colon \mathbb{Z} \to \mathbb{Z}_2, \qquad x \mapsto x \bmod 2$$

ordnet also jeder ganzen Zahl x ihren 2-er Rest zu, der angibt, ob die ganze Zahl x gerade oder ungerade ist. Die Operationen in $\mathbb{Z}_2$ sind gerade so erklärt, dass diese Projektion mit der Addition und der Multiplikation verträglich ist. Ein Vergleich mit den Wahrheitstabellen aus der Boole'schen Algebra zeigt, dass es sich bei der Addition um die exklusive Oder-Verknüpfung (XOR) und bei der Multiplikation um die Oder-Verknüpfung (OR) handelt. Entscheidend ist, dass man mit Hilfe dieser beiden Operationen in $\mathbb{Z}_2$ weitgehend so rechnen kann, wie man sich das von der üblichen rationalen Zahlen aus $\mathbb{Q}$ gewohnt ist, solange man sich auf die Grundoperationen beschränkt.

Im Körper $\mathbb{Z}_2$ gelten aber auch neue, und vielleicht etwas gewöhungsbedürftige Eigenschaften. Man beachtet beispielsweise, dass für alle Elemente $r \in \mathbb{Z}_2$ die Gleichung $-r = r$ gilt. Für $r = 0$ gilt sie in jedem Körper und die ungewohnte Gleichung $-1 = 1$ gilt in $\mathbb{Z}_2$, weil das Element $r = 1$ tatsächlich die Gleichung $r + 1 = 0$ erfüllt. Daher stimmen im Körper $\mathbb{Z}_2$ Addition und Subtraktion überein, was das Leben dort zusätzlich vereinfacht. Durch Einsetzen sämtlicher Elemente aus dem endlichen Körper bestätigt man auch leicht, dass in $\mathbb{Z}_2$ die Gleichung

$$(x + y)^2 = x^2 + y^2$$

gilt, die man einen Schüler lange Zeit vergeblich auszutreiben versucht hat.

Andererseits wird man in endlichen Körper gewisse irrationale Elemente nicht finden. Beispielsweise hat das quadratische Polynom $f(x) = x^2 + x + 1$, dessen Koeffizienten zwar aus $\mathbb{Z}_2 = \{0, 1\}$ stammen und das beim Studium der Fibonacci-Rekursion eine zentrale Rolle spielt, keine Nullstelle in $\mathbb{Z}_2$, wie man durch Ausprobieren der beiden einzigen Möglichkeiten erkennt. ◯

Über endlichen Körpern werden algebraische Gleichungen in der Regel keine oder zu wenig Lösungen haben. Braucht man beispielsweise als Eigenwerte von linearen dynamischen Systemen gewisse Nullstellen von Polynomen, muss man die fehlenden Elemente durch Übergang zu einem geeigneten grösseren Oberkörper $\mathbb{Z}_2 \subset \mathbb{K}_q$ konstruieren. Man findet sogar immer einen gewissen endlichen Oberkörper, in dem es sogar vollständig in Linearfaktoren zerfällt.

Beispiel. Über dem Körper $\mathbb{R}$ der reellen Zahlen hat das quadratische Polynom $f(x) = x^2 + x - 1$ die beiden Nullstellen

$$x_1 = \frac{-1 + \sqrt{5}}{2}, \qquad x_2 = \frac{-1 - \sqrt{5}}{2}$$

die beide irrational sind. Wir werden bald eine endliche Körpererweiterung $\mathbb{Z}_2 \subset \mathbb{K}_4$ konstruieren, in der dieses Polynom zwei verschiedene Nullstellen besitzt. ◯

Allgemeiner definieren wir für eine beliebige Primzahl $p \in \mathbb{Z}$ den *Primkörper* mit p Elementen

$$\mathbb{Z}_p = \{0, 1, \ldots, p-1\}.$$

Addition und Multiplikation in $\mathbb{Z}_p$ sind die Addition und Multiplikation ganzer Zahlen, deren Ergebnis dann modulo p reduziert werden. Wie im Fall $p = 2$ liefert der Rest nach Division durch p die kanonische Projektion

$$\pi_p \colon \mathbb{Z} \to \mathbb{Z}_p, \qquad x \mapsto x \bmod p$$

Diese Projektion ist mit der Addition und der Multiplikation ganzer Zahlen verträglich, weil die beiden Gleichungen

$$(x+y) \bmod p = x \bmod p + y \bmod p, \qquad (x \cdot y) \bmod p = x \bmod p \cdot y \bmod p$$

gelten, die das Rechnen mit Restklassen dominieren.

Beispiel. Im Primkörper $\mathbb{Z}_3 = \{0, 1, 2\}$ rechnet man mit den Operationen

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

x	$-x$
0	0
1	2
2	1

x	x^{-1}
0	–
1	1
2	2

In diesem Körper vereinfacht sich die Binom'sche Formel zur Gleichung

$$(x+y)^3 = x^3 + y^3$$

wie man wiederum leicht kontrollieren kann. ◯

Beispiel. Entsprechend benutzt man im Körper $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ die Operationen

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\cdot$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

x	$-x$
0	0
1	4
2	3
3	2
4	1

x	x^{-1}
0	–
1	1
2	3
3	2
4	4

Im Körper $\mathbb{Z}_5$ gilt für das multiplikative Inverse von 2 die Beziehung $\frac{1}{2} = 3$, weil tatsächlich die Gleichung $3 \cdot 2 = 1$ gilt. ◯

Wir wollen noch ein Beispiel für einen etwas grösseren Modul betrachten, dessen Reste aber einfach mit Hilfe des kleinen Einmaleins und der Wechselsumme bzw. durch Subtraktion der nächst kleiner Schnapszahl berechnet werden können. Dabei werden wir etwas tiefer in die Struktur endlicher Körper vordringen.

Beispiel. Im Primkörper $\mathbb{Z}_{11} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ gelten die Additionsregeln

$+$	0	1	2	3	4	5	6	7	8	9	10
0	0	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10	0
2	2	3	4	5	6	7	8	9	10	0	1
3	3	4	5	6	7	8	9	10	0	1	2
4	4	5	6	7	8	9	10	0	1	2	3
5	5	6	7	8	9	10	0	1	2	3	4
6	6	7	8	9	10	0	1	2	3	4	5
7	7	8	9	10	0	1	2	3	4	5	6
8	8	9	10	0	1	2	3	4	5	6	7
9	9	10	0	1	2	3	4	5	6	7	8
10	10	0	1	2	3	4	5	6	7	8	9

x	$-x$
0	0
1	10
2	9
3	8
4	7
5	6
6	5
7	4
8	3
9	2
10	1

Ein Blick auf diese Tabelle zeigt, das die Struktur der Addition einfach ist weil jeweils die nächste Zeile durch eine Linksverschiebung aus der vorangegangene hervorgeht. Daher ist auch das additive Inverse einfach zu berechnen. Für die Multiplikationsregeln erhalten wir

$\cdot$	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	10
2	0	2	4	6	8	10	1	3	5	7	9
3	0	3	6	9	1	4	7	10	2	5	8
4	0	4	8	1	5	9	2	6	10	3	7
5	0	5	10	4	9	3	8	2	7	1	6
6	0	6	1	7	2	8	3	9	4	10	5
7	0	7	3	10	6	2	9	5	1	8	4
8	0	8	5	2	10	7	4	1	9	6	3
9	0	9	7	5	3	1	10	8	6	4	2
10	0	10	9	8	7	6	5	4	3	2	1

x	x^{-1}
0	–
1	1
2	6
3	4
4	3
5	9
6	2
7	8
8	7
9	5
10	10

Weniger durchsichtig ist in dieser Darstellung die multiplikative Struktur endlicher Körper. Immerhin beobachten wir, dass jedes von 0 verschiedene Körperelement in jeder Zeile und jeder Spalte genau einmal vorkommt. Das hängt damit zusammen, dass diese Elemente die Einheitengruppe $\mathbb{Z}_{11}^{\times}$ bilden. Über diese Gruppe kann man mehr sagen. Um ihre zyklische Struktur etwas zu beleuchten, bilden wir von jedem Element der Einheitengruppe $r \in \mathbb{Z}_{11}^{\times}$ solange seine Potenzen, bis das multiplikative Neutralelement 1 herauskommt. Bezeichnen wir für ein Element $r \in \mathbb{Z}_{11}^{\times}$ den kleinste Exponent $k \geq 1$, für den $r^k = 1$ ist, als *Ordnung* $\mathrm{ord}_{11}(r)$ des betreffenden Elementes, erhalten wir die Werte

folgender Tabelle:

r	r^2	r^3	r^4	r^5	r^6	r^7	r^8	r^9	r^{10}	$\mathrm{ord}_{11}(r)$
1										1
2	4	8	5	10	9	7	3	6	1	10
3	9	5	4	1						5
4	5	9	3	1						5
5	3	4	9	1						5
6	3	7	9	10	5	8	4	2	1	10
7	5	2	3	10	4	6	9	8	1	10
8	9	6	4	10	3	2	5	7	1	10
9	4	3	5	1						5
10	1									2

Ein Blick auf diese Tabelle lässt vermuten, dass die Ordnung jedes Elemente ein Teiler von $p-1=10$ ist. Diese Exponententeilbarkeitseigenschaft lässt sich noch verfeinern. Bestimmen wir für jeden Teiler d von $p-1$ die Anzahl $\varphi(d)$ natürlicher Zahlen, die kleiner als d aber zu d teilerfremd sind, erhalten wir die Werte

d	1	2	5	10
$\varphi(d)$	1	1	4	4

Damit lässt sich vermuten, dass die Anzahl Elemente in $\mathbb{Z}_p$ der Ordnung d gerade $\varphi(d)$ beträgt, falls d ein Teiler von $p-1$ ist.

Daraus folgt für den Teiler $d=p-1$ insbesondere, dass es in $\mathbb{Z}_p$ insgesamt $\varphi(p-1)$ Elemente der maximalen Ordnung $p-1$ gibt. Weil $\varphi(p-1)\geq 1$ ist gibt im endlichen Körper $\mathbb{Z}_p$ immer mindestens ein Element der maximalen Ordnung $p-1$. Solche Elemente maximaler Ordnung heissen Generatoren, erzeugende oder primitive Elemente. Im Körper $\mathbb{Z}_{11}$ sind also die $\varphi(10)=4$ fett gesetzten Elemente $2,6,7,8$ primitiv, weil sie die maximal Ordung 10 haben.

Ist α einer dieser Generatoren, so kann also jedes von Null verschieden Körperelement $r\in\mathbb{Z}_p$ auf eindeutige Art in der Potenzform $r=\alpha^n$ geschrieben werden. Es ist also

$$\mathbb{Z}_p=\{0,\alpha,\alpha^2,\alpha^3,\ldots,\alpha^{p-1}\}$$

Den Exponenten $n=\log_\alpha(r)\in\mathbb{Z}_{p-1}$ bezeichnet man als *diskreten Logarithmus* von r zur Basis α im Körper $\mathbb{Z}_p$. Der Logarithmus $\log_\alpha(0)$ ist nicht definiert.

Beispielsweise hat $\mathbb{Z}_{11}$ den Generator $\alpha=2$ und obige Tabelle liefert die Potenzen

n	1	2	3	4	5	6	7	8	9	10
α^n	2	4	8	5	10	9	7	3	6	1

Daraus erhalten wir durch Invertieren die Logarithmentabelle für $\alpha\in\mathbb{Z}_p$.

r	1	2	3	4	5	6	7	8	9	10
$\log_\alpha(r)$	10	1	8	2	4	9	7	3	6	5

Wie in der üblichen Arithmetik vereinfacht der diskrete Logarithmus das Rechnen im Körper $\mathbb{Z}_p$, weil er eine Multiplikation in eine Addition überführt. Auf Grund des Exponentialgesetzes gelten nämlich in $\mathbb{Z}_{p-1}$ die üblichen Logarithmengesetze

$$\log_\alpha(r_1\cdot r_2)=\log_\alpha(r_1)+\log_\alpha(r_2),\qquad \log_\alpha(r^k)=k\cdot\log_\alpha(r)$$

wobei sorgfältig auf den Modul $p-1=10$ zu achten ist! Weil der Logarithmus nicht additiv ist, ist unklar, welcher Potenz α^n die Summe zweier Potenzen $\alpha^i+\alpha^j$ entspricht. Für den diskreten Logarithmus $\log_\alpha(r)$ gilt also eine fundamentale Asymmetrie bezüglich Addition und Multiplikation. Potenzen r^n eines Körperelementes können mit dem Verdoppelungsverfahren effizient berechnet werden. Hingegen scheint die Berechnung des diskreten Logarithmus schwierig zu sein. Diese Asymmetrie in der Komplexität nutzt man in der Kryptologie systematisch aus.

Eine weitere praktische Schwierigkeit im Umgang mit Generatoren der Einheitengruppe endlicher Körper besteht darin, dass unklar ist, wie man in der Einheitengruppe $\mathbb{K}_q$ eines endlichen Körpers einen Generator α findet. Für kleine q kann man einfach die in Frage kommenden Werte durchprobieren. Im allgemeinen versteht man die Frage, welche Elemente eines endlichen Körpers als Generatoren in Frage kommen, nicht gut. Das Distributivgesetz fügt die additive und die multiplikative Struktur eines endlichen Körpers auf subtile Art zusammen. ◯

Man beachte, dass der Modul p tatsächlich eine Primzahl sein muss, damit $\mathbb{Z}_p$ ein Körper sein kann. Für eine beliebige natürliche Zahl n wird $\mathbb{Z}_n$ zwar die Struktur eines kommutativen Rings haben und es gibt eine kanonische Projektion

$$\pi_n\colon \mathbb{Z}\to\mathbb{Z}_n, \qquad x\mapsto x \bmod n$$

die mit der Addition und der Multiplikation verträglich (Ringhomomorphismus) ist. In der Regel werden wir allerdings nicht garantieren können, dass jedes von 0 verschiedene Element in $\mathbb{Z}_n$ ein multiplikatives Inverses hat.

Beispiel. Im Ring $\mathbb{Z}_4=\{0,1,2,3\}$ gelten die Operationen

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	4
2	2	3	4	0
3	3	4	0	1

$\cdot$	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

x	$-x$
0	0
1	3
2	2
3	1

x	x^{-1}
0	–
1	1
2	–
3	3

Aus der Gleichung $2\cdot 2=0$ entnehmen wir, dass das Element $2\in\mathbb{Z}_4$ ein Nullteiler ist und dieser Ring daher sicher kein Körper sein kann, weil das Element 2 kein multiplikatives Inverses haben kann. ◯

Beispiel. Entsprechend existieren im Ring $\mathbb{Z}_6=\{0,1,2,3,4,5\}$ mit den Operationen

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

x	$-x$
0	0
1	5
2	4
3	3
4	2
5	1

x	x^{-1}
0	–
1	1
2	–
3	–
4	–
5	5

Nullteiler, weil beispielsweise die Gleichung $2\cdot 3=0$ gilt, aus der folgt, dass 2 und 3 in $\mathbb{Z}_6$ keine multiplikativen Inverse haben. Entsprechendens gilt für $4\in\mathbb{Z}_6$

wegen der Gleichung $3 \cdot 4 = 0$. Offenbar haben nur gerade diejenigen Elemente $k \in \mathbb{Z}_6$ ein multiplikatives Inverses, für die $\mathrm{ggT}(k, 6) = 1$ gilt. ◯

Die einzige Art, wie der Ring $\mathbb{Z}_n$ keinen Nullteiler haben kann, besteht darin, dass sich der Modul n nicht auf nichttriviale Art in ein Produkt $n = k \cdot l$ zerlegen kann. Daher muss n eine Primzahl sein, damit es sich tatsächlich um einen Körper handelt.

Neben den angetroffenen Primkörpern $\mathbb{Z}_p$ gibt es noch weitere endliche Körper $\mathbb{K}_q$, die in der Praxis deshalb benötigt werden, weil sich über den Primkörpern nicht alle algebraischen Gleichungen lösen lassen. Die Konstruktion eines endlichen Körpers $\mathbb{K}_q$ mit q Elementen ist allerdings etwas anspruchsvoller. Bevor wir uns damit befassen, überlegen wir uns, dass ein solcher Köper eng mit einem gewissen Primköper verknüpft ist und daher seine additive Struktur einfach zu durchschauen ist und für die Anzahl Elemente q eines solchen Körper eine gewisse Einschränkung gilt.

Zur Formulierung erinnern wir an einen Begriff aus der Ringtheorie. In einem kommutativen Ring $(R, +, \cdot, 0_R, 1_R)$ bilden die ganzzahligen Vielfachen des Einselementes 1_R einen Teilring, der zu $\mathbb{Z}$ oder zu $\mathbb{Z}_n$ isomorph ist.

Definition. Für einen kommutativen $(R, +, \cdot, -, 0_R, 1_R)$ verstehen wir unter seiner *Charakteristik* die kleinste ganzzahlige Anzahl Vielfache von $1_R \in R$, die 0_R ergibt. Es ist also

$$n1_R = \underbrace{1_R + 1_R + \cdots + 1_R}_{n} = 0_R$$

Falls keine solche ganze Zahl n existiert, hat der Ring die Charakteristik 0.

Wir können diesen Begriff konzeptioneller formulieren, indem wir beobachten, dass wegen der Freiheit von $\mathbb{Z}$ ein kanonischer Ring-Homomorphismus

$$\varphi\colon \mathbb{Z} \to R, \qquad 1 \mapsto 1_R$$

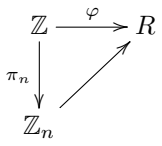
existiert. Für jede natürliche Zahl $n \in \mathbb{Z}$ gilt also

$$\varphi(n) = \underbrace{1_R + 1_R + \cdots + 1_R}_{n}$$

Für eine negative ganze Zahl n ist $\varphi(n) = -\varphi(-n)$. Die Homomorphieeigenschaften $\varphi(m + n) = \varphi(m) + \varphi(n)$ und $\varphi(m \cdot n) = \varphi(m) \cdot \varphi(n)$ und $\varphi(0) = 0_R$ sind mit Hilfe der Axiome eines kommutativen Rings leicht zu überprüfen. Es gibt nun zwei Fälle:

1. Falls φ injektiv ist, so ist das Bild $\varphi(\mathbb{Z})$ isomorph zu $\mathbb{Z}$. In diesem Fall sagt man, dass der Ring R die *Charakteristik* $\mathrm{char}(R) = 0$ hat.

2. Falls φ nicht injektiv ist, so ist sein Kern $\mathrm{Ker}(\varphi) = (n) \subseteq \mathbb{Z}$ ein nicht triviales Ideal und wir erhalten eine Faktorisierung



Dann ist das Bild $\varphi(\mathbb{Z})$ isomorph zum endlichen Ring $\mathbb{Z}_n$. In diesem Fall sagt man, dass der Ring R die *Charakteristik* $\text{char}(R) = n$ hat.

Falls R ein Integritätsbereich ist, dessen Charakteristik nicht 0 ist, so muss seine Charakteristik n eine Primzahl p sein. Gäbe es nämlich eine Faktorisierung $n = p \cdot q$ mit $1 < p, q < n$, so wäre $\varphi(p) \neq 0$ und $\varphi(q) \neq 0$, aber $\varphi(p) \cdot \varphi(q) = \varphi(p \cdot q) = \varphi(n) = 0$. Daher hätte der Ring R Nullteiler. Einen solchen Integritätsbereich R mit der Charakteristik $p \geq 2$ können wir also immer als Erweiterung

$$\mathbb{Z}_p \subseteq R$$

über seinem Primkörper $\mathbb{Z}_p$ mit $p = \text{char}(R)$ auffassen. Insbesondere sind die Addition und die Multiplikation in $\mathbb{Z}_p$ gerade die Ringoperationen, die in R erklärt sind. Betrachtet man nun nur die Produkte λr, wobei $\lambda \in \mathbb{Z}_p$ und $r \in R$ ist, so folgen aus den Rechengesetzen im Ring R die Eigenschaften

1. $(R, +, -, 0_R)$ ist eine kommutative Gruppe.
2. $\lambda(r_1 + r_2) = \lambda r_1 + \lambda_2 r_2$.
3. $(\lambda + \mu)r + \lambda r + \mu r$.
4. $\lambda(\mu r) = (\lambda\mu)r$.
5. $1r = r$.

Sie besagen gerade, dass R eine Vektorraum über dem Primköper $\mathbb{Z}_p$ ist.

Körper sind Integritätsbereiche und daher ist die Charakteristik eines Körpers 0 oder eine Primzahl p. Falls nun der Körper $\mathbb{K}_q$ endlich viele Elemente hat, kommt der erste Fall nicht in Frage und seine Charakteristik muss eine Primzahl p sein. Einen solchen Körper können wir also als Erweiterung

$$\mathbb{Z}_p \subseteq \mathbb{K}_q$$

über einem gewissen Primkörper $\mathbb{Z}_p$ mit $p = \text{char}(\mathbb{K}_q)$ auffassen. Dieser Primkörper ist der kleinste Teilkörper von $\mathbb{K}_q$.

Wir werden also den endlichen Körper $\mathbb{K}_q$ konstruieren, indem wir von seinem eindeutig bestimmten kleinsten Teilkörper d.h. von der Körpererweiterung $\mathbb{Z}_p \subseteq \mathbb{K}_q$ ausgehen. Im Fall eines endlichen Körpers $\mathbb{K}_q$ kann dieser Körper als Vektorraum über seinem Primkörper $\mathbb{Z}_p$ aufgefasst werden. Dieser Vektorraum hat dann q viele — also insbesondere endlich viele — Elemente und ist isomorph zum Vektorraum $\mathbb{Z}_p^d$ für eine gewisse gazzahlige Dimension $d = \dim(\mathbb{K}_q)$. Weil dieser Vektorraum insgesamt p^d viele Elemente hat, muss $q = p^d$, d.h. eine Primzahlpotenz sein. Über die Struktur endlicher Köper erhalten wir also folgendes fundamentales Resultat.

Satz. Die Anzahl Elemente eines endlichen Körpers $\mathbb{K}_q$ hat die Form $q = p^d$, wobei die *Charakteristik* p eine beliebige Primzahl und der Erweiterungsgrad $d \geq 1$ eine strikt positive natürliche Zahl ist.

Beispiel. Es kann keinen Körper mit $q = 6$ Elementen geben. Im Gegensatz dazu könnten Körper mit $q = 2^2 = 4$, mit $q = 2^3 = 8$ oder mit $q = 2^4 = 16$ Elementen existieren. Wir werden solche Körper als Vektorräume $\mathbb{Z}_2^2$, $\mathbb{Z}_2^3$, oder

als $\mathbb{Z}_2^4$ d.h. als Erweiterungen vom Grad $d = 2$, $d = 3$ oder $d = 4$ über dem Primköprer $\mathbb{Z}_2$ auch wirklich konstruieren und daran die Theorie der endlichen Körper exemplarisch vorführen. ◯

Weil der endliche Körper $\mathbb{K}_q$ ein Vektorraum der Dimension d über seinem Primköper $\mathbb{Z}_p$ ist, können wir seine Elemente mit den Vektoren aus $\mathbb{Z}_p^d$ identifizieren. Um die in Frage kommenden Körper zu konstruieren, brauchen wir zusätzlich ein Verfahren um die Multiplikation solcher Vektoren zu beschreiben. Das gelingt mit der Beobachtung, dass die Einheitengruppe $\mathbb{K}_q^\times$ eines solchen Körpers $q-1$ Elemente hat und alle diese Einheiten die Gleichung $x^{q-1} = 1$ erfüllen müssen. Zusammen mit dem Element 0 müssen also die q Elemente des endlichen Körpers $\mathbb{K}_q$ die Nullstellen des Polynoms $f(x) = x^q - x$ sein. Weil dieses Polynom vom Grad q höchstens q verschiedene Nullstellen hat, muss es über $\mathbb{K}_q$ in der Form

$$f(x) = x^q - x = (x - r_1) \cdot (x - r_2) \cdots (x - r_q) = \prod_{r_j \in \mathbb{K}_q} (x - r_j), \quad r_j \in \mathbb{K}_q$$

in Linearfaktoren zerfallen. Wir werden allerdings den Körper $\mathbb{K}_q$ nicht, wie in der Literatur üblich, als Zerfällungskörper des Polynoms $f(x) = x^q - x$ über dem Primkörper $\mathbb{Z}_p$ beschreiben, weil dieser Zugang Verständnis für Gruppentheorie, Äquivalenzrelationen und Quotientenringe erfordert und deshalb für die meisten Anfänger zu abstrakt ist. Wir werden versuchen, mit elementarer Matrizenrechnung auszukommen, die heute auch in den Schulsack jedes hartgesottenen Theorienmuffels gehören. Dabei wird sich direkt eine Beschreibung der endlichen Körper ergeben, die sich leicht durch ein Computerprogramm implementieren lässt.

Um eine Idee zu erhalten, wie vorzugehen ist, erinnern wir uns an die Konstruktion des Körpers $\mathbb{C}$ der komplexen Zahlen aus dem Körper $\mathbb{R}$ der reellen Zahlen und benutzen dann in unserer Situation ein analoges Vorgehen. Damals hatte die quadratische Gleichung $x^2 = -1$ keine Lösung bzw. das reelle Polynom $f(x) = -1 - x^2 \in \mathbb{R}[x]$ vom Grad $d = 2$ keine Nullstelle. Wir haben dann eine solche Nullstelle konstruiert, indem wir seine Begleitermatrix

$$\Omega = -I = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathbb{R}^{2,2}$$

verwendet haben. Sie erfüllt die charakteristische Gleichung

$$f(\Omega) = -E - \Omega^2 = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0$$

vom Grad $d = 2$. Sie drückt aus, wie sich die d-te Potenz von Ω als reelle Linearkombination der echt kleineren Potenzen kombinieren lässt. Im vorliegenden Fall ist

$$\Omega^2 = -E$$

Bilden wir schliesslich alle reellen Linearkombinationen dieser kleineren Potenzen, d.h. die Ausdrücke der Form

$$a_0 E + a_1 \Omega = \begin{pmatrix} a_0 & a_1 \\ -a_1 & a_0 \end{pmatrix}, \qquad a_0, a_1 \in \mathbb{R}$$

erhalten wir sämtliche komplexen Zahlen. Die Addition komplexer Zahlen ist komponentenweise

$$(a_0E + a_1\Omega) + (b_0 + b_1\Omega) = (a_0 + b_0)E + (a_1 + b_1)\Omega$$

definiert und für die Multiplikation erfüllt die magische Formel

$$(a_0E + a_1\Omega) \cdot (b_0 + b_1\Omega) = (a_0b_0 - a_1b_1)E + (a_0b_1 + a_1b_0)\Omega$$

Kürzen wir die Linearkombination $a_0E + a_1\Omega$ durch ihre Koeffizienten (a_0, a_1) ab, erhalten wir den Körper der komplexen Zahlen

$$\mathbb{C} = \{(a_0, a_1) \mid a_0, a_1 \in \mathbb{R}\}$$

in der üblichen Darstellung mit Hilfe von Vektoren, die sich als Punkte in der Ebene $\mathbb{R}^2$ geometrisch interpretieren lassen. Die Addition in $\mathbb{C}$ entspricht dann der Vektoraddition. Die Multiplikation komplexer Zahlen entspricht dann der durch die magischen Formeln beschriebenen Abbildung

$$\mu\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \qquad \big((a_0, a_1), (b_0, b_1)\big) \mapsto (c_0, c_1)$$

deren Komponenten gegeben sind durch

$$c_0 = a_0b_0 - a_1b_1, \qquad c_1 = a_0b_1 + a_1b_0$$

In beiden Fällen handelt es sich um Bilinearformen, die sich man sich zweckmässigerweise in matrizieller Form merkt. Mit Hilfe der Vektoren

$$\vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \qquad \vec{c} = \begin{pmatrix} c_0 \\ c_1 \end{pmatrix}$$

lässt sich nämlich der Produktvektor

$$\mu\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \qquad (\vec{a}, \vec{b}) \mapsto \mu(\vec{a}, \vec{b}) = \vec{c}$$

als Matrizenprodukt

$$\mu(\vec{a}, \vec{b}) = \vec{c} = \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} a_0 & -a_1 \\ a_1 & a_0 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = f(\vec{a}) \cdot \vec{b}$$

ausdrücken. Die Multiplikationstabelle der komplexen Zahlen steckt also in der Matrizenmultiplikation mit der quadratischen Matrix

$$f(\vec{a}) = \begin{pmatrix} a_0 & -a_1 \\ a_1 & a_0 \end{pmatrix} = a_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = a_0A_0 + a_1A_1,$$

die sich wiederum mit den beiden Matrizen des Systems $\{A_0, A_1\}$ linear kombinieren lässt, wobei wir die beiden Matrizen

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

definiert haben. Für die beiden Komponenten des Produktvektors gilt also

$$c_0 = \langle \vec{e}_1, f(\vec{a}) \cdot \vec{b} \rangle = \langle f(\vec{a})^T \cdot \vec{e}_1, \vec{b} \rangle, \qquad c_1 = \langle \vec{e}_2, f(\vec{a}) \cdot \vec{b} \rangle = \langle f(\vec{a})^T \cdot \vec{e}_2, \vec{b} \rangle$$

Beschreiben wir die Spalten der Matrix $f(\vec{a})^T$ ebenfalls als Matrizenprodukte

$$f(\vec{a})^T \cdot \vec{e}_1 = S_1^T \cdot \vec{a}, \qquad f(\vec{a})^T \cdot \vec{e}_2 = S_2^T \cdot \vec{a}$$

mit den beiden sogn. *Strukturmatrizen*

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

erhalten wir für die Komponenten des Produktvektors

$$\begin{aligned} \mu(\vec{a},\vec{b})_1 = c_0 &= \langle f(\vec{a})^T \cdot \vec{e}_1, \vec{b}\rangle = \langle S_1^T \cdot \vec{a}, \vec{b}\rangle = \langle \vec{a}, S_1 \cdot \vec{b}\rangle \\ \mu(\vec{a},\vec{b})_2 = c_1 &= \langle f(\vec{a})^T \cdot \vec{e}_2, \vec{b}\rangle = \langle S_2^T \cdot \vec{a}, \vec{b}\rangle = \langle \vec{a}, S_2 \cdot \vec{b}\rangle \end{aligned}$$

Die Elemente der Strukturmatrizen, die das Produkt ebenfalls vollständig beschreiben, sind durch die Produkte der Standardbasisvektoren

$$(S_1)_{ij} = \langle \vec{e}_i, S_1 \cdot \vec{e}_j\rangle = \mu(\vec{e}_i,\vec{e}_j)_1, \qquad (S_2)_{ij} = \langle \vec{e}_i, S_2 \cdot \vec{e}_j\rangle = \mu(\vec{e}_i,\vec{e}_j)_2$$

definiert und die Produkte der Standardbasisvektoren sind als Linearkombinationen

$$\mu(\vec{e}_i,\vec{e}_j) = \sum_{k=1}^{2} \mu(\vec{e}_i,\vec{e}_j)_k \vec{e}_k = \sum_{k=1}^{2} (S_k)_{ij} \vec{e}_k$$

zerlegbar, wobei die $2^3 = 8$ sogn. *Strukturkonstanten* $\mu(\vec{e}_i,\vec{e}_j)_k = (S_k)_{ij}$ die Koeffizienten der $n = 2$ Strukturmatrizen S_1 und S_2 sind.

Die Darstellung der komplexen Zahlen mit Hilfe reeller Matrizen übertragen wir nun sinngemäss von $\mathbb{R}$ auf den Primkörper $\mathbb{Z}_p$, um damit die endlichen Körper der Charakteristik p zu konstruieren.

Zur Konstruktion eines Körpers $\mathbb{K}_q$ mit einer Primzahlpotenz $q = p^d$ Elementen gehen wir vom Primkörper $\mathbb{Z}_p$ aus und adjungieren dann die Nullstellen eines geschickt gewählten Polynoms $f(x) \in \mathbb{Z}_p[x]$ vom minimalen Grad d. Dieses Polynom wählen wir so, dass es in $\mathbb{Z}_p$ keine Nullstellen hat, aber in einem grösseren Körper $\mathbb{Z}_p \subseteq \mathbb{K}_q$ insgesamt d verschieden Nullstellen haben sollte. Konkret gehen wir wie folgt vor:

1. Wähle ein normiertes, irreduzibles[44] Polynom

   $$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_{d-1} x^{d-1} + x^d \in \mathbb{Z}_p[x]$$

   vom Grad d.

2. Konstruiere eine Nullstelle dieses Polynoms als Begleitermatrix

   $$\Omega \in \mathbb{Z}_p^{d,d}$$

[44] Ein Polynom $f(x) \in \mathbb{Z}_p[x]$ von echt positivem Grad heisst *reduzibel*, falls zwei Polynome $f_1(x)$ und $f_2(x)$ von echt positivem Grad existieren, so dass

$$f(x) = f_1(x) \cdot f_2(x)$$

gilt. Falls keine solche Gleichung gilt, heisst $f(x)$ irreduzibel. Irreduzible Polynome entsprechen also im Polynomring $\mathbb{Z}_p[x]$ den Primelementen.

Sie erfüllt die charakteristische Gleichung $f(\Omega) = 0$ aus der folgt, wie sich die Potenz Ω^d als Linearkombination

$$\Omega^d = -\big(c_0 E + c_1\Omega + c_2\Omega^2 + \cdots + c_{d-1}\Omega^{d-1}\big)$$

der kleineren Potenzen ausdrücken lässt.

3. Die Menge der Linearkobinationen

$$a_0 E + a_1\Omega + a_2\Omega^2 + \cdots + a_{d-1}\Omega^{d-1}, \qquad a_0, a_1, \ldots, a_{d-1} \in \mathbb{Z}_p$$

bilden die $p^d = q$ Elemente der gesuchte Körpererweiterung $\mathbb{Z}_p \subseteq \mathbb{K}_q$. Sie lassen sich auch als Vektoren

$$\mathbb{K}_q = \{(a_0, a_1, \ldots, a_{d-1}) | a_j \in \mathbb{Z}_p\} = \mathbb{Z}_p^d$$

beschreiben. Die Additon im Körper $\mathbb{K}_q$ entspricht der Vektoraddition.

4. Das Multiplikationgesetz solcher Vektoren beschreiben wir mit Hilfe von Strukturmatrizen $S_1, \ldots, S_d \in \mathbb{Z}_p^{d,d}$ bzw. mit den zugehörigen d^3 Strukturkonstanten aus $\mathbb{Z}_p$.

Beispiel. Den Körper $\mathbb{K}_4$ mit $4 = 2^2$ Elementen konstruiert man, indem ein normiertes, irreduzibles Polynom $f(x) \in \mathbb{Z}_2[x]$ vom Grad $d = 2$ bestimmt. Im vorliegenden Fall können wir das normierte Polynom $f(x) = 1 + x + x^2 \in \mathbb{Z}_2[x]$ benutzen. Um zu zeigen, dass es wirklich irreduzibel ist, müssen wir untersuchen, dass in $\mathbb{Z}_2[x]$ keine Faktorisierung der Art

$$1 + x + x^2 = (x - a) \cdot f_2(x)$$

in Polynome echt kleinern Grades existiert. Gäbe es aber eine solche Faktorisierung, so müsste das Polynom $f(x)$ in $\mathbb{Z}_2$ eine Nullstelle haben, was nicht der Fall ist, wie man durch Ausprobieren der beiden Elemente leicht einsieht. Die restlichen drei normierten Polynome aus $\mathbb{Z}_2[x]$ kommen zur Konstruktion des endlichen Körpers $\mathbb{K}_4$ nicht in Frage, da sie alle reduzibel sind, wie die Faktorisierungen

$$x^2 = x \cdot x, \qquad 1 + x^2 = (1 + x)(1 + x), \qquad x + x^2 = x(x + 1)$$

zeigen. Sie haben tatsächlich alle eine Nullstelle in $\mathbb{Z}_2$.

Nun benötigen wir eine Nullstelle dieses irreduziblen Polynoms. Dazu können wir seine Begleitermatrix

$$\Omega = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{Z}_2^{2,2}$$

benutzen. Sie erfüllt tatsächlich ihre charakteristische Gleichung

$$f(\Omega) = E + \Omega + \Omega^2 = 0, \qquad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Der gesuchte Körper

$$\mathbb{K}_4 = \{a_0 E + a_1\Omega \mid a_0, a_1 \in \mathbb{Z}_2\}$$

wird also durch die $\mathbb{Z}_2$-Linearkombinationen d.h. durch die vier Matrizen

$$0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad E + \Omega = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

gebildet. Als Addition und Multiplikation in diesem Körper benutzen wir die üblichen Operationen in $\mathbb{Z}_2^{2,2}$. Die Arithmetik im endlichen Körper $\mathbb{K}_4$ ist damit auf die Arithmetik der Matrizen mit Elementen im Primkörper $\mathbb{Z}_2$ reduziert. Im vorliegenden Fall erhalten wir die folgenden Additionstabelle

$+$	0	E	Ω	$E+\Omega$
0	0	E	Ω	$E+\Omega$
E	E	0	$E+\Omega$	Ω
Ω	Ω	$E+\Omega$	0	E
$E+\Omega$	$E+\Omega$	Ω	E	0

x	$-x$
0	0
E	E
Ω	Ω
$E+\Omega$	$E+\Omega$

Um sich diese Additionstabelle einfach zu merken, kann man jedes Körperelement — wie seinerzeit bei den komplexen Zahlen — statt als Linearkombination $a_0E + a_1\Omega$ von Matrizen kompakter durch den Vektor seiner Komponenten $(a_0, a_1) \in \mathbb{Z}_2^2$ beschreiben und erhält damit folgende äquivalente Vektordarstellung der Körperelemente

Matrix	Vektor
0	$(0,0)$
E	$(1,0)$
Ω	$(0,1)$
$E+\Omega$	$(1,1)$

In der Vektorform entspricht die Addition im Körper $\mathbb{K}_4$ der komponentenweisen Vektoraddition in $\mathbb{Z}_2^2$. Beispielsweise ist $(1,1) + (1,0) = (0,1)$.

Für die Multiplikationstabelle erhalten wir durch Matrizenmultiplikation in $\mathbb{Z}_2^{2,2}$ analog

$\cdot$	0	E	Ω	$E+\Omega$
0	0	0	0	0
E	0	E	Ω	$E+\Omega$
Ω	0	Ω	$E+\Omega$	E
$E+\Omega$	0	$E+\Omega$	E	Ω

x	x^{-1}
0	$-$
E	E
Ω	$E+\Omega$
$E+\Omega$	Ω

Um diese Tabelle zu erhalten, kann man auch ohne Matrizenrechnung auskommen und merkt sich dazu die charakteristische Gleichung

$$\Omega^2 = E + \Omega$$

von Ω und reduziert damit höhere Potenzen von Ω. Im vorliegenden Fall ist

$$\Omega^3 = \Omega \cdot \Omega^2 = \Omega \cdot (E + \Omega) = \Omega + \Omega^2 = \Omega + (E + \Omega) = E$$

Die multiplikativen Inversen lassen sich in dieser Matrizendarstellung des Körpers $\mathbb{K}_4$ leicht durch Invertieren der repräsentierenden Matrizen bestimmen. Beispielsweise wird das multiplikative Inverse von

$$\Omega = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{Z}_2^{2,2}$$

durch ihre inverse Matrix

$$\Omega^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = E + \Omega \in \mathbb{Z}_2^{2,2}$$

gegeben, weil $\Omega \cdot (E + \Omega) = \Omega + \Omega^2 = \Omega + (E + \Omega) = E$ gilt.
Die Multiplikation zweier Körperelemente aus $\mathbb{K}_4$ erfüllt also die Bedingung

$$(a_0 E + a_1 \Omega) \cdot (b_0 E + b_1 \Omega) = (a_0 b_0 + a_1 b_1) E + (a_0 b_1 + a_1 b_0 + a_1 b_1) \Omega$$

Wie bei den komplexen Zahlen lässt sich auch dieses Multiplikationsgesetz

$$\mu \colon \mathbb{Z}_2^2 \times \mathbb{Z}_2^2 \to \mathbb{Z}_2^2, \qquad (\vec{a}, \vec{b}) \mapsto \mu(\vec{a}, \vec{b}) = \vec{c}$$

mit den magischen Komponenten

$$c_0 = a_0 b_0 + a_1 b_1, \qquad c_1 = a_0 b_1 + a_1 b_0 + a_1 b_1$$

durch Multiplikation $\mu(\vec{a}, \vec{b}) = \vec{c} = f(\vec{a}) \cdot \vec{b}$ mit der quadratischen Matrix

$$f(\vec{a}) = \begin{pmatrix} a_0 & a_1 \\ a_1 & a_0 + a_1 \end{pmatrix} = a_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = a_0 A_0 + a_1 A_1$$

beschreiben, die sich mit Hilfe der Matrizen des Systems $\{A_0, A_1\}$ linear kombinieren lässt, wobei die beiden Matrizen durch

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

definiert sind. Die Produkte der Standardbasisvektoren lassen sich als Linearkombination

$$\mu(\vec{e}_i, \vec{e}_j) = \sum_{k=1}^{2} \mu(\vec{e}_i, \vec{e}_j)_k \vec{e}_k = \sum_{k=1}^{2} (S_k)_{ij} \vec{e}_k$$

ausdrücken. Das Multiplkationsgesetz liefert die Werte in folgender Tabelle

$$\begin{array}{ll} \mu(\vec{e}_1, \vec{e}_1) = (1,0) & \mu(\vec{e}_1, \vec{e}_2) = (0,1) \\ \mu(\vec{e}_2, \vec{e}_1) = (0,1) & \mu(\vec{e}_2, \vec{e}_2) = (1,1) \end{array}$$

die in der Matrixdarstellung der Tabelle

$$\begin{array}{ll} \mu(\vec{e}_1, \vec{e}_1) = E \cdot E = E & \mu(\vec{e}_1, \vec{e}_2) = E \cdot \Omega = \Omega \\ \mu(\vec{e}_2, \vec{e}_1) = \Omega \cdot E = \Omega & \mu(\vec{e}_2, \vec{e}_2) = \Omega \cdot \Omega = E + \Omega \end{array}$$

entspricht, die man selbstverständlich auch direkt durch Reduktion der Potenzen von Ω mit Hilfe der charakteristischen Gleichung $\Omega^2 = E + \Omega$ erhalten kann.
Die Information dieser Tabellen kodiert man zweckmässigerweise mit Hilfe der $2^3 = 8$ Strukturkonstanten $\mu(\vec{e}_i, \vec{e}_j)_k = (S_k)_{ij}$, die als Elemente in den Strukturmatrizen

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

auftreten. Damit ist die Arithmetik im Körper $\mathbb{K}_4$ vollständig auf die Arithmetik in $\mathbb{Z}_2$ zurückgeführt.

Wie bei den komplexen Zahlen, wo das Rechnen mit Hilfe der Darstellung durch Linearkombinationen d.h. in Normalform mühsam wird, sobald man multipliziert, ist die Multiplikation endlicher Körper in der Darstellung durch Linearkombinationen bzw. in der Vektordarstellung nicht einfach zu durchschauen.

Deshalb sucht man sich zweckmässigerweise hier, wie dort mit der Exponentialform $z = re^{i\varphi}$, eine exponentielle Darstellungsart, weil dann die Multplikation der Körperelemente zur Addition der Exponenten wird. Im Fall unseres endlichen Körper beobachtet man überrascht, dass sich jedes von Null verschiedene Körperelement von $\mathbb{K}_4$ als Potenz des *primitiven Elementes*

$$\alpha = \Omega = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{Z}_2^{2,2}$$

schreiben lässt. Bildet man nämlich seine Potenzen, erhält man die Matrizen

$$\alpha^2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = E + \Omega, \qquad \alpha^3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = E \in \mathbb{Z}_2^{2,2}$$

Bezeichnet man die Einheitsmatrix kurz mit $1 = \alpha^0$ und mit 0 die Nullmatrix, haben die vier Matrizen des Körpers

$$\mathbb{K}_4 = \{0, 1, \alpha, \alpha^2\}$$

die folgende äquivalente Exponentialform

Matrix	Vektor	Potenz	Ordnung
0	$(0,0)$	0	$-$
E	$(1,0)$	1	1
Ω	$(0,1)$	α	3
$E+\Omega$	$(1,1)$	α^2	3

Der kleinste Exponent $k \geq 1$, für den $x^k = E$ ist, wird als *Ordnung* des betreffenden Elementes $x \in \mathbb{K}_4$ bezeichnet. Insbesondere ist die Ordnung $\text{ord}(\alpha) = 3$ maximal und daher α ein primitives Element (Generator) im Körper $\mathbb{K}_4$.

In der Exponentialform nehmen die Verknüpfungstabellen folgende Form

$+$	0	1	α	α^2
0	0	1	α	α^2
1	1	0	α^2	α
α	α	α^2	0	1
α^2	α^2	α	1	0

$\cdot$	0	1	α	α^2
0	0	0	0	0
1	0	1	α	α^2
α	0	α	α^2	1
α^2	0	α^2	1	α

an. Aus ihnen erkennt man sofort, dass der endliche Körper $\mathbb{K}_4$ den Primkörper $\mathbb{Z}_2 = \{0, 1\}$ enthält, von dem wir ausgegangen sind.

Die Multiplikationstabelle ist nun durch das Potenzgesetz

$$\alpha^i \cdot \alpha^j = \alpha^{i+j}, \qquad \alpha^3 = 1$$

einfach zu merken, wobei der Exponent modulo 3 zu reduzieren ist. Dafür wird in dieser exponentiellen Darstellung das Addieren undurchschaubarer.

Weil jedes von Null verschieden Körperelement $r \in \mathbb{K}_4$ auf eindeutige Art in der Exponentialform $r = \alpha^n$ geschrieben werden kann, bezeichnet man die natürliche Zahl n als *diskreten Logarithmus* von r zur Basis α im Körper $\mathbb{K}_4$. Für den Generator $\alpha \in \mathbb{K}_4$ liefert obige Tabelle die Potenzen

n	1	2	3
α^n	α	α^2	1

und daher die Logarithmentabelle

r	1	α	α^2
$\log_\alpha(r)$	3	1	2

Man beachte, dass auch im Körper $\mathbb{K}_4$ der Generator α nicht eindeutig bestimmt ist. Man hätte statt $\alpha = \Omega$ auch das Körperelement $\beta = E + \Omega$ als primitives Element wählen können. Wegen

$$\beta^2 = \Omega, \qquad \beta^3 = 1$$

ist seine Ordnung ebenfalls maximal. Weil die Gruppe der Einheiten $\mathbb{K}_q^\times$ in einem endlichen Körper mit q Elementen zyklisch ist, gibt es $\varphi(q-1)$ viele Generatoren. Ist allgemeiner t ein Teiler von $q-1$, so gibt es $\varphi(t)$ viele Elemente der Ordnung t in $\mathbb{K}_q$. Insbesondere existiert wegen $\varphi(q-1) \geq 1$ immer mindestens ein Generator.

Wir können unseren endlichen Körper also auch in der gleichwertigen Form

$$\mathbb{K}_4 = \{0, 1, \beta, \beta^2\}$$

beschreiben. In dieser neuen Beschreibungsform nehmen die Additions- und Multiplikationstabellen die Form

$+$	0	1	β	β^2
0	0	1	β	β^2
1	1	0	β^2	β
β	β	β^2	0	1
β^2	β^2	β	1	0

$\cdot$	0	1	β	β^2
0	0	0	0	0
1	0	1	β	β^2
β	0	β	β^2	1
β^2	0	β^2	1	β

an. Man beachte, dass unter dem Ring-Isomorphismus

$$\mathbb{K}_4 \to \mathbb{K}_4 \qquad \alpha^j \mapsto \beta^j$$

die Verküpfungstabellen im Wesentlichen übereinstimmen und es sich bloss um eine Umbezeichnung der Elemente handelt. Deshalb gelten isomorphe Körper als im Wesentlichen gleich.

In einem endlichen Körper haben also die von Null verschiedenen Elemente Exponentialdarstellungen, die von der Wahl eines primitiven Elementes (Generators) abhängen. Weil auch das Element $\beta \in \mathbb{K}_4$ Nullstelle des irreduziblen Polynoms $f(x) = 1 + x + x^2$ ist, von dem wir ausgegangen sind, hat diesen Polynom im Körper $\mathbb{K}_4$ zwei verschiedene Nullstellen α, β. Daher gilt in diesem Körper die Faktorisierung

$$1 + x + x^2 = (x - \alpha)(x - \beta)$$

wie man durch Ausmultiplizieren und den Beziehungen $\alpha + \beta = 1 = \alpha \cdot \beta$ leicht bestätigt.

Man beachte, dass sich die drei invertierbaren Körperelemente aus der Einheitengruppe $\mathbb{K}_4^\times = \{1, \alpha, \alpha^2\}$ wie die dritten komplexen Einheitswurzeln

$$\zeta_3^k = e^{ik\frac{2\pi}{3}} = \cos\left(k\,\frac{2\pi}{3}\right) + i\sin\left(k\frac{2\pi}{3}\right) \in \mathbb{C}^\times, \qquad (0 \le k \le 2)$$

benehmen, die auf den Ecken eines gleichseitigen Dreiecks liegen und die zum Beispiel beim zyklischen Shift oder bei der diskreten Fourier-Transformation eine zentrale Rolle spielen und die polynomiale Gleichung $x^3 - 1 = 0$ lösen.

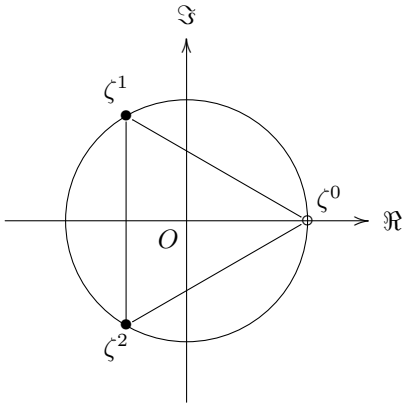


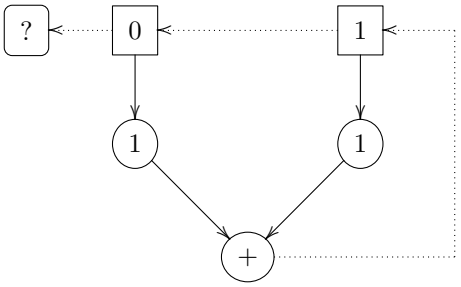
Abbildung 2.40: Lage der drei Eigenwerte in der komplexen Ebene $\mathbb{C}$.

Will man solch diskrete Methoden auch über endlichen Körpern zur Verfügung haben, benötigt man also in der Regel endliche Erweiterungskörper.

Auch im Körper $\mathbb{K}_4$ gelten für den Schüler ungewohnte Eigenschaften, wie die vereinfachte Binom'sche Formel

$$(x + y)^2 = x^2 + y^2$$

zeigt. Durch einen Blick auf die Additionstabelle von $\mathbb{K}_4$ bestätigt man, dass auch in diesem Erweiterungskörper für jedes Element die Gleichung $-r = r$ gilt und daher Addition und Subtraktion übereinstimmen. Daher hat in diesem Körper auch die Gleichung des goldenen Schnittes $x^2 + x - 1 = 0$ zwei verschiedene Lösungen, wie man durch Einsetzen von α oder β bestätigt. Wie über den reellen Zahlen kann man auch über diesem Körper die Fibonacci-Folge bzw. das zugehörige Schieberegister durch ein explizite Formel beschreiben. Möchte man über dem endlichen Körper $\mathbb{Z}_2$ die Fibonacci-Folge mit dem Schieberegister



und der linearen Rekursionsgleichung

$$x_{k+2} = x_k + x_{k+1}, \qquad x_0 = 0, x_1 = 1$$

und der periodischen Binärfolge

$$\mathbf{011}011011011\ldots$$

der Periodenlänge 3 durch eine explizite Formel beschreiben, braucht man für die Eigenwerte zwei verschiedene Lösungen der charakteristischen Gleichung

$$\lambda^2 = \lambda + 1$$

Wie wir wissen, gibt es in $\mathbb{Z}_2$ keine solchen Lösungen. Hingegen finden wir in der Körpererweiterung $\mathbb{K}_4$ die beiden Eigenwerte

$$\lambda_1 = \alpha, \qquad \lambda_2 = \beta$$

mit denen wir in gewohnter Weise die beiden Basislösungen $x_1(k) = \alpha^k$ und $x_2(k) = \beta^k$ erhalten. Um daraus die gesuchte Folge linear kombinieren zu knnnen, machen wir für die allgmeine Lösung der Rekursionsgleichung den Ansatz

$$x_k = c_1 x_1(k) + c_2 x_2(k) = c_1\alpha^k + c_2\beta^k$$

und müssen nun die Koeffizienten c_1, c_2 so bestimmen, dass die Anfangsbedingungen dass die Anfangsbedingung $x_0 = 0$ und $x_1 = 1$ erfüllt sind. Der Ansatz liefert für $k = 0$ und $k = 1$ das lineare Gleichungssystem

$$\begin{cases} k = 0: & c_1 + c_2 = 0 \\ k = 1: & \alpha c_1 + \beta c_2 = 1 \end{cases}$$

mit der eindeutigen Lösung in $\mathbb{K}_4$

$$c_1 = c_2 = 1$$

Daher hat die Fibonacci-Folge über $\mathbb{K}_4$ die explizite Beschreibung

$$x_k = \lambda_1^k + \lambda_2^k = \alpha^k + \beta^k$$

wie man durch Einsetzen in die Rekursionsgleichung verifiziert. Entsprechende Elemente gibt es im Unterkörper $\mathbb{Z}_2$ keine und man muss deshalb unter Umständen den Körper vergrössern, wenn man gewisse bekannte Methoden diskretisieren will.

Wir erinnern nochmals daran, dass der soeben konstruierte Körper $\mathbb{K}_4$ nicht mit dem kommutativen Ring $\mathbb{Z}_4$ verwechselt werden darf, der ja keine Körperstruktur hat, wie wir gesehen haben. Ein weiterer Blick auf die Additionstabelle des Körpers $\mathbb{K}_4$ zeigt, dass seine additive Gruppe isomorph zu Vierergruppe $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ ist, in der jedes der vertauschbaren Elemente die Ordnung 2 hat und es keine Elemente der Ordnung 4 gibt. Die Vierergruppe kann geometrisch als Symmetriegruppe eines Rechtecks oder allgemeiner eines Rhombus interpretiert werden.

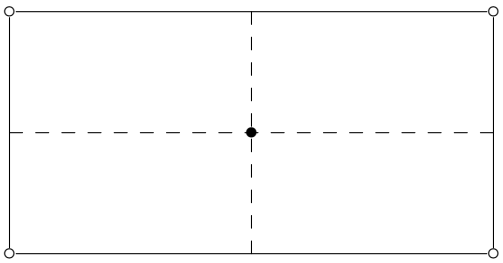


Abbildung 2.41: Symmetriegruppe eines Rechtecks.

Ihre vier Elemente $0, 1, \alpha, \alpha^2$ entsprechen der Identität Id, der horizontalen H und der vertikalen V Spiegelung sowie der Punktspiegelung P am Symmetriezentrum, wie ein Vergleich der Multiplikationstabelle von $\mathbb{K}_4$ mit jener der Symmetrien des Rechtecks

$\circ$	Id	H	V	P
Id	Id	H	V	P
H	H	Id	P	V
V	V	P	Id	H
P	P	V	H	Id

zeigt. Der kommutativen Struktur dieser Gruppe ist es zu verdanken, dass die allgemeine Gleichung vierten Grades $a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 = 0$ ausnahmsweise mit Hilfe von Radikalen gelöst werden kann.

Analog entspricht die Galois-Gruppe der allgemeinen Gleichung fünften Grades der Symmetriegruppe eines regulären Ikosaeders, deren verwickeltere Struktur es nicht mehr erlaubt, solche Gleichungen mit Hilfe von Radikalen zu lösen. ◯

Wie die Körpererweiterung $\mathbb{Z}_2 \subset \mathbb{K}_4$ können auch die anderen endlichen Köper $\mathbb{K}_q$ mit $q = p^d$ Elementen als Erweiterungskörper der Primkörper $\mathbb{Z}_p$ konstruiert werden.

Beispiel. Zur Konstruktion des Körpers $\mathbb{K}_8$ mit $q = 2^3 = 8$ Elementen, gehen wir vom normierten Polynom $f(x) = 1 + x + x^3 \in \mathbb{Z}_2[x]$ vom Grad $d = 3$ aus. Um zu zeigen, dass es irreduzibel ist, beachten wir, dass es sonst die Form

$$f(x) = (x - a) \cdot f_2(x)$$

und damit eine Nullstelle in $\mathbb{Z}_2$ haben müsste, was nicht der Fall ist.
Zum Polynom $f(x)$ gehört die Begleitermatrix

$$\Omega = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \in \mathbb{Z}_2^{3,3}$$

als Nullstelle. Sie erfüllt also die charakteristische Gleichung

$$f(\Omega) = E + \Omega + \Omega^3 = 0, \qquad \Omega^3 = E + \Omega$$

Die Elemente des gesuchten Körpers

$$\mathbb{K}_8 = \{a_0E + a_1\Omega + a_2\Omega^2 | a_0, a_1, a_2 \in \mathbb{Z}_2\}$$

sind also neben der Nullmatrix 0, der Einheitsmatrix E und der Matrix Ω die $\mathbb{Z}_2$-Linearkombinationen

$$\Omega^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad E+\Omega = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \Omega+\Omega^2 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$E+\Omega+\Omega^2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad E+\Omega^2 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Diese acht Matrizen aus $\mathbb{Z}_2^{3,3}$ bilden den gesuchten Körper

$$\mathbb{K}_8 = \{0, E, \Omega, \Omega^2, E+\Omega, \Omega+\Omega^2, E+\Omega+\Omega^2, E+\Omega^2\}$$

Auch hier geschieht die Addition komponentenweise. Die höheren Potenzen lassen sich mit Hilfe der charakteristischen Gleichung $\Omega^3 = E + \Omega$ reduzieren. In diesem Beispiel gilt

$$\Omega^4 = \Omega \cdot \Omega^3 = \Omega \cdot (E+\Omega) = \Omega + \Omega^2$$

Daher genügt die Multiplikation in $\mathbb{K}_8$ der magischen Formel

$$(a_0 + a_1\Omega + a_2\Omega^2) \cdot (b_0 + b_1\Omega + b_2\Omega^2) = (a_0b_0 + a_1b_2 + a_2b_1)E + (a_0b_1 + a_1b_0 + a_1b_2 + a_2b_1 + a_2b_2)\Omega + (a_0b_2 + a_1b_1 + a_2b_0 + a_2b_2)\Omega^2$$

wie man nach der Reduktion sieht. Um dieses Multiplikationsgesetz

$$\mu\colon \mathbb{Z}_2^3 \times \mathbb{Z}_2^3 \to \mathbb{Z}_2^3, \qquad (\vec{a}, \vec{b}) \mapsto \mu(\vec{a}, \vec{b}) = \vec{c}$$

mit den drei Komponenten

$$\begin{aligned} c_0 &= a_0b_0 + a_1b_2 + a_2b_1 \\ c_1 &= a_0b_1 + a_1b_0 + a_1b_2 + a_2b_1 + a_2b_2 \\ c_2 &= a_0b_2 + a_1b_1 + a_2b_0 + a_2b_2 \end{aligned}$$

durch Multiplikation $\mu(\vec{a}, \vec{b}) = \vec{c} = f(\vec{a}) \cdot \vec{b}$ mit der quadratischen Matrix

$$f(\vec{a}) = \begin{pmatrix} a_0 & a_2 & a_1 \\ a_1 & a_0 + a_2 & a_1 + a_2 \\ a_2 & a_1 & a_0 + a_2 \end{pmatrix} = a_0A_0 + a_1A_1 + a_2A_2$$

zu beschreiben, die sich mit Hilfe der Matrizen des Systems $\{A_0, A_1, A_2\}$ linear kombinieren lässt, definieren wir diese drei Matrizen durch

$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Die Produkte der Standardbasisvektoren lassen sich als Linearkombination

$$\mu(\vec{e}_i, \vec{e}_j) = \sum_{k=1}^{3} \mu(\vec{e}_i, \vec{e}_j)_k \vec{e}_k = \sum_{k=1}^{3} (S_k)_{ij} \vec{e}_k$$

ausdrücken. Sie können wahlweise mit Hilfe der Komponenten des Multiplikationsgesetzes oder durch Reduktion der höheren Potenzen von Ω berechnetet werden. In der Matrixschreibweise erhalten wir die Produkte in folgender symmetrischen Tabelle.

$$\begin{array}{lll} \mu(\vec{e}_1,\vec{e}_1)=E & \mu(\vec{e}_1,\vec{e}_2)=\Omega & \mu(\vec{e}_1,\vec{e}_3)=\Omega^2 \\ \mu(\vec{e}_2,\vec{e}_1)=\Omega & \mu(\vec{e}_2,\vec{e}_2)=\Omega^2 & \mu(\vec{e}_2,\vec{e}_3)=E+\Omega \\ \mu(\vec{e}_3,\vec{e}_1)=\Omega^2 & \mu(\vec{e}_3,\vec{e}_2)=E+\Omega & \mu(\vec{e}_3,\vec{e}_3)=\Omega+\Omega^2 \end{array}$$

Die $3^3 = 27$ Strukturkonstanten $\mu(\vec{e}_i,\vec{e}_j)_k = (S_k)_{ij}$ entnimmt man der zugehörigen Vektordarstellung

$$\begin{array}{lll} \mu(\vec{e}_1,\vec{e}_1)=(1,0,0) & \mu(\vec{e}_1,\vec{e}_2)=(0,1,0) & \mu(\vec{e}_1,\vec{e}_3)=(0,0,1) \\ \mu(\vec{e}_2,\vec{e}_1)=(0,1,0) & \mu(\vec{e}_2,\vec{e}_2)=(0,0,1) & \mu(\vec{e}_2,\vec{e}_3)=(1,1,0) \\ \mu(\vec{e}_3,\vec{e}_1)=(0,0,1) & \mu(\vec{e}_3,\vec{e}_2)=(1,1,0) & \mu(\vec{e}_3,\vec{e}_3)=(0,1,1) \end{array}$$

Die Strukturkonstanten treten als Elemente in den Strukturmatrizen

$$S_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \qquad S_3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

auf. Damit ist die Arithmetik im Körper $\mathbb{K}_8$ vollständig auf die Arithmetik in $\mathbb{Z}_2$ zurückgeführt.

Wie in jedem endlichen Körper ist auch hier die multiplikative Gruppe der Einheiten $\mathbb{K}_8^\times$ zyklisch und es muss deshalb einen Generator $\alpha \in \mathbb{K}_8$ geben, dessen Potenzen die von Null verschiedenen Körperelemente erzeugt. Sukzessives weiteres Potenzieren der Matrix $\alpha = \Omega$ liefert die Potenzen

$$\alpha^2 = \Omega^2, \alpha^3 = E + \Omega, \alpha^4 = \Omega + \Omega^2, \alpha^5 = E + \Omega + \Omega^2, \alpha^6 = E + \Omega, \alpha^7 = E$$

und zeigt, dass in diesem Fall $\alpha = \Omega$ als Generator verwendet werden kann. In der Exponentialdarstellung ist das Multiplizieren einfach, weil das Potenzgesetz

$$\alpha^i \cdot \alpha^j = \alpha^{i+j \bmod 7}$$

die Multiplikation auf eine Addition in $\mathbb{Z}_7$ zurückführt. Beispielsweise gilt

$$\alpha^4 \cdot \alpha^5 = \alpha^9 = \alpha^7 \cdot \alpha^2 = \alpha^2, \qquad \alpha^6 \cdot \alpha^5 = \alpha^{11} = \alpha^7 \cdot \alpha^4 = \alpha^4$$

Die multiplikativen Inversen können mit Hilfe von negativen Exponenten bestimmt werden. Beispielsweise ist

$$(\alpha^5)^{-1} = \alpha^2, \qquad \alpha^2 \cdot \alpha^5 = \alpha^7 = 1$$

Weil das Produkt zweier Potenzen von α wieder eine solche Potenz ergibt, hat der kommutative Ring $\mathbb{K}_4$ keine Nullteiler und es handelt sich tatsächlich um einen Körper.

Damit kann auch in diesem Körper jedes Element wahlweise als Linearkombination der Matrizen 0, E und Ω, als Vektor in $\mathbb{Z}_2^3$ oder als Potenz des Generators α ausgedrückt werden und wir erhalten die Zuordnungsvorschrift zwischen den

üblichen Darstellungsormen gemäss folgender Tabelle:

Matrix	Vektor	Potenz	Ordnung
0	$(0,0,0)$	0	$-$
E	$(1,0,0)$	1	1
Ω	$(0,1,0)$	α	7
Ω^2	$(0,0,1)$	α^2	7
$E+\Omega$	$(1,1,0)$	α^3	7
$\Omega+\Omega^2$	$(0,1,1)$	α^4	7
$E+\Omega+\Omega^2$	$(1,1,1)$	α^5	7
$E+\Omega^2$	$(1,0,1)$	α^6	7

Die Darstellung der Körperelemente als Linearkombination bzw. als Vektor ist bequem im Umgang mit der Addition, die Exponentialdarstellung im Umgang mit der Multiplikation und die Vektordarstellung im Umgang mit dem Skalarprodukt.

Das Einsundeins in diesem Körper ist durch die Matrizenaddition in $\mathbb{Z}_2^{3,3}$ bzw. durch Vektoraddition in $\mathbb{Z}_2^3$ gegeben und liefert folgende Additionstabelle, die wir mit Hilfe des gewählten Generators α in Exponentialdarstellung angeben, die wir mit Hilfe obiger Zuordnungstabelle erhalten.

$+$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	1	α	α^2	α^3	α^4	α^5	α^6
1	1	0	α^3	α^6	α	α^5	α^4	α^2
α	α	α^3	0	α^4	1	α^2	α^6	α^5
α^2	α^2	α^6	α^4	0	α^5	α	α^3	1
α^3	α^3	α	1	α^5	0	α^6	α^2	α^4
α^4	α^4	α^5	α^2	α	α^6	0	1	α^3
α^5	α^5	α^4	α^6	α^3	α^2	1	0	α
α^6	α^6	α^2	α^5	1	α^4	α^3	α	0

x	$-x$
0	0
1	1
α	α
α^2	α^2
α^3	α^3
α^4	α^4
α^5	α^5
α^6	α^6

Entsprechend ist das Einmaleins in diesem Körper durch die folgende Tabelle der Matrizenprodukte gegeben.

$\cdot$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	0	0	0	0	0	0	0
1	0	1	α	α^2	α^3	α^4	α^5	α^6
α	0	α	α^2	α^3	α^4	α^5	α^6	1
α^2	0	α^2	α^3	α^4	α^5	α^6	1	α
α^3	0	α^3	α^4	α^5	α^6	1	α	α^2
α^4	0	α^4	α^5	α^6	1	α	α^2	α^3
α^5	0	α^5	α^6	1	α	α^2	α^3	α^4
α^6	0	α^6	1	α	α^2	α^3	α^4	α^5

x	x^{-1}
0	$-$
1	1
α	α^6
α^2	α^5
α^3	α^4
α^4	α^3
α^5	α^2
α^6	α

Aus der Multiplikationstabelle entnimmt man leicht, dass jedes Element aus der Einheitengruppe $\mathbb{K}_8^\times$ ausser dem multiplikativen Neutralelement 1 die maximale Ordnung 7 hat und daher jedes dieser $\varphi(q-1) = \varphi(7) = 6$ Elementen als Generator hätte gewählt werden können.

Der endliche Körper $\mathbb{K}_8$ enthält offensichtlich den Primkörper $\mathbb{Z}_2 = \{0,1\}$, von dem wir ausgegangen sind. Im Gegensatz dazu kann er den früher konstruierten

Körper $\mathbb{K}_4$ mit vier Elementen nicht enthalten, weil er kein Element der Ordnung 3 enthält.

Die Logarithmen bezüglich des Generators α werden im Körper $\mathbb{K}_8$ in der üblichen Art bestimmt. Aus der Tabelle der Potenzen des Generators

n	1	2	3	4	5	6	7
α^n	Ω	Ω^2	$E+\Omega$	$\Omega+\Omega^2$	$E+\Omega+\Omega^2$	$E+\Omega$	E

erhalten wir durch Invertieren die Logarithmentabelle für $r \in \mathbb{K}_8$ bezüglich der Basis α.

r	Ω	Ω^2	$E+\Omega$	$\Omega+\Omega^2$	$E+\Omega+\Omega^2$	$E+\Omega^2$	E
$\log_\alpha(r)$	1	2	3	4	5	6	7

Auf Grund des Exponentialgesetzes gelten in $\mathbb{Z}_{p-1}$ die üblichen Logarithmengesetze

$$\log_\alpha(r_1 \cdot r_2) = \log_\alpha(r_1) + \log_\alpha(r_2), \qquad \log_\alpha(r^k) = k \cdot \log_\alpha(r)$$

wobei auch hier sorgfältig auf den Modul $p-1=7$ zu achten ist! Man beachte, dass auch im Körper $\mathbb{K}_8$ die Binom'sche Formel

$$(x+y)^2 = x^2 + y^2$$

gilt, die besagt, dass die Potenzierung mit der Charakteristik $p=2$ und die Addition ausnahmsweise verträglich sind.

Der aufmerksame Leser wird sich gefragt haben, was passiert wäre, wenn wir seinerzeit ein anderes irreduzibles Polynom $g(x) = 1 + x^2 + x^3 \in \mathbb{Z}_2[x]$, das ja in $\mathbb{Z}_2$ auch keine Nullstelle hat, gewählt hätten, um damit den endlich Körper $\tilde{\mathbb{K}}_8$ zu konstruieren. Wir wären dann von seiner Begleitermatrix

$$\Xi = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \in \mathbb{Z}_2^{3,3}$$

ausgegangen. Sie erfüllt die charakteristische Gleichung

$$g(\Xi) = E + \Xi^2 + \Xi^3 = 0, \qquad \Xi^3 = E + \Xi^2$$

Die Elemente des gesuchten Körpers

$$\tilde{\mathbb{K}}_8 = \{a_0 E + a_1 \Xi + a_2 \Xi^2 | a_, a_1, a_2 \in \mathbb{Z}_2\}$$

wären dann neben der Nullmatrix 0, der Einheitsmatrix und der Matrix Ξ die $\mathbb{Z}_2$-Linearkombinationen

$$\Xi^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad E+\Xi = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \Xi+\Xi^2 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$E+\Xi+\Xi^2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad E+\Xi^2 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

gewesen. Diese acht Matrizen bilden den Körper

$$\tilde{\mathbb{K}}_8 = \{0, E, \Xi, \Xi^2, E+\Xi, \Xi+\Xi^2, E+\Xi+\Xi^2, E+\Xi^2\}$$

mit acht Elementen.

Ein Vergleich dieser Matrizen mit den früher konstruierten zeigt, dass nur gerade die beiden Matrizen 0 und E des Primkörpers in beiden Darstellungen vorkommen. Daher stimmen die beiden Körper $\mathbb{K}_8$ und $\tilde{\mathbb{K}}_8$ nicht überein.

Die Suche nach einem Generator in $\tilde{\mathbb{K}}_8$ hätte gezeigt, dass auch dieser Körper $\varphi(7) = 6$ primitive Elemente hat und man etwa

$$\beta = \Xi$$

als primitives Element wählen könnte. Sukzessives Potenzieren der Matrix $\beta = \Xi$ liefert die Potenzen

$$\beta^2 = \Xi^2, \beta^3 = E+\Xi^2, \beta^4 = E+\Xi+\Xi^2, \beta^5 = E+\Xi, \beta^6 = \Xi+\Xi^2, \beta^7 = E$$

Damit hätte die Addtionstabelle die Form

$+$	0	1	β	β^2	β^3	β^4	β^5	β^6
0	0	1	β^1	β^2	β^3	β^4	β^5	β^6
1	1	0	β^5	β^3	β^2	β^6	β^1	β^4
β	β^1	β^5	0	β^6	β^4	β^3	1	β^2
β^2	β^2	β^3	β^6	0	1	β^5	β^4	β^1
β^3	β^3	β^2	β^4	1	0	β^1	β^6	β^5
β^4	β^4	β^6	β^3	β^5	β^1	0	β^2	1
β^5	β^5	β^1	1	β^4	β^6	β^2	0	β^3
β^6	β^6	β^4	β^2	β^1	β^5	1	β^3	0

x	$-x$
0	0
1	1
β	β
β^2	β^2
β^3	β^3
β^4	β^4
β^5	β^5
β^6	β^6

angenommen und die Multiplikationstabelle hätte

$\cdot$	0	1	β	β^2	β^3	β^4	β^5	β^6
0	0	0	0	0	0	0	0	0
1	0	1	β	β^2	β^3	β^4	β^5	β^6
β	0	β	β^2	β^3	β^4	β^5	β^6	1
β^2	0	β^2	β^3	β^4	β^5	β^6	1	β
β^3	0	β^3	β^4	β^5	β^6	1	β	β^2
β^4	0	β^4	β^5	β^6	1	β	β^2	β^3
β^5	0	β^5	β^6	1	β	β^2	β^3	β^4
β^6	0	β^6	1	β	β^2	β^3	β^4	β^5

x	x^{-1}
0	$-$
1	1
β	β^6
β^2	β^5
β^3	β^4
β^4	β^3
β^5	β^2
β^6	β

gelautet. Ein Vergleich dieser beiden Tabellen mit den früher bestimmten zeigt, dass die beiden Multiplikationstabellen selbstverständlich bis auf Bezeichnung übereinstimmen, da sowohl α also auch β Generatoren der Ordnung 7 sind. Weniger offensichtlich ist, ob die beiden Körper $\mathbb{K}_8$ und $\tilde{\mathbb{K}}_8$ nicht vielleicht trotz der unterschiedlichen Form ihrer Additionstabellen isomorph sind. Dieser Unterschied kommt daher, dass die beiden verwendeten Generatoren α und β in den beiden Köpern andere Gleichungen erfüllen. Genaueres Hinsehen zeigt jedoch, dass das Element $E + \Omega = \alpha^3 \in \mathbb{K}_8$ die charakteristische Gleichung

$$g(E+\Omega) = g(\alpha^3) = 1 + (\alpha^3)^2 + (\alpha^3)^3 = 1 + \alpha^6 + \alpha^9 = 1 + \alpha^6 + \alpha^2 = 0$$

von $\beta = \Xi \in \tilde{\mathbb{K}}_8$ erfüllt. Entsprechend erfüllt das Element β^5 die charakteristische Gleichung

$$f(\beta^5) = 1 + \beta^5 + (\beta^5)^3 = 0$$

von $\alpha = \Omega$.

Daher liefert die Abbildung

$$T\colon \tilde{\mathbb{K}}_8 \to \mathbb{K}_8, \quad \beta^j \mapsto (E + \Omega)^j = (\alpha^3)^j$$

einen Ringisomorphismus, d.h. einen Ringhomomorphismus, der mit der Addition und der Multiplikation und den Neutralelementen 0 und 1 verträglich ist und die inverse Abbildung

$$T^{-1}\colon \mathbb{K}_8 \to \tilde{\mathbb{K}}_8, \quad \alpha^k \mapsto (\beta^5)^k$$

hat. Tatsächlich ist

$$T^{-1}(T(\beta)) = T^{-1}(\alpha^3) = \left(\beta^5\right)^3 = \beta^{15} = \beta^{14} \cdot \beta = \beta$$

und

$$T(T^{-1}(\alpha)) = T(\beta^5) = \left(\alpha^3\right)^5 = \alpha^{15} = \alpha^{14} \cdot \alpha = \alpha$$

Daher sind die beiden Abbildungen in der Tat invers zueinander.

Man beachte, dass der Isomorphismus zwischen diesen beiden Körpern nicht eindeutig bestimmt ist, weil auch $g(\alpha^5) = 0$ und $g(\alpha^6) = 0$ ist, gibt es zwei weitere Isomorphismen $\tilde{\mathbb{K}}_8 \to \mathbb{K}_8$. Ihre Inversen verdankt man den Gleichungen $f(\beta^3) = 0$ und $f(\beta^6) = 0$. Damit haben die beiden Polynome die Faktorisierungen

$$f(x) = (x - \beta^5)(x - \beta^3)(x - \beta^6), \qquad g(x) = (x - \alpha^3)(x - \alpha^5)(x - \alpha^6)$$

in Linearfaktoren. Der Körper $\mathbb{K}_8$ kann als Zerfällungskörper der Polynoms

$$x^8 - x = (x - 0)(x - 1)(x - \beta^5)(x - \beta^3)(x - \beta^6)(x - \alpha^3)(x - \alpha^5)(x - \alpha^6)$$

aufgefasst werden. Weil das selbe für den Körper $\mathbb{K}_8$ gilt, sind die beiden Körper (nicht kanonisch) isomorph. ◯

Beispiel. Um einen Körper $\mathbb{K}_{2^4}$ mit $q = 2^4 = 16$ Elementen zu konstruieren, brauchen wir zunächst in $\mathbb{Z}_2[x]$ ein normiertes, irreduzibles Polynom vom Grad $d = 4$. Dazu kann man beispielsweise das Polynom $f(x) = 1 + x + x^4$ verwenden. Um zu zeigen, dass es irreduzibel ist, beachten wir zunächst, dass dieses Polynom im Körper $\mathbb{Z}_2$ keine Nullstelle hat, da $f(0) = f(1) = 1$ ist. Deshalb kann es also keinen Faktor vom Grad 1 haben. Dieses Polynom hat aber auch keinen Faktor vom Grad 2 d.h. keine Faktorisierung der Art

$$f(x) = (1 + ax + x^2) \cdot (1 + bx + x^2),$$

weil sonst nach dem Ausmultiplizieren in $\mathbb{Z}_2[x]$ die Gleichung

$$1 + x + x^4 = 1 + (a + b)x + (ab)x^2 + (a + b)x^3 + x^4$$

gelten müsste. Ein Koeffizientenvergleich zeigt, dass das evident nicht der Fall ist.

Zum Poynom $f(x)$ gehört die Begleitermatrix

$$\Omega = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \in \mathbb{Z}_2^{4,4}$$

als Nullstelle. Sie erfüllt also die charakteristische Gleichung

$$f(\Omega) = E + \Omega + \Omega^4 = 0, \qquad \Omega^4 = E + \Omega$$

Die Elemente des gesuchten Köpers

$$\mathbb{K}_{16} = \{a_0 E + a_1\Omega + a_2\Omega^2 + a_3\Omega^3 | a_0, a_1, a_2, a_3 \in \mathbb{Z}_2\}$$

sind Linearkombinationen der Matrizen E, Ω, Ω^2 und Ω^3. Sie können auch als Polynome vom Grad echt kleiner als 4 bzw. als Vektoren mit den Koeffizienten als Komponenten ausgedrückt werden. Diese 16 Elemente haben also wahlweise eine Darstellung, zwischen denen mit folgender Tabelle umgerechnet werden kann.

Name	Matrix	Polynom	Vektor	Potenz	Ordnung
0	0	0	(0000)	–	–
1	E	1	(1000)	γ^0	1
a	Ω	x	(0100)	γ^1	15
b	$E + \Omega$	$1 + x$	(1100)	γ^4	15
c	Ω^2	x^2	(0010)	γ^2	15
d	$E + \Omega^2$	$1 + x^2$	(1010)	γ^8	15
e	$\Omega + \Omega^2$	$x + x^2$	(0110)	γ^5	3
f	$E + \Omega + \Omega^2$	$1 + x + x^2$	(1110)	γ^{10}	3
g	Ω^3	x^3	(0001)	γ^3	5
h	$E + \Omega^3$	$1 + x^3$	(1001)	γ^{14}	15
i	$\Omega + \Omega^3$	$x + x^3$	(0101)	γ^9	5
j	$E + \Omega + \Omega^3$	$1 + x + x^3$	(1101)	γ^7	15
k	$\Omega^2 + \Omega^3$	$x^2 + x^3$	(0011)	γ^6	5
l	$E + \Omega^2 + \Omega^3$	$1 + x^2 + x^3$	(1011)	γ^{13}	15
m	$\Omega + \Omega^2 + \Omega^3$	$x + x^2 + x^3$	(0111)	γ^{11}	15
n	$E + \Omega + \Omega^2 + \Omega^3$	$1 + x + x^2 + x^3$	(1111)	γ^{12}	5

Die Additionstafel berechnen wir also äquivalent durch komponentenweise Ad-

dition der Linearkombinationen, Polynome oder Vektoren.

$+$	0	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n
0	0	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n
1	1	0	b	a	d	c	f	e	h	g	j	i	l	k	n	m
a	a	b	0	1	e	f	c	d	i	j	g	h	m	n	k	l
b	b	a	1	0	f	e	d	c	j	i	h	g	n	m	l	k
c	c	d	e	f	0	1	a	b	k	l	m	n	g	h	i	j
d	d	c	f	e	1	0	b	a	l	k	n	m	h	g	j	i
e	e	f	c	d	a	b	0	1	m	n	k	l	i	j	g	h
f	f	e	d	c	b	a	1	0	n	m	l	k	j	i	h	g
g	g	h	i	j	k	l	m	n	0	1	a	b	c	d	e	f
h	h	g	j	i	l	k	n	m	1	0	b	a	d	c	f	e
i	i	j	g	h	m	n	k	l	a	b	0	1	e	f	c	d
j	j	i	h	g	n	m	l	k	b	a	1	0	f	e	d	c
k	k	l	m	n	g	h	i	j	c	d	e	f	0	1	a	b
l	l	k	n	m	h	g	j	i	d	c	f	e	1	0	b	a
m	m	n	k	l	i	j	g	h	e	f	c	d	a	b	0	1
n	n	m	l	k	j	i	h	g	f	e	d	c	b	a	1	0

Die höheren Potenzen von Ω lassen sich mit Hilfe der charakteristischen Gleichung $\Omega^4 = E + \Omega$ reduzieren. In diesem Beispiel erhalten wir rekursiv

$$\Omega^5 = \Omega \cdot (E + \Omega) = \Omega + \Omega^2, \qquad \Omega^6 = \Omega \cdot (\Omega + \Omega^2) = \Omega^2 + \Omega^3$$

Zur vollständigen Beschreibung der Multiplikation

$$\mu\colon \mathbb{Z}_2^4 \times \mathbb{Z}_2^4 \to \mathbb{Z}_2^4, \qquad (\vec{a}, \vec{b}) \mapsto \mu(\vec{a}, \vec{b}) = \vec{c}$$

benötigen wir die Produkte der Standardbasisvektoren

$$\begin{array}{llll}
\mu(\vec{e}_1, \vec{e}_1) = E & \mu(\vec{e}_1, \vec{e}_2) = \Omega & \mu(\vec{e}_1, \vec{e}_3) = \Omega^2 & \mu(\vec{e}_1, \vec{e}_4) = \Omega^3 \\
\mu(\vec{e}_2, \vec{e}_1) = \Omega & \mu(\vec{e}_2, \vec{e}_2) = \Omega^2 & \mu(\vec{e}_2, \vec{e}_3) = \Omega^3 & \mu(\vec{e}_2, \vec{e}_4) = E + \Omega \\
\mu(\vec{e}_3, \vec{e}_1) = \Omega^2 & \mu(\vec{e}_3, \vec{e}_2) = \Omega^3 & \mu(\vec{e}_3, \vec{e}_3) = E + \Omega & \mu(\vec{e}_3, \vec{e}_4) = \Omega + \Omega^2 \\
\mu(\vec{e}_4, \vec{e}_1) = \Omega^3 & \mu(\vec{e}_4, \vec{e}_2) = E + \Omega & \mu(\vec{e}_4, \vec{e}_3) = \Omega + \Omega^2 & \mu(\vec{e}_4, \vec{e}_4) = \Omega^2 + \Omega^3
\end{array}$$

Daraus entnehmen wir die $4^3 = 64$ Strukturkonstanten $\mu(\vec{e}_i, \vec{e}_j)_k = (S_k)_{ij}$, die wir als Elemente in den 4 Strukturmatrizen

$$S_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \qquad S_4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

zusammenfassen. Damit ist die Arithmetik im Körper $\mathbb{K}_{16}$ vollständig auf die Arithmetik in $\mathbb{Z}_2$ zurückgeführt. Der Produktvektor $\mu(\vec{a}, \vec{b}) = \vec{c}$ besitzt die vier

Komponenten

$$\begin{aligned}
c_0 = \langle \vec{a}, S_1 \cdot \vec{b} \rangle &= a_0b_0 + a_1b_1 + a_2b_2 + a_3b_3 \\
c_1 = \langle \vec{a}, S_2 \cdot \vec{b} \rangle &= a_3b_0 + a_0b_1 + a_3b_1 + a_1b_2 + a_2b_3 \\
c_2 = \langle \vec{a}, S_3 \cdot \vec{b} \rangle &= a_2b_0 + a_2b_1 + a_3b_1 + a_0b_2 + a_3b_2 + a_1b_3 \\
c_3 = \langle \vec{a}, S_4 \cdot \vec{b} \rangle &= a_1b_0 + a_1b_1 + a_2b_1 + a_2b_2 + a_3b_2 + a_0b_3 + a_3b_3
\end{aligned}$$

Dieses Multiplikationsgesetz $\mu(\vec{a}, \vec{b}) = \vec{c} = f(\vec{a}) \cdot \vec{b}$ lässt sich auch durch Multiplikation mit der quadratischen Matrix

$$f(\vec{a}) = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_3 & a_0 + a_3 & a_1 & a_2 \\ a_2 & a_2 + a_3 & a_0 + a_3 & a_1 \\ a_1 & a_1 + a_2 & a_2a_3 & a_0 + a_3 \end{pmatrix} = a_0A_0 + a_1A_1 + a_2A_2 + a_3A_3$$

beschreiben, die sich mit Hilfe der Matrizen des Systems $\{A_0, A_1, A_2, A_3\}$ linear kombinieren lässt.

Für die vollständige Multiplikationstafel von $\mathbb{K}_{16}$ gilt

$\cdot$	0	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a	0	a	c	e	g	i	k	m	b	1	f	d	j	h	n	l
b	0	b	e	d	k	n	i	h	j	g	l	m	f	c	1	a
c	0	c	g	k	b	f	j	n	e	a	m	i	d	1	l	h
d	0	d	i	n	f	a	l	g	m	j	c	1	h	k	b	e
e	0	e	k	i	j	l	f	1	d	b	h	n	m	g	a	c
f	0	f	m	h	n	g	1	e	l	i	b	c	a	d	k	j
g	0	g	b	j	e	m	d	l	k	c	n	f	i	a	h	1
h	0	h	1	g	a	j	b	i	c	l	d	k	e	n	f	m
i	0	i	f	l	m	c	h	b	n	d	g	a	1	j	e	k
j	0	j	d	m	i	1	n	c	f	k	a	h	l	e	g	b
k	0	k	j	f	d	h	m	a	i	e	1	l	n	b	c	g
l	0	l	h	c	1	k	g	d	a	n	j	e	b	m	i	f
m	0	m	n	1	l	b	a	k	h	f	e	g	c	i	j	d
n	0	n	l	a	h	e	c	j	1	m	k	b	g	f	d	i

Jedes Element aus $\mathbb{K}_{16}^{\times}$ kann als Potenz des primitiven Elementes $\gamma = \Omega$ geschrieben werden, und es gilt $\gamma^{15} = E$. In der Potenzschreibweise ist das Multiplizieren zweier Elemente speziell einfach. Es gilt nämlich das Potenzgesetz

$$\gamma^p \cdot \gamma^q = \gamma^{p+q \bmod 15},$$

In diesem Additionstheorem für $\mathbb{K}_{16}$ sind also die Exponenten modulo 15 zu reduzieren.

Jedes Element von $\mathbb{K}_{16}^{\times}$ hat die Ordnung $1, 3, 5$ oder 15, weil das die Teiler von 15 sind. Das einzige Element der Ordnung 1 ist das multiplikative Neutralelement 1. Die $\varphi(3) = 2$ Elemente e, f haben die Ordnung 3, die $\varphi(5) = 4$ Elemente g, i, k, n haben die Ordnung 5 und wegen $\varphi(15) = 8$ hat dieser Körper insgesamt die 8 Generatoren a, b, c, d, h, j, l, m.

Wegen $\varphi(5) = 4$ gibt es insgesamt 4 primitive 5-te Einheitswurzeln in $\mathbb{K}_{16}$. Es kommen die Elemente aus $\mathcal{P}_5 = \{g, i, k, n\}$ in Frage. Man beachte, dass wir den Körper $\mathbb{K}_4$ ebenfalls als Unterkörper von $\mathbb{K}_{16}$ auffassen können. Dazu benötigen wir die Fixelemente von $x \mapsto x^4$. Es sind $0, 1, e, f$. Als Einbettung $\mathbb{K}_4 \subset \mathbb{K}_{16}$ können wir also $\alpha \mapsto e = \gamma^5$ und $\alpha^2 \mapsto f = \gamma^{10}$ benutzen. Im Gegensatz dazu kann $\mathbb{K}_8$ nicht als Unterkörper von $\mathbb{K}_{16}$ aufgefasst werden.

Zur Konstruktion des Köpers $\mathbb{K}_{16}$ hätten wir auch eines der irreduziblen Polynome $f_2(x) = 1+x^3+x^4$ oder $f_3(x) = 1+x+x^2+x^3+x^4$ verwenden können. Um zu zeigen, dass sie tatsächlich irreduzibel sind, beachten wir zuächst, dass diese Polynome im Körper $\mathbb{Z}_2$ keine Nullstelle haben da in beiden Fällen $f(0) = f(1) = 1$ ist. Deshalb können sie also keinen Faktor vom Grad 1 haben. Diese Polynome haben aber auch keinen Faktor vom Grad 2 d.h. keine Faktorisierung der Art

$$(1 + ax + x^2) \cdot (1 + bx + x^2),$$

weil sonst nach dem Ausmultiplizieren in $\mathbb{Z}_2[x]$ die Gleichung

$$f(x) = 1 + (a + b)x + (ab)x^2 + (a + b)x^3 + x^4$$

gelten müsste. Ein Koeffizientenvergleich zeigt, dass das für $f_2(x)$ evident nicht der Fall ist. Für $f_3(x)$ müsste das nicht lineare Gleichungsystem

$$\left\{ \begin{array}{rcl} a + b & = & 1 \\ ab & = & 1 \end{array} \right.$$

erfüllt sein muss, was in $\mathbb{Z}_2$ evident nicht der Fall ist.

Aus ihren Nullstellen in $\mathbb{K}_{16}$ erhalten wir für die drei erwähnten irreduziblen Polynome die Zerlegungen

$$\begin{array}{rcl} f_1(x) & = & (x-\gamma^1)(x-\gamma^2)(x-\gamma^4)(x-\gamma^8) = 1 + x + x^4 \\ f_2(x) & = & (x-\gamma^7)(x-\gamma^{11})(x-\gamma^{13})(x-\gamma^{14}) = 1 + x^3 + x^4 \\ f_3(x) & = & (x-\gamma^3)(x-\gamma^6)(x-\gamma^9)(x-\gamma^{12}) = 1 + x + x^2 + x^3 + x^4 \end{array}$$

in Linearfaktoren. Das bereits im Körper $\mathbb{K}_4$ vollständig faktorisierte quadratische Polynom hat im Körper $\mathbb{K}_{16}$ die Faktorisierung

$$g(x) = (x - \gamma^5)(x - \gamma^{10}) = 1 + x + x^2$$

Daher kann der Körper $\mathbb{K}_{16}$ als Zerfällungskörper des Polynoms

$$x^{16} - x = (x-0)(x-1)(1+x+x^2)(1+x+x^4)(1+x^3+x^4)(1+x+x^2+x^3+x^4)$$

aufgefasst werden. ◯

Zur Illustration eines weiteren Phänomens soll noch ein endlicher Körper einer Charakteristik $p = 3$ konstruiert werden.

Beispiel. Der Körper $\mathbb{K}_9$ mit $9 = 3^2$ Elementen ist eine Erweiterung des Primkörpers $\mathbb{Z}_3$. Um ihn zu konstruieren, gehen wir vom irreduziblen, normierten Polynom $f(x) = 1 + x^2 \in \mathbb{Z}_3[x]$ vom Grad $d = 2$ aus. Es ist tatsächlich irreduzibel, weil es in $\mathbb{Z}_3$ keine Nullstelle besitzt.

Als Nullstelle wählen wir seine Begleitermatrix

$$\Omega = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} \in \mathbb{Z}_3^{2,2}$$

Sie erfüllt die charakteristische Gleichung

$$f(\Omega) = E + \Omega^2 = 0, \qquad \Omega^2 = -E = 2E$$

Die Elemente des gesuchten Körpers lassen sich wahlweise als $\mathbb{Z}_3$-Linearkombinationen der Basismatrizen E und Ω, als Polynome in $\mathbb{Z}_3[x]$ oder als Vektoren in $\mathbb{Z}_3^2$ darstellen und man erhält die Übersetzungstabelle

Matrix	Polynom	Vektor	Potenz	Ordnung
0	0	$(0,0)$	0	$-$
E	1	$(1,0)$	δ^0	1
$2E$	2	$(2,0)$	δ^4	2
Ω	x	$(0,1)$	δ^6	4
2Ω	$2x$	$(0,2)$	δ^2	4
$E+\Omega$	$1+x$	$(1,1)$	δ^1	8
$2E+\Omega$	$2+x$	$(2,1)$	δ^7	8
$E+2\Omega$	$1+2x$	$(1,2)$	δ^3	8
$2E+2\Omega$	$2+2x$	$(2,2)$	δ^5	8

Der gesuchte Körper hat also die $\mathbb{Z}_3$-Linearkombinationen

$$\mathbb{K}_9 = \{0, E, 2E, \Omega, 2\Omega, E+\Omega, 2E+\Omega, E+2\Omega, 2E+2\Omega\}$$

als Elemente. Sie entsprechen den 9 Matrizen aus $\mathbb{Z}_3^{2,2}$.

$$0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad 2E = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$
$$\Omega = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \quad 2\Omega = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}, \quad E+\Omega = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix},$$
$$2E+\Omega = \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix}, \quad E+2\Omega = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, \quad 2E+2\Omega = \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix}.$$

Zur vollständigen Beschreibung der Multiplikation

$$\mu\colon \mathbb{Z}_3^2 \times \mathbb{Z}_3^2 \to \mathbb{Z}_3^2, \qquad (\vec{a}, \vec{b}) \mapsto \mu(\vec{a}, \vec{b}) = \vec{c}$$

benötigen wir die Produkte der Standardbasisvektoren

$$\begin{array}{ll} \mu(\vec{e}_1, \vec{e}_1) = E & \mu(\vec{e}_1, \vec{e}_2) = \Omega \\ \mu(\vec{e}_2, \vec{e}_1) = \Omega & \mu(\vec{e}_2, \vec{e}_2) = 2E \end{array}$$

Sie nehmen in der Vektordarstellung die Form

$$\begin{array}{ll} \mu(\vec{e}_1, \vec{e}_1) = (1,0) & \mu(\vec{e}_1, \vec{e}_2) = (0,1) \\ \mu(\vec{e}_2, \vec{e}_1) = (0,1) & \mu(\vec{e}_2, \vec{e}_2) = (2,0) \end{array}$$

an. Daraus lesen wir die $2^3 = 8$ Strukturkonstanten $\mu(\vec{e}_i, \vec{e}_j)_k = (S_k)_{ij}$ ab, die wir als Elemente in den beiden Strukturmatrizen

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

zusammenfassen. Damit ist die Arithmetik im Körper $\mathbb{K}_9$ vollständig auf die Arithmetik in $\mathbb{Z}_3$ zurückgeführt. Der Produktvektor $\mu(\vec{a},\vec{b}) = \vec{c}$ besitzt die beiden Komponenten

$$\begin{aligned} c_0 = \langle \vec{a}, S_1 \cdot \vec{b} \rangle &= a_0 b_0 + 2 a_1 b_1 \\ c_1 = \langle \vec{a}, S_2 \cdot \vec{b} \rangle &= a_1 b_0 + a_0 b_1 \end{aligned}$$

Dieses Multiplikationsgesetz $\mu(\vec{a},\vec{b}) = \vec{c} = f(\vec{a}) \cdot \vec{b}$ lässt sich auch durch Multiplikation mit der quadratischen Matrix

$$f(\vec{a}) = \begin{pmatrix} a_0 & 2a_1 \\ a_1 & a_0 \end{pmatrix} = a_0 A_0 + a_1 A_1$$

beschreiben, die sich mit Hilfe der Matrizen des Systems $\{A_0, A_1\}$ linear kombinieren lässt.

Man beachte, dass diesmal die Begleitermatrix Ω des gewählten irreduziblen Polynoms kein Generator ist. Weil bereits die Potenz $\Omega^4 = E$ ist, hat Ω die Ordnung 4. Dieses Phänomen führt zu einem weiteren Begriff.

Definition. Das normierte, irreduzible Polynom $f(x) \in \mathbb{Z}_p[x]$ heisst *primitiv*, falls seine Begleitermatrix $\Omega \in \mathbb{Z}_p^{n,n}$ ein primitives Element ist.

Wir könnten nun zur Konstruktion der Körpers $\mathbb{K}_9$ eines der irreduziblen Polynom $g(x) = 2 + x + x^2$ oder $h(x) = 2 + 2x + x^2$ aus $\mathbb{Z}_3[x]$ benutzen, die beide primitiv sind. Weil dieser Körper $\varphi(8) = 4$ viele Elemente der maximalen Ordnung 8 d.h. primitive Elemente besitzen sollte, zeigt Versuch und Irrtum andereseits, dass wir diesmal etwa

$$\delta = E + \Omega = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$$

als Generator wählen können, weil dessen Ordnung 8 ist. Dieser Körper hat $\varphi(2) = 1$ Element der Ordnung 2 und $\varphi(4) = 2$ Elemente der Ordnung 2, wie man durch Berechnung der einzelnen Ordnungen bestätigt.

In der Exponentialdarstellung bezüglich δ hat die Additionstabelle die Form

$+$	0	1	δ	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7
0	0	1	δ^1	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7
1	1	δ^4	δ^7	δ^3	δ^5	0	δ^2	δ^1	δ^6
δ	δ^1	δ^7	δ^5	1	δ^4	δ^6	0	δ^3	δ^2
δ^2	δ^2	δ^3	1	δ^6	δ^1	δ^5	δ^7	0	δ^4
δ^3	δ^3	δ^5	δ^4	δ^1	δ^7	δ^2	δ^6	1	0
δ^4	δ^4	0	δ^6	δ^5	δ^2	1	δ^3	δ^7	δ^1
δ^5	δ^5	δ^2	0	δ^7	δ^6	δ^3	δ^1	δ^4	1
δ^6	δ^6	δ^1	δ^3	0	1	δ^7	δ^4	δ^2	δ^5
δ^7	δ^7	δ^6	δ^2	δ^4	0	δ^1	1	δ^5	δ^3

x	$-x$
0	0
1	δ^4
δ	δ^5
δ^2	δ^6
δ^3	δ^7
δ^4	1
δ^5	δ^1
δ^6	δ^2
δ^7	δ^3

und die Multiplikationstabelle lautet

$\cdot$	0	1	δ	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7
0	0	0	0	0	0	0	0	0	0
1	0	1	δ^1	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7
δ	0	δ^1	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7	1
δ^2	0	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7	1	δ^1
δ^3	0	δ^3	δ^4	δ^5	δ^6	δ^7	1	δ^1	δ^2
δ^4	0	δ^4	δ^5	δ^6	δ^7	1	δ^1	δ^2	δ^3
δ^5	0	δ^5	δ^6	δ^7	1	δ^1	δ^2	δ^3	δ^4
δ^6	0	δ^6	δ^7	1	δ^1	δ^2	δ^3	δ^4	δ^5
δ^7	0	δ^7	1	δ^1	δ^2	δ^3	δ^4	δ^5	δ^6

x	x^{-1}
0	–
1	1
δ	δ^7
δ^2	δ^6
δ^3	δ^5
δ^4	δ^4
δ^5	δ^3
δ^6	δ^2
δ^7	δ^1

Aus ihren Nullstellen in $\mathbb{K}_9$ erhalten wir für die drei erwähnten irreduziblen Polynome die Zerlegungen

$$f(x) = (x-\delta^2)(x-\delta^6), \quad g(x) = (x-\delta^1)(x-\delta^3), \quad h(x) = (x-\delta^5)(x-\delta^7)$$

in Linearfaktoren. Daher kann der Körper $\mathbb{K}_9$ als Zerfällungskörper des Polynoms

$$x^9 - x = (x-0)(x-1)(x-2)(x-\delta^2)(x-\delta^6)(x-\delta^1)(x-\delta^3)(x-\delta^5)(x-\delta^7)$$

aufgefasst werden. ◯

Primitive Polynome spielen in anderen Anwendungen eine Rolle.

Beispiel. Als Zufallszahl wird das Ergebnis von speziellen Zufallsexperimenten bezeichnet. Zufallszahlen werden bei verschiedenen statistischen Methoden benötigt, z. B. bei der Auswahl einer Stichprobe aus einer Grundgesamtheit, bei der Verteilung einer Population auf verschiedene Versuchsgruppen (Randomisierung), bei der Monte-Carlo-Simulation. Zur Erzeugung von Zufallszahlen gibt es verschiedene Verfahren. Diese werden als Zufallszahlengeneratoren bezeichnet. Ein entscheidendes Kriterium für Zufallszahlen ist, ob das Ergebnis der Generierung als unabhängig von früheren Ergebnissen angesehen werden kann oder nicht. Für andere Zwecke, z. B. bei der Erzeugung kryptographischer Schlüssel für den One-Time-Pad, werden hingegen echte Zufallszahlen benötigt.

Echte Zufallszahlen werden mit Hilfe physikalischer Phänomene erzeugt: Münzwurf, Würfel, Roulette, Rauschen elektronischer Bauelemente, radioaktive Zerfallsprozesse oder quantenphysikalische Effekte. Diese Verfahren heissen physikalische Zufallszahlengeneratoren, sind jedoch zeitlich oder technisch recht aufwändig.

In realen Anwendungen genügt häufig eine Folge von Pseudozufallszahlen, das sind scheinbar zufällige Zahlen, die nach einem festen, reproduzierbaren Verfahren erzeugt werden. Sie sind also nicht wirklich zufällig, da sie sich vorhersagen lassen, haben aber ähnliche statistische Eigenschaften (gleichmässige Häufigkeitsverteilung, geringe Korrelation) wie echte Zufallszahlenfolgen. Solche Verfahren nennt man Pseudozufallszahlengeneratoren. ◯

Nach so viel konkreter aber im Grunde genommen langweiliger Rechnerei, die man in der Praxis einer Maschine überlässt, wollen wir die wichtigsten Beobachtungen für beliebige endliche Köprer formulieren und dabei die meisten

Resultate für Primkörper aus der Zahlentheorie auf beliebige endliche Köper aus der Algebra verallgemeinern. Zentral ist der Umstand, dass sich die Teilbarkeitseigenschaften gewisser Binomialkoeffizienten vereinfachen.

Satz. Für jede Primzahl p und jede natürliche Zahl $1 \leq r \leq p-1$ gilt

$$\binom{p}{r} \equiv 0 \bmod p$$

Diese Binomialkoeffizienten sind also alle durch p teilbar.

Beweis. Für diesen Binomialkoeffizienten gilt

$$\binom{p}{r} = \frac{p \cdot (p-1) \cdots (p-r+1)}{r!}$$

und daher ist

$$r! \cdot \binom{p}{r} = p \cdot (p-1) \cdots (p-r+1)$$

Weil die rechte Seite evident durch p teilbar ist, muss es auch die linke sein und es gilt

$$r! \cdot \binom{p}{r} \equiv 0 \bmod p$$

Da aber jeder Faktor von

$$r! = 1 \cdot 2 \cdots r$$

strikt kleiner als p ist, kann die Primzahl p dieses Produkt nicht teilen und muss nach dem Hauptsatz der Arithmetik den anderen Faktor $\binom{p}{r}$ teilen, was behauptet wurde. □

Als Folge dieser Teilbarkeitseigenschaft nimmt der Binomialsatz über einem endlichen Körper eine unerwartet einfache Gestalt an.

Satz. In einem endlichen Körper $\mathbb{K}_q$ der Charakteristik p gilt für alle Elemente x, y der *vereinfachte Binomialsatz*

$$(x+y)^p = x^p + y^p$$

Beweis. In der üblichen Binomschen Formel, die in jedem Körper gilt,

$$(x+y)^p = x^p + \binom{p}{1} x^{p-1} y + \cdots + \binom{p}{p-1} x y^{p-1} + y^p$$

verschwinden nach der Teilbarkeitseigenschaft der Binomialkoeffizienten sämtliche Summanden mit Ausnahme des ersten und des letzten, woraus die Behauptung folgt. □

Verallgemeinern wir den vereinfachten Binomialsatz auf mehr als zwei Elemente, erhalten wir die Formel

$$(x_1 + x_2 + \cdots + x_n)^p = x_1^p + x_2^p + \cdots + x_n^p$$

Die vereinfachte Binomsche Formel besagt, dass sich in einem endlichen Körper $\mathbb{K}_q$ der Charakteristik p das Potenzieren $x \mapsto x^p$ nicht nur mit der Multiplikation, sondern auch mit der Additon verträgt und liefert daher einen Ringhomomorphismus $\mathbb{K}_q \to \mathbb{K}_q, x \mapsto x^p$. Er ist, wie jeder nicht triviale Ringhomomorphismus zwischen Körpern, injektiv. Weil die Körper zusätzlich endlich sind, ist jede injektive Abbildung auch surjektiv und deshalb ist dieser Ringhomomorphismus sogar umkehrbar, wie folgender konzeptionell wichtige Satz über endliche Körper besagt.

Satz. In jedem endlichen Köper $\mathbb{K}_q$ mit $q = p^d$ Elementen gibt es einen Automorphismus

$$\mathbb{K}_q \to \mathbb{K}_q, \qquad x \mapsto x^p$$

Dieser sogn. Frobenius-Automorphismus spielt im Umgang mit endlichen Körpern eine zentrale Rolle, weil er sämtliche Automorphismen auf übersichtliche Art erzeugt.

Satz. Es sei $\mathbb{K}_q$ ein endlicher Körper mit $q = p^m$ Elementen. Dann ist die Automorphismengruppe von $\mathbb{K}_q$ zyklisch der Ordnung m und wird vom Frobenius-Automorphismus erzeugt.

In der Zahlentheorie und ihren Anwendungen spielt der kleine Satz von Fermat eine zentrale Rolle. Er besagt, dass für jede ganze Zahl a die Beziehung

$$a^p \equiv a \bmod p$$

gilt. Das ist gleichwertig damit, dass für jedes Element im Primkörper $x \in \mathbb{Z}_p$ die Beziehung $x^p = x$ gilt. Der entsprechende Satz gilt in jedem endlichen Körper.

Satz. Die Einheitengruppe $\mathbb{K}_q^\times$ eines endlichen Körpers ist zyklisch der Ordnung $q - 1$. Insbesondere existiert ein Generator $\alpha \in \mathbb{K}_q$ so dass

$$\alpha^{q-1} = 1, \qquad \alpha^k \neq 1, \text{ für } 1 \leq k \leq q-2$$

ist. Der Körperelemente können also in der Exponentialform

$$\mathbb{K}_q = \{0, 1, \alpha, \alpha^2, \ldots, \alpha^{q-2}\}$$

aufgelistet werden.

Satz. In jedem endlichen Köper $\mathbb{K}_q$ mit $q = p^d$ Elementen gilt für jedes Element $r \in \mathbb{K}_q$ die Beziehung $r^q = r$.

Beweis. Da jedes Element $r \in \mathbb{K}_q$ die Form $r = \alpha^k$ hat, gilt nach der Definition des Generators

$$r^{q-1} = (\alpha^k)^{q-1} = (\alpha^{q-1})^k = 1$$

Multiplikation dieser Gleichung mit r liefert die behautptete Gleichung. □

Diese Gleichung wird oft als Ausgangspunkt der Theorie der endlichen Körper verwendet. Im endlichen Körper $\mathbb{K}_q = \{r_1, r_2, \ldots, r_q\}$ hat also das Polynom $f(x) = x^q - x$ vom Grad q genau q viele verschiedene Nullstellen. Deshalb

ist dieser Körper gross genug, dass dort dieses Polynom vollständig in Linearfaktoren zerfällt.

$$f(x) = x^q - x = (x - r_1) \cdot (x - r_2) \cdots (x - r_q) = \prod_{r_j \in \mathbb{K}_q} (x - r_j), \quad r_j \in \mathbb{K}_q$$

Wie erwähnt, wird in der Literatur zur Konstruktion des endlichen Körpers $\mathbb{K}_q$ meistens vom Polynom $f(x) = x^q - x$ ausgegangen und dann der endliche Körper $\mathbb{K}_q$ als Zerfällungskörper dieses Polynoms konstruiert. Dabei benötigt man die Theorie der Zerfällungskörper und muss mit Hilfe der Theorie der Polynome überprüfen, dass dieses Polynom keine mehrfachen Nullstellen hat. Daraus folgt dann, dass der endliche Körper bis auf Isomorphie, d.h. im Wesentlichen eindeutig bestimmt ist. Es gilt also folgender Klassifikationssatz endlicher Köper. Er besagt, dass zwei Köper mit gleich vielen Elementen im Wesentlichen gleich sind, d.h. ihre Additions- und Multiplikationstabellen stimmen — ev. nach einer geeigneten Umbezeichnung der Elemente — überein.

Satz. Je zwei endliche Körper mit gleich vielen Elementen sind isomorph.

Diese Satz ist insofern nicht konstruktiv als sich der behauptete Isomorphismus nicht aus den gewählten normierten, irreduziblen Polynomen ablesen lässt, die zur Konstruktion verwendet wurden.

Der Klassifikationssatz endlicher Köper rechtfertigt die Bezeichnungsweise $\mathbb{K}_q$ für einen endlichen Köper, der im Wesentlichen nur von der Anzahl $q = p^d$ seiner Elemente abhängt.

Aus der Eigenschaft $f(r_j) = 0$ für alle $r_j \in \mathbb{K}_q$ folgt, dass für das Polynoms

$$g(x) = 1 + f(x) = 1 + \prod_{r_j \in \mathbb{K}_q} (x - r_j) \in \mathbb{K}_q[x]$$

die Bedingung $g(r_j) = 1$. Dieses Beispiel zeigt, dass es über jedem endlichen Körper $\mathbb{K}_q$ Polynome gibt, die dort keine Nullstelle haben. Möchte man sicher stellen, dass über dem endlichen Körper $\mathbb{K}_q$ für $q = p^d$ *jedes* Polynom in Linearfaktoren zerfällt, muss man zum *algebraischen Abschluss* $\overline{\mathbb{Z}}_p$ übergehen. Um ihn zu konstruieren beachtet man, dass der endliche Körper $\mathbb{K}_{p^{d_1}}$ genau dann als Teilkörper von $\mathbb{K}_{p^{d_2}}$ aufgefasst werden kann, wenn d_1 ein Teiler von d_2 ist. In diesem Fall erhalten wir also die Körpererweiterung $\mathbb{K}_{p^{d_1}} \subseteq \mathbb{K}_{p^{d_2}}$. Die Vereinigung dieses Systems ineinander geschachtelter endlicher Körper liefert den gesuchten algebraischen Abschluss $\overline{\mathbb{Z}}_p$, der als Körper allerdings nicht mehr endlich ist, da er alle Körper der Form $\mathbb{K}_{p^d}$ der Charakteristik p enthält. Weil der Frobenius-Automorphismus mit den Körpererweiterungen vertauschbar ist, lässt er sich auf den algebraischen Abschluss zum Automorphismus

$$F\colon \overline{\mathbb{Z}}_p \to \overline{\mathbb{Z}}_p$$

erweitern und der endliche Körper $\mathbb{K}_{p^d}$ ergibt sich als Fixpunktmenge der d-ten Potenz F^d.

Auch der Existenzsatz endlicher Körper mit $q = p^d$ vielen Elementen ist nicht ganz einfach. Algebraiker konstruieren solche Körper, wie erwähnt, als Zerfällungskörper des Polynoms $f(x) = x^q - x$. Bei unserem matriziellen Vorgehen müssen wir sicherstellen, dass ein irreduzibles Polynom $f(x) \in \mathbb{Z}_q[x]$ vom

Grad d existiert. Konventionell wird ein solches Polynom als Minimalpolynom des endlichen Körpers über seinem Primkörper gewählt. Dabei muss man allerdings voraussetzen, dass der betreffende Körper bereits existiert und benötigt die Idealtheorie des Polynomrings $\mathbb{Z}_p[x]$. Weil wir nicht in dieser allgmeinen Art vorgegangen sind und um Zirkelschlüsse zu vermeiden, benötigen wir ein unabhängiges Argument.

Weil unsere Konstruktion endlicher Köper von der Wahl eines normierten, irreduziblen Polynom in $f(x) \in \mathbb{Z}_p[x]$ abhängt, stellen sich unmittelbar zwei Fragen.

1. Wie entscheidet man, ob ein Polynom $f(x) \in \mathbb{Z}_p[x]$ irreduzibel ist?
2. Gibt es solche Polynome immer?

Die Antwort auf die erste Frage, d.h. nach der Faktorzerlegung von Polynomen ist, wie beim Faktorisieren ganzer Zahlen, kein einfaches Geschäft. Es gibt zwar Algorithmen zur Faktorisierung von Polynomen aus $\mathbb{Q}[x]$ und aus $\mathbb{Z}_p[x]$ — wie jene von Kronecker und Berlekamp. Sie sind aber nicht effizient und liefern nicht die gewünschten Einsichten. Ihre nicht ganz einfach Beschreibung findet man in der Literatur.

Immerhin kann die Existenz der benötigten irreduziblen Polynome garantiert werden. Dazu verwendet man den Umstand, dass man die irreduziblen Polynome von einem festen Grad über einem Primkörper abzählen kann.

Zur Formulierung des kombinatorischen Resultates benötigt man die Möbius-Funktion $\mu\colon \mathbb{N} \to \{-1, 0, 1\}$. Sie liefert den Wert $\mu(n) = 1$, falls n ein Produkt einer geraden Anzahl verschiedener Primfaktoren ist und den Wert -1 falls n eine Produkt einer ungeraden Anzahl verschiedener Primfaktoren ist. Ferner ist $\mu(n) = 0$, falls n mehrfache Primfaktoren hat. Sie dient zum Invertieren einer Funktion

$$g\colon \mathbb{N} \to \mathbb{C}, \qquad n \mapsto g(n) = \sum_{t|n} f(t)$$

wobei $f\colon \mathbb{N} \to \mathbb{C}$ irgend eine Funktion und t ein Teiler von n ist. Dann gilt nämlich die Inversionsformel

$$f(n) = \sum_{t|n} \mu(t) g\Big(\frac{n}{t}\Big)$$

Damit erhält man das gewünschte Resultat.

Satz. Die Anzahl normierter, irreduzibler Polynome vom Grad r im Polynomring $\mathbb{Z}_p[x]$ über einem Primkörper beträgt

$$N(r, q) = \frac{1}{r} \sum_{t|r} \mu(t) p^{\frac{r}{t}}$$

Von diesen Polynomen sind $\frac{\varphi(p^r-1)}{r}$ viele primitiv.

Zur Illustration dieses Satzes gehen wir nochmals die betrachteten Beispiele durch.

Beispiel. Im Polynomring $\mathbb{Z}_2[x]$ gibt es für $r = 2$ genau

$$N(2, 2) = \frac{1}{2}\Big(\mu(1)2^2 + \mu(2)2\Big) = \frac{1}{2}\Big(2^2 - 2\Big) = 1$$

normiertes irreduzibles quadratische Polynom. Aus der Faktorisierung von

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

in $\mathbb{Z}_2[x]$ lesen wir das primitive Polynom $f(x) = 1 + x + x^2$ ab, das wir oben zur Konstruktion des Körpers $\mathbb{K}_4$ benutzt haben. ◯

Beispiel. Die Anzahl kubischer, normierter irreduzibler Polynome im Polynomring $\mathbb{Z}_2[x]$ beträgt mit $r = 3$

$$N(2,3) = \frac{1}{3}\Big(\mu(1)2^3 + \mu(3)2\Big) = \frac{1}{3}\Big(2^3 - 2\Big) = 2$$

Aus der Faktorisierung

$$x^8 - x = x(x+1)(1+x+x^3)(1+x^2+x^3)$$

lesen wir die beiden primitiven Polynome

$$f_1(x) = 1 + x + x^3, \qquad f_2(x) = 1 + x^2 + x^3$$

ab. Wir haben sie oben zu zwei isomorphen Konstruktionen des endlichen Körpers $\mathbb{K}_8$ benutzt. ◯

Beispiel. Die Anzahl normierter, irreduzibler Polynome vierten Grades im Polynomring $\mathbb{Z}_2[x]$ beträgt mit $r = 4$

$$N(4,2) = \frac{1}{4}\Big(\mu(1)2^4 + \mu(2)2^2 + \mu(4)2\Big) = \frac{1}{4}\Big(2^4 - 2^2\Big) = 3$$

Aus der Faktorisierung

$$x^{16} - x = x(1+x)(1+x+x^2)(1+x+x^4)(1+x^3+x^4)(1+x+x^2+x^3+x^4)$$

entnimmt man die drei irreduziblen Polynome vierten Grades

$$f_1(x) = 1 + x + x^4, \qquad f_2(x) = 1 + x^3 + x^4, \qquad f_3(x) = 1 + x + x^2 + x^3 + x^4$$

Wir haben $f_1(x)$ zur Konstruktion des endlichen Körpers $\mathbb{K}_{16}$ benutzt. Das Polynom $f_3(x)$ ist nicht primitiv, weil aus der charakteristischen Gleichung

$$x^4 = 1 + x + x^2 + x^3$$

die Beziehung

$$x^5 = xx^4 = x(1+x+x^2+x^3) = x+x^2+x^3+x^4 = x+x^2+x^3+(1+x+x^2+x^3) = 1$$

folgt. Die beiden anderen Polynome sind primitiv. ◯

Beispiel. Im Polynomring $\mathbb{Z}_3[x]$ gibt es für $r = 2$ genau

$$N(2,3) = \frac{1}{2}\Big(\mu(1)3^2 + \mu(2)3\Big) = \frac{1}{2}\Big(3^2 - 3\Big) = 3$$

normierte, irreduzibles quadratische Polynome. Aus der Faktorisierung

$$(x^9 - x) = x(x+1)(x+2)(1+x^2)(2+x+x^2)(2+2x+x^2)$$

erhält man

$$f(x) = 1 + x^2, \qquad g(x) = 2 + x + x^2, \qquad h(x) = 2 + 2x + x^2$$

Wir haben $f(x)$ zur Konstruktion des endlichen Körpers $\mathbb{K}_9$ benutzt. Die anderen beiden sind primitiv. ◯

Beispiel. Im Polynomring $\mathbb{Z}_2[x]$ gibt es für $r = 16$ genau

$$\begin{aligned} N(16,2) &= \frac{1}{16}\Big(\mu(1)2^{16} + \mu(2)2^8 + \mu(4)2^4 + \mu(8)2^2 + \mu(16)2^1\Big) \\ &= \frac{1}{16}(2^{16} - 2^8) = 4080 \end{aligned}$$

normierte, irreduzible Polynome vom Grad 16. Davon sind $\frac{\varphi(2^{16}-1)}{16} = 2048$ primitiv. ◯

Beispiel. Einen expliziten Isomorphismus

$$T\colon \mathbb{K}_{65'536} \to \mathbb{K}_{65'536}$$

zu konstruieren, der zu den beiden irreduziblen Polynomen

$$\begin{aligned} f_1(x) &= x^{16} + x^{15} + x^{13} + x^4 + 1 \\ f_2(x) &= x^{16} + x^{14} + x^{13} + x^{12} + x^{11} + x^{10} + x^9 + x^7 + x^3 + x + 1 \end{aligned}$$

vom Grad 16 aus dem Polynomring $\mathbb{Z}_2[x]$ gehört, ist allerdings eine andere Sache. Insbesondere scheint es in der Menge der normierten irreduziblen Polynomen keinen kanonischen Vertreter zu geben. Zur Reduzierung des Rechenaufwands ist es vorteilhaft, ein irreduzibles Polynom mit einer minimalen Anzahl von Null verschiedenen Koeffizienten zu wählen. ◯

Um ein irreduzibles Polynom in $\mathbb{Z}_p[x]$ vom Grad r zu konstruieren, kann man sich für grossen Grad r nicht auf sein Glück verlassen. Aus obigem Satz folgt nämlich, dass die Wahrscheinlichkeit, dass ein beliebiges Polynom aus $\mathbb{Z}_p[x]$ vom Grad r irreduzibel ist, nahe bei $\frac{1}{r}$ ist. Um ein normiertes, irreduzibles Polynom vom Grad r in $\mathbb{Z}_p[x]$ zu finden, braucht man also etwa $O(r)$ Versuche. Mit den bekannten Faktorisierungsalgorithmen bedeutet es im allgemeinen einen gewaltigen Aufwand, ein gegebenes normiertes Polynom $f(x) \in \mathbb{Z}_p[x]$ vom Grad r auf Irreduzibilität zu testen. Braucht man irreduzible Polynome für grossen Grad r, muss man die Fachliteratur zu Rate ziehen.

Aus der bekannten Anzahl irreduzibler Polynome ergibt sich folgende Abschätzung mit Hilfe der geometrischen Reihe.

Korollar. Für die Anzahl normierter, irreduzibler Polynome vom Grad r im Polynomring $\mathbb{Z}_p[x]$ über einem Primkörper gilt

$$N(r,q) \geq \frac{1}{n}\Big(q^n - \frac{q^n - 1}{q - 1}\Big) > 0$$

Dieses Resultat garantiert die Existenz der benötigten Polynome.

Satz. Für jede natürliche Zahl d gibt es mindestens ein irreduzibles Polynom vom Grad d in $\mathbb{Z}_p[x]$.

Insbesondere folgt aus diesem Satz, dass für jede Primzahlpotenz $q = p^d$ tatsächlich ein endlicher Körper $\mathbb{K}_q$ mit q Elementen existiert.

Satz. Zu jeder Primzahl p und jeder natürlichen Zahl $d \geq 1$ gibt es einen endlichen Körper mit $q = p^d$ Elementen.

Dieser Existenzsatz sagt nicht, wie ein solcher Körper mit Hilfe eines irreduziblen Polynoms $f(x) \in \mathbb{Z}_p[x]$ vom Grad d effizient konstruiert werden kann. Um nämlich ein solches Polynom zu finden, bleibt einem nicht viel anderes übrig, als einen der mehr oder weniger effizienten Algorithmen zum Faktorisieren von Polynomen auf das Polynom $x^q - x$ anzuwenden.

Kapitel 3

Lineare Gleichungssysteme

3.1 Die Struktur der Lösungsräume

Nachdem wir die Begriffe der Matrizenalgebra kennen, wenden wir uns nun der vertieften Untersuchung linearer Gleichungssysteme mit m Gleichungen in n Unbekannten zu. Als Abkürzung eines solche Systems

$$\left\{\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array}\right. \qquad \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right)$$

benutzen wir selbstverständlich die erweiterte Matrix und setzen den Matrizen-Kalkül gleich nutzbringend ein.

Zunächst betrachten wir einige konkrete Beispiele und schauen uns die Form der Lösung genau an, ohne uns primär um die Frage zu kümmern, wie man sie findet. Wenn wir dann wissen, was wir wollen, werden wir leicht ein effizientes Verfahren (Algorithmus) angeben können, mit dem wir das Gewünschte in endlich vielen, genau definierten, Schritten erreichen.

Um in den Beispielen nicht zu viele Indizes schreiben zu müssen und damit den Überblick zu verlieren, benutzen wir für kleine Systeme mit wenig Unbekannten Symbole wie x, y, z für die Unbekannten.

Beispiel. Das System

$$\left\{\begin{array}{l} 4x - y + 3z = -1 \\ 3x + y + 9z = -4 \end{array}\right. \qquad \left(\begin{array}{ccc|c} 4 & -1 & 3 & -1 \\ 3 & 1 & 9 & -4 \end{array}\right)$$

kürzen wir mit der rechts daneben stehenden erweiterten Matrix ab. Es besitzt $x = 1$, $y = 2$, $z = -1$ als Lösung, da diese Werte *beide* Gleichungen des Systems erfüllen, wie man durch Einsetzen leicht überprüft. Aber auch die drei Zahlen $x = -11$, $y = -25$, $z = 6$ sind Lösungen des Systems. Hingegen bilden die Zahlen $x = 1$, $y = 8$, $z = 1$, die zwar die erste Gleichung erfüllen, keine Lösung des Systems. Durch Einsetzen überprüft man ebenfalls leicht, dass die drei Terme $x = 1 - 12t$, $y = 2 - 27t$, $z = -1 + 7t$ Lösungen des Systems sind. Weil damit das System für irgend einen Wert des Parameters $t \in \mathbb{R}$ gelöst wird, hat dieses System also unendlich viele numerische Lösungen, die alle durch die

drei Formeln beschrieben werden. Wir werden feststellen, dass *alle* numerischen Lösungen dieses Systems so parametrisiert werden können. Die formale Lösung lässt sich in vektorieller Form schreiben.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + t \begin{pmatrix} -12 \\ -27 \\ 7 \end{pmatrix} \qquad (t \in \mathbb{R})$$

Für $t = 0$ und $t = 1$ ergeben sich die beiden erwähnten numerischen Spezialfälle. Im allgemeinen besteht die Lösung eines linearen Gleichungssystems also aus gewissen *Formeln* für die Unbekannten und diese endlich vielen Formeln hängen von gewissen Parametern ab und sind vom ersten Grad. Deshalb können sie vektoriell geschrieben und interpretiert werden. Nur in den Spezialfällen aus der Schule, wo meistens sogn. reguläre Gleichungssysteme aufgetaucht sind, bestehen diese Lösungsformeln aus einzelnen Zahlen.

Geometrisch beschreiben die beiden ursprünglichen Gleichungen je eine Ebene im Raum $\mathbb{R}^3$. Die Lösungsmenge des linearen Gleichungssystems kann als Schnittgerade dieser beiden Ebenen interpretiert werden und die angegebene Vektorgleichung beschreibt diese Gerade als Parametergleichung. Die Lösung des Gleichungssystems läuft darauf hinaus, dass wir von der Beschreibung der Schnittgeraden mit Hilfe von Koordinatengleichungen zur Beschreibung mit Hilfe von Parametergleichungen übergegangen sind. ◯

Nicht jedes lineare System besitzt Lösungen.

Beispiel. Zum Beispiel widersprechen sich die beiden folgenden Gleichungen:

$$\begin{cases} x + y = 4 \\ x + y = 3 \end{cases} \qquad \left(\begin{array}{cc|c} 1 & 1 & 4 \\ 1 & 1 & 3 \end{array}\right)$$

Daher ist dieses System *unlösbar*. Ein System, das keine Lösung besitzt, heisst *inkonsistent*.

Geometrisch interpretiert, beschreibt dieses Gleichungssystem zwei parallele Geraden in der Ebene $\mathbb{R}^2$ oder zwei parallele Ebenen im Raum $\mathbb{R}^3$. Die algebraische Tatsache, dass das Gleichungssystem keine Lösungen hat, widerspiegelt sich geometrisch im Umstand, dass die Geraden bzw. die Ebenen parallel sind, d.h. keine Schnittpunkte haben. ◯

Die folgenden linearen Gleichungssysteme sind konsistent.

Definition. Sind in einem linearen Gleichungssystem alle Konstanten 0, so heisst das System *homogen*; anderenfalls heisst es *inhomogen*[1]. Ersetzt man in einem linearen Gleichungssystem alle Konstanten durch 0, so nennt man das entstehende lineare Gleichungssystem das *zugehörige homogene* System. Es lautet also im allgemeinen Fall in Komponenten ausgeschrieben

$$\begin{cases} a_{11}z_1 + a_{12}z_2 + \cdots + a_{1n}z_n = 0 \\ a_{21}z_1 + a_{22}z_2 + \cdots + a_{2n}z_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}z_1 + a_{m2}z_2 + \cdots + a_{mn}z_n = 0 \end{cases} \qquad \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} & 0 \end{array}\right)$$

[1] und nicht etwa heterogen.

In der Matrizensprache lautet das zum linearen Gleichungssystem $A\vec{x} = \vec{b}$ gehörende homogene lineare System $A\vec{z} = \vec{0}$.

Ein homogenes System ist konsistent, da es die Lösung hat, die aus lauter Nullen besteht, wie man durch Einsetzen von $z_1 = 0$, $z_2 = 0, \cdots, z_n = 0$ bestätigt. Der Nullvektor ist also immer Lösung eines homogenen Systems, da $A \cdot \vec{0} = \vec{0}$ ist. Diese Lösung heisst die *triviale* Lösung des homogenen Systems. Gibt es noch weitere Lösungen des homogenen Systems, so werden sie als *nichttrivial* bezeichnet. Nichttriviale Lösungen sind in diesem Zusammenhang interessant. Es ist aber in der Regel nicht trivial zu entscheiden, ob ein gegebenes homogenes Gleichungssystem solche nichttrivialen Lösungen hat.

Beispiel. In unserem bereits oben gelösten Beispiel lautet das zugehörige homogene lineare Gleichungssystem

$$\left\{ \begin{array}{l} 4x - y + 3z = 0 \\ 3x + y + 9z = 0 \end{array} \right. \qquad \left(\begin{array}{ccc|c} 4 & -1 & 3 & 0 \\ 3 & 1 & 9 & 0 \end{array} \right)$$

Es hat neben der trivialen Lösung noch die drei formalen Lösungen:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = t \begin{pmatrix} -12 \\ -27 \\ 7 \end{pmatrix} \qquad (t \in \mathbb{R})$$

da die drei Terme $x = -12t$, $y = -27t$, $z = 7t$ tatsächlich Lösungen des zugehörigen homogenen Systems sind. Für $t = 0$ erhalten wir die triviale Lösung und jede Wahl $t \neq 0$ liefert eine nichttriviale numerische Lösung. Damit haben wir in unserem Beispiel unendlich viele nichttriviale Lösungen des zugehörigen homogenen Systems gefunden.

Aus geometrischer Sicht, beschreibt das zugehörige homogene Gleichungssystem zwei Ebenen, die zu den ursprünglichen beiden parallel sind, aber den Ursprung enthalten. Entsprechend handelt es sich bei der zugehörigen Lösungsmenge um die Parameterdarstellung einer Gerade durch den Ursprung, die zur Schnittgeraden der beiden ursprünglichen Ebenen parallel ist. Umgekehrt geht die Lösungsmenge des früher gelösten inhomogenen Systems aus der Lösung des zugehörigen homogenen Systems durch Parallelverschieben um eine Partikulärlösung hervorgeht. ◯

Bereits dieses eine Beispiel lässt erahnen, dass zwischen einem linearen Gleichungssystem und seinem zugehörigen homogenen System ein enger Zusammenhang besteht. Wir wollen diesen Zusammenhang genau untersuchen, da er theoretische Einsichten liefert und beim Studium linearer Differentialgleichungen eine wichtige Rolle spielt und dort für den Anfänger vor lauter Analysis weniger durchsichtig ist. Dazu gehen wir von einem linearen Gleichungssystem $A \cdot \vec{x} = \vec{b}$ aus. Für seinen Lösungsraum schreiben wir $L(A, \vec{b})$. Entsprechend ist $L(A, \vec{0})$ der Lösungsraum des zugehörigen homogenen Gleichungssystems $A \cdot \vec{x} = \vec{0}$.

Ein inhomogenes lineares Gleichungssystem braucht nicht immer eine Lösung zu haben. Aber *falls* es eine Lösung hat, kann man alle Lösungen gut beschreiben. Dieser Sachverhalt folgt aus den folgenden beiden Sätzen, die zwar simpel zu beweisen, aber zentral für die Struktur der Lösungsmenge linearer Gleichungssysteme sind.

Satz. Falls $\vec{x}$ Lösung eines linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ und $\vec{z}$ eine Lösung des zugehörigen homogenen Gleichungssystems $A \cdot \vec{z} = \vec{0}$ ist, so ist die Summe $\vec{z} + \vec{x}$ Lösung des linearen Gleichungssystems.

Beweis. Nach Voraussetzung gelten die beiden Gleichungen $A \cdot \vec{z} = \vec{0}$ und $A \cdot \vec{x} = \vec{b}$. Addiert man diese beiden Gleichungen, so ergibt sich mit den Regeln für das Rechnen mit Matrizen $A \cdot \vec{z} + A \cdot \vec{x} = A \cdot (\vec{z} + \vec{x}) = \vec{0} + \vec{b} = \vec{b}$, was die Behauptung zeigt. □

Satz. Falls $\vec{x}$ und $\vec{y}$ Lösungen eines linearen Gleichungssystems sind, so ist die Differenz $\vec{x} - \vec{y}$ Lösung des zugehörigen homogenen Gleichungssystems.

Beweis. Nach Voraussetzung gilt $A \cdot \vec{x} = \vec{b}$ und $A \cdot \vec{y} = \vec{b}$. Subtrahieren wir diese beiden Gleichungen, ergibt sich: $A \cdot \vec{x} - A \cdot \vec{y} = A \cdot (\vec{x} - \vec{y}) = \vec{b} - \vec{b} = \vec{0}$, was die Behauptung zeigt. □

Diese beiden Sätze haben die wichtige Konsequenz, dass man sämtliche Lösungen eines linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ kennt, falls bekannt sind:

1. Eine einzelne[2] Lösung $\vec{x}_0$ des Systems $A \cdot \vec{x}_0 = \vec{b}$.
2. Sämtliche Lösungen $\vec{z}$ des zugehörigen homogenen Systems $A \cdot \vec{z} = \vec{0}$.

Dann sind nämlich alle Lösungen des Systems $A \cdot \vec{x} = \vec{b}$ von der Form $\vec{z} + \vec{x}_0$. Für die Lösungsräume gilt also die Beziehung

$$L(A, \vec{b}) = \vec{x}_0 + L(A, \vec{0})$$

Die Lösungsräume $L(A, \vec{b})$ linearer Gleichungssysteme sind also Translate des Lösungsraumes $L(A, \vec{0})$ des zugehörigen homogenen Systems.

Korollar. Ist ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ konsistent, so sind äquivalent:

1. Das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ ist eindeutig lösbar.
2. Das zugehörige homogene System $A \cdot \vec{x} = \vec{0}$ hat nur die triviale Lösung.

In erster Linie müssen wir also homogene lineare Gleichungssysteme untersuchen und einen Weg finden, um einzelne partikuläre Lösungen zu finden. Diese Resultate werden beim Studium linearer Differential- und Integralgleichungen in der Analysis eine zentrale Rolle spielen und erlauben neben einer Einsicht in die Stuktur der Lösungsmenge eines Gleichungssystems Fragen nach der Existenz und der Eindeutigkeit der Lösung voneinander zu trennen.

Beispiel. Die läppisch einfache Idee der letzten beiden Sätze hat wichtige Konsequenzen für das Studium von linearen Netzwerken. Dort entsteht das zugehörige homogene Gleichungungssystem dadurch, dass man die Energiequellen abschaltet. Hat man einmal dieses einfachere Netzwerk verstanden, braucht man dann nur noch eine einzige partikuläre Lösung für das gegebene Netzwerk zu suchen, was mit den zu besprechenden Methoden mindestens dann nicht all zu

[2] sogn. partikuläre Lösung.

schwierig ist, wenn die Spannungsquellen des Netzwerkes entweder Gleichstromquellen (Gleichstromkreis) sind oder aber wenn die Spannungsquellen Wechselspannungen, d.h. allgemeine Kosinus-Funktionen (Wechselstromkreis) sind. ◯

Als nächstes stellen wir fest, dass homogene lineare Gleichungssysteme angenehmer zu untersuchen sind als beliebige lineare Gleichungssysteme, weil man für sie leicht aus bekannten Lösungen neue herstellen kann. Die Lösungsräume homogener linearer Gleichungssysteme haben eine für die lineare Algebra fundamentale Struktur.

Satz. Für die Lösungen eines homogenen linearen Gleichungssystems gilt:

1. Falls $\vec{x}$ und $\vec{y}$ zwei Lösungen eines homogenen linearen Gleichungssystems sind, so ist auch die Summe $\vec{x} + \vec{y}$ Lösung dieses homogenen linearen Gleichungssystems.

2. Falls $\vec{x}$ eine Lösung eines homogenen linearen Gleichungssystems ist, dann ist auch das Vielfache $r \cdot \vec{x}$ Lösung des homogenen Gleichungssystems. Dabei steht r für irgend einen Skalar.

Man fasst diese beiden Eigenschaften dadurch zusammen, dass man sagt, der Lösungsraum $L(A, \vec{0})$ eines homogenen linearen Gleichungssystems bilde einen *Vektorraum*.

Beweis. Das Resultat ergibt sich ebenfalls leicht durch Anwenden der Rechenregeln für Matrizen: Falls $A \cdot \vec{x} = \vec{0}$ und $A \cdot \vec{y} = \vec{0}$ gilt, so ist auch $A(\vec{x} + \vec{y}) = A \cdot \vec{x} + A \cdot \vec{y} = \vec{0}$, was die erste Behauptung zeigt. Ferner ist auch $A \cdot (r\vec{x}) = r(A \cdot \vec{x}) = \vec{0}$, was die zweite Behauptung zeigt. □

Beispiel. Die Beweisidee dieses Satzes hat eine wichtige Konsequenz für das Studium von linearen Wechselstromkreisen und allgemeiner für das Lösen von linearen Differential- und Integralgleichungen. Mit etwas Erfahrung und geeigneten Ansätzen kann man sich dort in der Regel leicht einige nichttriviale Basis-Lösungen des zugehörigen homogenen Systems verschaffen. Auf Grund unseres Satzes kann man daraus unendlich viele weitere Lösungen dieses Systems zusammenbauen — die Anwender reden von der Superposition von Basislösungen. Mit etwas Glück stellt sich dann sogar heraus, dass dies bereits alle Lösungen des betreffenden homogenen Systems sind. ◯

Aus diesem Resultat folgt übrigens, dass der Lösungsraum eines homogenen linearen Gleichungssystems wirklich sämtliche Regeln eines Vektorraumes erfüllt. Die Überprüfung der Axiome sei dem Leser als Übung empfohlen. Man überlege sich an einem Beispiel, warum die entsprechende Aussage für ein nicht homogenes linearen Gleichungssystem falsch ist.

Als Folge dieser Sätze ergibt sich sofort, dass die im letzten Abschnitt vorgekommenen Lösungsverhalten von linearen Gleichungssystemen alle möglichen Fälle abdecken.

Zunächst ergibt sich aus dem letzten Satz, dass ein homogenes lineares Gleichungssystems mit *einer* nicht trivialen Lösung gleich unendlich viele nicht triviale Lösungen haben muss, falls uns unendlich viele Skalare zur Verfügung

stehen. Wir können ja eine solche Lösung mit irgend einem der Skalare multiplizieren und erhalten eine neue Lösung.
Daraus folgt aber mit den ersten beiden Sätzen, dass ein lineares Gleichungssystem, das zwei verschiedene Lösungen hat, gleich unendlich viele Lösungen haben muss. Die Differenz der beiden Lösungen ist nämlich eine nicht triviale Lösung des zugehörigen homogenen linearen Gleichungssystems.

Satz. Besitzt ein lineares Gleichungssystem zwei verschiedene Lösungen, so besitzt es unendlich viele Lösungen, falls unendlich viele Skalare zur Verfügung stehen.

Das Lösungsverhalten linearer Gleichungssystem wird also durch einen der folgenden drei disjunkten Fälle beschrieben.

Korollar. Ein lineares Gleichungssystem hat entweder:

1. Keine Lösung.
2. Genau eine Lösung.
3. Unendlich viele Lösungen.

falls unendlich viele Skalare zur Verfügung stehen.

Zum Schluss dieses Überblickes über das Lösungsverhalten linearer Gleichungssysteme wollen wir noch überlegen, dass man in gewissen Fällen Partikulärlösungen von bereits gelösten linearen Gleichungssystemen zu Partikulärlösungen von weiteren linearen Gleichungssystemen zusammenbauen kann.

Satz. Für ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ gilt:

1. Falls $\vec{x}_0$ eine Partikulärlösung des linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}_1$ und $\vec{y}_0$ eine Partikulärlösung des linearen Gleichungssystems $A \cdot \vec{y} = \vec{b}_2$ ist, so ist $\vec{x}_0 + \vec{y}_0$ eine Partikulärlösung des linearen Gleichungssystems $A \cdot \vec{z} = \vec{b}_1 + \vec{b}_2$.
2. Falls $\vec{x}_0$ eine Partikulärlösung des linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ ist, so ist $r \cdot \vec{x}_0$ eine Partikulärlösung des linearen Gleichungssystems $A \cdot \vec{z} = r\vec{b}$.

Beweis. Eine weiter Anwendung der Rechenregeln für Matrizen zeigt auch diese beiden Behauptungen. Falls $A \cdot \vec{x}_0 = \vec{b}_1$ und $A \cdot \vec{y}_0 = \vec{b}_2$ gilt, so ist $A \cdot (\vec{x}_0 + \vec{y}_0) = A \cdot \vec{x}_0 + A \cdot \vec{y}_0 = \vec{b}_1 + \vec{b}_2$, was die erste Behauptung zeigt.
Falls $A \cdot \vec{x}_0 = \vec{b}$ ist, so ist $A \cdot (r \cdot \vec{x}_0) = r \cdot (A \cdot \vec{x}_0) = r \cdot \vec{b}$, was die zweite Aussage beweist. □

Beispiel. Auch die Idee dieses Resultates kann in der Netzwerktheorie benutzt werden. Kennt man die Reaktion eines Wechselstromkreises auf Wechselspannungen d.h. auf allgemeine Kosinus-Funktionen, so kann man damit auch die Reaktion auf beliebige zeitabhängige Spannungsquellen berechnen. Dazu baut man die beschreibende Funktionen der Spannungsquellen als Linearkombination der grundlegenden allgemeinen Kosinus-Funktionen zusammen. Das gelingt mit Hilfe der Fourier-Theorie. ◯

Der Leser ist explizit aufgefordert, sich die soeben aufgelisteten Resultate über das Lösungsverhalten linearer Gleichungssysteme geometrisch zu veranschaulichen und an einfachen numerischen Beispielen zu überprüfen.

3.2 Stufenform

Nachdem wir wissen, wie die Struktur der Lösung eines linearen Gleichungssystems aussieht, machen wir uns daran, diese Lösung algorithmisch zu bestimmen. Dieser Algorithmmus basiert, wie viele andere, auf einer natürlichen Idee folgender Art: Man beginnt mit einer endlichen Beschreibung (Präsentierung) einer in der Regel unendlichen algebraischen Struktur und modifiziert dann diese Beschreibung durch eine endliche Anzahl von *Operationen*, unter denen die algebraische Struktur *invariant* bleibt. Der Algorithmus besteht aus einer endlichen Liste solcher Operationen, die in einer ganz bestimmten Reihenfolge durchgeführt werden. Dabei wird die ursprüngliche Beschreibungsform der algebraischen Struktur eine einfachere Form annehmen, so dass wir am Schluss die gewünschten Antworten über die algebraische Struktur direkt aus der vereinfachten Beschreibungsform ablesen können. Die *Korrektheit* eines solchen Algorithmus ergibt sich also aus der Invarianz der Operationen und sein *Abbruch* in endlich vielen Schritten ergibt sich aus den gewählten Operationen. Die *Effizienz* ergibt sich aus einer möglichst einfacher Liste solcher Operationen.

Die grundlegende algorithmische Lösungsmethode zum Lösen beliebiger linearer Gleichungssysteme basiert auf der Idee, dass man das gegebene System durch ein neues ersetzt, dessen Lösungsmenge mit der des ursprünglichen Systems übereinstimmt, aber leichter zu bestimmen ist.

Offenbar stellen sich zwei grundlegende Fragen:

1. Was darf man mit linearen Gleichungssystemen tun, ohne die Lösung zu ändern?

2. Wie sehen Gleichungssysteme aus, deren Lösung leicht zu bestimmen ist?

Die Antwort auf die erste Frage wird uns als Nebeneffekt ein systematisches Verfahren zum Abändern linearer Gleichungssystemen liefern, das die gesuchte Information konserviert. Die Antwort auf die zweite Frage wird uns zeigen, in welche Richtung wir das System abändern müssen und garantieren, dass dieses Verfahren nach endlich vielen Schritten abbricht. Am Schluss werden wir einen Algorithmus zur Lösung linearer Gleichungssysteme haben.

Die Schwierigkeit beim Lösen linearer Gleichungssysteme ist die gute Organisation des Lösungs- d.h. des Vereinfachungsprozesses. Man muss darauf achten, dass nicht Gleichungen mehrfach benutzt werden d.h. dass man nicht „im Kreis herum rechnet“. Wir müssen insbesondere sicher sein, dass das Verfahren im Laufe der Zeit abbricht. Eine gute Organisation des Lösungsvorganges ist auch für die maschinelle Lösung linearer Gleichungssysteme unverzichtbar. Sie liefert ferner Einsichten geometrischer und theoretischer Natur und erlaubt es, die Komplexität, d.h. den maximalen Rechenaufwand abzuschätzen, der zur Lösung eines linearen Gleichungssystemes erforderlich ist. Aus diesen Gründen müssen wir über die Methoden hinausgehen, die in der Schule aus Zeitgründen besprochen werden (Einsetzen, Gleichsetzen und was es dort an Mystik mehr gibt) und die nur für sehr kleine (bis drei Variablen) Gleichungssysteme brauchbar sind.

Wir beginnen mit der zweiten Frage und wollen an Hand ausgewählter Beispiele erkennen, was mit „leichter“ gemeint sein könnte.

Beispiel. Die Lösung des linearen Gleichungssystems

$$\left\{\begin{array}{rrrr} 3x+4y-5z=50 \\ 2y+7z=20 \\ 5z=10 \end{array}\right. \qquad \left(\begin{array}{ccc|c} 3 & 4 & -5 & 50 \\ 0 & 2 & 7 & 20 \\ 0 & 0 & 5 & 10 \end{array}\right)$$

kann man durch *Rückwärtseinsetzen* gleich ablesen. Die dritte Gleichung liefert $z = 2$. Setzen wir diesen, nun bekannten, Wert in die zweite Gleichung ein, können wir den Wert von y berechnen. Aus der Gleichung $2y + 14 = 20$ folgt $y = 3$. In analoger Weise ergibt das Einsetzen dieser beiden Werte in die erste Gleichung $3x + 12 - 10 = 50$. Daraus ergibt sich der Wert $x = 16$. Der einzige Lösungsvektor für dieses System ist

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 16 \\ 3 \\ 2 \end{pmatrix}$$

Geometrisch handelt es sich um einen 0-dimensionalen Punkt im $\mathbb{R}^3$, der als Schnittpunkt der drei Ebenen aufgefasst werden kann. ◯

Beispiel. Auch bei folgendem Beispiel ist sofort ersichtlich, wie es durch Rückwärtseinsetzen zu lösen ist:

$$\left\{\begin{array}{rrrr} x & +4u=-1 \\ y & +2u=6 \\ z & +3u=2 \end{array}\right. \qquad \left(\begin{array}{cccc|c} 1 & 0 & 0 & 4 & -1 \\ 0 & 1 & 0 & 2 & 6 \\ 0 & 0 & 1 & 3 & 2 \end{array}\right)$$

Wir nennen hier die erste von Null verschiedene Zahl einer Zeile den *führenden Koeffizienten* und die zugehörige Variable die führende Variable. Lösen wir nach den führenden Variablen auf, ergibt sich durch Rückwärtseinsetzen der Reihe nach

$$\begin{array}{rr} x= & -1-4u \\ y= & 6-2u \\ z= & 2-3u \end{array}$$

Da u einen beliebigen Wert t annehmen kann — man sagt, die Variable u sei *frei* — so hat das System unendlich viele Lösungen. Für die Lösung ergibt sich $x = -1 - 4t$, $y = 6 - 2t$, $z = 2 - 3t$, $u = t$. Man beachte, dass die Lösung eines linearen Gleichungssystems im allgemeinen nicht aus einer Liste numerischer Zahlen, sondern allgemeiner aus einem Liste symbolischer linearer Terme besteht. In vektorieller Form lautet diese Lösung

$$\begin{pmatrix} x \\ y \\ z \\ u \end{pmatrix} = \begin{pmatrix} -1 \\ 6 \\ 2 \\ 0 \end{pmatrix} + t \begin{pmatrix} -4 \\ -2 \\ -3 \\ 1 \end{pmatrix}$$

Geometrisch handelt es sich in diesem Beispiel beim Lösungsraum um eine 1-dimensionale Gerade im $\mathbb{R}^4$. Es ist die Schnittgerade der drei ursprünglich gegebenen Hyperebenen. ◯

Beispiel. Als weiteres Beispiel, wo wir die Lösung gleich ablesen können, betrachten wir das System

$$\left\{\begin{array}{rrrr} x+6y & +4v=-2 \\ z & +3v=1 \\ u & +5v=2 \end{array}\right. \qquad \left(\begin{array}{ccccc|c} 1 & 6 & 0 & 0 & 4 & -2 \\ 0 & 0 & 1 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 5 & 2 \end{array}\right)$$

Die führenden Variablen sind hier x, z, u. Die restlichen Variablen y, v sind frei. Lösen wir nach den führenden Variablen auf, so ergibt sich

$$\begin{aligned} x &= -2 - 6y - 4v \\ z &= \quad 1 \qquad - 3v \\ u &= \quad 2 \qquad - 5v \end{aligned}$$

Nun kommen auf der rechten Seite nur noch freie Variablen vor. Diese freien Variablen y und v können beliebige Werte s und t annehmen. Die Lösung lautet $x = -2 - 6s - 4t$, $y = s$, $z = 1 - 3t$, $u = 2 - 5t$, $v = t$. Sie kann vektoriell in der Form

$$\begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} + t \begin{pmatrix} -4 \\ 0 \\ -3 \\ -5 \\ 1 \end{pmatrix} + s \begin{pmatrix} -6 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

geschrieben werden. Man beachte, dass der erste Vektorsummand eine Partikulärlösung des ursprünglichen Gleichungssystems und die beiden anderen Vektoren Lösungen des zugehörigen homogenen Gleichungssystems sind. Geometrisch handelt es sich um eine 2-dimensionale Ebene im $\mathbb{R}^5$. ◯

Wir haben soeben gesehen, wie leicht es ist, ein lineares Gleichungssystem zu lösen, wenn seine Matrix eine spezielle Form hat. Das Spezielle besteht darin, dass in der erweiterten Matrix für jede Zeile gilt: Sind die ersten s Elemente der Zeile Null, so sind für alle folgenden Zeilen mindestens die ersten $s+1$ Elemente Null. Solchen linearen Gleichungssystemen bzw. den zugehörigen Matrizen geben wir einen Namen.

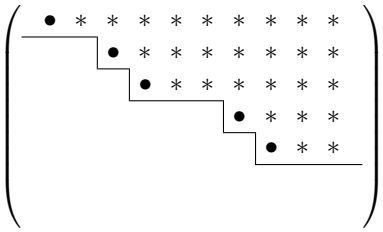
Definition. Ein lineares Gleichungssystem ist in *Stufenform*, wenn gilt:

1. In zwei aufeinanderfolgenden Zeilen, die nichtverschwindende Elemente besitzen, steht der führende Koeffizient der unteren Zeile rechts vom führenden Koeffizient der oberen Zeile.

2. Alle Zeilen, die ausschliesslich Nullen haben, stehen am unteren Rand des Systems.

Wir nennen die Anzahl Zeilen, die nicht Null sind, den *Rang* des linearen Gleichungssystems bzw. der Matrix.

Das folgende Schema soll eine Matrix in Stufenform andeuten:



Elemente unterhalb der Treppen-Stufen sind Null. Die mit einem schwarzen Punkt angedeuteten führenden Elemente sind verschieden von Null, und die mit einem Stern markierten Elemente sind beliebig.

Ein Blick auf die Koeffizientenmatrix der soeben erfolgreich gelösten Gleichungssysteme zeigt, dass dort noch mehr gilt.

Definition. Hat die Stufenform die zusätzliche Eigenschaft

3. Sämtliche führenden Koeffizienten sind 1.

sagt man, das lineare Gleichungssystem sei in *normierter Stufenform*.

Das zugehörige Schema hat dann also die Form

$$\left(\begin{array}{cccccccccc} 1 & * & * & * & * & * & * & * & * & * \\ & 1 & * & * & * & * & * & * & * & * \\ & & 1 & * & * & * & * & * & * & * \\ & & & & & & 1 & * & * & * \\ & & & & & & & 1 & * & * \\ & & & & & & & & & \end{array}\right)$$

in der alle mit einem schwarzen Punkt versehenen Elemente durch eine 1 ersetzt sind. Selbstverständlich kann eine Stufenform normiert werden, indem man jede Zeile durch den führenden Koeffizienten (der ja nicht 0 ist!) dividiert. Beim Normieren ändert sich die Lösungsmenge nicht.

Ein weiterer Blick auf die Koeffizientenmatrix der soeben erfolgreich gelösten Gleichungssysteme zeigt, dass dort noch mehr gilt.

Definition. Hat die Stufenform die zusätzliche Eigenschaft

4. Eine Spalte, die ein führendes Element enthält, hat keine weiteren von Null verschiedenen Einträge.

sagt man, das lineare Gleichungssystem sei in *reduzierter Stufenform*.

Das zugehörige Schema sieht diesmal wie folgt aus:

$$\left(\begin{array}{cccccccccc} \bullet & * & 0 & 0 & * & * & 0 & 0 & * & * \\ & \bullet & 0 & * & * & 0 & 0 & * & * & \\ & & \bullet & * & * & 0 & 0 & * & * & \\ & & & & & & \bullet & 0 & * & * \\ & & & & & & & \bullet & * & * \\ & & & & & & & & & \end{array}\right)$$

Selbstverständlich lassen sich die beiden Bedingungen kombinieren. Man sagt dann, dass Gleichungssystem sei in *normierter, reduzierter Stufenform*.

Das zugehörige Schema sieht diesmal wie folgt aus:

$$\left(\begin{array}{cccccccccc} 1 & * & 0 & 0 & * & * & 0 & 0 & * & * \\ & 1 & 0 & * & * & 0 & 0 & * & * & \\ & & 1 & * & * & 0 & 0 & * & * & \\ & & & & & & 1 & 0 & * & * \\ & & & & & & & 1 & * & * \\ & & & & & & & & & \end{array}\right)$$

Die entsprechenden Begriffe verwendet man auch für die zugehörigen erweiterten Matrizen.

1. Stufenform: Das führende Element der Zeile $i+1$ steht rechts vom führenden Element der Zeile i.

2. Normierte Stufenform: Das führende Element jeder Zeile ist 1.

3. Reduzierte Stufenform: Alle Elemente oberhalb eines führenden Elementes sind Null.

Wie bei jedem neuen Muster muss der Leser jetzt sein Gehirn für diese speziellen Matrizenformen konditionieren. Dazu dienen die folgenden Beispiele.

Beispiel. Folgende Matrizen haben Stufenform:

$$\begin{pmatrix} 3 & 4 & -5 & 50 \\ 0 & 2 & 7 & 20 \\ 0 & 0 & 5 & 10 \end{pmatrix}, \quad \begin{pmatrix} 2 & 3 & -8 & 5 & 1 & -2 \\ 0 & 4 & 7 & 2 & 0 & 6 \\ 0 & 0 & 0 & 0 & -2 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

und den Rang 3. Sie sind weder normiert noch reduziert. Man zeichne die Treppenstruktur ein. Mit unterschiedlichen Farben markiere man die führenden und die freien Variablen. ◯

Beispiel. Folgende Matrizen sind in reduzierter aber nicht normierter Stufenform:

$$\begin{pmatrix} 2 & 0 & 0 & 4 \\ 0 & 3 & 0 & 7 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & -5 & -2 & 0 & 1 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Man zeichne die Treppenstruktur ein und markiere die führenden und die freien Elemente mit unterschiedlichen Farben. Dann hebe man die Nullen oberhalb der führenden Koeffizienten hervor. ◯

Beispiel. Folgende Matrizen sind in normierter Stufenform:

$$\begin{pmatrix} 1 & 0 & 3 & 7 \\ 0 & 1 & 6 & 2 \\ 0 & 0 & 1 & 5 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 1 & -1 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{pmatrix}$$

Die rechte ist zusätzlich reduziert hingegen die linke nicht. Man hebe in jedem Beispiel die spezielle Struktur wie in den vorangehenden Beispiel hervor. ◯

Wie unsere Beispiele gezeigt haben, kann man ein lineares Gleichungssystem, das in Stufenform vorliegt, durch Rückwärtseinsetzen einfach lösen. Dazu geht man nach folgendem (provisorischen) Algorithmus für das Rückwärtseinsetzen vor.

1. Falls freie Variablen vorkommen, weise ihnen neue Symbole (Parameter) zu.

2. Löse die Gleichungen nach den führenden Variablen auf.

3. Setze von unten nach oben fortschreitend jede Gleichung in die übrigen ein.

Die beliebigen Werte, die den freien Variablen zugewiesen werden, wollen wir zur besseren Übersicht von den Variablen unterscheiden und mit den Buchstaben $r, s, t \ldots$ bezeichnen. Allerdings kann jeder Buchstabe, der nicht schon als Variablenname vergeben ist, verwendet werden.

Beispiel. Gehen wir von der erweiterten Matrix in Stufenform aus.

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Man beachte, dass diese Matrix nicht reduziert, aber normiert ist.

Das zugehörige lineare Gleichungssystem lautet:

$$\left\{\begin{array}{rrrrrrr} x_1 + 3x_2 - 2x_3 & & + 2x_5 & & = 0 \\ & x_3 + 2x_4 & & + 3x_6 & = 1 \\ & & & x_6 & = \frac{1}{3} \end{array}\right.$$

Die führenden Variablen sind x_1, x_3, x_6. Die freien Variablen sind x_2, x_4, x_5. Wir bezeichnen sie von nun an zur besseren Übersicht mit den neuen Symbolen $x_2 = r, x_4 = s, x_5 = t$. Damit nimmt das Gleichungssystem die Form

$$\left\{\begin{array}{rrrrrrr} x_1 + 3r - 2x_3 & & + 2t & & = 0 \\ & x_3 + 2s & & + 3x_6 & = 1 \\ & & & x_6 & = \frac{1}{3} \end{array}\right.$$

an. Löst man nach den führenden Variablen auf, ergibt sich:

$$\begin{array}{rrrrrr} x_1 = & -3r + 2x_3 & & - 2t & \\ x_3 = 1 & & - 2s & & - 3x_6 \\ x_6 = \frac{1}{3} & & & & \end{array}$$

Man beachte, dass auf der rechten Seite immer noch Variablen vorkommen, die nicht frei sind. Das hängt damit zusammen, dass die Koeffizientenmatrix nicht reduziert ist. Das Ziel des Rückwärtseinsetzen besteht darin, alle Variablen auf der rechten Seite durch freie Variablen auszutauschen.

Einsetzen der untersten Zeile x_6 in alle oberen Gleichungen liefert:

$$\begin{array}{rrrr} x_1 = & -3r + 2x_3 & - 2t & \\ x_3 = & & - 2s & \\ x_6 = \frac{1}{3} & & & \end{array}$$

Immer noch kommen auf der rechten Seiten Variablen vor, die nicht frei sind. Einsetzen von x_3 in die erste Gleichung liefert

$$\begin{array}{rrrr} x_1 = & -3r - 4s - 2t & & \\ x_3 = & - 2s & & \\ x_6 = \frac{1}{3} & & & \end{array}$$

Jetzt kommen auf der rechten Seite nur noch freie Variablen vor und wir haben das vorliegende Gleichungssystem vollständig gelöst.

Bringt man die rechte Seite auf die linke Seite, so lautet das entstande Gleichungssystem

$$\begin{cases} x_1 + 3r \quad + 4s + 2t \quad = 0 \\ \qquad\quad x_3 + 2s \qquad\quad = 0 \\ \qquad\qquad\qquad\quad x_6 = \frac{1}{3} \end{cases}$$

Man beachte, dass die erweiterte Matrix dieses Gleichungssystem

$$\left(\begin{array}{cccccc|c} 1 & 3 & 0 & 4 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

nun reduzierte Stufenform hat. Rückwärtseinsetzen führt also eine Matrix in Stufenform in reduzierte Stufenform über.

Die allgemeine Lösung unseres Gleichungssystems lautet nun

$$\begin{aligned} x_1 &= -3r - 4s - 2t \\ x_2 &= r \\ x_3 &= -2s \\ x_4 &= s \\ x_5 &= t \\ x_6 &= \tfrac{1}{3} \end{aligned}$$

und kann in vektorieller Form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix} + r \begin{pmatrix} -3 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -4 \\ 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Geometrisch handelt es sich um einen 3-dimensionalen Teilraum im $\mathbb{R}^6$. ◯

3.3 Das Eliminationsverfahren

Wir haben gesehen, wie man ein lineares Gleichungssystem lösen kann, dessen erweiterte Matrix Stufenform hat. Es stellt sich also die Frage, ob wir ein beliebiges lineares Gleichungssystem stets so umformen können, dass gilt:

1. Der Lösungsraum ändert sich beim Umformen nicht.
2. Das neue System hat Stufenform.

Das ist in der Tat möglich. Bevor wir den Algorithmus angeben können, müssen wir überlegen, welche Operationen man mit einem linearen Gleichungssystem durchführen darf, ohne dass sich der Lösungsraum ändert. Eine solche Operation muss selbstverständlich strukturverträglich d.h. in unserem Fall *linear* sein. Zusätzlich muss sie *umkehrbar* sein, weil sich sonst der Lösungsraum ändern würde.

Als grundlegende lineare Operationen kommen in Frage:

Satz. Der Lösungsraum eines linearen Gleichungssystems ändert sich nicht, wenn man eine der folgenden drei Operationen durchführt:

I. *Vertauschen* von zwei Gleichungen.

II. *Addition* eines Vielfachen einer Gleichung zu einer anderen Gleichung und Ersetzen dieser Gleichung durch die Summe.

III. *Multiplikation* einer Gleichung mit einem von Null verschieden Skalar.

Da die Zeilen der erweiterten Matrix den Gleichungen des zugehörigen Systems entsprechen, liefern diese Umformungen die zugehörigen *elementare Zeilenumformungen* der erweiterten Matrix.

I. Vertauschen von zwei Zeilen.

II. Addition eines Vielfachen einer Zeile zu einer anderen Zeile und Ersetzen dieser Zeile durch die Summe.

III. Multiplikation einer Zeile mit einem von Null verschieden Skalar.

Dieser Satz erfordert entweder blindes Nachrechnen oder etwas Verständnis. Wir werden ihn später in einem allgemeineren Rahmen beweisen, nachdem wir uns von seiner Nützlichkeit überzeugt haben. Er ist der Schlüssel zur Lösung linearer Gleichungssysteme.

Man beachte, dass es bei der Zeilenumformung vom Typ II im Gegensatz zu jener vom Typ III keine Rolle spielt, ob der Faktor verschieden von 0 ist. Das bewirkt, dass die Zeilenumformungen vom Typ II in dieser Hinsicht angenehmer sind, weil man nämlich nicht immer Sonderfälle berücksichtigen muss, falls die Koeffizientenmatrix Parameter enthält.

Für das Lösen linearer Gleichungssysteme gilt das fundamentale Resultat[3]:

Satz. (Elimination)

1. Jede Matrix $A \in \mathbb{R}^{m,n}$ lässt sich durch eine endliche Folge von Zeilenoperationen vom Typ I und II in reduzierte Stufenform bringen.

2. Jede Matrix $A \in \mathbb{R}^{m,n}$ lässt sich durch eine endliche Folge von Zeilenoperationen vom Typ I,II,III in normierte, reduzierte Stufenform $S(A)$ bringen.

[3] sogn. Gauss'sche Elimination; Die Chinesen kannten offenbar im 2. Jhd. v. Chr. diesen Algorithmus bereits. In einem Lehrbuch wird die Mathematik jener Zeit zusammengefasst. Darin findet man auch einen Algorithmus für das Lösen linearer Gleichungssysteme mit beliebig vielen Gleichungen: Methode des Fang-Cheng.

3. Die normierte, reduzierte Stufenform $S(A)$ ist eindeutig bestimmt d.h. unabhängig von der Wahl der Zeilenoperationen.

Wir wollen den Satz nicht formal beweisen, sondern den Beweis in Form eines Algorithmus an einigen typischen Beispielen praktisch vorführen.

Beispiel. Wir gehen vom linearen Gleichungssystem

$$\begin{cases} x_1 + 3x_2 - 2x_3 + 2x_5 = 0 \\ 2x_1 + 6x_2 - 5x_3 - 2x_4 + 4x_5 - 3x_6 = -1 \\ 5x_3 + 10x_4 + 15x_6 = 5 \\ 2x_1 + 6x_2 + 8x_4 + 4x_5 + 18x_6 = 6 \end{cases}$$

aus. Die zugehörige erweiterte Matrix lautet:

$$(A \mid \vec{b}) = \left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 2 & 6 & -5 & -2 & 4 & -3 & -1 \\ 0 & 0 & 5 & 10 & 0 & 15 & 5 \\ 2 & 6 & 0 & 8 & 4 & 18 & 6 \end{array}\right)$$

Die Koeffizientenmatrix hat noch nicht Stufenform. In der Vorwärtsphase des Algorithmus stellen wir zunächst Stufenform her und führen über jede durchgeführte Zeilenoperation in eckigen Klammern rechts daneben sorgfältig Buch. Dafür benützen wir folgende Abkürzungen:

I. Vertauschen der i-ten und der j-ten Zeile: T_{ij}

II. Addition des r-fachen der i-ten Zeile zur j-ten Zeile: $Z_{ij}(r)$

III. Multiplikation der Zeile i mit dem Faktor $r \neq 0$: $S_i(r)$

Addition des (-2)-fachen der ersten Zeile zur zweiten und vierten Zeile liefert:

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & -3 & -1 \\ 0 & 0 & 5 & 10 & 0 & 15 & 5 \\ 0 & 0 & 4 & 8 & 0 & 18 & 6 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ \\ Z_{14}(-2) \end{array}\right]$$

Hier ist automatisch ein Treppenstufe der Breite 2 entstanden. Addition des 5-fachen der zweiten Zeile zur dritten und des 4-fachen der zweiten Zeile zur vierten Zeile liefert eine weitere Trepppenstufe, die diesmal die Breite 3 hat.

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 2 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(5) \\ Z_{24}(4) \end{array}\right]$$

Damit die entstandene Nullzeile am unteren Rand steht, vertauschen wir die dritte und die vierte Zeile und erhalten die gesuchte Stufenform

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 6 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ T_{43} \\ T_{34} \end{array}\right]$$

Die Vorwärtsphase des Algorithmus ist abgeschlossen. Wir können die führenden und die freien Variablen identifizieren. Der Rang der Matrix ist 3.

Statt wie in der Literatur üblich, nun zu normieren und uns dabei Brüche einzuhandeln, mit denen wir dann Bruch rechnen, sorgen wir nur dafür, dass die ganzen Zahlen in jedem Schritt betragsmässig so klein wie möglich werden. Das erreichen wir dadurch, dass wir jede Zeile durch ihren grössten gemeinsamen Teiler dividieren.

Multiplikation der dritten Zeile mit $\frac{1}{2}$ liefert die ganzzahlige Stufenform

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ S_3(\frac{1}{2}) \\ \\ \end{array}\right]$$

Hätten wir die Stufenform normiert, indem wir die zweite Zeile mit (-1) und die dritte Zeile mit $\frac{1}{3}$ multipliziert hätten, würde die normierte Stufenform

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ S_2(-1) \\ S_3(\frac{1}{3}) \\ \\ \end{array}\right]$$

Brüche erhalten. Um nicht Bruchrechnen zu müssen, verschieben wir das Normieren so lange wie möglich.

Weil es sich bei der letzten Matrix um die selbe Matrix handelt, von der wir im letzten Abschnitt ausgegangen sind, könnte man man jetzt das lineare Gleichungssystem, wie besprochen, durch Rückwärtseinsetzen fertig lösen. Es ist unangenehm, wenn man für die Rückwärtsphase — wie oben — mit Gleichungssystemen handieren und symbolisch rechnen muss. Das lässt sich aber zum Glück vermeiden. Wir wollen den Datentyp der Matrizen beibehalten und das Rückwärtseinsetzen auch matriziell durchführen. Der Leser vergleiche die folgende Berechnung der reduzierten Stufenform mit dem seinerzeit durchgeführten Vorgehen beim Rückwärtseinsetzen. Wir kehren zur ganzzahligen Stufenform zurück.

Addition der dritten Zeile zur zweiten zerstört die Stufenform nicht und liefert

$$\left(\begin{array}{cccccc|c} 1 & 3 & -2 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{32}(1) \\ \\ \end{array}\right]$$

Addition des (-2)-fachen der zweiten Zeile zur ersten liefert schliesslich die ganzzahlige reduzierte Stufenform

$$\left(\begin{array}{cccccc|c} 1 & 3 & 0 & 4 & 2 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(-2) \\ \\ \\ \end{array}\right]$$

Jetzt ist es immer noch früh genug, die erhaltene ganzzahlige, reduzierte Stufenform zu normieren. Multiplikation der zweiten Zeile mit (-1) und der dritten Zeile mit $\frac{1}{3}$ liefert die gesuchte normierte, reduzierte Stufenform

$$\left(\begin{array}{cccccc|c} 1 & 3 & 0 & 4 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ S_2(-1) \\ S_3(\frac{1}{3}) \\ \\ \end{array}\right]$$

die dem Gleichungssystem

$$\left\{\begin{array}{rrrrr} x_1 + 3x_2 & & + 4x_4 + 2x_5 & & = 0 \\ & & x_3 + 2x_4 & & = 0 \\ & & & & x_6 = \frac{1}{3} \end{array}\right.$$

entspricht. Die letzte Gleichung liefert keine Information, da sie immer erfüllt ist. Deshalb haben wir sie weggelassen. Durch Auflösen der Gleichungen nach den führenden Variablen ergibt wie früher die gesuchte Lösung

$$\begin{array}{rl} x_1 = & -3x_2 - 4x_4 - 2x_5 \\ x_3 = & \quad -2x_4 \\ x_6 = & \frac{1}{3} \end{array}$$

Diese Lösung kann wie dort vektoriell geschrieben werden.

Mit einer verfeinerten Version des in diesem Abschnitt besprochenen Eliminationsverfahrens findet man in diesem Beispiel für die verwendete Koeffizientenmatrix A die ganzzahlige Faktorisierung

$$U \cdot A \cdot V = S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

wobei die beiden unimodularen Transformationsmatrizen U, V gegeben sind durch

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ -10 & 4 & 0 & 1 \\ -10 & 5 & 1 & 0 \end{pmatrix}, \qquad V = \begin{pmatrix} 1 & 2 & -6 & -3 & -2 & -4 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -3 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Aus dieser ganzzahligen Normalform von A folgt, dass das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ keine ganzzahlige Partikulärlösung $\vec{x}$ haben kann. Für sie müsste nämlich

$$\vec{x} = V \cdot \vec{y}, \qquad S \cdot \vec{y} = U \cdot \vec{b} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ -10 & 4 & 0 & 1 \\ -10 & 5 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -1 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}$$

gelten, was offensichtlich ganzzahlig nicht möglich ist. In der Tat muss

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \frac{1}{3} \\ 0 \\ 0 \\ 0 \end{pmatrix} + t_1 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + t_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

gelten, was bis auf Permutation der berechnete Lösung

$$\vec{x} = V \cdot \vec{y} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix} + t_1 \begin{pmatrix} -3 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t_3 \begin{pmatrix} -4 \\ 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

des Systems $A \cdot \vec{x} = \vec{b}$ entspricht. ◯

In Zukunft werden wir also das Rückwärtseinsetzen, wie soeben gezeigt, matriziell durchführen. Im nächsten Beispiel soll der vollständige Algorithmus an einem simplen Beispiel vorgeführt werden.

Beispiel. Wir gehen vom linearen Gleichungssystem

$$\begin{cases} x + \ y + 2z = 9 \\ 2x + 4y - 3z = 1 \\ 3x + 6y - 5z = 0 \end{cases}$$

aus. Die zugehörige erweiterte Matrix lautet

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 2 & 4 & -3 & 1 \\ 3 & 6 & -5 & 0 \end{array}\right)$$

Addition des (-2)-fachen der ersten Zeile zur zweiten und des (-3)-fachen zur dritten Zeile liefert:

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 2 & -7 & -17 \\ 0 & 3 & -11 & -27 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ Z_{13}(-3) \end{array}\right]$$

Zur Lösung mit dem Taschenrechner würde man vielleicht normieren und dazu die Zeile mit $\frac{1}{2}$ multiplizieren und

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 1 & -\frac{7}{2} & -\frac{17}{2} \\ 0 & 3 & -11 & -27 \end{array}\right) \qquad \left[\begin{array}{c} \\ S_2(\frac{1}{2}) \\ \end{array}\right]$$

erhalten. Addition des (-3)-fachen der zweiten Zeile zur dritten liefert die Stufenform

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 1 & -\frac{7}{2} & -\frac{17}{2} \\ 0 & 0 & -\frac{1}{2} & -\frac{3}{2} \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(-3) \end{array}\right]$$

Multiplikation der dritte Zeile mit (-2) liefert die normierte Stufenform

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 1 & -\frac{7}{2} & -\frac{17}{2} \\ 0 & 0 & 1 & 3 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ S_3(-2) \end{array}\right]$$

Sie ist noch nicht reduziert. Das Rückwärtseinsetzen erledigen wir durch das Berechnen der reduzierte Stufenform. Addition des $\frac{7}{2}$-fachen der dritten Zeile zur zweiten und des (-2)-fachen der dritten zur ersten Zeile ergibt

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right) \qquad \left[\begin{array}{c} Z_{31}(-2) \\ Z_{32}(\frac{7}{2}) \\ \end{array}\right]$$

Addition des (-1)-fachen der zweiten Zeile zur ersten Zeile ergibt schliesslich die gewünschte normirete, reduzierte Srufenform

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(-1) \\ \\ \end{array}\right]$$

Daraus kann nun die gesuchte Lösung des System sofort abgelesen werden. Die eindeutig bestimmte Lösung ist hier $x = 1$, $y = 2$, $z = 3$.

Geometrisch haben wir diesmal den eindeutig bestimmten Schnittpunkt von drei Ebenen bestimmt. ◯

Man beachte, dass beim Eliminations-Algorithmus im Verlauf der Rechnung unangenehme Brüche auftreten können, wenn man zu früh normiert, auch wenn das Ergebnis schliesslich ganzzahlig ist. Das Bruchrechnen kann vermieden werden, wenn man die Normierung bis zum Schluss hinausschiebt. Für die Handrechnung und für viele Anwendungen, bei denen es auf numerische Präzision ankommt, ist es zweckmässiger, ganzzahlig zu rechnen und so lange wie möglich mit betragsmässig möglichst kleinen ganzen Zahlen zu arbeiten. Dazu dividieren wir vor jeder elementaren Zeilenumformung jede Zeile durch ihren grössten gemeinsamen Teiler und verschieben das Normieren der Zeilen auf den Schluss, weil dann mit den Brüchen nicht mehr weitergerechnet werden muss.

Beispiel. Wir kehren zum letzten Beispiel zurück und gehen von der letzen Matrix aus, die ganzzahlige Einträge hatte.

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 2 & -7 & -17 \\ 0 & 3 & -11 & -27 \end{array}\right)$$

Statt nun zu normieren, schlagen wir diesmal einen anderen Weg ein. Wir addieren das (-3)-fache der zweiten Zeile zum 2-fachen der dritten Zeile. Man beachte, dass wir dabei zwei der Elementeroperationen zu einer neuen Operation zusammenfassen.

IV. Addition eines Vielfachen einer Zeile zu einem von Null verschieden Vielfachen einer anderen Zeile und Ersetzen dieser anderen Zeile durch die Summe.

Auch diese gemischte Operation kürzen wir ab.

IV. Addition des r-fachen der i-ten Zeile zum s-fachen der j-ten Zeile: $ZS_{ij}(r,s)$, falls $s \neq 0$ ist.

Man beachte, dass hier nur der zweite Faktor s von Null verschieden sein muss.
Die entstehende ganzzahlige Matrix lautet in unserem Fall

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 9 \\ 0 & 2 & -7 & -17 \\ 0 & 0 & -1 & -3 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ ZS_{23}(-3,2) \end{array}\right]$$

Auch hier normieren wir die zweite Zeile nun nicht. Addition des (-7)-fachen der dritten Zeile zur zweiten und des 2-fachen zur ersten Zeile liefert die Matrix

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & -1 & -3 \end{array}\right) \qquad \left[\begin{array}{c} Z_{31}(2) \\ Z_{32}(-7) \\ \end{array}\right]$$

Damit die Zahlen so klein wie möglich bleiben, dividieren jeweils die erhaltene Zeile durch den grössten gemeinsamen Teiler ihrer Elemente. In unserem Fall dividieren wir die zweite Zeile also durch $\mathrm{ggT}(2,4) = 2$. Dazu brauchen wir offenbar einen efffizienten Algorithmus zum Berechnen des grössten gemeinsame Teilers zweier ganzen Zahlen m, n. Das lässt sich mit dem Euklid'schen Algorithmus erreichen.
Zusätzlich normieren wir die dritte Zeile, indem wir sie (-1) multiplizieren und erhalten die ganzzahlige Matrix

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right) \qquad \left[\begin{array}{c} \\ S_2(\frac{1}{2}) \\ S_3(-1) \end{array}\right]$$

Addition des (-1)-fachen der zweiten Zeile zur ersten Zeile liefert die selbe normierte, reduzierte Stufenform wie oben.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(-1) \\ \\ \end{array}\right]$$

Durch dieses Vorgehen wird das Rechnen mit Brüchen vollständig vermieden. Das wird man selbstverständlich nur in jenen Fällen tun können, in denen das Resultat auch wirklich ganzzahlig ist. Im allgemeinen Fall wird man im letzten Schritt noch durch die führenden Elemente dividieren müssen, um die einzelen Zeilen zu normieren. Dabei muss man jedoch mit Brüche nicht wirklich rechnen, da sie erst im letzten Schritt entstehen. ◯

Dieses Phänomen gilt allgemein. Zwar kann eine Matrix verschiedene Stufenformen und verschiedene reduzierte Stufenformen haben, weil man verschiedene Wege einschlagen kann. Ihre normierte, reduzierte Stufenform dagegen ist

eindeutig bestimmt. Man kann also auf eine Matrix verschiedene Folgen von Zeilenoperationen anwenden und erhält immer die selbe normierte reduziere Stufenform. Insbesondere folgt daraus, dass der Rang einer Matrix eindeutig bestimmt ist, d.h. nicht von der Wahl der Zeilenoperationen abhängt.

Beispiel. Wir gehen vom linearen Gleichungssystem

$$\begin{cases} 2x - 3y + 5z = 2 \\ 4x - 6y + 10z = 3 \end{cases}$$

mit der zugehörigen erweiterten Matrix:

$$\left(\begin{array}{ccc|c} 2 & -3 & 5 & 2 \\ 4 & -6 & 10 & 3 \end{array}\right)$$

aus. Addition des (-2)-fachen der ersten Zeile zur zweiten Zeile liefert die reduzierte Stufenform.

$$\left(\begin{array}{ccc|c} 2 & -3 & 5 & 2 \\ 0 & 0 & 0 & -1 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \end{array}\right]$$

Aus ihrer zweiten Zeile liest man ab, dass dieses lineare Gleichungssystem keine Lösung hat, was man im vorliegenden einfachen Fall auch „vom Schiff aus“ hätte sehen können. Geometrisch schneiden sich also die durch die beiden linearen Gleichungen beschriebenen Ebenen nicht.

Im Laufe der Zeit werden wir sehen, wie wichtig es ist, von einem linearen Gleichungssystem $A \cdot \vec{x} = \vec{b}$ entscheiden zu können, dass es keine Lösungen hat. Auch diese Information lässt sich bei grossen Gleichungssystemen in der Regel nicht vom Schiff aus sehen. ◯

Sehr oft kommen in der Praxis lineare Gleichungssysteme vor, die für verschiedene Konstantenvektoren $\vec{b}$ bei gleicher Koeffizientenmatrix A gelöst werden müssen. Insbesondere stellt sich dann die Frage, für welche Konstantenvektoren $\vec{b}$ ein solches lineares Gleichungssystem überhaupt lösbar ist. Zur Beantwortung dieser und anderer im Lauf der Zeit zum Vorschein kommender Fragen formuliert man lineare Gleichungssystem mit beliebigem Konstantenvektor und rechnet (vorläufig) symbolisch.

Wir demonstrieren sämtliche Ideen an einem umfangreichen Beispiel.

Beispiel. Wir gehen vom linearen Gleichungssystem

$$\begin{cases} \quad x_3 - x_4 + 3x_5 + x_6 - 6x_7 = b_1 \\ -x_3 + x_4 - 3x_5 - x_6 + 7x_7 = b_2 \\ 2x_1 + 4x_2 - x_4 + 2x_5 + 4x_6 + 5x_7 = b_3 \\ 3x_1 + 6x_2 - 3x_3 + \frac{3}{2}x_4 - 6x_5 + 3x_6 + \frac{51}{2}x_7 = b_4 \\ -4x_1 - 8x_2 + 2x_3 + 2x_5 - 9x_6 - 19x_7 = b_5 \\ 7x_1 + 14x_2 - 5x_3 + \frac{3}{2}x_4 - 8x_5 + 14x_6 + \frac{85}{2}x_7 = b_6 \end{cases}$$

aus. Mit seiner Koeffizientenmatrix $A \in \mathbb{R}^{6,7}$ und dem beliebigen Konstanten-

vektor $\vec{b} \in \mathbb{R}^6$ erhalten wir die erweiterte Matrix

$$(A \mid \vec{b}) = \left(\begin{array}{rrrrrrr|l} 0 & 0 & 1 & -1 & 3 & 1 & -6 & b_1 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 3 & 6 & -3 & \frac{3}{2} & -6 & 3 & \frac{51}{2} & b_4 \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 & b_5 \\ 7 & 14 & -5 & \frac{3}{2} & -8 & 14 & \frac{85}{2} & b_6 \end{array}\right)$$

Damit der führende Koeffizient der ersten Zeile verschieden von 0 wird, vertauschen wir in einem Vorbereitunsschritt die erste und die dritte Zeile. Um mit ganzen Zahlen statt mit Brüchen rechnen zu können, multiplizieren wir die Zeilen mit den kleinsten gemeinsamen Vielfachen ihrer Nenner, d.h. die vierte und die sechste Zeile je mit 2, und erhalten die erweiterte Matrix mit ganzzahliger Koeffizientenmatrix

$$\left(\begin{array}{rrrrrrr|l} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 1 & -1 & 3 & 1 & -6 & b_1 \\ 6 & 12 & -6 & 3 & -12 & 6 & 51 & 2b_4 \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 & b_5 \\ 14 & 28 & -10 & 3 & -16 & 28 & 85 & 2b_6 \end{array}\right) \qquad \left[\begin{array}{l} T_{31} \\ \\ T_{13} \\ S_4(2) \\ \\ S_6(2) \end{array}\right]$$

Addition des (-3)-fachen der ersten Zeile zur vierten, des 2-fachen zur fünften und des (-7)-fachen zur sechsten liefert

$$\left(\begin{array}{rrrrrrr|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 1 & -1 & 3 & 1 & -6 & b_1 \\ 0 & 0 & -6 & 6 & -18 & -6 & 36 & -3b_3 + 2b_4 \\ 0 & 0 & 2 & -2 & 6 & -1 & -9 & 2b_3 + b_5 \\ 0 & 0 & -10 & 10 & -30 & 0 & 50 & -7b_3 + 2b_6 \end{array}\right) \qquad \left[\begin{array}{l} \\ \\ \\ Z_{14}(-3) \\ Z_{15}(2) \\ Z_{16}(-7) \end{array}\right]$$

Man beachte, dass hier automatisch eine Treppenstufe der Breite 2 entstanden ist. Addition der zweiten zur dritten Zeile und des (-6)-fachen der zweiten Zeile zur vierten Zeile sowie des 2-fachen zur fünften und des (-10)-fachen zur sechsten Zeile liefert

$$\left(\begin{array}{rrrrrrr|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1 + b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -6b_2 - 3b_3 + 2b_4 \\ 0 & 0 & 0 & 0 & 0 & -3 & 5 & 2b_2 + 2b_3 + b_5 \\ 0 & 0 & 0 & 0 & 0 & 10 & -20 & -10b_2 - 7b_3 + 2b_6 \end{array}\right) \left[\begin{array}{l} \\ \\ \\ Z_{23}(1) \\ Z_{24}(-6) \\ Z_{25}(2) \\ Z_{26}(-10) \end{array}\right]$$

Hier ist eine weitere Treppenstufe der Breite 3 entstanden. Vertauschen der dritten und fünften Zeile ergibt

$$\left(\begin{array}{rrrrrrr|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 0 & 0 & 0 & -3 & 5 & 2b_2 + 2b_3 + b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -6b_2 - 3b_3 + 2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1 + b_2 \\ 0 & 0 & 0 & 0 & 0 & 10 & -20 & -10b_2 - 7b_3 + 2b_6 \end{array}\right) \qquad \left[\begin{array}{l} \\ \\ T_{53} \\ \\ T_{35} \\ \end{array}\right]$$

Um ganzzahlig weiterrechnen zu können, addieren wir das 10-fache der dritten Zeile zum 3-fachen der sechsten Zeile und erhalten

$$\left(\begin{array}{ccccccc|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 0 & 0 & 0 & -3 & 5 & 2b_2+2b_3+b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -6b_2-3b_3+2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1+b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -10 & -10b_2-b_3+10b_5+6b_6 \end{array}\right) \left[\begin{array}{c} \\ \\ \\ \\ \\ Z_{36}(10,3) \end{array}\right]$$

Um die Rechnung etwas zu vereinfachen und die 1 in der siebten Spalte auszunützen, vertauschen wir die vierte und fünfte Zeile und erhalten

$$\left(\begin{array}{ccccccc|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 0 & 0 & 0 & -3 & 5 & 2b_2+2b_3+b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1+b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -6b_2-3b_3+2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & -10 & -10b_2-b_3+10b_5+6b_6 \end{array}\right) \left[\begin{array}{c} \\ \\ \\ \\ T_{43} \\ T_{34} \end{array}\right]$$

Addition des 6-fachen der vierten Zeile zur fünften und des 10-fachen zur sechsten Zeile ergibt

$$\left(\begin{array}{ccccccc|c} 2 & 4 & 0 & -1 & 2 & 4 & 5 & b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & b_2 \\ 0 & 0 & 0 & 0 & 0 & -3 & 5 & 2b_2+2b_3+b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1+b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6b_1-3b_3+2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10b_1-b_3+10b_5+6b_6 \end{array}\right) \left[\begin{array}{c} \\ \\ \\ \\ Z_{45}(6) \\ Z_{46}(10) \end{array}\right]$$

Die erreichte Matrix hat nun Stufenform und die Vorwärtsphase des Algorithmus ist abgeschlossen. Aus der gefundenen Stufenform lesen wir zunächst den Rang der Koeffizientenmatrix A ab. Es ist $r = \mathrm{Rang}(A) = 4$. Aus der Stufenform lesen wir ferner ein *Lösbarkeitskriterium* ab. Für das gegebene lineare Gleichungssystem existiert genau dann eine Lösung, wenn die Komponenten b_i des Konstantenvektors $\vec{b}$ so beschaffen sind, dass die letzten beiden Zeilen Nullzeilen sind, d.h. falls das homogene Gleichungssystem

$$\left\{\begin{array}{rcl} 6b_1 - 3b_3 + 2b_4 & & = 0 \\ 10b_1 - b_3 & + 10b_5 + 6b_6 & = 0 \end{array}\right.$$

mit $6 - r = 2$ Zeilen erfüllt ist. Mit seiner Koeffizientenmatrix

$$B = \left(\begin{array}{cccccc} 6 & 0 & -3 & 2 & 0 & 0 \\ 10 & 0 & -1 & 0 & 10 & 6 \end{array}\right) \in \mathbb{R}^{2,6}$$

in der wir die fehlende Variable b_2 durch 0 markiert haben, um mit der Nummerierung der Variablen kein Durcheinander zu erhalten, können wir nun also leicht kontrollieren, für welche Vektoren $\vec{b} \in \mathbb{R}^6$ das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ eine Lösung hat. Das ist genau dann der Fall, falls die Lösbarkeitsbedingung

$$B \cdot \vec{b} = \vec{0}$$

gilt. Mit dem Lösbarkeitskriterium können genau diejenigen Vektoren $\vec{b} \in \mathbb{R}^6$ beschrieben werden, für die das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ eine Lösung hat. Insbesondere lassen sich mit B jene Vektoren $\vec{b} \in \mathbb{R}^6$ beschreiben, für die das System *keine* Lösung hat. Negative Information ist mathematisch genauso von Bedeutung, wie die Information, dass der Tb-Test negativ verlaufen ist, für den Patienten eine erfreuliche Mitteilung beinhaltet. Das Lösbarkeitskriterium hat etwas mit der logischen Negation zu tun.

Wir erwarten also, dass das homogene Gleichungssystem der Lösbarkeitskeitbedingung bzw. seine Koeffizientenmatrix $B \in \mathbb{R}^{2,6}$ in einer sehr engen Beziehung zur Koeffizientenmatrix $A \in \mathbb{R}^{6,7}$ steht, die durch die kurze exakte Folge

$$\mathbb{R}^7 \xrightarrow{A} \mathbb{R}^6 \xrightarrow{B} \mathbb{R}^2$$

ausgedrückt wird. Tatsächlich rechnet man die *Kettenbedingung*

$$B \cdot A = 0$$

nach, die auch eine gewissen Kontrolle der bisherigen Rechnungen liefert. Sie besagt, dass die Spalten der Matrix A und die Zeilen von B orthogonal sind.

Wir lösen nun das Gleichungssystem fertig und stellen zunächst fest, dass die Variablen x_2, x_4, x_5 frei und x_1, x_3, x_6, x_7 gebunden sind. Um nach den übrigen Variablen aufzulösen, überführen wir die Matrix nun in der Rückwärtsphase in normierte, reduzierte Stufenform. Dazu addieren wir das (-5)-fache der vierten zur dritten Zeile, das (-7)-fache zur zweiten und das (-5)-fache zur ersten Zeile.

$$\left(\begin{array}{ccccccc|c} 2 & 4 & 0 & -1 & 2 & 4 & 0 & -5b_1 - 5b_2 + b_3 \\ 0 & 0 & -1 & 1 & -3 & -1 & 0 & -7b_1 - 6b_2 \\ 0 & 0 & 0 & 0 & 0 & -3 & 0 & -5b_1 - 3b_2 + 2b_3 + b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1 + b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6b_1 - 3b_3 + 2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10b_1 - b_3 + 10b_5 + 6b_6 \end{array}\right) \qquad \left[\begin{array}{c} Z_{41}(-5) \\ Z_{42}(-7) \\ Z_{43}(-5) \\ \\ \\ \\ \end{array}\right]$$

Um ganzzahlig weiterrechnen zu können, addieren wir die dritte Zeile zum (-3)-fachen der zweiten Zeile. Ferner addieren wir das 4-fache der dritten Zeile zum 3-fachen der ersten Zeile und erhalten die ganzzahlige, reduzierte Stufenform

$$\left(\begin{array}{ccccccc|c} 6 & 12 & 0 & -3 & 6 & 0 & 0 & -35b_1 - 27b_2 + 11b_3 + 4b_5 \\ 0 & 0 & 3 & -3 & 9 & 0 & 0 & 16b_1 + 15b_2 + 2b_3 + b_5 \\ 0 & 0 & 0 & 0 & 0 & -3 & 0 & -5b_1 - 3b_2 + 2b_3 + b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1 + b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6b_1 - 3b_3 + 2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10b_1 - b_3 + 10b_5 + 6b_6 \end{array}\right)$$

Man beachte, das wir bis hierhin ganzzahlig rechnen konnten. Um nun die führenden Koeffizienten zu normieren, dividieren wir durch sämtliche führenden Koeffizienten, d.h. die erste Zeile durch 6, die zweite durch 3 und die dritte durch -3. Dabei werden nun halt Brüche entstehen. Die normierte, reduzierte

Stufenform lautet also

$$\left(\begin{array}{ccccccc|c} 1 & 2 & 0 & -\frac{1}{2} & 1 & 0 & 0 & -\frac{35}{6}b_1 - \frac{9}{2}b_2 + \frac{11}{6}b_3 + \frac{2}{3}b_5 \\ 0 & 0 & 1 & -1 & 3 & 0 & 0 & \frac{16}{3}b_1 + 5b_2 + \frac{2}{3}b_3 + \frac{1}{3}b_5 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{5}{3}b_1 + b_2 - \frac{2}{3}b_3 - \frac{1}{3}b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & b_1 + b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6b_1 - 3b_3 + 2b_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10b_1 - b_3 + 10b_5 + 6b_6 \end{array}\right) \quad \left[\begin{array}{l} S_1(\frac{1}{6}) \\ S_2(\frac{1}{3}) \\ S_3(-\frac{1}{3}) \\ \\ \\ \\ \end{array}\right]$$

Auflösen nach den führenden Variablen ergibt schliesslich im Fall, wo das Lösbarkeitskriterium erfüllt ist, die Lösung

$$\begin{aligned} x_1 &= -\tfrac{35}{6}b_1 - \tfrac{9}{2}b_2 + \tfrac{11}{6}b_3 + \tfrac{2}{3}b_5 - 2x_2 + \tfrac{1}{2}x_4 - x_5 \\ x_3 &= \tfrac{16}{3}b_1 + 5b_2 + \tfrac{2}{3}b_3 + \tfrac{1}{3}b_5 + x_4 - 3x_5 \\ x_6 &= \tfrac{5}{3}b_1 + b_2 - \tfrac{2}{3}b_3 - \tfrac{1}{3}b_5 \\ x_7 &= b_1 + b_2 \end{aligned}$$

Diese Lösung soll nun, wie üblich, in vektorieller Form geschrieben werden. Setzen wir für die freien Variablen $r = x_2, \tilde{s} = x_4, t = x_5$, so gilt:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} -\frac{35}{6}b_1 - \frac{9}{2}b_2 + \frac{11}{6}b_3 + \frac{2}{3}b_5 \\ 0 \\ \frac{16}{3}b_1 + 5b_2 + \frac{2}{3}b_3 + \frac{1}{3}b_5 \\ 0 \\ 0 \\ \frac{5}{3}b_1 + b_2 - \frac{2}{3}b_3 - \frac{1}{3}b_5 \\ b_1 + b_2 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \tilde{s} \begin{pmatrix} \frac{1}{2} \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

Um den Lösungsraum durch ganzzahlige Vektoren aufspannen zu können, wählen wir als Parmeter $\tilde{s} = 2s$ und erhalten schliesslich als Parameterdarstellung des Lösungsraumes:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} -\frac{35}{6}b_1 - \frac{9}{2}b_2 + \frac{11}{6}b_3 + \frac{2}{3}b_5 \\ 0 \\ \frac{16}{3}b_1 + 5b_2 + \frac{2}{3}b_3 + \frac{1}{3}b_5 \\ 0 \\ 0 \\ \frac{5}{3}b_1 + b_2 - \frac{2}{3}b_3 - \frac{1}{3}b_5 \\ b_1 + b_2 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

Wir stellen also fest, dass das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ unter der Voraussetzung, dass die Lösungsbarkeitsbedingung $B \cdot \vec{b} = \vec{0}$ erfüllt ist, unendlich viele Lösungen hat. Insbesondere ist der Lösungsraum $L(A, \vec{b})$ des Gleichungssystems 3-dimensional, da er durch drei freie Variablen parametrisiert wird. Das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ schneidet aus dem umgebenden 7-dimensionalen Raum $\mathbb{R}^7$ diesen 3-dimensionalen Raum $L(A, \vec{b})$ aus.

Man kann die Lösung eines linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ mit beliebigem Konstantenvektor und das Lösbarkeitskriterium auch finden, ohne symbolisch

rechnen zu müssen. Das ist für die Hand- und die maschinelle Rechnung mit Hilfe einer konventionellen Programmiersprache angenehmer. Dazu denkt man sich die Komponenten des Konstantenvektors $\vec{b}$ als zusätzliche Variablen, die man mit Hilfe einer weiteren Matrix kodiert. Das erreicht man durch die Gleichung $\vec{b} = E_m \cdot \vec{b}$. Wir schreiben also die Einheitsmatrix E_m rechts neben A und erhalten die Blockmatrix $(A \mid E_m)$, auf die wir geeignete Elementaroperationen anwenden, bis der linke Block normierte, reduzierte Stufenform $S(A)$ hat; gleichzeitig ergibt sich für den rechten Block eine Matrix M, so dass wir schliesslich die Blockmatrix $(S(A) \mid M)$ erhalten. Das umgeformte Gleichungssystem lautet $S(A) \cdot \vec{x} = M \cdot \vec{b}$. Weil der linke Block $S(A)$ normierte, reduzierte Stufenform hat, kann man dann die gesuchten Lösungen problemlos ablesen.

In dem bereits behandelten Beispiel beginnen wir also mit der Blockmatrix

$$(A \mid E_6) = \left(\begin{array}{rrrrrrr|cccccc} 0 & 0 & 1 & -1 & 3 & 1 & -6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & -1 & 2 & 4 & 5 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 6 & -3 & \frac{3}{2} & -6 & 3 & \frac{51}{2} & 0 & 0 & 0 & 1 & 0 & 0 \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 & 0 & 0 & 0 & 0 & 1 & 0 \\ 7 & 14 & -5 & \frac{3}{2} & -8 & 14 & \frac{85}{2} & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Führen wir nun wortwörtlich die selben Elementaroperationen wie oben beschrieben durch, erhalten wir schliesslich die Blockmatrix:

$$(S(A) \mid M) = \left(\begin{array}{rrrrrrr|rrrrrr} 1 & 2 & 0 & -\frac{1}{2} & 1 & 0 & 0 & -\frac{35}{6} & -\frac{9}{2} & \frac{11}{6} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 & -1 & 3 & 0 & 0 & \frac{16}{3} & 5 & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{5}{3} & 1 & -\frac{2}{3} & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & -3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & -1 & 0 & 10 & 6 \end{array}\right)$$

Durch Auflösen nach den führenden Variablen und korrekte Interpretation der Elemente der Matrix M im rechten Block erhalten wir das selbe Lösbarkeitskriterium und die selbe Lösung wie oben. Aus dem nächsten Abschnitt wird sich ergeben, was es mit dieser Matrix M auf sich hat. Der neugierige Leser kann rechnerisch überprüfen, dass M invertierbar ist und die Gleichung $M \cdot A = S(A)$ bzw. $A = M^{-1} \cdot S(A)$ gilt.

Soll etwa das lineare Gleichungssystem mit gegebenem Konstantevektor

$$\left\{\begin{array}{rrrrrrrrr} & & & x_3 - & x_4 + 3x_5 + & x_6 - & 6x_7 = & 6 \\ & & - & x_3 + & x_4 - 3x_5 - & x_6 + & 7x_7 = & 2 \\ 2x_1 + & 4x_2 & & - & x_4 + 2x_5 + & 4x_6 + & 5x_7 = & 6 \\ 3x_1 + & 6x_2 & - 3x_3 + & \frac{3}{2}x_4 - & 6x_5 + & 3x_6 + & \frac{51}{2}x_7 = & -9 \\ -4x_1 - & 8x_2 & + 2x_3 & & + 2x_5 - & 9x_6 - & 19x_7 = & 3 \\ 7x_1 + & 14x_2 & - 5x_3 + & \frac{3}{2}x_4 - & 8x_5 + & 14x_6 + & \frac{85}{2}x_7 = & -14 \end{array}\right.$$

gelöst werden, untersuchen wir zunächst für die numerischen Werte $b_1 = 6$, $b_2 = 2$, $b_3 = 6$, $b_4 = -9$, $b_5 = 3$, $b_6 = -14$, ob die Lösbarkeitbedingung erfüllt

ist. Das ist tatsächlich der Fall und der Lösungsraum des Gleichungssystems lässt sich durch die Parameterdarstellung

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} -31 \\ 0 \\ 47 \\ 0 \\ 0 \\ 7 \\ 8 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

beschreiben. Die ganzzahligen Lösungen des Systems ergeben sich allgemein aus der ganzzahligen Normalform

$$S = U \cdot A \cdot V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

der Koeffizientenmatrix des ganzzahlig gemachten Gleichungssystems und den beiden unimodularen Transformationsmatrizen

$$U = \begin{pmatrix} 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 5 & 1 & -1 & 0 & 3 & 1 \\ 6 & 0 & -3 & 1 & 0 & 0 \\ 10 & 0 & -1 & 0 & 10 & 3 \end{pmatrix}, \qquad V = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 12 & 3 & 2 & -1 & 4 \\ 1 & 0 & 5 & 4 & 2 & 2 & 4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Das zugehörige homogene System $A \cdot \vec{x} = \vec{0}$ lässt sich lösen, weil der Nullvektor die Lösbarkeitsbedingung $B \cdot \vec{0} = \vec{0}$ erfüllt. Die allgemeine Lösung des zugehörigen homogenen Systems $A \cdot \vec{x} = \vec{0}$ erhält man also dadurch, dass man $b_1 = \cdots = b_6 = 0$ setzt. Ein beliebiger Vektor des entsprechendenen Lösungsraumes $L(A, \vec{0})$ kann in vektorieller Form als

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

beschrieben werden. Zur Kontrolle beachten wir die bemerkenswerte Eigenschaft, dass jeder dieser Basisvektoren von $L(A, \vec{0})$ auf sämtlichen Zeilenvektoren der Koeffizientenmatrix A senkrecht steht.

Der Lösungsraum $L(A, \vec{0})$ des zugeordneten homogenen Systems, für den man $\mathrm{Ker}(A)$ schreibt, heisst *Kern* der Matrix A. Es sind genau diejenigen Vektoren aus $\mathbb{R}^7$, die beim Multiplizieren mit der Matrix A auf den Nullvektor abgebildet

werden. Für seine Dimension gilt ebenfalls $\dim(\mathrm{Ker}(A)) = 3$. Wir haben 3 Basisvektoren in $\mathbb{R}^7$ bestimmt, die den Kern von A aufspannen.

Geometrisch interpretiert handelt es sich beim Kern $\mathrm{Ker}(A)$ um den Teilraum von $\mathbb{R}^7$, den man erhält, wenn man den Lösungsraum $L(A, \vec{b})$ des allgemeinen Systems parallel in den Ursprung verschiebt.

Im Zusammenhang mit einem linearen Gleichungssystem $A \cdot \vec{x} = \vec{b}$ spielt neben dem Kern ein zweiter Teilraum eine zentrale Rolle. Dabei handelt es sich in unserem Fall um den Teilraum von $\mathbb{R}^6$ jener Vektoren, die in der Form $A \cdot \vec{x}$ für einen gewissen Vektor $\vec{x}$ erhalten werden können. Das *Bild* der Matrix $A \in \mathbb{R}^{m,n}$ besteht aus allen Vektoren der Form $A \cdot \vec{x}$. Weil ein solches Matrizenprodukt gerade die Linearkombinationen der Spaltenvektoren $\vec{a}_1, \ldots, \vec{a}_n \in \mathbb{R}^m$ der Matrix A, d.h. die Vektoren der Form

$$x_1\vec{a}_1 + \cdots + x_n\vec{a}_n = \vec{b}$$

sind, wird das Bild von A gelegentlich auch als Spaltenraum von A bezeichnet. Die Lösung des Systems liefert eine Beschreibung des konstanten Vektors $\vec{b}$ als Linearkombination der Spaltenvektoren der Koeffizientenmatrix A.

Als Referenz fassen wir die beiden fundamentalen Teilräume in folgender Definition zusammen.

Definition. Es sei $A \in \mathbb{R}^{m,n}$ eine Matrix und $f_A\colon \mathbb{R}^n \to \mathbb{R}^m$ die zugehörige lineare Abbildung.

Unter dem *Kern* von A verstehen wir den Teilraum

$$\mathrm{Ker}(A) = \{\vec{x} \in \mathbb{R}^n \mid A \cdot \vec{x} = \vec{0}\} \subseteq \mathbb{R}^n$$

und unter dem *Bild* von A verstehen wir den Teilraum

$$\mathrm{Im}(A) = \{b \in \mathbb{R}^m \mid \vec{b} = A \cdot \vec{x} \text{ für einen gewissen Vektor } \vec{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$$

Der Kern von A ist also der Lösungsraum des zugehörigen homogenen Gleichungssystems $A \cdot \vec{x} = \vec{0}$ und das Bild von A besteht genau aus jenen Vektoren $\vec{b}$, für die das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ lösbar ist.

Weil das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ definitionsgemäss genau dann lösbar ist, falls der Vektor $\vec{b} \in \mathbb{R}^m$ zum Bild von A gehört und diese Vektoren durch die Lösbarkeitsbedingung $B \cdot \vec{b} = \vec{0}$ charakterisiert werden können, bleibt in unserem Beispiel noch die detailierte Untersuchung der Lösbarkeitsbedingung $B \cdot \vec{b} = \vec{0}$. Man beachte, dass es sich dabei um ein gewisses *homogenes* lineares Gleichungssystem handelt. In der Folge linearer Abbildungen

$$\mathbb{R}^7 \xrightarrow{A} \mathbb{R}^6 \xrightarrow{B} \mathbb{R}^2$$

gilt also die fundamentale *Exaktheit*

$$\mathrm{Im}(A) = \mathrm{Ker}(B)$$

Die eine Inklusion $\mathrm{Im}(A) \subseteq \mathrm{Ker}(B)$ drückt sich durch bereits festgestellte Kettenbedingung $B \cdot A = 0$ aus. Sie besagt, dass jeder Konstantenvektor, für den das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ gelöst werden kann, die Lösbarkeitsbedingung

$B \cdot \vec{b} = \vec{0}$ erfüllen muss. Die andere Inklusion $\mathrm{Ker}(B) \subseteq \mathrm{Im}(A)$ besagt umgekehrt, dass das Gleichungssystem für jeden Vektor, der die Lösbarkeitsbedingung erfüllt, auch wirklich gelöst werden kann. Wegen der Exaktheit kann der selbe Vektorraum $\mathcal{L} = \mathrm{Im}(A) = \mathrm{Ker}(B)$ auf zwei unterschiedliche Arten beschrieben werden. Seine Beschreibung als Bild von A ist dann vorteilhaft, wenn wenn man Vektoren in diesem Raum produzieren will. Die andere Beschreibung als Kern von B ist dann vorteilhaft, wenn man von einem Vektor entscheiden will, ob er zum Raum gehört.

Algorithmisch lässt sich die Berechnung des Bildes von A auf die Berechnung des Kerns der Matrix B reduzieren. Das lineare Gleichungssystem der Lösbarkeitsbedingung $B \cdot \vec{b} = \vec{0}$ lässt sich aber mit Hilfe von Elimination lösen. In der Koeffizientenmatrix

$$B = \begin{pmatrix} 6 & 0 & -3 & 2 & 0 & 0 \\ 10 & 0 & -1 & 0 & 10 & 6 \end{pmatrix}$$

ist in der ersten Spalte $\mathrm{ggT}(6, 10) = 2$. Da wir mit betragsmässig möglichst kleinen ganzen Zahlen weiterrechnen wollen, addieren wir das $10 \div 2 = 5$-fache der ersten Zeile zum $-(6 \div 2) = -3$-fachen der zweiten Zeile und erhalten

$$\begin{pmatrix} 6 & 0 & -3 & 2 & 0 & 0 \\ 0 & 0 & -12 & 10 & -30 & -18 \end{pmatrix} \qquad \begin{bmatrix} \\ Z_{12}(5, -3) \end{bmatrix}$$

Um die Zahlen weiterhin so klein wie möglich zu halten, dividieren wir die erhaltene zweite Zeile durch ihren grössten gemeinsamen Teiler 2 und erhalten die Matrix in Stufenform

$$\begin{pmatrix} 6 & 0 & -3 & 2 & 0 & 0 \\ 0 & 0 & -6 & 5 & -15 & -9 \end{pmatrix} \qquad \begin{bmatrix} \\ S_2(\frac{1}{2}) \end{bmatrix}$$

Selbstverständlich existiert für homogene lineare Gleichungssysteme keine Lösbarkeitsbedingung, da sie immer mindestens die triviale Lösung haben. Um die reduzierte Stufenform zu finden, addieren wir die zweite Zeile zum (-2)-fachen der ersten Zeile und erhalten die reduzierte Stufenform

$$\begin{pmatrix} -12 & 0 & 0 & 1 & -15 & -9 \\ 0 & 0 & -6 & 5 & -15 & -9 \end{pmatrix} \qquad \begin{bmatrix} Z_{21}(1, -2) \\ \\ \end{bmatrix}$$

Die normierte, reduziete Stufenform erhalten wir schliesslich, indem wir die erste Zeile durch (-12) und die zweite durch (-6) dividieren.

$$\begin{pmatrix} 1 & 0 & 0 & -\frac{1}{12} & \frac{5}{4} & \frac{3}{4} \\ 0 & 0 & 1 & -\frac{5}{6} & \frac{5}{2} & \frac{3}{2} \end{pmatrix} \qquad \begin{bmatrix} S_1(-\frac{1}{12}) \\ S_2(-\frac{1}{6}) \end{bmatrix}$$

Auflösen nach den führenden Variablen ergibt

$$\begin{aligned} b_1 &= \tfrac{1}{12}b_4 - \tfrac{5}{4}b_5 - \tfrac{3}{4}b_6 \\ b_3 &= \tfrac{5}{6}b_4 - \tfrac{5}{2}b_5 - \tfrac{3}{2}b_6 \end{aligned}$$

Diese Lösung soll, wie üblich, in vektorieller Form geschrieben werden. Setzen

wir für die freien Variablen $r = b_2$, $\tilde{s} = b_4$, $\tilde{t} = b_5$, $\tilde{u} = b_6$, erhalten wir:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = r \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \tilde{s} \begin{pmatrix} \frac{1}{12} \\ 0 \\ \frac{5}{6} \\ 1 \\ 0 \\ 0 \end{pmatrix} + \tilde{t} \begin{pmatrix} -\frac{5}{4} \\ 0 \\ -\frac{5}{2} \\ 0 \\ 1 \\ 0 \end{pmatrix} + \tilde{u} \begin{pmatrix} -\frac{3}{4} \\ 0 \\ -\frac{3}{2} \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Um den Lösungsraum durch ganzzahlige Vektoren aufspannen zu können, wählen wir als Parmeter $\tilde{s} = 12s$, $\tilde{t} = 4t$, $\tilde{u} = 4u$, und erhalten schliesslich für das Bild $\mathrm{Im}(A)$ die Parameterdarstellung

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = r \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 10 \\ 12 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -5 \\ 0 \\ -10 \\ 0 \\ 4 \\ 0 \end{pmatrix} + u \begin{pmatrix} -3 \\ 0 \\ -6 \\ 0 \\ 0 \\ 4 \end{pmatrix}$$

Das Bild hat also die Dimension $\dim(\mathrm{Im}(A)) = 4$ und wir haben 4 Basisvektoren in $\mathbb{R}^6$ bestimmt, die das Bild von A aufspannen. Das Bild von A beschreibt genau jene Vektoren $\vec{b} \in \mathbb{R}^6$, für die das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mindestens eine Lösung hat. Die vollständige Lösung des linearen Gleichungssystems liefert also für die Koeffizientenmatrix $A \in \mathbb{R}^{6,7}$ die beiden Räume $\mathrm{Ker}(A) \subseteq \mathbb{R}^7$ der Dimension 3 und das Bild $\mathrm{Im}(A) \subseteq \mathbb{R}^6$ der Dimension 4.

Der Eliminationsalgorithmus hat für den Kern $\mathrm{Ker}(A) \subseteq \mathbb{R}^7$ die Basis

$$\begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^7$$

und für das Bild $\mathrm{Im}(A) \subseteq \mathbb{R}^6$ die Basis

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 10 \\ 12 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -5 \\ 0 \\ -10 \\ 0 \\ 4 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -3 \\ 0 \\ -6 \\ 0 \\ 0 \\ 4 \end{pmatrix} \in \mathbb{R}^6$$

geliefert. Mit Ihnen kann jeder Vektor des Kern bzw. des Bildes auf eindeutige Art linear kombiniert werden. ◯

Mit Hilfe des Eliminations-Algorithmus sind wir in der Lage, eine Parameterdarstellung für den Lösungsraum $L(A, \vec{b})$ eines linearen Gleichungssystems zu

finden. Wir können den Algorithmus auch so interpretieren, dass er aus einer impliziten Darstellung des geometrischen Objektes $L(A,\vec{b})$ eine explizite Darstellung berechnet. Auch das umgekehrte Problem spielt eine zentrale Rolle. Aus der Parameterdarstellung eines Raumes, der vektoriell durch die Gleichung

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} + t_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + \cdots + t_m \begin{pmatrix} a_{1m} \\ \vdots \\ a_{nm} \end{pmatrix}$$

gegeben ist soll ein lineares Gleichungssystem gesucht werden, das diesen Raum als Lösungsraum hat. Gleichbedeutend damit ist, dass Skalare $t_1,\ldots,t_m$ existieren, so dass die Gleichung

$$t_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + \cdots + t_m \begin{pmatrix} a_{1m} \\ \vdots \\ a_{nm} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} x_1 - b_1 \\ \vdots \\ x_n - b_n \end{pmatrix}$$

gilt. Das ist aber genau dann der Fall, wenn dieses lineare Gleichungssystem eine Lösung hat. Um diese Frage zu untersuchen, haben wir nur das Lösbarkeitskriterium dieses Gleichungssystems zu berechnen.

Beispiel. Wir sind im letzten Beispiel vom linearen Gleichungssystem

$$\left\{\begin{array}{rrrrrrrrr}
 & & & x_3 - & x_4 + 3x_5 + & x_6 - & 6x_7 & = 6 \\
 & & - & x_3 + & x_4 - 3x_5 - & x_6 + & 7x_7 & = 2 \\
2x_1 + & 4x_2 & & - & x_4 + 2x_5 + & 4x_6 + & 5x_7 & = 6 \\
3x_1 + & 6x_2 & - 3x_3 + & \tfrac{3}{2}x_4 - & 6x_5 + & 3x_6 + & \tfrac{51}{2}x_7 & = -9 \\
-4x_1 - & 8x_2 & + 2x_3 & & + 2x_5 - & 9x_6 - & 19x_7 & = 3 \\
7x_1 + & 14x_2 & - 5x_3 + & \tfrac{3}{2}x_4 - & 8x_5 + & 14x_6 + & \tfrac{85}{2}x_7 & = -14
\end{array}\right.$$

ausgegangen und haben seinen Lösungsraum durch die Parameterdarstellung

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} -31 \\ 0 \\ 47 \\ 0 \\ 0 \\ 7 \\ 8 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

beschrieben. Nun wollen wir umgekehrt diesen Lösungsraum durch ein geeignetes lineares Gleichungssystem d.h. durch Koordinatengleichungen beschreiben. Dazu müssen wir untersuchen, unter welchen Bedingungen das lineare Gleichungssystem

$$r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 + 31 \\ x_2 \\ x_3 - 47 \\ x_4 \\ x_5 \\ x_6 - 7 \\ x_7 - 8 \end{pmatrix}$$

Lösungen hat. Um nicht symbolisch rechnen zu müssen, kodieren wir den Konstantenvektor wie früher mit Hilfe einer geeigneten Matrix. Wir gehen also von der Blockmatrix

$$\left(\begin{array}{ccc|cccccccc} -2 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 31 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & -3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -47 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -8 \end{array}\right)$$

aus und wenden auf sie geeignete Elementarumformungen an, bis die linke Seite Stufenform hat. Dann lesen wir das gesuchte Lösbarkeitskriterium ab. Zur Vereinfachung vertauschen wir zunächst die erste und die zweite Zeile

$$\left(\begin{array}{ccc|cccccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 31 \\ 0 & 2 & -3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -47 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -8 \end{array}\right) \qquad \left[\begin{array}{c} T_{21} \\ T_{12} \\ \\ \\ \\ \\ \end{array}\right]$$

Nun addieren wir das 2-fache der ersten Zeile zur zweiten Zeile und erhalten:

$$\left(\begin{array}{ccc|cccccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 31 \\ 0 & 2 & -3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -47 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -8 \end{array}\right) \qquad \left[\begin{array}{c} Z_{12}(2) \\ \\ \\ \\ \\ \\ \end{array}\right]$$

Als nächstes addieren wir das (-2)-fache der zweiten Zeile zur dritten und vierten Zeile und erhalten:

$$\left(\begin{array}{ccc|cccccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 31 \\ 0 & 0 & -1 & -2 & -4 & 1 & 0 & 0 & 0 & 0 & -109 \\ 0 & 0 & 2 & -2 & -4 & 0 & 1 & 0 & 0 & 0 & -62 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -8 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(-2) \\ Z_{24}(-2) \\ \\ \\ \end{array}\right]$$

Schliesslich addieren wir das 2-fache der dritten Zeile zur vierten Zeile und die dritte Zeile zur fünften Zeile und erhalten:

$$\left(\begin{array}{ccc|cccccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 31 \\ 0 & 0 & -1 & -2 & -4 & 1 & 0 & 0 & 0 & 0 & -109 \\ 0 & 0 & 0 & -6 & -12 & 2 & 1 & 0 & 0 & 0 & -280 \\ 0 & 0 & 0 & -2 & -4 & 1 & 0 & 1 & 0 & 0 & -109 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -8 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{34}(2) \\ Z_{35}(1) \\ \\ \\ \end{array}\right]$$

Der linke Block hat nun Stufenform und wir können das gesuchte Lösbarkeitskriterium ablesen. Es hat die Form des folgenden linearen Gleichungssystems

$$\left\{\begin{array}{rrrrrrrrr} -6x_1 & -12x_2 & +2x_3 & +x_4 & & & & -280 & =0 \\ -2x_1 & -4x_2 & +x_3 & & +x_5 & & & -109 & =0 \\ & & & & & x_6 & - & 7 & =0 \\ & & & & & & x_7 - & 8 & =0 \end{array}\right.$$

oder indem wir, wie üblich, die Konstanten auf die rechte Seite nehmen:

$$\left\{\begin{array}{rrrrrrrrr} 6x_1 & +12x_2 & -2x_3 & -x_4 & & & & =-280 \\ 2x_1 & +4x_2 & -x_3 & & -x_5 & & & =-109 \\ & & & & & x_6 & = & 7 \\ & & & & & & x_7 = & 8 \end{array}\right.$$

Dieses Gleichungssystem hat als Lösungsraum die ursprünglich gewählte Parameterdarstellung.

Man beachte, dass dieses Gleichungssystem aus vier Gleichungen besteht. Das ursprüngliche Gleichungssysteme, das nach Konstruktion den selben Lösungsraum hat, besitzt aber zwei Gleichungen mehr. Offenbar waren ursprünglich also 2 Gleichungen von den anderen abhängig und damit überflüssig. Unsere Rechnung beseitigt diese lineare Abhängigkeit und führt damit zu einer Datenkompression. Man beachte ferner, dass die vier am Schluss entstandenen Gleichungen voneinander unabhängig sind, da die letzten vier Variablen in Stufenform vorliegen. Daher ist keine der vier Gleichungen mehr überflüssig und keine weitere Datenkompression mehr möglich. ◯

Die Idee der Datenkompression lässt sich nicht nur auf ein System linearer Gleichungen, sondern auch auf ein System von Vektoren $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^n$ anwenden. Dazu kann man vom Unterraum $\mathcal{L}(\vec{v}_1, \ldots, \vec{v}_k) \subseteq \mathbb{R}^n$ ausgehen, der durch diese Vektoren aufgespannt wird. Ein beliebiger Vektor $\vec{x}$ dieses Raumes lässt sich leicht durch die Parameterdarstellung

$$\vec{x} = t_1\vec{v}_1 + \cdots + t_k\vec{v}_k$$

darstellen. Mit Hilfe des soeben beschriebenen Verfahrens findet man ein lineares Gleichungssystem, das diesen Unterraum aufspannt. Falls man dieses Gleichungssystem nun nach dem beschriebenen Verfahren auflöst und die Lösung matriziell anschreibt, so hat man dadurch ein System von Vektoren gefunden, dass zwar immer noch den selben Unterraum aufspannt, aber in der Regel kleiner ist.

Beispiel. Wir gehen vom System der Vektoren

$$\vec{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 3 \\ 1 \\ -6 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ -3 \\ -1 \\ 7 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} 2 \\ 4 \\ 0 \\ -1 \\ 2 \\ 4 \\ 5 \end{pmatrix}$$

$$\vec{v}_4 = \begin{pmatrix} 3 \\ 6 \\ -3 \\ \frac{3}{2} \\ -6 \\ 3 \\ \frac{51}{2} \end{pmatrix}, \qquad \vec{v}_5 = \begin{pmatrix} -4 \\ -8 \\ 2 \\ 0 \\ 2 \\ -9 \\ -19 \end{pmatrix}, \qquad \vec{v}_6 = \begin{pmatrix} 7 \\ 14 \\ -5 \\ \frac{3}{2} \\ -8 \\ 14 \\ \frac{85}{2} \end{pmatrix}$$

in $\mathbb{R}^7$ aus.

Statt des beschriebenen aufwändigen Kreisgangs zur Bestimmung einer Basis des von diesen Vektoren aufgespannten Teilraums d.h. der Linearkombinationen

$$\vec{x} = t_1\vec{v}_1 + t_2\vec{v}_2 + t_3\vec{v}_3 + t_4\vec{v}_4 + t_5\vec{v}_5 + t_6\vec{v}_6$$

gehen wir etwas anders und schneller vor und fassen dazu diese Vektoren als *Zeilen*vektoren zu einer Matrix zusammen. In unserem Beispiel erhalten wir

$$A = \begin{pmatrix} 0 & 0 & 1 & -1 & 3 & 1 & -6 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 \\ 2 & 4 & 0 & -1 & 2 & 4 & 5 \\ 3 & 6 & -3 & \frac{3}{2} & -6 & 3 & \frac{51}{2} \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 \\ 7 & 14 & -5 & \frac{3}{2} & -8 & 14 & \frac{85}{2} \end{pmatrix}$$

d.h. genau die Koeffizientenmatrix des im letzten Beispiel untersuchten linearen Gleichungssystems. Diese Matrix überführen wir nun durch Zeilenoperationen in normierte, reduzierte Stufenform und erhalten, wie seinerzeit, die Matrix

$$S(A) = \begin{pmatrix} 1 & 2 & 0 & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Aus ihr lesen wir ab, dass die ersten vier nicht verschwindenden Zeilenvektoren wegen der Stufenform linear unabhängig sind. Entscheidend ist nun die Beobachtung, dass sich unter den Zeilenoperationen der von den Zeilenvektoren aufgespannte Teilraum, den man auch als *Zeilenraum* der Matrix A bezeichnet, nicht ändert. Weil nämlich sämtliche Zeilenoperationen linear sind, ist jeder Zeilenvektor von $S(A)$ eine Linearkombination der Zeilenvektoren von A und liegt daher im Zeilenraum von A. Daher ist der Zeilenraum von $S(A)$ im Zeilenraum von A enthalten. Ferner ist jede Elementaroperation umkehrbar und ihre Umkehrung wiederum linear. Wir können also die inversen Operationen auf $S(A)$ anwenden und erhalten A zurück. Daher ist der Zeilenraum von A im Zeilenraum von $S(A)$ enthalten und die Zeilenräume von A und von $S(A)$ stimmen tatsächlich überein. In der Sprache des übernächsten Abschnittes gibt es eine invertierbare Matrix M mit der Eigenschaft, dass

$$M \cdot A = S(A), \qquad A = M^{-1} \cdot S(A)$$

ist. Daher ändern sich beim Eliminieren zwar die Zeilen von A in jene von $S(A)$. Die Zeilenräume bleiben aber unverändert.

Daher bilden die vier, wieder als Spalten geschriebenen Vektoren

$$\vec{w}_1 = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -\frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{w}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{w}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \vec{w}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

eine Basis des ursprünglichen Zeilenraums. ◯

3.4 Der Hauptsatz

Nach der Rechnerei im umfangreichen Beispiel des letzten Abschnittes ist es an der Zeit, das Erreichte etwas aus Distanz zu betrachten, d.h. zusammenzufassen und etwas abstrakter zu formulieren. Das Eliminationsverfahren für ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mit der Koeffizientenmatrix $A \in \mathbb{R}^{m,n}$ bzw. etwas abstrakter für die zugehörige lineare Abbildung

$$f_A \colon \mathbb{R}^n \to \mathbb{R}^m, \qquad \vec{x} \mapsto A \cdot \vec{x}$$

liefert nämlich einen effizienten Zugang zu theoretischen Aussagen über vier Teilräume, die paarweise als Kern von A (geschrieben $\mathrm{Ker}(A)$) und Bild von A (geschrieben $\mathrm{Im}(A)$) und dual für die transponierte Matrix $A^T \in \mathbb{R}^{n,m}$ als $\mathrm{Ker}(A^T)$ und $\mathrm{Im}(A^T)$ auftreten und deren Dimensionen vom Rang r von A abhängen. Sie sind als *fundamentale Teilräume* von A bekannt. Das Ziel der algorithmischen linearen Algebra besteht darin, für diese vier Teilräume geeignete Basen zu bestimmen und ihre Dimensionen zu berechnen.

1. $\mathrm{Ker}(A) \subseteq \mathbb{R}^n$: Der *Kern* von A. (Der Nullraum von A) Es handelt sich um einen Teilraum von $\mathbb{R}^n$ der Dimension $n - r$.

2. $\mathrm{Im}(A) \subseteq \mathbb{R}^m$: Das *Bild* von A. (Der Spaltenraum von A) Es handelt sich um einen Teilraum von $\mathbb{R}^m$ der Dimension r.

3. $\mathrm{Ker}(A^T) \subseteq \mathbb{R}^m$: Der Kern von A^T. (Der Kokern von A) Es handelt sich um einen Teilraum von $\mathbb{R}^m$ der Dimension $m - r$.

4. $\mathrm{Im}(A^T) \subseteq \mathbb{R}^n$: Das Bild von A^T. (Der Zeilenraum von A) Es handelt sich um einen Teilraum von $\mathbb{R}^n$ der Dimension r.

Der Vollständigkeit halber einnern wir an ihre formalen Definitionen.

Definition. Es sei $A \in \mathbb{R}^{m,n}$ eine beliebige Matrix mit der zugehörigen linearen Abbildung $f_A \colon \mathbb{R}^n \to \mathbb{R}^m$. Unter ihrem *Kern* versteht man den Teilraum

$$\mathrm{Ker}(A) = \{\vec{x} \in \mathbb{R}^n \mid A \cdot \vec{x} = \vec{0}\} \subseteq \mathbb{R}^n$$

Der Kern der Matrix A kann als Hinderniss für die Injektivität der zugehörigen linearen Abbildung $f_A\colon \mathbb{R}^n \to \mathbb{R}^m$ interpretiert werden. Diese Abbildung ist genau dann injektiv, wenn der Kern so klein wie möglich d.h. $\mathrm{Ker}(A) = \{\vec{0}\}$ ist.

Unter dem *Bild* von A versteht man den Teilraum

$$\begin{aligned} \mathrm{Im}(A) &= \{\vec{b} \in \mathbb{R}^m \mid \vec{b} = A \cdot \vec{x} \text{ für einen gewissen Vektor } \vec{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m \\ &= \{A \cdot \vec{x} \in \mathbb{R}^m \mid \vec{x} \in \mathbb{R}^n\} \end{aligned}$$

Das Bild der Matrix A kann als Hinderniss für die Surjektivität der zugehörigen linearen Abbildung $f_A\colon \mathbb{R}^n \to \mathbb{R}^m$ interpretiert werden. Diese Abbildung ist genau dann surjektiv, wenn das Bild so gross wie möglich, d.h. $\mathrm{Im}(A) = \mathbb{R}^m$ ist. Das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ ist genau dann lösbar, wenn $\vec{b}$ im Bild von A liegt.

Man beachte, dass es sich beim Raumpaar $\mathrm{Ker}(A)$ und $\mathrm{Im}(A)$ im allgemeinen um Teilräume in verschiedenen Räumen handelt wie aus dem folgenden schematischen Bild hervorgeht, mit dem die vier fundamentalen Räume einer Matrix A dargestellt werden.

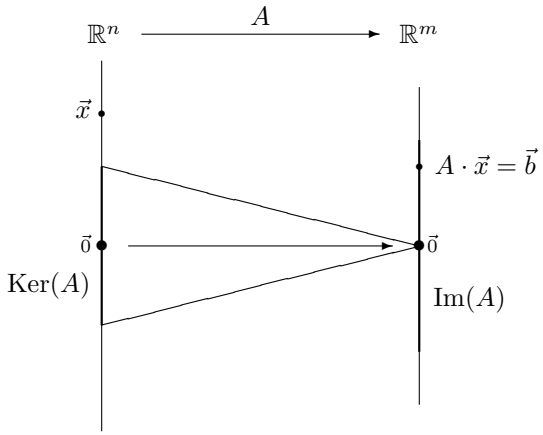


Abbildung 3.1: Der Kern und das Bild einer Matrix als Hindernisse für die Injektivität und Surjektivität der zugehörigen linearen Abbildung.

In dieser Figur sind die Streckenlängen proportional zu dem im ausführlich besprochenen Beispiel gefundenen Dimensionen gewählt.

Aus Dimensionsgründen muss diese Figur selbstverständlich schematisch ausfallen. Sie soll dazu dienen, in den folgenden Erläuterungen die Orientierung zu erleichtern und den Überblick zu ermöglichen.

Beispiel. Man kehre nochmals zum im letzten Abschnitt besprochenen ausführlichen Beispiel zurück. Dort war $m = 6$, $n = 7$ und wir haben dort die Teilräume $\mathrm{Ker}(A) \subseteq \mathbb{R}^7$ und $\mathrm{Im}(A) \subseteq \mathbb{R}^6$ identifiziert. Im Laufe der Rechnung haben wir $\dim(\mathrm{Im}(A)) = r = 4$ und $\dim(\mathrm{Ker}(A)) = k = 3$ gefunden und explizite Basen für $\mathrm{Ker}(A)$ und $\mathrm{Im}(A)$ bestimmt. Der Leser möge nun übungshalber die entsprechende Rechnungen für die transponierte Matrix A^T durchführen und sich die duale Informationen über die Teilräume $\mathrm{Ker}(A^T) \subseteq \mathbb{R}^6$ und $\mathrm{Im}(A^T) \subseteq \mathbb{R}^7$

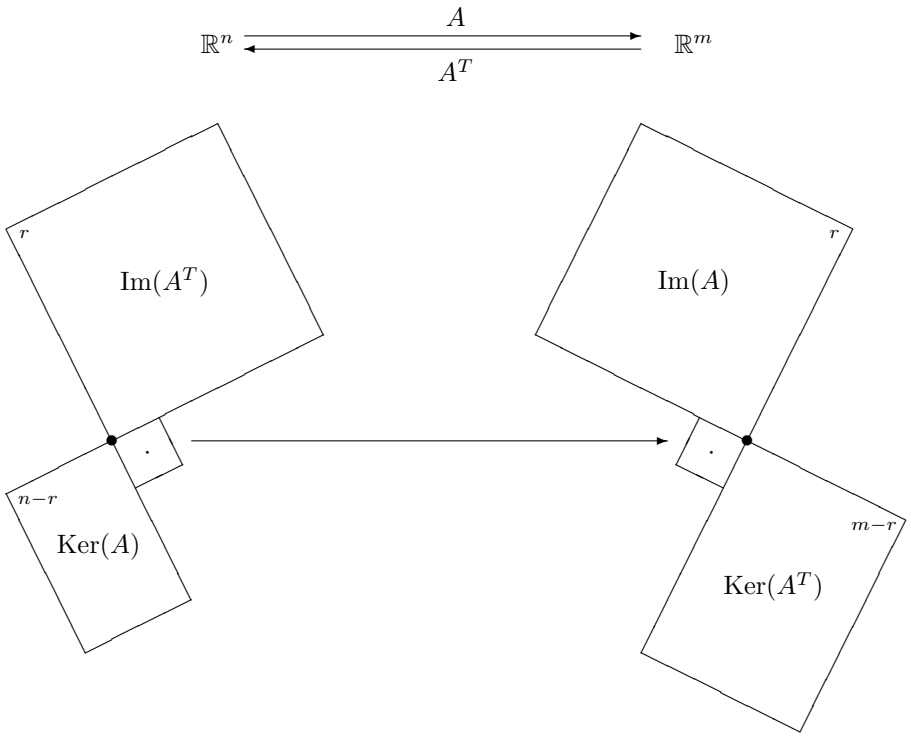


Abbildung 3.2: Der Hauptsatz: Zwei Paare orthogonaler Räume einer Matrix.

beschaffen, um die nun zu besprechenenden allgemeinen Zusammenhänge an diesem konkreten Beispiel detailiert verfolgen zu können. ◯

Aus dem Elimationsverfahren ergeben sich eine Reihe interessanter Beziehungen zwischen diesen vier fundamentalen Teilräumen. Das folgende fundamentale Resultat beschreibt eine Beziehung zwischen den beiden Räumen eines Paares. Es wir manchmal als *Hauptsatz der linearen Algebra* bezeichnet.

Satz. Für eine Matrix $A \in \mathbb{R}^{m,n}$ mit den Teilräumen $\mathrm{Ker}(A) \subseteq \mathbb{R}^n$ und $\mathrm{Im}(A) \subseteq \mathbb{R}^m$ und dem zugehörigen linearen Gleichungssystem $A \cdot \vec{x} = \vec{b}$ bezeichnen wir mit $r = \dim(\mathrm{Im}(A))$ den Rang von A und mit $k = \dim(\mathrm{Ker}(A))$ die Anzahl freier Variablen. Dann gilt die *Dimensionsformel*

$$n = \dim(\mathrm{Ker}(A)) + \dim(\mathrm{Im}(A)), \qquad n = k + r$$

Das Lösbarkeitskriterium besteht aus $m - r$ vielen homogenen Gleichungen.

Beweis. Die normierte, reduzierte Stufenform $S(A)$ hat nach Voraussetzung r Zeilen, die nicht Null sind. Jede besitzt eine 1 als führende Variable. Also gibt es r führende Variablen. Insgesamt gibt es aber n viele Variablen, also müssen $n - r$ viele frei sein. Daher ist $k = n - r$, was die erste Behauptung zeigt. Die Behauptungen über die Anzahl Gleichungen des Lösbarkeitskriteriums sind klar. □

Wir können dieses Resultat geometrisch so interpretieren, dass die Dimension des Lösungsraumes $L(A, \vec{b})$ eines Gleichungssystems mit m Gleichung in n Unbekannten $k = n - r$ ist. Damit liefert also k einen Überblick über die „Anzahl" der Lösungen eines linearen Gleichungssystems. Man beachte insbesondere, dass die Dimension des Lösungsraumes in der Regel *nicht* $n - m$ beträgt, obwohl er von m Gleichungen beschrieben wird und man intuitiv erwarten könnte, dass durch jede Gleichung, d.h. durch jede zusätzliche Bedingung, die Dimension um 1 verringert wird. Der Grund für dieses Phänomen liegt darin, dass nur die *unabhängigen* Gleichungen die Dimension um 1 verringern. Und solche unabhängigen Lösungen gibt es eben nur so viele, wie der Rang r angibt, der aber echt kleiner als m sein kann. Mit Matrizenrechnung kann also auch der wichtige Begriff der linearen Unabhängigkeit genauer untersucht werden.

Die zu einer quadratischen Matrix $A \in \mathbb{R}^{n,n}$ gehörende lineare Abbildung $f_A: \mathbb{R}^n \to \mathbb{R}^n$ ist genau dann bijektiv, wenn sie injektiv und surjektiv ist. Das ist wegen der Dimensionsformel genau dann der Fall, wenn sie injektiv oder surjektiv ist. Sie ist nämlich genau dann injektiv, wenn $\dim(\mathrm{Ker}(A)) = 0$ ist. Das ist aber auf Grund der Dimensionsformel genau dann der Fall, wenn $\dim(\mathrm{Im}(A)) = n - \dim(\mathrm{Ker}(A)) = n$ ist. Genau dann ist sie aber surjektiv. Solchen Matrizen, deren zugehörige lineare Abbildung gleichzeitig injektiv und surjektiv, d.h. bijektiv ist, widmen wir den nächsten Abschnitt.

Als unmittelbare Folgerung des Hauptsatzes ergibt sich ein Kriterium dafür, ob ein lineares Gleichungssystem überhaupt eine Lösung hat.

Satz. Das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ ist genau dann lösbar, wenn der Rang der Koeffizientenmatrix A durch Hinzufügen der Spalte $\vec{b}$ nicht grösser wird d.h. falls $\mathrm{Rang}(A)$=$\mathrm{Rang}(A \mid \vec{b})$ gilt.

Beweis. Das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ kann genau dann gelöst werden, falls $\vec{b}$ die Lösbarkeitsbedingung erfüllt. Das ist aber genau dann der Fall, wenn sich beim Dazunehmen des Konstantenvektors $\vec{b}$ der Rang nicht vergrössert. □

Als unmittelbare Folgerung können wir damit die Frage beantworten, wann ein lineares Gleichungssystem lösbar ist.

Korollar. Ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mit einer Koeffizientenmatrix A vom Typ $m \times n$ vom Rang r hat genau dann mindestens eine Lösung, wenn der Vektor $\vec{b}$ das Lösbarkeitskeitskriterium erfüllt. Falls $r = m$ ist, ist das Lösbarkeitskriterium immer erfüllt und damit existiert mindestens eine Lösung.

Wir können unter gewissen Voraussetzungen garantieren, dass ein homogenes Gleichungssystem nichttriviale Lösungen hat.

Satz. Jedes homogene Gleichungssystem $A \cdot \vec{x} = \vec{0}$ mit echt weniger Gleichungen als Unbekannte[4] besitzt eine nichttriviale Lösung.

Dieser Satz kann geometrisch so interpretiert werden, dass für $n > m$ keine injektive lineare Abbildung $f: \mathbb{R}^n \to \mathbb{R}^m$ existiert.

Beweis. Die Koeffizientenmatrix A des Systems habe m Zeilen und n Spalten mit $n > m$. Nach dem Fundamentalsatz ist $n = r + k$. Die Voraussetzung liefert

[4] sogn. unterbestimmtes Gleichungssystem.

$m < r + k$. Weil $r \leq m$ ist, erhalten wir $m < r + k \leq m + k$. Das geht aber nur, falls $k > 0$ ist. Daher existiert mindestens eine freie Variable. □

Korollar. Jedes konsistente lineare Gleichungssystem mit echt weniger Gleichungen als Unbekannte besitzt eine zweite Lösung.

Auch das folgende duale Resultat spielt eine wichtige Rolle:

Satz. Jedes Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mit echt weniger Unbekannten als Gleichungen[5] besitzt für einen gewissen Konstantenvektor $\vec{b}$ keine Lösung.

Dieser Satz kann geometrisch so interpretiert werden, dass für $n < m$ keine surjektive lineare Abbildung $f: \mathbb{R}^n \to \mathbb{R}^m$ existiert.

Beweis. Die Koeffizientenmatrix A des Systems habe m Zeilen und n Spalten mit $n < m$. Wir müssen zeigen, dass das zugehörige Existenzkriterium aus einem homogenen Gleichungssystem besteht, das mindestens eine Gleichung hat, die nicht Null ist. Solche Gleichungen können aber nie Null sein, wie obiges Verfahren zeigt: auf der rechten Seite stand ja am Anfang die Einheitsmatrix. Das zugehörige Existenzkriterium besteht aus einem homogenen Gleichungssystem mit $m - r$ Gleichungen. Wir haben also noch zu zeigen, dass $m - r > 0$ bzw. $m > r$ ist. Nach dem Fundamentalsatz ist $r + k = n$. Die Voraussetzung liefert also $r + k < m$. Weil selbstverständlich $k \geq 0$ ist, folgt daraus in der Tat $r < m$, wie behauptet wurde. □

Im Verlauf des besprochenen Beispiels haben wir gesehen, dass nicht nur der Kern und das Bild einer Matrix $A \in \mathbb{R}^{m,n}$ in einer interessanten Beziehung stehen, sondern dass diese Teilräume ferner in einer interessanten Beziehung zu den Dimensionen der entsprechenden Räumen der transponierten Matrix $A^T \in \mathbb{R}^{n,m}$ stehen. Wir erinnern daran, dass der Zeilenraum, d.h. der Unterraum von $\mathbb{R}^n$, der von den Zeilen der Matrix A aufgespannt wird, durch Transponieren als Spaltenraum der transponierten Matrix A^T d.h. als Bild $\mathrm{Im}(A^T)$ aufgefasst werden kann.

Man beachte, dass der Spaltenraum von A, d.h. das Bild $\mathrm{Im}(A)$ ein Teilraum von $\mathbb{R}^m$ ist. Daher liegen der Zeilenraum $\mathrm{Im}(A^T)$ und der Spaltenraum $\mathrm{Im}(A)$ von A in unterschiedlichen Räumen! Um so bemerkenswerter ist es, dass die Dimensionen dieser beiden Teilräume d.h. des Zeilen- und des Spaltenraumes von A trotz der scheinbaren Unvergleichbarkeit übereinstimmen. Diese Gleichheit zwischen den Dimensionen des Zeilen- und Spaltenräumen d.h. zwischen den Bildern von zwei transponieten Matrizen A und A^T kann als *Zusatz zum Hauptsatzes der linearen Algebra* angesehen werden.

Satz. Für eine Matrix $A \in \mathbb{R}^{m,n}$ mit dem Rang r gilt für die Dimension der Bildes und des Zeilenraumes die Beziehung

$$\dim(\mathrm{Im}(A)) = r = \dim(\mathrm{Im}(A^T))$$

Für die Dimensionen der beiden anderen involvierten Räume Kern und Kokern gilt

$$\dim(\mathrm{Ker}(A)) = n - r, \qquad \dim(\mathrm{Ker}(A^T)) = m - r$$

[5] sogn. überbestimmtes Gleichungssystem.

Für den Rang von A gilt die Abschätzung

$$0 \leq r \leq \min(m, n)$$

Man formuliert die erste Behauptung dieses Satzes oft so, dass man sagt, dass der Spaltenrang und der Zeilenrang einer Matrix A übereinstimmen.

Beweis. Um die Gleichheit zwischen den Dimensionen von Zeilen- und Spaltenraum einzusehen, denken wir uns die Matrix A in die normierte, reduzierte Stufenform $S(A)$ übergeführt. In dieser Form sind die Dimensionen der beiden Räume leicht zu bestimmen. Der Rang von A ist definitionsgemäss die Anzahl Zeilen $r = \text{Rang}(A)$ von $S(A)$, die nicht verschwinden.

Um die linke Gleichheit einzusehen, beachten wir, dass das Bild $\text{Im}(A)$, d.h. der Spaltenraum von A definitionsgemäss von den Spaltenvektoren von A aufgespannt wird. Man beachte, dass die Matrizen A und $S(A)$ nicht etwa die selben Spaltenräume haben! Behauptet wird hier nur, dass die *Dimensionen* dieser Räume mit dem Rang r übereinstimmen! Das folgt aber aus der Tatsache, dass eine Linearkombination der Spaltenvektoren von A genau dann den Nullvektor ergibt, wenn die selbe Linearkombination der Spaltenvektoren von $S(A)$ den Nullvektor ergibt. In matrizieller Sprache heisst das, dass $A \cdot \vec{x} = \vec{0}$ ist, genau wenn $S(A) \cdot \vec{x} = \vec{0}$ gilt. Weil die r Spalten mit den führenden Variabeln eine Basis des Bildes von $S(A)$ bilden, gilt die linke Beziehung $\dim(\text{Im}(A)) = r = \text{Rang}(A)$.

Um die rechte Gleichheit einzusehen, erinnern wir uns, dass der Zeilenraum von A idefinitionsgemäss der Teilraum ist, der von den Zeilenvektoren von A aufgespannt wird. Weil er sich unter Elementaroperationen nicht ändert, kann seine Dimension auch an Hand der Matrix $S(A)$ bestimmt werden. Man erkennt aber wegen der speziellen Stufen-Gestalt der normierten, reduzierten Stufenform leicht, dass die ersten r Zeilen der Matrix $S(A)$ linear unabhängig sind. Sie erzeugen aber bereits den ganzen Zeilenraum, weil die verbleibenden Nullzeilen nichts hinzufügen. Daher bilden diese r Zeilen eine Basis des Zeilenraumes und seine Dimension ist tatsächlich r. Damit ist die rechte Beziehung $\dim(\text{Im}(A^T)) = r = \text{Rang}(A)$ gezeigt.

Die Dimensionen der beiden anderen involvierten Räume $\text{Ker}(A)$ und $\text{Ker}(A^T)$ ergeben sich unmittelbar aus dem Fundamentalsatz der linearen Algebra zusammen mit dem soeben gezeigten Zusatz.

Um die Abschätzung für den Rang einzusehen, beachten wir zunächst, dass selbstverständlich die Ungleichung $r \leq m$ gilt. Das duale Resultat und der soeben gezeigte Zusatz liefern die Ungleichung $r \leq n$. Zusammengefasst liefern diese beiden Ungleichungen die behauptete Abschätzung $r \leq \min(m, n)$ für den Rang von A. □

Im ausführlich besprochenen Beispiel haben wir noch etwas mehr gesehen und erkannt, dass der Kern mit der Dimension k und der Zeilenraum mit der Dimension r einer Matrix $A \in \mathbb{R}^{m,n}$ nicht bloss durch die Dimensionsbedingung $n = k + r$ verknüpft sind, sondern durch das Skalarprodukt noch enger miteinander zusammenhängen. Wir haben nämlich gesehen, dass jeder Basisvektor des Kerns $\text{Ker}(A)$ auf jedem Zeilenvektor der Koeffizientenmatrix A senkrecht steht. Diese Eigenschaft besagt, dass der Kern $\text{Ker}(A)$ und der Zeilenraum $\text{Im}(A^T)$ von A Orthogonalräume sind.

Definition. Für einen beliebigen Teilvektorraum $L \subseteq \mathbb{R}^l$ verstehen wir unter dem *Orthogonalraum* von L den Teilvektorraum

$$L^\perp = \{\vec{x} \in \mathbb{R}^l \mid \langle \vec{x}, \vec{v} \rangle = 0, \vec{v} \in L\} = \{\vec{x} \in \mathbb{R}^l \mid \vec{x} \perp \vec{v} \text{ für alle } \vec{v} \in L\}$$

Er ist orthogonal zu L, da definitonsgemäss jedes seiner Elemente $\vec{x}$ zu jedem Vektor $\vec{v}$ aus L orthogonal ist.

Der Orthogonalraum hat eine Reihe von Eigenschaften, die man von einem Komplement erwarten würde.

Satz. Es sei $L \subseteq \mathbb{R}^l$ ein Teilvektorraum. Dann gilt:

1. Der Orthogonalraum $L^\perp$ ist ein Teilraum von $\mathbb{R}^l$.
2. Es ist $\left(\mathbb{R}^l\right)^\perp = \{\vec{0}\}$.
3. Für die Dimensionen gilt $\dim(L) + \dim(L^\perp) = \dim(\mathbb{R}^l) = l$.
4. Es ist $\left(L^\perp\right)^\perp = L$.
5. Es ist $\dim(L^\perp) = n - \dim(L)$.

Die vielleicht erwartete weitere Bedingung $L \cap L^\perp = \{\vec{0}\}$ eines Komplements haben wir nicht formuliert. Sie gilt zwar über den Körpern der reellen und komplexen Zahlen, weil dort das Skalarprodukt nicht entartet ist und deshalb aus der Bedingung $\langle \vec{a}, \vec{a} \rangle = 0$ die Eigenschaft $\vec{a} = \vec{0}$ folgt. Wenn aber in L ein Vektor $\vec{a} \neq \vec{0}$ existieren würde, der auf sich selber orthogonal ist, d.h. für den $\langle \vec{a}, \vec{a} \rangle = 0$ gilt, so kann $L \cap L^\perp \neq \{\vec{0}\}$ sein. Genau solche Vektoren — man nennt sie *isotrop* — für die also $\vec{a} \perp \vec{a}$ ist, existieren aber in der Regel über endlichen Körpern und man muss mit dieser Bedingung, die besagt, dass $\mathbb{R}^l$ eine direkte Summe von L und $L^\perp$, d. h. $L \oplus L^\perp = \mathbb{R}^l$ ist, etwa in der Kodierungstheorie aufpassen. Dort spielen selbstorthogonale Teilvektorräume, für die $L^\perp = L$ ist, eine wichtige Rolle.

Mit dieser Bezeichnung formulieren wir die *Verschärfung des Hauptsatzes der linearen Algebra*.

Satz. Für ein Matrix $A \in \mathbb{R}^{m,n}$ mit den zugehörigen fundamentalen Teilräumen $\mathrm{Ker}(A) \subseteq \mathbb{R}^n$ und $\mathrm{Im}(A) \subseteq \mathbb{R}^m$ gilt in $\mathbb{R}^n$ die Orthogonalitätsbeziehung $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^T)$. Der Zeilenraum von A ist also der Orthogonalraum seines Kerns.

Ersetzen wir A durch A^T, erhalten wir die dualen Teilräume $\mathrm{Ker}(A^T) \subseteq \mathbb{R}^m$ und $\mathrm{Im}(A^T) \subseteq \mathbb{R}^n$ und damit in $\mathbb{R}^m$ die duale Orthogonalitätsbedingung $\mathrm{Im}(A)^\perp = \mathrm{Ker}(A^T)$. Der Kokern von A ist also der Orthogonalraum seines Bildes.

Beweis. Um einzusehen, warum jeder Vektor $\vec{x} \in \mathrm{Ker}(A)$ orthogonal zu jedem Zeilenvektor sein muss, erinnern wir uns, dass für einen Vektor $\vec{x}$ im Kern definitionsgemäss $A \cdot \vec{x} = \vec{0}$ gilt. Die Elemente des Zeilenraumes $\mathrm{Im}(A^T)$ lassen sich definitionsgemäss in der Form $A^T \cdot \vec{y}$ beschreiben. Für das Skalarprodukt dieser beiden Vektoren gilt wegen seiner Adjunktionseigenschaft

$$\langle \vec{x}, A^T \cdot \vec{y} \rangle = \langle A \cdot \vec{x}, \vec{y} \rangle = \langle \vec{0}, \vec{y} \rangle = 0$$

die behauptete Orthogonalitätsbeziehung. □

Hinter diesem Sachverhalt steckt die bekannte Tatsache, dass das Matrixprodukt $A \cdot \vec{x}$ mit Hilfe von Skalarprodukten der Zeilenvektoren von A mit dem Vektor $\vec{x}$ interpretiert werden kann. Falls $\vec{x}$ im Kern von A liegt, so verschwinden alle diese Skalarprodukte. Das heisst aber gerade, dass $\vec{x}$ orthogonal zu sämtlichen Zeilenvektoren von A ist. Dieser Sachverhalt dürfte aus der Vektorgeometrie bekannt sein, wo man ihm im Spezialfall begegnet ist, wo die Matrix A aus einem einzigen Zeilenvektor besteht.

Beispiel. Eine Hyperebene wird durch eine einzige lineare Gleichung

$$\langle \vec{a}, \vec{x} \rangle = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b$$

mit der erweiterten Matrix $(\vec{a}^T, b)$ beschrieben, wobei die Koeffizientenmatrix durch den Zeilenvektor

$$\vec{a}^T = (a_1, a_2, \ldots, a_n)$$

gegeben ist. Die Hyperebene ist also definitionsgemäss der Lösungsraum $L(\vec{a}^T, b)$. Der Lösungsraum des zugehörigen homogenen Systems ist der Kern $L(\vec{a}^T, 0) = \mathrm{Ker}(\vec{a}^T)$ und liegt parallel zur gegebenen Hyperebene. Wegen $\mathrm{Ker}(\vec{a}^T)^\perp = \mathrm{Im}(\vec{a})$ ist sein Orthogonalraum der Zeilenraum $\mathrm{Im}(\vec{a})$, der vom einzigen Vektor $\vec{a} \in \mathbb{R}^n$ aufgespannt wird. Dieser Vektor $\vec{a}$ kann daher als Normalenvektor der Hyperebene aufgefasst werden. ◯

3.5 Gleichungssysteme mit Parametern

In den Anwendungen kommen in einem linearen Gleichungssystem oft neben den Variablen, nach denen aufzulösen ist, noch andere unbekannte Hilfsgrössen vor. Solche konstante aber unbestimmt gelassene Zahlen nennt man *Parameter*. Aus dem Kontext ist meistens klar, welches die Parameter und welches die Lösungsvariablen sind. Mit Parametern im Konstantenvektor können wir bereits umgehen. Hier sollen also vor allem Parameter in der Koeffizientenmatrix ins Auge gefasst werden, mit denen viel vorsichtiger umgegangen werden muss.

Beim Lösen von linearen Gleichungssystemen mit Parametern muss man *vor* jeder Elementaroperation sorgfältig überprüfen, ob man eine umkehrbare Operation durchführen kann. Insbesondere verwendet man zweckmässigerweise nach Möglichkeit die Operationen vom Typ I und II, die immer umkehrbar sind. Bei der Verwendung der Operation vom Typ III bzw. bei ganzzahliger Rechnung vom Typ IV, muss man mit Hilfe von Fallunterscheidungen dafür sorgen, dass die entsprechende Operation umkehrbar ist. Wir illustrieren das Vorgehen an einem kleinen Beispiel, an dem die typische Schwierighkeit aber bereits ersichtlich ist.

Beispiel. Gegeben ist das Gleichungssystem mit Parameter d:

$$\begin{cases} x + dy = 6 \\ dx + y = 6 \end{cases}$$

Die zugehörige erweiterte Matrix lautet:

$$\left(\begin{array}{cc|c} 1 & d & 6 \\ d & 1 & 6 \end{array}\right)$$

Addition des $(-d)$-fachen der ersten Zeile zur zweiten Zeile ist eine Operation vom Typ II und deshalb spielt es es keine Rolle ob $d = 0$ ist oder nicht. Also ist keine Fallunterscheidung nötig und die umgeformte Matrix lautet:

$$\left(\begin{array}{cc|c} 1 & d & 6 \\ 0 & 1-d^2 & 6-6d \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-d) \\ \\ \end{array}\right]$$

Um zu einer reduzierten Stufenform zu kommenen, müssen wir das $(-d)$-fache der zweiten Zeile zum $(1-d^2)$-fachen der ersten Zeile addieren. Dieser Schritt ist aber nur dann unbedenklich, wenn $1-d^2 \neq 0$ ist. Daher sind wir nun gezwungen, Fälle zu unterscheiden. Wir weisen darauf hin, dass für die Fallunterscheidungen mit *polynomialen* Gleichungen in den Parametern zu rechnen ist. Solche Gleichungen erfordern zur Lösung Methoden, die in der algebraischen Geometrie oder in der Analysis untersucht werden.

1. Normalfall: $1 - d^2 \neq 0$. In diesem Fall ist der gewünschte Schritt erlaubt und wir erhalten die reduzierte Stufenform

$$\left(\begin{array}{cc|c} 1-d^2 & 0 & 6-6d \\ 0 & 1-d^2 & 6-6d \end{array}\right) \qquad \left[\begin{array}{c} \\ ZS_{12}(1-d^2) \\ \\ \end{array}\right]$$

   Um die normierte Stufenform zu erhalten, müssten wir nun beide Zeilen durch $1 - d^2$ dividieren, was nach unserer Voraussetzung erlaubt ist. Die normierte Stufenform lautet damit

$$\left(\begin{array}{cc|c} 1 & 0 & \frac{6}{1+d} \\ 0 & 1 & \frac{6}{1+d} \end{array}\right) \qquad \left[\begin{array}{c} S_1(1-d^2) \\ S_2(1-d^2) \end{array}\right]$$

   Für die Lösung ergibt sich also in diesem Fall:

$$x = \frac{6}{1+d}, \qquad y = \frac{6}{1+d}$$

   In diesem Fall hat das System also eine eindeutige Lösung. Offenbar müssen wir als Lösung eines linearen Gleichungssystems im allgemeinen mit rationalen Ausdrücken in den Parametern rechnen. Das Beispiel zeigt, dass es nicht genügt, diese rationalen Ausdrücke zu bestimmen und anschliessend deren Nenner zu untersuchen. Im vorliegenden Fall würde man so nur auf den Sonderfall $d = -1$ stossen.

2. Sonderfall $1 - d^2 = 0$. Diese quadratische Gleichung hat zwei Lösungen, die entsprechend zwei Unterfälle erfordern.

   2.1 $d = 1$. Setzen wir diesen speziellen Wert in die letzte Matrix vor der betreffenden Fallunterscheidung ein, so lautet sie

$$\left(\begin{array}{cc|c} 1 & 1 & 6 \\ 0 & 0 & 0 \end{array}\right)$$

   Das zugehörige System $x + y = 6$ hat in diesem Fall unendlich viele Lösungen der Form $x = 6 - t, y = t$.

2.2 $d = -1$. Setzen wir diesen speziellen Wert in die Matrix ein, die wir erhalten haben, bevor wir zur Fallunterscheidung gezwungen wurden, erhalten wir diesmal

$$\left(\begin{array}{cc|c} 1 & -1 & 6 \\ 0 & 0 & 12 \end{array}\right)$$

Das zugehörige System hat in diesem Fall also keine Lösung.

Der Leser mache sich die zugehörigen Verhältnisse an einer Figur klar. ◯

Bereits aus diesem einfachen Beispiel wird deutlich, dass die verschiedenen Fälle in der Regel von den Nullstellen polynomialer Gleichungen abhängen und daher diese Fallunterscheidungen eigentlich nicht mehr zur linearen Algebra gehören, wo Systeme *linearer* Gleichungen untersucht werden. Die algebraische Geometrie, wo die Lösungsmengen von allgemeineren Systemen polynomialer Gleichungen untersucht werden, ist deutlich anspruchsvoller und weniger effektiv als die lineare Algebra und liegt jenseits unserer derzeitigen Möglichkeiten.

Man hüte sich vor einer — unter gewissen „Praktikern“ recht schicken — Theorienfeindlichkeit. Zwar mag Mathematik manchmal etwas abstrakt aussehen und die Anmarschroute zum Ziel beschwerlich sein; irrelevant ist sie deshalb noch lange nicht. So trifft es zwar zu, dass die Sonderfälle statistisch gesehen selten auftreten. Das macht sie aber nicht minder sondern im Gegenteil erst recht interessant. Sehr oft ist man nämlich ausschliesslich an diesen Sonderfällen interessiert, denn dort passiert, physikalisch interpretiert, etwas Gewünschtes, etwas Gefährliches oder etwas Neues. Bei der Netzwerksynthese müssen Komponenten so dimensioniert werden, dass sich das Netzwerk in einer gewünschten Art benimmt. Eine Suchmaschine muss die Treffer so sortieren, dass die gewünschte Information möglichst weit oben steht. Durch Berechnen der Eigenwerte findet man Eigenfrequenzen eines schwingenden Systems. Eigenwerte liefern in der Mechanik kritische Drehzahlen von rotierenden Wellen, in der Elektrotechnik Eigenfrequenzen von Schaltkreisen, die man je nach Anwendung vermeiden oder treffen möchte, auf jeden Fall aber kennen muss. Sie geben auch über die Stabilität eine Systems Auskunft. Mit Hilfe der Eigenwerte lässt sich das Langzeitverhalten von dynamischen Systemen bzw. das Gleichgewicht von Prozessen vorhersagen. In der Quantemechanik spielen die Eigenwerte die Rolle von Messwerten, die dann etwa in Form von Spektrallinien beobachtet werden können.

Sonderfälle sind also nicht einfach als mathematische Pathologien abzutun, sondern als Indiz dafür zu werten, dass an dieser Stelle die Voraussetzungen des Normalfalls nicht erfüllt sind und das System einen „Phasenübergang“, d.h. qualitativ ein ganz anderes Verhalten zeigen wird. Nichts ist so praktisch, wie eine gute Theorie! Komplexe Sachverhalte müssen halt auch in einer komplexen Sprache formuliert werden, die den diversen Verästelungen Rechnung trägt. Nur Fix und Foxy sind in der Lage, ihre Welt mit wenigen simplen Worten vollständig zu beschreiben. Wenn man dem Berufsmilitär General Eisenhower nachsagt, dass er sich stets geweigert habe, auf Fragen einzutreten, die nicht auf 130 Worte komprimiert waren, kann man leicht ausrechnen, welches der maximale geistige Tiefgang eines Generals ist. Man vergleiche zum Kontrast die weniger beschränkte Haltung des Wissenschaftlers A. Einstein, der einem Journalisten, der ihn gebeten hatte, ihm die Relativitätstheorie einfach zu erklären, erwiderte: „So einfach wie möglich; aber nicht einfacher.“

Mit dem wichtgen Spezialfall des Eigenwertproblems werden wir uns später eingehend befassen.

3.6 Elementarmatrizen

Das Eliminationsverfahren basiert auf der Tatsache, dass die elementaren Zeilenumformungen den Lösungsraum eines linearen Gleichungssystems nicht ändern. Dieses Faktum bleibt einzusehen. Dazu holen wir etwas aus und überlegen uns, dass elementare Zeilenumformungen durch Multiplikation mit gewissen quadratischen Matrizen entstehen. Dadurch werden wir weitere Einsichten in den Algorithmus erhalten.

Definition. Eine $m \times m$-Matrix, die aus der Einheitsmatrix E_m durch eine einzige Zeilenoperation hervorgeht, heisst *Elementarmatrix*.

Entsprechend den drei Typen von elementaren Zeilenoperationen gibt es drei Typen von Elementarmatrizen:

I. Vertauschen von zwei Zeilen liefert die *Vertauschungsmatrizen*. Vertauschen der i-ten und der j-ten Zeile liefert die folgende Vertauschungsmatrix:

$$T_{ij} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & & & & 1 & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & 1 & & & & 0 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{bmatrix} \\ \\ \\ \\ T_{ji} \\ \\ \\ T_{ij} \\ \\ \\ \end{bmatrix}$$

Ausser den eingetragenen Matrixelemente sind dabei alle Elemente Null. Auf der Diagonalen stehen Einsen ausser an den Stellen i und j, wo die Einsen durch Null ersetzt sind. An den Stellen (i, j) und (j, i) sind dafür zwei Einsen hinzugefügt worden.

II. Addition des Vielfachen einer Zeile zu einer anderen Zeile liefert die *Scherungen*. Addition der r-fachen der i-ten zur j-ten Zeile liefert die folgende Scherungsmatrix:

$$Z_{ij}(r) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & r & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \quad \left[\; Z_{ij}(r) \;\right]$$

Sämtliche Diagonalelemente einer solchen Matrix sind Einsen. Es gibt genau einen Eintrag r an der Stelle (i, j) ausserhalb der Diagonalen.

III. Multiplikation einer Zeile mit einem von Null verschieden Skalar liefert die *Streckungen*. Multiplikation der i-ten Zeile mit dem Faktor r liefert die folgende Diagonalmatrix:

$$S_i(r) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & r & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \quad \left[\; S_i(r) \;\right]$$

Sämtliche Diagonalelemente einer solchen Matrix sind Einsen. Es gibt genau einen Eintrag $r \neq 0$ auf der Diagonalen.

Beispiel. Die Elementarmatrizen vom Typ 2×2 sehen wie folgt aus:

$$T_{12} = T_{21} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Z_{21}(r) = \begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix}, \quad Z_{12}(r) = \begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix}$$

$$S_1(r) = \begin{pmatrix} r & 0 \\ 0 & 1 \end{pmatrix}, \quad S_2(r) = \begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix}$$

Um von den zugehörigen Elementaroperationen eine gewisse Vorstellung zu erhalten, machen wir uns je ein geometrisches Bild dieser Matrizen. Jede unserer Elementarmatrizen liefert nämlich eine lineare Abbildung $\mathbb{R}^2 \to \mathbb{R}^2$ der Ebene in sich selbst. ◯

Beispiel. Durch die Elementarmatrix T_{12} wird die Zuordnung

$$T_{12}\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x \end{pmatrix}$$

beschrieben. Dabei wird der Punkt $P = (x, y)$ in den Punkt $P^* = (y, x)$ abgebildet. Die zugehörige geometrische Transformation ist eine Geradenspiegelung an der ersten Winkelhalbierenden. Das typische Transformationsverhalten kann an folgender Figur abgelesen werden:

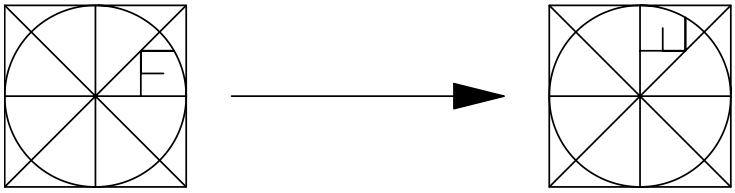


Abbildung 3.3: Interpretation der Elementaroperation vom Typ I.

Man beachte, dass unter einer Spiegelung Längen, Winkel und Flächeninhalte erhalten bleiben. Entsprechend ist die Matrix T_{12} orthogonal. ◯

Beispiel. Als nächstes betrachten wir die Elementarmatrix $Z_{21}(1)$. Durch sie wird die Zuordnung

$$Z_{21}(1)\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y \\ y \end{pmatrix}$$

beschrieben. Dabei wird der Punkt $P = (x, y)$ in den Punkt $P^* = (x + y, y)$ abgebildet. Die zugehörige geometrische Transformation ist eine Scherung in Richtung der ersten Achse mit dem Scherungsfaktor 1. Das typische Transformationsverhalten kann an folgender Figur abgelesen werden:

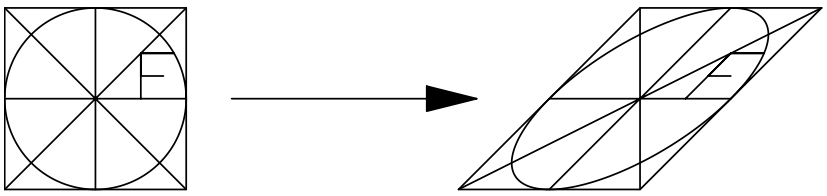


Abbildung 3.4: Interpretation der Elementaroperation vom Typ II.

Man beachte, dass unter einer Scherung Längen und Winkel nicht erhalten bleiben. Sie erhalten allerdings Flächeninhalte und spielen deshalb als sogn.

Flächenverwandlungen eine Rolle in der Schulmathematik. Entsprechend ist die Matrix $Z_{21}(1)$ nicht orthogonal, aber ihre Determinante ist 1. ◯

Beispiel. Als nächstes betrachten wir die Elementarmatrix $S_1(2)$. Durch sie wird die Zuordnung

$$S_1(2)\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x \\ y \end{pmatrix}$$

beschrieben. Dabei wird der Punkt $P = (x, y)$ in den Punkt $P^* = (2x, y)$ abgebildet. Die zugehörige geometrische Transformation ist eine Streckung um den Faktor 2 in Richtung der ersten Achse. Das typische Transformationsverhalten kann an folgender Figur abgelesen werden:

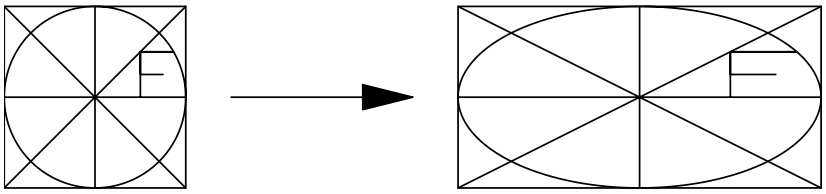


Abbildung 3.5: Interpretation der Elementaroperation vom Typ III.

Unter einer Streckung bleiben weder Längen, Winkel noch Flächeninhalte erhalten. Sie erhalten allerdings Geraden und sind Parallelen- und Teilverhältnistreu. Streckungen spielen beim Skalieren einer Figur eine Rolle. Entsprechend ist die Matrix nicht orthogonal und ihre Determinante ist 2. ◯

Wie wir sehen werden, spielen Spiegelungen, Scherungen und Streckungen eine fundamentale Rolle als Grundbausteine aller invertierbarer linearer Abbildungen. Das hängt damit zusammen, dass die Multiplikation einer Matrix A von *links* mit einer Elementarmatrix M einer elementaren Zeilenoperation entspricht. Daher können sämtliche elementaren Zeilenoperationen als Multiplikation mit geeigneten Elementarmatrizen von links aufgefasst werden.

Satz. Die Elementarmatrix M sei aus der Einheitsmatrix E_m durch eine elementare Zeilenoperation hervorgegangen. Ist nun A eine $m \times n$-Matrix, so ist das Produkt $M \cdot A$ gerade die Matrix, die durch Anwenden derselben Zeilenoperation auf A entsteht.

Statt eines formalen Beweises dieses Satzes, der als Übungsaufgabe in Matrizenrechnung betrachtet werden kann, geben wir in jedem Fall ein illustrierendes Beispiel.

Beispiel. Wir gehen von der konkreten Matrix

$$A = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix}$$

aus und betrachten die folgenden Matrizenprodukte:

I. Das Produkt $T_{23} \cdot A$ ist die Matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 1 & 4 & 4 & 0 \\ 2 & -1 & 3 & 6 \end{pmatrix} \quad \begin{bmatrix} \\ T_{32} \\ T_{23} \end{bmatrix}$$

die aus A durch Vertauschen der zweiten mit der dritten Zeile entsteht.

II. Das Produkt $Z_{13}(3) \cdot A$ ist die Matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 4 & 4 & 10 & 9 \end{pmatrix} \quad \begin{bmatrix} \\ \\ Z_{13}(3) \end{bmatrix}$$

die aus A durch Addition des Dreifachen der ersten zur dritten Zeile entsteht.

III. Das Produkt $S_2(5) \cdot A$ ist die Matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 10 & -5 & 15 & 30 \\ 1 & 4 & 4 & 0 \end{pmatrix} \quad \begin{bmatrix} S_2(5) \\ \end{bmatrix}$$

die aus A durch Multiplikation des Fünffachen der zweiten Zeile entsteht.

In jedem Fall lässt sich die elementare Zeilenoperation durch Multiplikation mit der betreffenden Elementarmatrix von links bewerkstelligen. ◯

Für Elementarmatrizen gilt ein fundamentales Resultat.

Satz. Jede Elementarmatrix ist invertierbar. Ihre Inverse ist wieder eine Elementarmatrix vom betreffenden Typ.

Beweis. Wir gehen die einzelnen Typen der Reihe nach durch und geben in jedem Fall eine Inverse an. Das formale Nachrechnen überlassen wir wieder den Mathematik-Studenten.

I. Es ist $T_{ij}^{-1} = T_{ij}$.

II. Es ist $Z_{ij}^{-1}(r) = Z_{ij}(-r)$.

III. Es ist $S_i^{-1}(r) = S_i(\frac{1}{r})$.

Selbstverständlich muss im letzten Fall $r \neq 0$ sein, damit $S_i(\frac{1}{r})$ existiert. In der Tat ist $S_i(r)$ sonst nicht invertierbar. □

Dieser Satz lässt sich mit elementaren Zeilenoperationen umformulieren. Da diese Operationen nämlich der Multiplikation mit gewissen Elementarmatrizen entsprechen, folgt daraus, dass die elementaren Zeilenoperationen umkehrbar sind, d.h. zu jeder Zeilenoperation gibt es eine inverse Zeilenoperation vom selben Typ. Es gilt:

I. Vertauschen von i-ter und j-ter Zeile lässt sich umkehren. Die Umkehrung besteht aus der Vertauschung von i-ter und j-ter Zeile.

II. Addition des r-fachen der i-ten zur j-ten Zeile lässt sich umkehren. Die Umkehrung besteht aus der Addition des $-r$-fachen der i-ten zur j-ten Zeile.

III. Multiplikation der i-ten Zeile mit $r \neq 0$ lässt sich umkehren. Die Umkehrung besteht aus der Multiplikation der i-ten Zeile mit $\frac{1}{r}$.

Es sei nicht verschwiegen, dass wir seit einiger Zeit etwas zu viel arbeiten. Es stellt sich heraus, dass die Elementaroperationen vom Typ I aus den beiden anderen Typen erzeugt werden können. Man kann nämlich das folgende Resultat beweisen, das in der K-Theorie benutzt wird.

Satz. Es gilt $T_{ij} = S_i(-1) \cdot Z_{ij}(1) \cdot Z_{ji}(-1) \cdot Z_{ij}(1)$.

Auch diesen Satz veranschaulichen wir uns nur an einem konkreten Beispiel.

Beispiel. Wir überprüfen für $n = 2$ die Matrizengleichung:

$$T_{12} = S_1(-1) \cdot Z_{12}(1) \cdot Z_{21}(-1) \cdot Z_{12}(1)$$

Eine Übung in Matrizenrechnen zeigt in der Tat:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Der Leser mache sich diesen Sachverhalt an Hand eines Filmes, bestehend aus 4 Bildern, klar. ◯

Obwohl also die Elementaroperationen vom Typ I für unsere Zwecke streng genommen überflüssig sind, sind sie dermassen praktisch, dass wir nicht darauf verzichten wollen.

Wir können nun, wie angekündigt, beweisen, dass die Elementaroperationen die Lösungsmenge eines linearen Gleichungssystems nicht ändern.

Satz. Es sei $A \cdot \vec{x} = \vec{b}$ ein lineares Gleichungssystem und M eine invertierbare Matrix. Dann stimmt der Lösungsraum $L(A, \vec{b})$ des gegebenen Gleichungssystems mit dem Lösungsraum $L(M \cdot A, M \cdot \vec{b})$ des transformierten linearen Gleichungssystems $(M \cdot A) \cdot \vec{y} = M \cdot \vec{b}$ überein.

Beweis. Um zu zeigen, dass die beiden Lösungsmengen

$$L(A, \vec{b}) = L(M \cdot A, M \cdot \vec{b})$$

übereinstimmen, zeigen wir dass die eine in der anderen enthalten ist und umgekehrt.

Es sei $\vec{x}$ eine Lösung des gegebenen Gleichungssystems. Es gilt also $A \cdot \vec{x} = \vec{b}$. Multiplizieren wir die Gleichung von links mit der Matrix M, ergibt die Gleichung $(M \cdot A) \cdot \vec{x} = M \cdot \vec{b}$. Daher ist $\vec{x}$ auch Lösung des transformierten Systems.

Aus diesem Gleichungssystem kann wegen der Invertierbarkeit von M das alte zurückgewonnen werden. Ist nämlich $\vec{y}$ eine Lösung des transformierten Systems, d.h. gilt $(M \cdot A) \cdot \vec{y} = M \cdot \vec{b}$ und multiplizieren wir diese Gleichung von links mit M^{-1}, ergibt sich die Gleichung $A \cdot \vec{y} = \vec{b}$. Daher erfüllt $\vec{y}$ auch das erste Gleichungssystem. □

3.7 Reguläre lineare Gleichungssysteme

In diesem Abschnitt wollen wir eine häufig vorkommende Klasse linearer Gleichungssysteme untersuchen. Dazu gehen wir vom linearen Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mit der selben Anzahl Gleichungen wie Unbekannten aus.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

Seine Koeffizientenmatrix $A \in \mathbb{R}^{n,n}$ ist also quadratisch. Falls die Matrix A invertierbar ist, sagen wir, das Gleichungssystem sei *regulär*. Anderenfalls heisst es *singulär*. Die selbe Sprechweise benutzen wir auch für die zugehörigen Koeffizientenmatrix. Reguläre Matrizen sind also im Prinzip das selbe wie invertierbare Matrizen und die Untersuchung des regulären Falles entspricht der Untersuchung invertierbarer Matrizen.

3.7.1 Der Äquivalenzsatz

Wir wissen bereits, dass im regulären Fall die Lösung des Gleichungssystems $A \cdot \vec{x} = \vec{b}$ eindeutig bestimmt ist und in der Form $\vec{x} = A^{-1} \cdot \vec{b}$ ausgedrückt werden kann.

Für quadratische Matrizen gibt es einen engen Zusammenhang zwischen der Invertierbarkeit von A, der Lösbarkeit eines linearen Gleichungssystems, der reduzierten Stufenform und den Elementarmatrizen. Er wird im folgenden Satz ausgedrückt.

Satz. Für eine $n \times n$-Matrix A sind folgende Aussagen äquivalent:

1. Die Matrix A ist invertierbar.
2. Das homogene System $A \cdot \vec{x} = \vec{0}$ hat nur die triviale Lösung.
3. Die Matrix A lässt sich durch eine Folge von Elementaroperationen in die Einheitsmatrix E_n überführen.
4. Die Matrix A ist ein Produkt von Elementarmatrizen.
5. Es gibt eine Matrix X mit $X \cdot A = E$.
6. Es gibt eine Matrix Y mit $A \cdot Y = E$.
7. Das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ ist für jeden Vektor $\vec{b}$ eindeutig lösbar.
8. Das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ ist für jeden Vektor $\vec{b}$ lösbar.

Wegen der fünften und sechsten Aussage müssen wir in Zukunft für die Invertierbarkeit einer Matrix nur nur eine der beiden Bedingung $X \cdot A = E$ bzw. $A \cdot Y = E$ überprüfen. Die andere Bedingung ist also automatisch erfüllt.

Beweis. Zum Beweis des Satzes überprüfen wir die unorthodoxen Implikationsketten $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$; $1 \Rightarrow 5 \Rightarrow 2$; $5 \Rightarrow 6 \Rightarrow 1$ und $1 \Rightarrow 7 \Rightarrow 8 \Rightarrow 1$.

$1 \Rightarrow 2$: Es sei A invertierbar und $\vec{z}$ eine Lösung des homogenen Gleichungssystemes. Es ist also $A \cdot \vec{z} = \vec{0}$. Multiplikation dieser Gleichung mit A^{-1} von links liefert $A^{-1}(A\vec{z}) = \vec{0}$. Also ist $(A^{-1}A)\vec{z} = \vec{0}$. Daraus ergibt sich aber $E\vec{z} = \vec{z} = \vec{0}$, was die Behauptung zeigt.
$2 \Rightarrow 3$: Mit der quadratischen Matrix A bilden wir das homogene lineare Gleichungssystem $A \cdot \vec{x} = \vec{0}$.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = 0 \end{cases}$$

Seine erweiterte Matrix ist:

$$(A \mid \vec{0}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 \end{array}\right)$$

Diese Matrix werde durch Elementaroperationen auf reduzierte Stufenform gebracht. Wäre irgend eine der Stufenbreiten echt grösser als 1, so enthielte die Lösung eine freie Variable. Dann hätte aber das zugehörige Gleichungssystem eine nichttriviale Lösung, was der Voraussetzung widerspräche. Daher haben also sämtliche Stufen die Einheitsbreite. Da oberhalb der führenden 1 in der reduzierten Stufenformen nur 0 vorkommen dürfen, lautet die reduzierte Stufenform

$$\left(\begin{array}{cccc|c} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdot \\ 0 & 0 & \cdots & 1 & 0 \end{array}\right)$$

Die Koeffizientenmatrix zeigt, dass die gewählten Elementaroperationen die Matrix A tatsächlich in die Einheitsmatrix E_n übergeführt haben.
$3 \Rightarrow 4$: Nach Voraussetzung geht A nach einer Folge von Elementaroperationen in E_n über. Weil jede Elementaroperation durch Multiplikation von links mit einer gewissen Elementarmatrix ausgeführt werden kann, existieren also Elementarmatrizen $M_1, M_2, \ldots, M_k$ mit

$$M_k \cdots M_2 \cdot M_1 \cdot A = E_n$$

Elementarmatrizen sind invertierbar. Multiplizieren wir diese Gleichung von links mit $M_1^{-1} \cdot M_2^{-1} \cdots M_k^{-1}$, erhalten wir $A = M_1^{-1} \cdot M_2^{-1} \cdots M_k^{-1} \cdot E_n$. Weil die Inversen von Elementarmatrizen wieder Elementarmatrizen sind, kann daher A in der Tat als Produkt von Elementarmatrizen geschrieben werden.
$4 \Rightarrow 1$: Da Elementarmatrizen invertierbar sind und wir wissen, dass ein Produkt von invertierbaren Matrizen wiederum invertierbar ist, muss also die Matrix A, die ja nach Voraussetzung als Produkt von Elementarmatrizen darstellbar ist, invertierbar sein.
$1 \Rightarrow 5$: Diese Behauptung folgt aus der Definition der Invertierbarkeit.
$5 \Rightarrow 2$: Zum Beweis dieser Behauptung müssen wir überlegen, dass das homogene Gleichungssystem $A \cdot \vec{x} = \vec{0}$ nur die triviale Lösung hat. Es sei also $\vec{z}$ irgend eine Lösung d.h. es gilt $A \cdot \vec{z} = \vec{0}$. Multiplizieren wir beide Seiten dieser Gleichung von links mit X, so erhalten wir $XA \cdot \vec{z} = X \cdot \vec{0}$. Daraus ergibt sich aber

nach Voraussetzung $E\vec{z} = \vec{0}$ d.h. $\vec{z} = \vec{0}$ und damit hat das Gleichungssystem $A \cdot \vec{x} = \vec{0}$ in der Tat nur die triviale Lösung.
$5 \Rightarrow 6$: Weil nach der im letzten Schritt bewiesen Implikation $5 \Rightarrow 1$ gilt, ist A invertierbar. Multiplikation der Gleichung $X \cdot A = E$ von rechts mit der Matrix A^{-1} liefert die Gleichung $XAA^{-1} = EA^{-1}$ bzw. $X = A^{-1}$. Daher gilt in der Tat $A \cdot X = E$.
$6 \Rightarrow 1$: Falls eine Matrix Y existiert mit $A \cdot Y = E$, so ist nach der bewiesenen Implikation $5 \Rightarrow 1$ die Matrix Y invertierbar. Durch Multiplikation mit Y^{-1} von rechts ergibt sich $A = Y^{-1}$ d.h. A ist als Inverse einer invertierbaren Matrix in der Tat invertierbar.
$1 \Rightarrow 7$: Diese Aussage wurde früher bereits bewiesen.
$7 \Rightarrow 8$: Diese Behauptung ist trivial.
$8 \Rightarrow 1$: Da das Gleichungssystem $A \cdot \vec{x} = \vec{b}$ nach Voraussetzung für *jedes* n-Tupel $\vec{b}$ lösbar ist, gilt dies insbesondere für die Standard-Basis Vektoren:

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Wir lösen also die linearen Gleichungssysteme

$$A \cdot \vec{x}_1 = \vec{e}_1, \quad A \cdot \vec{x}_2 = \vec{e}_2, \quad \cdots \quad A \cdot \vec{x}_n = \vec{e}_n$$

Die Lösungen $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n$ bilden eine $n \times n$-Matrix Y, deren Spalten gerade aus diesen Lösungen bestehen d.h. Y hat die Gestalt $Y = (\vec{x}_1 \ \vec{x}_2 \ \cdots \ \vec{x}_n)$. Wir haben das Matrizenprodukt gerade so erklärt, dass $A \cdot Y$ aus den Spalten $A \cdot \vec{x}_1 \ A \cdot \vec{x}_2 \ \cdots \ A \cdot \vec{x}_n$ besteht. Es ist also:

$$A \cdot Y = (A \cdot \vec{x}_1 \ A \cdot \vec{x}_2 \ \cdots \ A \cdot \vec{x}_n) = (\vec{e}_1 \ \vec{e}_2 \ \cdots \ \vec{e}_n) = E_n$$

Aus der sechsten Behauptung folgt, dass A invertierbar ist. □

Als unmittelbare Folgerung dieses Satzes ergibt sich folgender *Alternativ-Satz* der linearen Algebra.

Korollar. Für ein lineares Gleichungssystem $A \cdot \vec{x} = \vec{b}$ mit quadratischer Koeffizientenmatrix A gilt genau eine der beiden folgenden Alternativen:

- Entweder: Das lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ hat für jeden Vektor $\vec{b}$ genau eine Lösung.
- Oder: Das zugehörige homogene lineare Gleichungssystem $A \cdot \vec{x} = \vec{0}$ besitzt eine nichttriviale Lösung.

Korollar. Jedes lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$, das mit Hilfe der Kirchhoff'schen Regeln aus einem elektrischen Netzwerk entsteht, hat genau eine Lösung.

Beweis. Setzt man alle Spannungsquellen im Netzwerk 0, so werden natürlich auch keine Ströme fliessen, da man sonst ein Perpetuum Mobile konstruiert

hätte. Beim Nullsetzen der Spannungen geht das lineare Gleichungssystem in das zugehörige homogene Gleichungssystem $A \cdot \vec{x} = \vec{0}$ über. Das zugehörige homogene Gleichungssystem hat also wegen dem Energie-Satz nur die triviale Lösung und wir sind im ersten Fall des Alternativ-Satzes. □

Korollar. Eine (obere oder untere) Dreiecksmatrix A ist genau dann invertierbar, wenn sämtliche Diagonalelemente von Null verschieden sind.

Beweis. Das zugehörige homogene Gleichungssystem $A \cdot \vec{x} = \vec{0}$ hat genau dann nur die triviale Lösung $\vec{x} = \vec{0}$, wenn alle Diagonalelemente ungleich Null sind. Die Behauptung folgt aus dem Äquivalenzsatz. □

3.7.2 Berechnung der inversen Matrix

Der Beweis des Äquivalenzsatzes bzw. der Eliminations-Algorithmus liefert als Nebenprodukt eine effektive Methode zur Berechnung der Inversen einer invertierbaren Matrix A. Da A durch Elementaroperationen auf die reduzierte Stufenform E_n gebracht werden kann, gibt es Elementarmatrizen $M_1, M_2, \cdots, M_k$ mit

$$M_k \cdots M_2 \cdot M_1 \cdot A = E_n$$

Da die Inverse einer inverbaren Matrix eindeutig bestimmt ist, gilt: $A^{-1} = M_k \cdots M_2 \cdot M_1$ oder

$$A^{-1} = M_k \cdots M_2 \cdot M_1 \cdot E_n$$

Wir erhalten also A^{-1}, indem wir E_n von links der Reihe nach mit den Elementarmatrizen $M_1, M_2 \cdots, M_k$ multiplizieren. Da jede dieser Multiplikationen einer Elementaroperation entspricht, folgt, dass die selbe Folge von Elementaroperationen, die A nach E_n überführt, gleichzeitig E_n in A^{-1} verwandelt. Halten wir diese Beobachtung fest:

Satz. Wir erhalten die Inverse einer invertierbaren Matrix A, indem wir eine Folge von Elementaroperationen bestimmen, die A zur Einheitsmatrix umformt, und diese Folge auf E_n anwenden.

Beispiel. Um die Inverse der Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix}$$

zu berechnen, verwandeln wir A durch Elementaroperationen zur Einheitsmatrix um und führen gleichzeitig mit denselben Operationen E_3 nach A^{-1} über. Dazu schreiben wir die Einheitsmatrix rechts neben A und erhalten die Blockmatrix $(A \mid E)$, auf die wir geeignete Elementaroperationen anwenden, bis die linke Seite die Gestalt von E hat; gleichzeitig ergibt sich auf der rechten Seite A^{-1}, so dass wir schliesslich die Blockmatrix $(E \mid A^{-1})$ erhalten. Im numerischen Beispiel gehen wir von der Blockmatrix

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 5 & 3 & 0 & 1 & 0 \\ 1 & 0 & 8 & 0 & 0 & 1 \end{array}\right)$$

aus. Addition des (-2)-fachen der ersten Zeile zur zweiten Zeile und des (-1)-fachen der ersten Zeile zur dritten Zeile liefert:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 1 & 0 \\ 0 & -2 & 5 & -1 & 0 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ Z_{13}(-1) \end{array}\right]$$

Addition der 2-fachen der zweiten Zeile zur dritten liefert:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 1 & 0 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(2) \end{array}\right]$$

Addition des (-3)-fachen der dritten Zeile zur zweiten und des 3-fachen der dritten Zeile zur ersten ergibt:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 0 & -14 & 6 & 3 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} Z_{31}(3) \\ Z_{32}(-3) \\ \end{array}\right]$$

Addition des (-2)-fachen der zweiten Zeile zur ersten liefert schliesslich im linken Block reduzierte Stufenform.

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -40 & 16 & 9 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & -1 & -5 & 2 & 1 \end{array}\right) \qquad \left[\begin{array}{c} Z_{21}(-2) \\ \\ \end{array}\right]$$

Zum Normieren wird nun noch die dritte Zeile mit (-1) multipliziert und wir erhalten die Matrix

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -40 & 16 & 9 \\ 0 & 1 & 0 & 13 & -5 & -3 \\ 0 & 0 & 1 & 5 & -2 & -1 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ S_3(-1) \end{array}\right]$$

in deren linkem Block die Einheitsmatrix und im rechten die gesuchte Inverse stehen.

Damit ergibt sich für die gesuchte Inverse:

$$A^{-1} = \left(\begin{array}{ccc} -40 & 16 & 9 \\ 13 & -5 & -3 \\ 5 & -2 & -1 \end{array}\right)$$

wie man durch Nachrechnen bestätigen kann. Man beachte, dass hier genau die selbe Rechnung durchgeführt wurde, die man hätte durchführen müssen, um das entsprechende lineare Gleichungssystem $A \cdot \vec{x} = \vec{b}$ für einen beliebigen Konstantenvektor $\vec{b}$ zu lösen. Das folgt aus der Tatsache, dass die Inverse von A als Lösung von gewissen linearen Gleichungssystemen mit den Standard-Basisvektoren als Konstanenvektoren und der Koeffizientenmatrix A aufgefasst werden kann. ◯

Es ist nicht von vorneherein klar, ob eine gegebene quadratische Matrix A invertierbar ist. Aus dem Äquivalenzsatz folgt aber, dass für eine nicht invertierbare $n \times n$-Matrix A keine Folge elementarer Zeilenoperationen existieren kann, die A in die Einheitsmatrix E_n überführt. Also muss in diesem Fall die Stufenform

mindestens eine Nullzeile enthalten. Wendet man also die beschriebene Methode zur Invertierung auf eine nicht invertierbare Matrix an, erhält man während der Rechnung eine Nullzeile im linken Block. Daraus kann man dann schliessen, dass die gegebene Matrix nicht invertierbar ist und die Prozedur abbrechen.

Beispiel. Um zu entscheiden, ob die Matrix

$$A = \begin{pmatrix} 1 & 6 & 4 \\ 2 & 4 & -1 \\ -1 & 2 & 5 \end{pmatrix}$$

invertierbar ist, gehen wir also analog zum letzten Beispiel vor. Wir gehen also von folgender Blockmatrix aus.

$$\left(\begin{array}{ccc|ccc} 1 & 6 & 4 & 1 & 0 & 0 \\ 2 & 4 & -1 & 0 & 1 & 0 \\ -1 & 2 & 5 & 0 & 0 & 1 \end{array}\right)$$

Addition des (-2)-fachen der ersten Zeile zur zweiten und Addition der ersten Zeile zur dritten Zeile liefert:

$$\left(\begin{array}{ccc|ccc} 1 & 6 & 4 & 1 & 0 & 0 \\ 0 & -8 & -9 & -2 & 1 & 0 \\ 0 & 8 & 9 & 1 & 0 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ Z_{13}(1) \end{array}\right]$$

Addition der zweiten Zeile zur dritten ergibt:

$$\left(\begin{array}{ccc|ccc} 1 & 6 & 4 & 1 & 0 & 0 \\ 0 & -8 & -9 & -2 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(1) \end{array}\right]$$

Weil der linke Block jetzt eine Nullzeile enthält, ist A nicht invertierbar. ◯

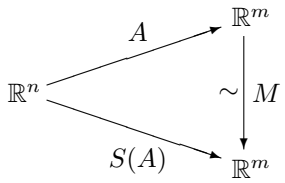
Das Lösbarkeitskriterium wird also hier zum Invertierbarkeitskriterium.

3.8 Normalform

Nach diesem Ausflug zu den invertierbaren Matrizen wollen wir uns jetzt matriziell klar machen, was wir seinerzeit mit dem Eliminations-Algorithmus eigentlich gemacht haben, als wir beim Lösen eines linearen Gleichungssystemen $A \cdot \vec{x} = \vec{b}$ mit der Koeffizientenmatrix $A \in \mathbb{R}^{m,n}$ und beliebigem Konstantenvektor $\vec{b} \in \mathbb{R}^m$ von der Blockmatrix $(A \mid E_m)$ ausgegangen sind und auf sie geeignete Elementaroperationen angewandt haben, bis die linke Seite reduzierte Stufenform $S(A)$ hatte; gleichzeitig ergab sich auf der rechten Seite eine gewisse Matrix M, so dass wir am Ende des Algorithmus die Blockmatrix $(S(A) \mid M)$ erhielten. Die Elementarmatrizen, die zu den verwendeten Elementaroperationen gehören, seien $M_1, M_2, \cdots, M_k$. Weil wir im rechten Block von der Einheitsmatrix E_m ausgegangen sind, gilt also

$$M = M_k \cdot M_{k-1} \cdots M_2 \cdot M_1 \cdot E_m$$

Wir stellen also fest, dass M gerade das Produkt der Elementarmatrizen ist, die wir verwendet haben, um die gegebenen Matrix A in ihre reduzierte Stufenform $S(A)$ überzuführen. Ferner ist M als Produkt von Elementarmatrizen invertierbar und es gilt $M \cdot A = S(A)$, so dass folgendes Diagramm kommutiert:



Das ist der konzeptionelle Gehalt des Eliminationsverfahrens. Dass man diesen Algorithmus zum Lösen eines linearen Gleichungssystemen der Form $A \cdot \vec{x} = \vec{b}$ verwenden kann sieht man, wenn man dieses System von links mit der invertierbaren Matrix M multipliziert. Dabei geht das gegebene Gleichungssystem in das transformierte Gleichungssystem

$$M \cdot A \cdot \vec{x} = M \cdot \vec{b} \qquad \text{d.h.} \qquad S(A) \cdot \vec{x} = M \cdot \vec{b}$$

über, dessen Lösungsraum wegen der Invertierbarkeit von M mit demjenigen des ursprünglichen System aber immer noch übereinstimmt. Weil die transformierte Koeffizientenmatrix $S(A)$ von A normierte, reduzierte Stufenform hat, lässt sich die Lösung des transformierten Systems sofort ablesen. Geometrisch wird dabei das ursprüngliche Koordinatensystem durch M in eine neue Lage transformiert, in der die Informationen der einzelnen Koordinatenachsen entflochten sind.

Beispiel. Im bereits ausführlich behandelten Beispiel begannen wir mit der Matrix

$$(A \mid E_6) = \left(\begin{array}{rrrrrrr|rrrrrr} 0 & 0 & 1 & -1 & 3 & 1 & -6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & -1 & 2 & 4 & 5 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 6 & -3 & \frac{3}{2} & -6 & 3 & \frac{51}{2} & 0 & 0 & 0 & 1 & 0 & 0 \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 & 0 & 0 & 0 & 0 & 1 & 0 \\ 7 & 14 & -5 & \frac{3}{2} & -8 & 14 & \frac{85}{2} & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

und der Eliminations-Algorithmus hat zur Matrix

$$(S(A) \mid M) = \left(\begin{array}{rrrrrrr|rrrrrr} 1 & 2 & 0 & -\frac{1}{2} & 1 & 0 & 0 & -\frac{35}{6} & -\frac{9}{2} & \frac{11}{6} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 & -1 & 3 & 0 & 0 & \frac{16}{3} & 5 & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{5}{3} & 1 & -\frac{2}{3} & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & -3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & -1 & 0 & 10 & 6 \end{array}\right)$$

geführt. Durch Nachrechnen bestätigt man hier tatsächlich die Matrizengleichung $M \cdot A = S(A)$ und die Tatsache, dass M invertierbar ist.

Für den bereits benutzten Konstantenvektor $\vec{b}$ ist

$$M \cdot \vec{b} = \begin{pmatrix} -\frac{35}{6} & -\frac{9}{2} & \frac{11}{6} & 0 & \frac{2}{3} & 0 \\ \frac{16}{3} & 5 & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ \frac{5}{3} & 1 & -\frac{2}{3} & 0 & -\frac{1}{3} & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 6 & 0 & -3 & 2 & 0 & 0 \\ 10 & 0 & -1 & 0 & 10 & 6 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ 2 \\ 6 \\ -9 \\ 3 \\ -14 \end{pmatrix} = \begin{pmatrix} -31 \\ 47 \\ 7 \\ 8 \\ 0 \\ 0 \end{pmatrix}$$

Die Gleichung $S(A) \cdot \vec{x} = M \cdot \vec{b}$ lässt sich nun wegen der reduzierten Stufenform von $S(A)$ leicht lösen und liefert die bereits gefundene Beschreibung für den Lösungsraum $L(A, \vec{b})$. ◯

Der für die Dualität sensible Leser wird sich bereits gefragt haben, wie die entsprechenden Resultate aussehen, wenn man mit einer Matrix A statt Zeilenoperationen die dazu dualen Spaltenoperationen durchführt. Statt der bisher ausschliesslich benutzten Zeilenumformungen lassen sich mit ihr nämlich die dualen *elementaren Spaltenumformungen* durchführen.

I^T. Vertauschen zweier Spalten.

II^T. Addition eines Vielfachen einer Spalte zu einer anderen Spalte und Ersetzen dieser Spalte durch die Summe.

III^T. Multiplizieren einer Spalte mit einem von Null verschiedenen Skalar.

Entsprechend den drei Typen elementarer Spaltenoperationen gibt es drei Typen Elementarmatrizen, die durch die jeweilige Spaltenoperation aus der Einheitsmatrix hervorgehen.

I^T. Vertauschen der i-ten und der j-ten Spalte liefert die Vertauschungsmatrix $T_{ij}^T = T_{ji}$.

II^T. Addition der r-fachen der i-ten zur j-ten Spalte liefert die Scherungsmatrix $Z_{ij}^T(r) = Z_{ji}(r)$

III^T. Multiplikation der i-ten Spalte mit dem Faktor $r \neq 0$ liefert die Diagonalmatrix $S_i^T(r) = S_i(r)$.

Man beachte, dass diese Matrizen alle invertierbar sind und ihre Inversen jeweils vom selben Typ sind.

Weil beim Transponieren einer Matrix A Spalten und Zeilen ausgetauscht werden, lassen sich die Spaltenoperationen mit der Matrix A durch Zeilenoperationen mit der transponierten Matrix A^T durchführen, wenn man anschliessend das Ergebnis transponiert. Da sich sämtliche Zeilenoperationen durch Multiplikationen mit einer gewissen Elementarmatrix M von links durchführen lassen und die Beziehung $(M \cdot A^T)^T = A \cdot M^T$ gilt, können die Spaltenoperationen dadurch ausgeführt werden, dass man die Matrix A von *rechts* mit der Transponierten der betreffenden Elementarmatrix d.h. mit der Matrix M^T multipliziert.

Satz. Die Elementarmatrix M^T sei aus der Einheitsmatrix E_m durch eine elementare Spaltenoperation hervorgegangen. Ist nun A eine $m \times n$-Matrix, so ist

das Produkt $A \cdot M^T$ gerade die Matrix, die durch Anwenden derselben Spaltenoperation auf A entsteht.

Wiederum ist es zweckmässig, die Situation statt eines formalen Beweises an einem genügend komplexen Beispiel zu illustrieren.

Beispiel. Wir gehen von der konkreten Matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 4 \\ 2 & 3 & 4 \\ 3 & 6 & 0 \end{pmatrix}$$

aus und betrachten die folgenden Matrizenprodukte:

I^T. Das Produkt $A \cdot T_{23}^T$ ist die Matrix,

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 4 \\ 2 & 3 & 4 \\ 3 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & -1 \\ 2 & 4 & 3 \\ 3 & 0 & 6 \end{pmatrix} \qquad \begin{bmatrix} \\ T_{32} \\ T_{23} \end{bmatrix}^T$$

die aus A durch Vertauschen der zweiten mit der dritten Spalte entsteht.

II^T. Das Produkt $A \cdot Z_{13}^T(3)$ ist die Matrix,

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 4 \\ 2 & 3 & 4 \\ 3 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 4 \\ 0 & -1 & 4 \\ 2 & 3 & 10 \\ 3 & 6 & 9 \end{pmatrix} \qquad \begin{bmatrix} \\ \\ Z_{13}(3) \end{bmatrix}^T$$

die aus A durch Addition des Dreifachen der ersten zur dritten Spalte entsteht.

III^T. Das Produkt $A \cdot S_2^T(5)$ ist die Matrix,

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 4 \\ 2 & 3 & 4 \\ 3 & 6 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 10 & 1 \\ 0 & -5 & 4 \\ 2 & 15 & 4 \\ 3 & 30 & 0 \end{pmatrix} \qquad \begin{bmatrix} \\ S_2(5) \\ \end{bmatrix}^T$$

die aus A durch Multiplikation des Fünffachen der zweiten Spalte entsteht.

In jedem Fall lässt sich die elementare Spaltenoperation durch Multiplikation mit der betreffenden Elementarmatrix von rechts bewerkstelligen. Man beachte, dass es sich um das selbe Beispiel handelt, das wir seinerzeit zur Illustration der entsprechenden Zeilenoperationen verwendet haben. Der Unterschied zwischen den beiden Beispielen besteht einzig darin, dass sämtlich Matrizen transponiert und ihre Reihenfolge vertauscht worden sind. ◯

Bisher haben wir ausschliesslich Zeilenoperationen benutzt, um eine gegebene Matrix A mit Hilfe einer invertierbaren Transformationsmatrix M in normierte, reduzierte Stufenform $S(A) = M \cdot A$ überzuführen, mit der wir das ursprüngliche Gleichungssystem $A \cdot \vec{x} = \vec{b}$ einfacher lösen konnten. Man fragt sich nun,

ob man man nicht auf die so entstande Stufenform $S(A)$ zusätzlich noch Spaltenoperationen anwenden kann, um auf diese Art die Stufenform noch weiter zur Normalform $\mathcal{N}(A)$ zu vereinfachen. Dabei werden wir eine invertierbare Transformationsmatrix N finden, so dass

$$\mathcal{N}(A) = S(A) \cdot N$$

gilt. Mit der bereits berechneten Stufenform gilt dann für die Normalform die Beziehung

$$\mathcal{N}(A) = M \cdot A \cdot N$$

Um eine Vorstellung zu erhalten, wie die Normalform einer Matrix aussieht, gehen wir von unserem mehrfach benutzten Beispiel aus.

Beispiel. Wir haben die Blockmatrix

$$(A \mid E_6) = \left(\begin{array}{rrrrrrr|cccccc} 0 & 0 & 1 & -1 & 3 & 1 & -6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -3 & -1 & 7 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & -1 & 2 & 4 & 5 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 6 & -3 & \frac{3}{2} & -6 & 3 & \frac{51}{2} & 0 & 0 & 0 & 1 & 0 & 0 \\ -4 & -8 & 2 & 0 & 2 & -9 & -19 & 0 & 0 & 0 & 0 & 1 & 0 \\ 7 & 14 & -5 & \frac{3}{2} & -8 & 14 & \frac{85}{2} & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

mit Hilfe von Zeilenoperationen in die ganzzahlige, reduzierte Stufenform

$$(B \mid M) = \left(\begin{array}{rrrrrrr|rrrrrr} 6 & 12 & 0 & -3 & 6 & 0 & 0 & -35 & -27 & 11 & 0 & 4 & 0 \\ 0 & 0 & 3 & -3 & 9 & 0 & 0 & 16 & 15 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -3 & 0 & -5 & -3 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & -3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & -1 & 0 & 10 & 6 \end{array}\right)$$

überführt. Für sie ist also $M \cdot A = B$. Aus typographischen Gründen führen wir mit der entstandene Matrix B nun nicht Spaltenoperationen durch, sondern gehen zur transponierten Matrix B^T über und führen Zeilenoperationen durch. Selbstverständlich müssen wir uns am Schluss daran erinnern, dass wir dann das Ergebniss wieder transponieren. Wie früher führen wir über die durchgeführten Operationen in gewohnter Manier Buch. Dabei wird die Blockmatrix $(B^T \mid E_7^T)$ in die normierte, reduzierte Sufenform $(\mathcal{N}(A)^T \mid N^T)$ übergeführt. Es gilt also $N^T \cdot B^T = \mathcal{N}(A)^T$. Durch Transponieren des Ergebnisses erhalten wir die Normalform $\mathcal{N}(A) = B \cdot N$. Mit der invertierbaren Transformationsmatrix N gilt die Gleichung $\mathcal{N}(A) = M \cdot A \cdot N$.

Im konkreten numerischen Beispiel gehen wir also von der Blockmatrix

$$(B^T \mid E_7^T) = \left(\begin{array}{rrrrrr|ccccccc} 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 6 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

aus und addieren nun das (-2)-fache der ersten Zeile zur zweiten, die erste Zeile zum 2-fachen der vierten und schliesslich das (-1)-fache der ersten Zeile zur fünften und erhalten die Matrix

$$\left(\begin{array}{cccccc|ccccccc} 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ \\ Z_{14}(1,2) \\ Z_{15}(-1) \\ \\ \\ \end{array}\right]$$

Weil die zweite Zeile im linken Block nun lauter Nullen hat, vertauschen wir sie mit sämtlichen Zeilen darunter, bis sie in der letzen Zeile steht und erhalten die Matrix

$$\left(\begin{array}{cccccc|ccccccc} 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ T_2 \\ \\ \\ \\ \\ \\ \end{array}\right]$$

Addition des 2-fachen der zweiten Zeile zur dritten und des (-3)-fachen zur vierten ergibt

$$\left(\begin{array}{cccccc|ccccccc} 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -3 & 0 & 1 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(2) \\ Z_{23}(-3) \\ \\ \\ \\ \end{array}\right]$$

Auch hier verschieben wir die beiden neu entstandenen dritte und vierte Nullzeilen im linken Block durch sukzessives Vertauschen mit den Zeilen darunter unterhalb derjenigen Zeilen, die nicht verschwinden und erhalten die Stufenform

$$\left(\begin{array}{cccccc|ccccccc} 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ T_3 \\ T_4 \\ \\ \\ \\ \end{array}\right]$$

Sie ist automatisch reduziert, weil von einer Matrix ausgegangen sind, deren Transformierte bereits Stufenform hatte. Zur Normierung dividieren wir nun noch die erste Zeile durch 6, die zweite durch 3 und die dritte durch -3 und

erhalten die gesuchte normierte Stufenform $(\mathcal{N}(A)^T, N^T)$.

$$\left(\begin{array}{cccccc|cccccc} 1&0&0&0&0&0&\frac{1}{6}&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&\frac{1}{3}&0&0&0&0\\ 0&0&1&0&0&0&0&0&0&0&0&-\frac{1}{3}&0\\ 0&0&0&1&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&-1&0&-3&0&1&0&0\\ 0&0&0&0&0&0&1&0&2&2&0&0&0\\ 0&0&0&0&0&0&-2&1&0&0&0&0&0 \end{array}\right) \qquad \left[\begin{array}{l} S_1(\frac{1}{6})\\ S_2(\frac{1}{3})\\ S_3(-\frac{1}{3}) \end{array}\right]$$

Für die berechnete Normalform der Matrix A und die beiden entstanden Transformationsmatrizen M und N gilt nun tatsächlich

$$\mathcal{N}(A) = M \cdot A \cdot N = \left(\begin{array}{cccc|ccc} 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0\\ \hline 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0 \end{array}\right)$$

wie man leicht durch Nachrechnen kontrolliert. Ein Blick auf die gefundene Normalform zeigt, dass sie eine sehr einfache Blockstruktur hat und ihre Elemente lauter Binärzahlen sind. ◯

Satz. Jede Matrix $A \in \mathbb{R}^{m,n}$ kann durch eine endliche Folge von Zeilenoperationen, deren elementare Matrizen sich zu einer invertierbaren Matrix $M \in \mathbb{R}^{m,m}$ zusammenfassen lassen und einer endlichen Folge von Spaltenoperationen, deren elementare Matrizen sich zu einer invertierbaren Matrrix $N \in \mathbb{R}^{n,n}$ zusammenfassen lassen, in die eindeutig bestimmte Normalform

$$\mathcal{N}(A) = M \cdot A \cdot N = \left(\begin{array}{c|c} E_r & 0_{r,n-r} \\ \hline 0_{m-r,r} & 0_{m-r,n-r} \end{array}\right)$$

überführen. Dabei ist $r = \mathrm{Rang}(A) = \mathrm{Rang}(A^T)$.

In der Blockmatrix der Normalform dürfen gewisse Blöcke leer sein. Das ist genau dann der Fall, wenn $r = 0$, $r = m$ oder $r = n$ ist.

Die Gleichung $\mathcal{N}(A) = M \cdot A \cdot N$ für die Normalform einer Matrix A entspricht dem kommutativen Diagramm

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{\quad A \quad} & \mathbb{R}^m \\ N \uparrow \sim & & \sim \downarrow M \\ \mathbb{R}^n & \xrightarrow[\quad \mathcal{N}(A) \quad]{} & \mathbb{R}^m \end{array}$$

Zwei Matrizen $A, B \in \mathbb{R}^{m,n}$ heissen *äquivalent*, falls die eine aus der anderen durch Anwendung von endlich vielen Zeilen- und Spaltenoperationen entsteht. Insbesondere ist also jede Matrix zu ihrer Normalform äquivalent. Allgemeiner sind zwei Matrizen vom selben Typ genau dann äquivalent, wenn sie die selben

Normalformen haben. Das ist genau dann der Fall, wenn sie den selben Rang haben. Nach dem Hauptsatz ist der Rang die einzige Invariante äquivalenter Matrizen. Das Diagramm besagt, dass äquivalente Matrizen als unterschiedliche Darstellungsformen derselben linearen Abbildung in verschiedenen Bezugssystemen, zwischen denen mit Hilfe von M und N transformiert werden kann, aufgefasst werden können.

Man beachte, dass der Rang von Matrizen nicht mit dem Matrizenprodukt verträglich ist. Im allgemeinen gilt nur

$$\mathrm{Rang}(A \cdot B) \leq \max\big(\mathrm{Rang}(A), \mathrm{Rang}(B)\big)$$

Um mit der Normalform $\mathcal{N}(A)$ und den beiden Transformationsmatrizen M, N, die zur Faktorisierung $\mathcal{N}(A) = M \cdot A \cdot N$ gehören, ein lineares Gleichungssystemen der Form $A \cdot \vec{x} = \vec{b}$ zu lösen, lösst man zunächst die Hilfsgleichung $\mathcal{N}(A) \cdot \vec{y} = M \cdot \vec{b}$ nach $\vec{y}$ auf, was wegen der einfachen Normalform simpel ist. Zur Lösung der ursprünglichen Gleichung berechnet man dann mit der gefunden Lösung den Vektor $\vec{x} = N \cdot \vec{y}$. Dann gilt nämlich insgesamt

$$A \cdot \vec{x} = (M^{-1} \cdot \mathcal{N}(A) \cdot N^{-1}) \cdot N \cdot \vec{y} = M^{-1} \cdot \mathcal{N}(A) \cdot \vec{y} = M^{-1} \cdot M \cdot \vec{b} = \vec{b}$$

und das ursprüngliche System kann tatsächlich so gelöst werden.

Beispiel. Im bereits behandelten Beispiel mit dem Konstantenvektor $\vec{b}$ ist

$$M \cdot \vec{b} = \begin{pmatrix} -35 & -27 & 11 & 0 & 4 & 0 \\ 16 & 15 & 2 & 0 & 1 & 0 \\ -5 & -3 & 2 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 6 & 0 & -3 & 2 & 0 & 0 \\ 10 & 0 & -1 & 0 & 10 & 6 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ 2 \\ 6 \\ -9 \\ 3 \\ -14 \end{pmatrix} = \begin{pmatrix} -186 \\ 141 \\ -21 \\ 8 \\ 0 \\ 0 \end{pmatrix}$$

Die Hilfsgleichung $\mathcal{N}(A) \cdot \vec{y} = M \cdot \vec{b}$ lässt sich nun wegen der simplen Form von $\mathcal{N}(A)$ sofort lösen und mit

$$\begin{aligned} \vec{x} &= N \cdot \vec{y} = \begin{pmatrix} \frac{1}{6} & 0 & 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{1}{3} & 0 & 0 & -3 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} -186 \\ 141 \\ -21 \\ 8 \\ t \\ s \\ r \end{pmatrix} \\ &= \begin{pmatrix} -31 \\ 0 \\ 47 \\ 0 \\ 0 \\ 7 \\ 8 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + r \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

erhält man die bekannte Beschreibung für den Lösungsraum $L(A, \vec{b})$. ◯

3.9 Determinante

In vielen alten Schulbüchern findet man ein Kriterium dafür, ob eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ invertierbar bzw. ein lineares Gleichungssystem regulär ist oder nicht. Die dabei entwickelten Methoden werden dann sogar zum Invertieren solcher Matrizen bzw. zum Lösen regulärer Gleichungssysteme benutzt und könnten beim Anfänger den falschen Eindruck erwecken, sie seien praktisch oder fundamental. Um den Leser mit diesen Methoden mindestens vertraut gemacht zu haben, erwähnen wir hier die Determinantenmethode, die zwar Formelsalat liefert, einem tieferen Verständnis der linearen Algebra aber nur im Weg steht. Insbesondere lassen sich damit beliebige Gleichungssysteme, die also nicht regulär sind, nicht übersichtlich behandeln, obwohl solche allgemeineren Systeme in der Praxis z.B. in der Kodierungstheorie sehr oft auftreten. Vom theoretischen Standpunkt aus lässt sich das Eliminationsverfahren im Gegensatz zum Determinantenkalkül relativ naheliegend auf beliebige Polynomgleichungen verallgemeinern. Dieser allgemeinere Gröbner-Algorithmus zur Behandlung von Systemen von Polynomgleichungen macht viele Probleme, etwa in der Robotik, erst algorithmisch behandelbar. Im Spezialfall der Polynome einer Unbestimmten enthält er die Spektraltheorie.

Wie wir noch sehen werden, sind die Determinantenmethoden zum Lösen linearer Gleichungssysteme auch vom praktischen Standpunkt des Informatikers her gesehen unbrauchbar, weil ausser für sehr kleine Gleichungssysteme (maximal 3 Gleichungen und Unbekannte) der Rechenaufwand unverantwortbar gross ist. Im Zeitalter der Sparmanie scheint es also mehrfach gerechtfertigt, sich von diesem alten Zopf zu trennen und gleich von Anfang an eine mathematische Sprache zu entwickeln, die trag- und ausbaufähig ist und erst noch tiefere Einsichten ermöglicht. Der Einwand, man benötige den Determinantenkalkül auf dieser Stufe zur Berechnung des charakteristischen Polynoms und zum Transformieren von Mehrfachintegralen, löst sich in Luft auf, wenn man beachtet, dass man im ersten Fall am Minimalpolynom interessiert ist und dieses einfach mit Hilfe des Eliminationsverfahren bestimmen kann, wie wir vorführen werden. Im zweiten Fall entwickelt man statt der überholten Vektoranalysis besser den Kalkül der Differentialformen, der sich unter Transformationen beliebiger Abbildungen natürlich verhält. Wenn schon gehören die Determinanten dort hin.

Nach dieser Philippika gegen den Missbrauch des Determinantenkalküls sei eingeräumt, dass er an gewissen, sorgfältig ausgewählten, Orten seine Berechtigung haben mag. Flächeninhalte von Dreiecken und Parallelogrammen bzw. Volumina von Tetraedern und Parallelepipeden (Spaten) können zweckmässig mit Hilfe der Determinante ausgedrückt und untersucht werden. Ausgehend von diesen scheinbar intuitiven geometrischen Konzepten untersucht man ihre charakteristischen Eigenschaften und stellt fest, dass man eine normierte, multilineare alternierende Funktion ihrer Vektorargumente vor sich hat. Das sind genau die Eigenschaften der Determinante. So gesehen richtet man also beim Studium der Determinanten sein Augenmerk besser auf die Geometrie als auf die Algebra. Zur Berechnung der Determinante verwendet man dann entsprechend nicht ihre Standardformel, sondern macht sich die geometrischen Einsichten zu Nutze, um das Eliminationsverfahren auch zur effizienten Berechnung von Determinanten einsetzen zu können.

Durch die erwähnten geometrischen Überlegungen motiviert, definiert man die Determinante einer quadratischen Matrix algebraisch wie folgt.

Definition. Es sei $A \in \mathbb{R}^{n,n}$ eine quadratische Matrix. Ihre *Determinante* definiert man wie folgt rekursiv:

1. Ist $n = 1$ und $A = (a)$ so ist $\det(A) = a$.
2. Ist $n \geq 2$ und $i, j \in \{1, \ldots, n\}$ so bezeichnen wir mit $A_{i,j} \in \mathbb{R}^{(n-1)\times(n-1)}$ die Untermatrix, die man aus A erhält, wenn man in A die i-te Zeile und die j-te Spalte streicht. Die Determinante von A kann nun wahlweise wie folgt mit Hilfe der Determinanten dieser Untermatrizen ausgedrückt werden:
   - (Entwickeln nach der i-ten Zeile) Für jedes $i \in \{1, \ldots, n\}$ ist
   $$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{i,j})$$
   - (Entwickeln nach der j-ten Spalte) Für jedes $j \in \{1, \ldots, n\}$ ist
   $$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(A_{i,j})$$

Man kann zeigen, dass diese Werte übereinstimmen und unabhängig von der Wahl der Zeile i bzw. der Spalte j sind.

Beispiel. Für die Determinante einer beliebigen 2×2-Matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

liefert die rekursive Entwicklung nach der ersten Spalte die bekannte Formel

$$\det(A) = ad - bc$$

Man kann zeigen, dass durch diese Formel der gewichtete Flächeninhalt eines Parallelogramms beschrieben wird, das durch die beiden Spaltenvektoren

$$\vec{v} = \begin{pmatrix} a \\ c \end{pmatrix}, \qquad \vec{w} = \begin{pmatrix} b \\ d \end{pmatrix}$$

aufgespannt wird. Das Vorzeichen ist genau dann positiv, wenn die beiden Vektoren im Gegenuhrzeigersinn orientiert sind. ◯

Beispiel. Für die Determinante einer beliebigen 3×3-Matrix

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

liefert die rekursive Entwicklung nach der ersten Spalte die Formel

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

die aus $3! = 6$ Summanden besteht, die sich ihrerseits je aus 3 Faktoren zusammensetzen. Ihr Betrag kann als Volumeninhalt des Spates interpretiert werden, da durch die drei Spaltenvektoren aufgespannt wird. Das Vorzeichen dieser Determinante ist genau dann positiv, wenn die drei Spaltenvektoren in dieser Reihenfolge ein Rechtssystem bilden. Zwischen rechts und links kann also mit Hilfe der Determinanten algebraisch unterschieden werden. ◯

Das erwähnte Invertierbarkeits-Kriterium, das man dem Äquivalenzsatz als zusätzliche Aussage beifügen könnte, lautet nun wie folgt:

Satz. Eine quadratische Matrix A ist genau dann invertierbar, wenn ihre Determinante nicht verschwindet.

Mit Hilfe der Determinante einer quadratischen Matrix kann man entscheiden, ob die Matrix invertierbar ist oder nicht. Daher stammt ihr Name. Mit Hilfe der Determinante lassen sich auch reguläre Gleichungssysteme $A \cdot \vec{x} = \vec{b}$ charakterisieren. Es ist genau dann singulär, wenn $\det(A) = 0$ gilt.

Dieser Satz ist nur scheinbar praktisch. Entwickelt man die Determinante einer $n \times n$-Matrix A nach dem rekursiven Muster, erhält man eine Summe mit $n!$ Summanden, die sich je aus n Faktoren zusammensetzen. Die Berechnung einer Determinante nach diesem Muster erfordert für $n \geq 4$ einen unvertretbaren Rechenaufwand. Zur effektiven numerischen Berechnung der Determinante einer Matrix A benutzt man also *nicht* ihre rekursive Definition, sondern führt mit A zweckmässigerweise das Eliminationsverfahren durch. Dabei kann man, wie wir gesehen haben, die Information, ob die Matrix invertierbar ist oder nicht, direkt aus dem Lösbarkeitskriterium ablesen. Die Determinante von A kann man, falls man sie aus geometrischen Gründen wirklich einmal benötigt, unterwegs sozusagen als Abfallprodukt des Eliminationsverfahrens, mitablesen. Das algorithmische Arbeitspferd der linearen Algebra ist eben das Eliminationsvefahren und nicht etwa der oft von Schulmeistern in den Vordergrund gestellte Determinantenkalkül!

Dieses, soeben erwähnte, effiziente Verfahren zu Berechnung der Determinante beruht auf der Tatsache, dass sich die Determinante einer Matrix unter Elementaroperationen sehr einfach mittransformiert und von einer Matrix in Stufenform unmittelbar abgelesen werden kann. Wie erwähnt, liefert das Volumen eines Spates eine normierte, multilineare schiefsymmetrische Funktion. Übersetzt man diese Eigenschaften in die Algebra, erhält man folgendes, für die Berechnung von Determinanten fundamentales, Resultat.

Satz. Die Determinante hat folgende charakteristischen Eigenschaften:

DI. Beim Vertauschen von zwei Zeilen wechselt die Determinante ihr Vorzeichen.

DII. Bei der Addition eines Vielfachen einer Zeile zu einer anderen Zeile ändert sich die Determinante nicht.

DIII. Bei der Multiplikation einer Zeile mit einem von Null verschiedenen Skalar multipliziert sich die Determinante mit diesem Skalar.

Durch Entwickeln bestätigt man leicht, dass die Determinante einer Dreiecksmatrix das Produkt ihrer Diagonalelemente ist.

Um also die Determinante einer quadratischen Matrix A effizient zu berechnen, führt man mit ihr die Vorwärtsphase des Eliminationsverfahrens durch und macht sich bei jeder Elementaroperation klar, von welchem Typ sie ist. Bei jeder Operation vom Typ I wechselt die Determinante das Vorzeichen. Also muss man am Schluss nur wissen, ob man eine gerade oder eine ungerade Anzahl solcher Operationen durchgeführt hat. Über die Faktoren der Operationen vom Typ III führt man sorgfälltig Buch. Am Schluss werden alle führenden Elemente in der Stufenform miteinander multipliziert und durch das Produkt der aus der Buchhaltung entnommenen Faktoren dividiert.

Beispiel. Kehren wir nochmals zu den beiden besprochenen Beispielen zur Berechnung der Inversen zurück. Im ersten Beispiel der Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix}$$

wurden lauter Operationen vom Typ II durchgeführt, um die Matrix in Stufenform

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & -1 \end{pmatrix}$$

überzuführen. Daher haben die ursprüngliche Matrix A und diese obere Dreiecksmatrix die selben Determinanten. Die Determeinante der oberen Dreiecksmatrix ist aber $1 \cdot 1 \cdot -1 = -1$. Also ist auch $\det(A) = -1$. Insbesondere ist A invertierbar. Natürlich wissen wir viel mehr: wir kennen sogar ihre Inverse.

Im zweiten Beispiel sind wir von der Matrix

$$A = \begin{pmatrix} 1 & 6 & 4 \\ 2 & 4 & -1 \\ -1 & 2 & 5 \end{pmatrix}$$

ausgegangen und haben sie durch Elementaroperationen vom Typ II in die Stufenform

$$\begin{pmatrix} 1 & 6 & 4 \\ 0 & -8 & -9 \\ 0 & 0 & 0 \end{pmatrix}$$

übergeführt. Die Determinante dieser Dreiecksmatrix ist 0 und daher ist auch $\det(A) = 0$. Die Matrix ist also nicht invertierbar. ◯

Das Volumen eines von n Vektoren aufgespannten Spates in $\mathbb{R}^n$ lässt sich mit Hilfe des Betrags der Determinante berechnen. Die Frage stellt sich, wie sich allgemeiner das Volumen des von k Vektoren $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k$ aus $\mathbb{R}^n$ aufgespannten Parallelotops

$$P(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k) = \{t_1\vec{a}_1 + t_2\vec{a}_1 + \cdots + t_k\vec{a}_k \mid 0 \leq t_j \leq 1\}$$

berechnen lässt. Das Vektorprodukt liefert eine Beziehung für dieses Volumen nur für den Spezialfall $n = 3$ und $k = 2$ und die Determinante für $n = k$.

Fasst man die k gegebenen Vektoren zu einer Matrix

$$A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k) \in \mathbb{R}^{n,k}$$

zusammen, erkennt man, dass sich die Determinante nicht unmittelbar darauf anwenden lässt, weil ja diese Matrix in der Regel nicht quadratisch ist. Die Methode wird aber durch folgende Definition auf den allgemeinen Fall verallgemeinert.

Definition. Das vom System der k Spaltenvektoren $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k \in \mathbb{R}^n$ der Matrix $A \in \mathbb{R}^{n,k}$ aufgespannte Parallelotop hat das k-dimensionale Volumen

$$\mathrm{Vol}_k(P(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k)) = \sqrt{\det(A^T \cdot A)}$$

Diese Definition ist sinnvoll, da $\det(A^T \cdot A) \geq 0$ ist und sie lautet wurzelfrei

$$\mathrm{Vol}_k^2(P(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k)) = \det(A^T \cdot A)$$

Die Determinante der Gram-Matrix $A^T \cdot A$ hängt mit dem Volumen zusammen.

Man beachte, dass in der für viele Anwendungen fundamentale sogn. Gram'schen Matrix $A^T \cdot A$ die transponierte Matrix links steht. Diese symmetrische Matrix enthält als Einträge sämtliche Skalarprodukte der gegebenen Vektoren. Weil das Skalarprodukt geometrisch nur von den Länge und vom Zwischenwinkel abhängt, lässt sich also das Volumen eines Parallelotops allein mit Hilfe der Längen der gegebenen Vektoren und ihren Zwischenwinkeln ausdrücken.

Beispiel. Im Spezialfall $k = 1$ ist das Parallelotop $P(\vec{a})$ ein Streckenstück und $A = \vec{a}$. Wegen $\vec{a}^T \cdot \vec{a} = \langle \vec{a}, \vec{a} \rangle$ liefert unsere Definition als Volumen dieser Strecke

$$\mathrm{Vol}_1(P(\vec{a})) = \sqrt{\langle \vec{a}, \vec{a} \rangle} = |\vec{a}|$$

die Länge von $\vec{a}$, die als 1-dimensionales Volumen interpretiert werden kann. ◯

Beispiel. Im Spezialfall $n = 3$ und $k = 2$ d.h. für zwei Vektoren $\vec{a}_1, \vec{a}_2 \in \mathbb{R}^3$ ist $P(\vec{a}_1, \vec{a}_2)$ ein Parallelogramm und eine direkte Rechnung zeigt, dass dann

$$\mathrm{Vol}_2(P(\vec{a}_1, \vec{a}_2)) = |\vec{a}_1 \times \vec{a}_2|$$

gilt. In diesem Sonderfall können wir also das 2-dimensionale Volumen als Flächeninhalt des Parallelogramms interpretieren.

Allgemeiner ergibt sich für $k = 2$ Vekoren $\vec{a}_1, \vec{a}_2 \in \mathbb{R}^n$ als Determinante der Gram'schen Matrix

$$\det(A^T \cdot A) = \det \begin{pmatrix} |\vec{a}_1|^2 & \langle \vec{a}_1, \vec{a}_2 \rangle \\ \langle \vec{a}_2, \vec{a}_1 \rangle & |\vec{a}_2|^2 \end{pmatrix} = |\vec{a}_1|^2 \cdot |\vec{a}_2|^2 - \langle \vec{a}_1, \vec{a}_2 \rangle^2$$

Ersetzen wir nun diesen algebraischen Ausdrücke durch die geometrische Version des Skalarproduktes $\langle \vec{a}_1, \vec{a}_2 \rangle = |\vec{a}_1| \cdot |\vec{a}_2| \cdot \cos(\varphi)$ mit Hilfe des Zwischenwinkels $\varphi = \angle(\vec{a}_1, \vec{a}_2)$ der beiden Vektoren, erhalten wir für die Gram'sche Determinante

$$\det(A^T \cdot A) = |\vec{a}_1|^2 \cdot |\vec{a}_2|^2 \cdot \big(1 - \cos^2(\varphi)\big) = |\vec{a}_1|^2 \cdot |\vec{a}_2|^2 \cdot \sin^2(\varphi)$$

und damit für den Flächeninhalt des von den beiden Vektoren aufgespannten Parallelogramms

$$\mathrm{Vol}_2(P(\vec{a}_1,\vec{a}_2)) = \sqrt{\det(A^T\cdot A)} = |\vec{a}_1|\cdot|\vec{a}_2|\cdot|\sin(\varphi)|$$

Diese Formel drückt den aus der Schule bekannten Sachverhalt aus, dass der Flächeninhalt eines Parallelogramms als Länge einer Seite $|\vec{a}_1|$ mal der Länge der Höhe $|\vec{a}_2|\cdot|\sin(\varphi)|$ berechnet werden kann. ◯

Beispiel. Die selbe Überlegung im Fall $k = 3$ liefert eine Formel für das Volumen eines von drei Vektoren $\vec{a}_1,\vec{a}_2,\vec{a}_3$ aufgespannten Spates mit Hilfe der drei Zwischenwinkel $\varphi_1 = \angle(\vec{a}_2,\vec{a}_3)$, $\varphi_2 = \angle(\vec{a}_1,\vec{a}_3)$ und $\varphi_3 = \angle(\vec{a}_1,\vec{a}_2)$. Diesmal erhalten wir für die symmetrische Gramsche Matrix

$$A^T\cdot A = \begin{pmatrix} |\vec{a}_1|^2 & \langle\vec{a}_1,\vec{a}_2\rangle & \langle\vec{a}_1,\vec{a}_3\rangle \\ \langle\vec{a}_2,\vec{a}_1\rangle & |\vec{a}_2|^2 & \langle\vec{a}_2,\vec{a}_3\rangle \\ \langle\vec{a}_3,\vec{a}_1\rangle & \langle\vec{a}_3,\vec{a}_2\rangle & |\vec{a}_3|^2 \end{pmatrix}$$

Ihre Determinante ist das Formel-Monster

$$\begin{aligned}&|\vec{a}_1|^2|\vec{a}_2|^2|\vec{a}_3|^2 + 2\langle\vec{a}_1,\vec{a}_2\rangle\langle\vec{a}_2,\vec{a}_3\rangle\langle\vec{a}_1,\vec{a}_3\rangle - |\vec{a}_1|^2\langle\vec{a}_2,\vec{a}_3\rangle^2 - |\vec{a}_2|^2\langle\vec{a}_1,\vec{a}_3\rangle^2 - |\vec{a}_3|^2\langle\vec{a}_2,\vec{a}_2\rangle^2\\ &\quad = |\vec{a}_1|^2|\vec{a}_2|^2|\vec{a}_3|^2\Big(1 + 2\cos(\varphi_1)\cos(\varphi_2)\cos(\varphi_3) - \big(\cos^2(\varphi_1)+\cos^2(\varphi_2)+\cos^2(\varphi_3)\big)\Big)\end{aligned}$$

Damit liefert ihre Quadratwurzel einen Ausdruck für das Volumen eines Spates, der den wenigsten von uns aus der Schule bekannt sein dürfte.

Schliessen drei Einheitsvektoren je einen Winkel von $\varphi_1 = \varphi_2 = \varphi_3 = \frac{\pi}{4}$ ein, so hat das von ihnen aufgespannte Spat ein Volumen

$$\mathrm{Vol}_3(P(\vec{a}_1,\vec{a}_2,\vec{a}_3)) = \sqrt{1 + 2\cos^3\big(\tfrac{\pi}{4}\big) - 3\cos^2\big(\tfrac{\pi}{4}\big)} = \sqrt{\frac{\sqrt{2}-1}{2}} \approx 0.455\ldots$$

das unabhängig von der Dimension und der Lage der drei Vektoren ist. ◯

Beispiel. Um das Volumen des von den 3 Spaltenvektoren der Matrix

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

augespannten Spates zu bestimmen, nimmt man nicht den Umweg über die Winkel sondern berechnet ihre Gram'sche Matrix und erhält dafür

$$A^T\cdot A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}\cdot\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Für ihre Determinante erhält man $\det(A^T\cdot A) = 4$ und deshalb ist das gesuchte Volumen 2. ◯

Beispiel. Im Spezialfall $k = n$ erhalten wir ein n-dimensionales Spat, dessen Volumen wir auf Grund des Determinantenmultiplikationssatzes für quadratische Matrizen

$$\det(A \cdot B) = \det(A) \cdot \det(B)$$

und der Verträglichkeit der Determinante mit der Transposition

$$\det(A^T) = \det(A)$$

wegen

$$\sqrt{\det(A^T \cdot A)} = \sqrt{\det(A^T) \cdot \det(A)} = \sqrt{(\det(A))^2} = |\det(A)|$$

mit Hilfe des Betrages der Determinante

$$\mathrm{Vol}_n(P(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)) = |\det(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)|$$

berechnen können. ◯

Kapitel 4

Spektraltheorie

In diesem Kapitel suchen wir für eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ eine äquivalente Beschreibung, die möglichst übersichtlich ist, um damit Potenzen A^k der betreffenden Matrix oder allgemeiner beliebige Funktionen und insbesondere Potenzreihen

$$f(A) = \sum_{k=0}^{\infty} a_k A^k$$

effizient berechnen zu können. Potenzen und Potenzreihen quadratischer Matrizen spielen bei der Untersuchung dynamischer Systeme eine zentrale Rolle. Bei diskreten Sichtweise geht man von einem System autonomer, linearer Differenzengleichungen aus.

$$\vec{y}(k+1) = A \cdot \vec{y}(k), \qquad \vec{y}(k) = A^k \cdot \vec{y}(0)$$

Zur Lösung benötigt man die Matrizenpotenz A^k. Im Gegensatz dazu geht man bei kontinuierlicher[1] Sichtweise von einem System autonomer, linearer Differenzialgleichungleichungen aus.

$$\vec{y}\,'(t) = A \cdot \vec{y}(t), \qquad \vec{y}(t) = e^{At} \cdot \vec{y}(0)$$

Zur Lösung benötigt man diesmal den Propagator e^{At}, den man als Exponentialfunktion von Matrizen berechnet. Wir werden uns hier ausschliesslich mit dem diskreten Problem befassen und den kontinuierlichen Fall der Analysis überlassen, obwohl er eigentlich auch zu linearen Algebra gehören würde. Der diskrete Fall approximiert den kontinuierlichen Fall genau so, wie die Zinseszinsrechnung im Grenzwert beliebig kleiner Zeitschritte in die stetige Verzinsung übergeht. Umgekehrt entsteht der diskrete Fall aus dem kontinuierlichen Fall, indem man eines der in der Analysis besprochenen Näherungsverfahren zum numerischen Lösen von Differentialgleichungen — etwa das Euler-Verfahren — verwendet. Insofern sind die beiden Vorgehensweisen äquivalent und dem Studenten, der beide Gebiete kennt, sei dringend empfohlen, sie miteinander zu kombinieren[2].

[1] und nicht etwa indiskreter!

[2] In der tausendjährigen Diskussion darüber, ob Raum und Zeit eine *kon*tinuierliche oder eine dis*krete* Struktur haben, scheint sich das Blatt in den letzten Jahrzehnten — mindestens seit der Quantenmechanik, in der sich etwa die Energie oder der Drehimpuls als diskret erwiesen haben und nicht messbare Konzepte, wie beispielweise Bahnkurven und damit Ge-

In einem ersten Schritt werden wir die Eigenwertgleichung

$$A \cdot \vec{x} = \lambda \vec{x}, \qquad (\vec{x} \neq \vec{0})$$

bzw. die äquivalente homogene Eigenwertgleichung

$$(A - \lambda E) \cdot \vec{x} = \vec{0}, \qquad (\vec{x} \neq \vec{0})$$

lösen und dazu das Minimalpolynom $\mu_A(\lambda)$ der Matrix A bestimmen müssen. Im Gegensatz zu den bisherigen Themen werden wir hier nicht mehr mit rationalen Zahlen als Matrizenelemente auskommen, sondern Matrizen wie die Koeffizientenmatrix der homogenen Eigenwertgleichung — die *charakteristische Matrix* von A

$$A - \lambda E \in \mathrm{Mat}_n(\mathbb{Q}[\lambda])$$

antreffen, deren Elemente Polynome im Polynomring $\mathbb{Q}[\lambda]$ sind. Nicht nur ist die Arithmetik mit Polynomen teurer als jene mit Skalaren, sondern wir müssen beim Rechnen mit Polynommatrizen sorgfältig darauf achten, dass wir keine der notwendigen Fallunterscheidungen übersehen, bzw. beim Rechnen klug genug vorgehen, um Fallunterscheidungen vermeiden zu können. Ferner können wir im Polynomring $\mathbb{Q}[\lambda]$ in der Regel nicht dividieren. Das ist nur für die sogn. Einheiten möglich. Das sind genau die nicht verschwindenden Konstanten. Zum Glück haben Polynomringe über Köpern viele Gemeinsamkeiten mit dem Ring $\mathbb{Z}$ der ganzen Zahlen, so dass wir uns dort umsehen können und nur die Arithmetik der ganzen Zahlen auf solche Polynome übertragen müssen. Heikler wird es im Umgang mit dem ganzzahligen Polynomring $\mathbb{Z}[x]$, der zwar noch faktoriell, aber kein Hauptidealbereich ist und sich der grösste gemeinsame Teiler zweier Elemente nicht mehr mit dem Euklid'schen Algorithmus als Linearkombination dieser beiden Elemente schreiben lässt. Dieses Phänomen erkennt man schon an den beiden teilerfremden Polynomen $2, x \in \mathbb{Z}[x]$, für die also $\mathrm{ggT}(2, x) = 1$ gilt.

Neben diesen organisatorischen Schwierigkeiten beim Bestimmen des Minimalpolynoms $\mu_A(\lambda)$ kommt dann die Schwierigkeit dazu, dass die gesuchten Eigenwerte, die Nullstellen dieses Polynoms sind, in der Regel nicht durch rationale Operationen beschrieben werden können. Weil wir polynomiale Gleichungen ab dem Grad 5 in der Regel nur näherungsweise lösen können, sind wir zum Ausweichen auf numerische Methoden gezwungen. Weil die charakteristische Gleichung $\mu_A(\lambda) = 0$ nicht linear ist, aber zur Bestimmung der für das Studium der Dynamik erforderlichen Normalform $\mathcal{J}(A)$ der Matrix A gelöst werden muss, ist die lineare Algebra als Theorie unvollständig.

In der numerischen Mathematik werden Methoden angegeben, die direkt auf iterativen Prozessen beruhen, um das Eigensystem einer Matrix über einem ungenauen Körper zu approximieren. Wer also auf solche numerischen Näherungen angewiesen ist, sollte sich in der zugehörigen Literatur [**?**] der numerischen Analysis umsehen. Er sei aber gewarnt, dass die Normalform $\mathcal{J}(A)$ nicht stabil ist, weil sie von den Lösungen nichtlinearer Gleichungen abhängt.

---

schwindigkeit und Beschleunigung keinen Sinn haben — zu Gunsten der diskreten Struktur zu wenden. Obwohl wir Zeitintervalle von Naonsekunden noch nicht genau messen können, messen unsere Uhren, wie genau sie auch gehen mögen, eine diskrete Information und das Kontinuum scheint eine (hartnäckige) Illusion zu sein, für die wir lebenslänglich trainiert wurden und die Filmemacher und Informatiker geschickt auszunützen wissen. Der Autor wäre nicht überrascht, wenn wir gelegentlich eine ganz neue *kon-krete* Struktur fänden und es sich zeigt, wie Raum und Zeit besser mit dieser Strukur beschrieben werden.

4.1 Lineare Vektorfolgen

Weil wir seinerzeit bei der Besprechung linearer Vektorfolgen das Eigensystem ohne viel Federlesen benutzt haben, müssen wir erklären, wie es auf natürliche Art beim Studium diskreter dynamischer Systeme auftaucht. Dazu gehen wir vom autonomen, linearen Anfangswertproblem der Form

$$\vec{y}(k+1) = A \cdot \vec{y}(k), \qquad \vec{y}(0) = \vec{a}$$

aus, dessen Zustand $\vec{y}(k)$ nach k Schritten durch die Beziehung

$$\vec{y}(k) = A^k \cdot \vec{a}, \qquad k \geq 0$$

beschrieben werden kann, wie man durch Einsetzen in die Differenzengleichung mit Hilfe der Rekursion der Potenzen leicht erkennt.

Offenbar führt das Problem, die Dynamik von Vektorfolgen zu verstehen, auf das Problem, die Potenzen A^k der quadratischen Matrix $A \in \mathbb{R}^{n,n}$ zu untersuchen. Zu diesem Zweck ist die seinerzeit bestimmte Normalform $\mathcal{N}(A)$ von A deshalb nicht brauchbar, weil sie nicht mit den Potenzen von Matrizen verträglich ist. Das liegt daran, dass sich die Äquivalenz von Matrizen nicht mit dem Matrizenprodukt verträgt.

Wie seinerzeit bei den Wünschen an das Eliminationsverfahren fragen wir uns auch jetzt, ob wir eine quadratische Matrix A so zu einer gewissen Matrix B umformen können, dass gilt:

1. Die Dynamik ändern sich beim Umformen nicht.
2. Die Dynamik der neuen Matrix ist einfach zu berechnen.

Wiederum überlegen wir uns zunächst, welche Operationen man mit einer quadratischen Matrix durchführen darf, ohne dass sich die Dynamik ändert. Eine solche Operation muss wie früher strukturverträglich d.h. in unserem Fall linear und umkehrbar sein. Solche umkehrbaren linearen Operationen werden bekanntlich durch eine invertierbare Matrix X beschrieben. Weil diesmal die Matrix A quadratisch ist und daher Quelle und Ziel des zugehörigen Prozesses übereinstimmen, bietet es sich an, die beiden Transformationen *simultan* durchzuführen. Wir suchen also eine invertierbare Matrix X mit der Eigenschaft, dass das Diagramm

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{\quad B \quad} & \mathbb{R}^n \\ X \,\big|\, \sim \downarrow & & \downarrow \sim \,\big|\, X \\ \mathbb{R}^n & \xrightarrow[\quad A \quad]{} & \mathbb{R}^n \end{array}$$

kommutiert, d.h. dass zwischen der gegebenen Matrix A und der transformierten Matrix B die Beziehung

$$A \cdot X = X \cdot B$$

Man nennt zwei Matrizen A und B, die in dieser Beziehung stehen, *ähnlich*. Die zu A ähnliche Matrix B wird mit der invertierbaren Matrix $X \in \mathrm{Gl}_n(\mathbb{R})$ konjugiert, so dass die Bedingung

$$A = X \cdot B \cdot X^{-1}$$

gilt. Ähnlich Matrizen beschreiben dieselbe lineare Abbildung bezüglich verschiedener Bezugssysteme. Früher hätte man in dieser Situation gesagt, die beiden Matrizen A und B gehen durch eine Koordinatentransformation mit Hilfe von X auseinander hervor.

Satz. Die Ähnlichkeit von Matrizen ist eine Äquivalenzrelation.

Beweis. Um zu erkennen, dass die Ähnlichkeit reflexiv ist, müssen wir zeigen dass jede Matrix A zu sich selber ähnlich ist. Das erkennt man aber leicht, wenn man für $X = E$ die Einheitsmatrix wählt.

Um zu zeigen, dass die Ähnlichkeit symmetrisch ist, müssen wir zeigen, dass wenn A ähnlich zu B ist, dass dann auch B ähnlich zu A ist. Nach Voraussetzung existiert also eine invertierbare Matrix X mit der Eigenschaft, dass $B = X^{-1} \cdot A \cdot X$ gilt. Dann gilt durch Auflösen nach A die symmetrische Beziehung

$$A = X \cdot B \cdot X^{-1} = (X^{-1})^{-1} \cdot B \cdot X^{-1}$$

Weil mit X auch X^{-1} invertierbar ist, zeigt sich, dass B in der Tat ähnlich zu A ist.

Dass die Ähnlichkeit transitiv ist, ergibt sich daraus, dass nach Voraussetzung A ähnlich zu B und B ähnlich zu C ist. Daher gibt es invertierbare Matrizen X und Y so, dass die Beziehungen

$$B = X^{-1} \cdot A \cdot X, \qquad C = Y^{-1} \cdot B \cdot Y$$

gelten. Einsetzen von B liefert damit die Beziehung

$$C = Y^{-1} \cdot (X^{-1} \cdot A \cdot X) \cdot Y = (Y^{-1} \cdot X^{-1}) \cdot A \cdot (X \cdot Y) = (X \cdot Y)^{-1} \cdot A \cdot (X \cdot Y)$$

Aus ihr folgt, dass A auch zu C ähnlich ist. □

Dass die Dynamik ähnlicher Matrizen die selben sind, erkennt man, wenn man die Differenzengleichung

$$\vec{y}(k+1) = A \cdot \vec{y}(k)$$

mit der invertierbaren Matrix X^{-1} transformiert. Erklärt man nämlich den transformierten Zustand durch $\vec{z}(k) = X^{-1} \cdot \vec{y}(k)$, so gilt $\vec{y}(k) = X \cdot \vec{z}(k)$. Damit wird aus der gegebenen Differenzengleichung

$$X^{-1} \cdot \vec{y}(k+1) = X^{-1} \cdot A \cdot \vec{y}(k) = (X^{-1} \cdot A) \cdot X \cdot \vec{z}(k)$$

Daher erfüllt der transformierte Zustand $\vec{z}(k)$ die Differenzengleichung

$$\vec{z}(k+1) = (X^{-1} \cdot A \cdot X) \cdot \vec{z}(k) = B \cdot \vec{z}(k), \qquad B = X^{-1} \cdot A \cdot X$$

mit einer ähnlichen Systemmatrix. Im kontinuierlichen Fall liefert eine Variablensubstitution in einer Differentialgleichung das entsprechende Ergebnis.

Ähnliche Matrizen haben ähnliche Potenzen, wie das Teleskopprodukt

$$A^n = \underbrace{(X \cdot B \cdot X^{-1}) \cdot (X \cdot B \cdot X^{-1}) \cdots (X \cdot B \cdot X^{-1})}_{n \text{ Faktoren}} = X \cdot B^n \cdot X^{-1}$$

zeigt.

Damit eine Matrix A in eine ähnliche Matrix $B = X^{-1} \cdot A \cdot X$ übergeht, kommen die folgenden *simultanen Elementaroperationen* in Frage.

SI. Vertauschen der i-ten mit der j-ten Zeile und in der entstandenen Matrix simultanes Vertauschen der j-ten Spalte mit der i-ten Spalte.

SII. Addition des r-fachen der i-ten Zeile zur j-ten Zeile und in der entstandenen Matrix simultane Addition des $(-r)$-fachen der j-ten Spalte zur i-ten Spalte.

SIII. Multiplikation der i-ten Zeile mit dem Faktor $r \neq 0$ und in der entstandenen Matrix simultane Multiplikation der i-ten Spalte mit dem Faktor $\frac{1}{r}$.

Vom Standpunkt der Matrizenrechnung aus suchen wir also für eine quadratische Matrix A eine ähnliche Normalform $\mathcal{J}(A)$, deren Potenzen einfach zu berechnen sind. Leider ist es beim Rechnen mit simultanen Operationen etwas schwieriger, den Überblick zu behalten, als beim Umgang mit Zeilenoperationen. Ferner bleibt die Frage zu klären, wie eine ähnliche Normalform $\mathcal{J}(A)$ aussehen könnte. Ihre Antwort ist Gegensatz zur seinerzeit benutzen Stufenform $S(A)$ bzw. der Normalform $\mathcal{N}(A)$ einer rechteckigen Matrix nicht ganz einfach. Der Grund für die Schwierigkeit liegt darin, dass wir nun bei einer quadratischen Matrix A nur eine einzige Matrix X zur Verfügung haben, die wir wählen können, um für A ein möglichst einfaches Bezugssystem zu finden. Im Gegensatz dazu konnten wir früher zwei Basen in der Quelle und im Ziel der Abbildung variieren. Nur eine simultane Transformation X lässt allerdings das dynamisch Verhalten unverändert, wie obiges Teleskopprodukt zeigt.

Es wird sich zeigen, dass die Untersuchung der ähnlichen Normalform $\mathcal{J}(A)$ einer quadratischen Matrix $A \in \mathbb{Q}^{n,n}$ mit der Untersuchung der Normalform einer gewissen Matrix — der charakteristischen Matrix $A - \lambda E \in \mathrm{Mat}_n(\mathbb{Q}[\lambda])$ von A zusammenhängt. Zwei quadratische Matrizen A und B sind nämlich genau dann ähnlich, wenn die charakteristischen Polynommatrizen $A - \lambda E$ und $B - \lambda E$ in $\mathrm{Mat}_n(\mathbb{Q}[\lambda])$ äquivalent sind. Daher werden wir uns also mit dem Berechnen der Normalformen von Polynommatrizen, deren Elemente also Polynome im Polynomring $\mathbb{Q}[\lambda]$ sind, befassen müssen.

Um mit dieser anspruchsvolleren Theorie Erfahrungen sammeln zu können, wollen wir sie nicht direkt anpacken, sondern uns zunächst mit einem (wichtigen!) Spezialfall befassen. Klar ist, dass die Potenzen einer *Diagonalmatrix* einfach zu berechnen sind, da die Information in einer Diagonalmatrix vollständig entkoppelt ist. Tatsächlich wird eine beliebige Diagonalmatrix $D = \mathrm{diag}(d_1, \ldots, d_n)$ potenziert, indem man sämtlich Diagonalelemente in die entsprechende Potenz erhebt. Analoges gilt damit für beliebige Potenzreihen von Diagonalmatrizen und wir erhalten nämlich für Diagonalmatrizen die einfache Beziehung

$$f(D) = \mathrm{diag}\big(f(d_1), \ldots, f(d_n)\big)$$

Falls es uns also gelingt, zur gegebenen Matrix A eine konjugierte Diagonalmatrix $D(A)$ als Normalform zu finden, ist unser Problem gelöst. Eine solche Diagonalisierung ist jedoch nicht für jede quadratische Matrix A erhältlich und wir brauchen deshalb später eine feinere Theorie um zu beschreiben, wie die Gestalt der sogn. Jordan'schen Normalform $\mathcal{J}(A)$ aussehen soll.

Zur Motivation der Diagonalisierung und der zugehörigen Eigenwertgleichung bieten sich grundsätzlich zwei Wege an. Einerseits könnten wir uns auf algebraischem Weg überlegen, wie wir zu einer quadratischen Matrix A eine ähnliche

Diagonalmatrix $D(A)$ als Normalform erhalten. Andererseits können wir zu den diskreten dynamischen Systemen zurückkehren und uns überlegen, wie sie zu lösen sind, indem wir uns gewisse einfache Lösungen aus den Fingern saugen. Wir schlagen in diesem Abschnitt zunächst den zweiten, dynamisch motiverten Weg ein, der in der Analysis üblich und deshalb wohl aus historischen Gründen vielen Anwendern geläufig ist. Im nächsten Abschnitt werden wir dann auf direktere, algebraische, Art vorgehen.

Um zu verstehen, in welche Richtung wir suchen müssen, kehren wir zum Problem der Vektorfolge zurück, die einem diskreten, autonomen linearen Differenzengleichung der Form

$$\vec{y}(k+1) = A \cdot \vec{y}(k)$$

genügt. Wir versuchen nun, einen möglichst grossen Vorrat an einfachen Basislösungen $\vec{y}_1(k), \vec{y}_2(k), \ldots, \vec{y}_n(k) \in \mathbb{R}^n$ für diese Differenzengleichung anzulegen, aus dem wir dann hoffentlich die gesuchte spezielle Lösung, die zusätzlich der Anfangsbedingung $\vec{y}(0) = \vec{a}$ genügt, linear zusammenbauen können.

Man beachte, dass der Nullvektor $\vec{y}(k) = \vec{0}$ zwar eine triviale Lösung der Differenzengleichung liefert, an der wir aber nicht interessiert sind, weil wir damit keine interessanten Anfangsbedingungen linear kombinieren können.

Weil uns die Differenzengleichung an die Rekursionsgleichung der geometrischen Folge $y(k+1) = a \cdot y(k)$ aus der Schule erinnert — es ist ihre höherdimensionale Verallgemeinerung — erwarten wir, dass sich gewisse ihrer Lösungen mit Hilfe der Exponentialfunktion beschreiben lassen. Um einen möglichst grossen Vorrat an Lösungen unserer Differenzengleichung anlegen zu können, machen wir für die gesuchte Lösung den *Exponentialansatz*

$$\vec{y}(k) = \lambda^k \vec{v}$$

für die vorläufig noch unbekannte Konstante λ und den unbekannten, von k unabhängigen Vektor $\vec{y}(0) = \vec{v} \neq \vec{0}$. Konstante und Vektor sind nun so zu bestimmen, dass der Exponentialansatz eine nichttriviale Lösung der Differenzengleichung liefert. Auf Grund des Ansatzes ist

$$\vec{y}(k+1) = \lambda^{k+1} \vec{v}$$

Setzen wir dies in die Differenzengleichung ein, erhalten wir

$$\lambda^{k+1} \vec{v} = \lambda^k A \cdot \vec{v}$$

bzw. nach Division durch λ^k die zeitunabhängige *Eigenwertgleichung*

$$A \cdot \vec{v} = \lambda \vec{v}$$

Wir suchen Bedingungen an λ, so dass diese Gleichung nicht triviale Lösungesvektoren $\vec{v}$ hat. Ein solches λ muss also *Eigenwert* von A und $\vec{v}$ muss der zugehörige *Eigenvektor* sein.

Jeder solche Eigenvektor $\vec{v} \neq \vec{0}$ zum Eigenwert λ liefert mit Hilfe des Exponentialansatzes eine Lösung $\vec{y}(k) = \lambda^k \vec{v}$ der Differenzengleichung. Um zu verstehen, welche spezielle Bedeutung diese *Basislösungen* für die Dynamik des Systems hatben, beachten wir ihre Eigenschaften.

1. Es gilt die Differenzengleichung $\vec{y}(k+1) = A \cdot \vec{y}(k)$.

2. Es gilt die Anfangsbedingung $\vec{y}(0) = \vec{v}$.

Die erste Eigenschaft ergibt sich nach Konstruktion sofort aus

$$\vec{y}(k+1) = \lambda^{k+1}\vec{v} = \lambda^k \cdot \lambda\vec{v} \overset{(*)}{=} \lambda^k \cdot A \cdot \vec{v} = A \cdot \lambda^k\vec{v} = A \cdot \vec{y}(k)$$

Dabei haben wir für die Gleichung $(*)$ die Voraussetzung verwendet, dass $\vec{v}$ Lösung der Eigenwertgleichung zum Eigenwert λ ist. Zur Kontrolle der angegebenen Anfangsbedingung setzt man im Exponentialansatz $k = 0$. Auf Grund der Eigenwertgleichung für den Eigenvektor $\vec{v}$ zum Eigenwert λ folgt durch wiederholte Multiplikation mit A die Beziehung

$$A^k \cdot \vec{v} = \lambda^k\vec{v}$$

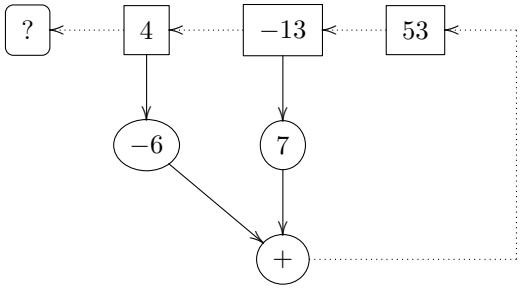
Daher ist $\vec{v}$ auch Eigenvektor jeder Potenz A^k zum Eigenwert λ^k.

Falls wir das System speziell präparieren und es in den Anfangszustand $\vec{v}$ versetzen, entwickelt es sich im Lauf der Zeit auf Grund des Exponentialansatzes

$$\vec{y}(k) = A^k \cdot \vec{y}(0) = A^k \cdot \vec{v} = \lambda^k\vec{v}$$

so, dass seine Zustände von $\vec{v}$ linear abhängig bleiben, d.h. das System bewegt sich dann ausschliesslich auf einer Ursprungsgeraden in Richtung von $\vec{v}$. Die Beträge der Zustände ändern sich im Laufe der Zeit exponentiell mit dem Wachstumsfaktor λ. Selbstverständlich wird es für unser System nicht all zu viele unabhängige solcher speziell überschaubarer Basislösungen geben. Aus Dimensionsgründen erwarten wir höchstens so viele Eigenzustände, wie der Zustandsraum Dimensionen hat. Falls es so viele Eigenzustände gibt, können wir sie als Basislösungen zur Beschreibung des Systems benutzen und aus ihnen die allgemeine Lösung des Systems als Superposition darstellen. Das entspricht genau dem Fall, wo die Systemmatrix A diagonalisierbar ist.

Beispiel. Das seinerzeit bei der Besprechung der Potenzen bereits eingehend untersuchte Schieberegister



gehört zum rekursiven Anfangswertproblem

$$x(k+3) = -6x(k) + 7x(k+1), \qquad x(0) = 4, x(1) = -13, x(2) = 53$$

und daher zum System mit der Systemmatrix und dem Anfangszustand

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & 7 & 0 \end{pmatrix}, \qquad \vec{a} = \begin{pmatrix} 4 \\ -13 \\ 53 \end{pmatrix}$$

Da es sich hier um eine Begleitermatrix handelt, lässt sich sein Eigensystem leicht bestimmen. Auf Grund unserer Erfahrungen vermuten wir, dass das Minimalpolynom

$$\mu_A(\lambda) = -6 + 7\lambda - \lambda^3$$

lautet und die Lösungen $\sigma_A = \{1, 2, -3\}$ hat. Auch die zugehörigen Eigenvektoren kennen wir bereits und erwarten die drei Vektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} 1 \\ -3 \\ 9 \end{pmatrix}$$

die zu den drei speziellen exponentiellen Lösungen $y_1(k) = 1^k$, $y_2(k) = 2^k$ und $y_3(k) = (-3)^k$ gehören, deren Basen die Eigenwerte sind.

Um zunächst die Vermutung bezüglich des Minimalpolynoms zu bestätigen, machen wir für das Minimalpolnom den Ansatz $\mu_A(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 - \lambda^3$ eines beliebigen normierten kubischen Polynoms. Einsetzen der Matrix A liefert die Linearkombination

$$\begin{aligned} & a_0E + a_1A + a_2A^2 \\ = \; & a_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & 7 & 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 & 0 & 1 \\ -6 & 7 & 0 \\ 0 & -6 & 7 \end{pmatrix} \\ = \; & \begin{pmatrix} a_0 & a_1 & a_2 \\ -6a_0a_2 & a_0 + 7a_2 & a_1 \\ -6a_1 & 7a_1 - 6a_2 & a_0 + 7a_2 \end{pmatrix} = \begin{pmatrix} -6 & 7 & 0 \\ 0 & -6 & 7 \\ -42 & 49 & -6 \end{pmatrix} = A^3 \end{aligned}$$

Man beachte, dass hier die obersten Zeilen wegen der speziellen Form von Begleitermatrizen eine einfache Struktur haben, die es sofort erlaubt die eindeutige Lösung abzulesen. Der Koeffizientenvergleich der obersten Zeilen zeigt sofort, dass die eindeutige Lösung hier $a_0 = -6, a_1 = 7, a_2 = 0$ lauten muss, was unsere Vermutung $\mu_A(\lambda) = -6 + 7\lambda - \lambda^3$ bestätigt.

Um zu demonstrieren, wie man die erwarteten Resultate ohne weitere Voraussetzung an die Matrix A durch blindes Rechnen erhält, lösen wir nun die Eigenwertgleichung $A \cdot \vec{x} = \lambda\vec{x}$ bzw. die durch Subtraktion erhaltene homogene Eigenwertgleichung

$$(A - \lambda E) \cdot \vec{x} = \vec{0}, \qquad (\vec{x} \neq \vec{0})$$

und bestimmen das zugehörigen Eigensystems. Dazu gehen wir von seiner Koeffizientenmatrix, d.h. von der charakteristischen Matrix

$$A - \lambda E_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & 7 & 0 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \\ -6 & 7 & -\lambda \end{pmatrix}$$

aus. Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die erste mit der dritten Zeile und erhalten

$$\begin{pmatrix} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ -\lambda & 1 & 0 \end{pmatrix} \qquad \begin{bmatrix} T_{13} \\ \\ T_{31} \end{bmatrix}$$

Addition des $(-\lambda)$-fachen der ersten Zeile zum 6-fachen der dritten Zeile ergibt

$$\left(\begin{array}{ccc} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ 0 & 6-7\lambda & \lambda^2 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ ZS_{13}(-\lambda,6) \end{array}\right]$$

Addieren wir nun das (-7)-fache der zweiten Zeile zur dritten, erhalten wir die Matrix

$$\left(\begin{array}{ccc} -6 & 7 & -\lambda \\ 0 & -\lambda & 1 \\ 0 & 6 & \lambda^2-7 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(-7) \end{array}\right]$$

Vertauschen der letzte beiden Zeilen führt zu einer Matrix, deren führende Elemente feste numerische Werte sind und wir deshalb keine unnötigen Fallunterscheidungen durchführen müssen.

$$\left(\begin{array}{ccc} -6 & 7 & -\lambda \\ 0 & 6 & \lambda^2-7 \\ 0 & -\lambda & 1 \end{array}\right) \qquad \left[\begin{array}{c} \\ T_{23} \\ T_{32} \end{array}\right]$$

Addition des λ-fachen der zweiten Zeile zum 6-fachen der dritten Zeile liefert die Stufenform

$$S(\lambda) = \left(\begin{array}{ccc} -6 & 7 & -\lambda \\ 0 & 6 & \lambda^2-7 \\ 0 & 0 & 6-7\lambda+\lambda^3 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ ZS_{23}(\lambda,6) \end{array}\right]$$

Damit nichttriviale Lösungen der Eigenwertgleichung $(A-\lambda E)\cdot\vec{x}=\vec{0}$ bzw. des äquivalenten Systems $S(\lambda)\cdot\vec{x}=\vec{0}$ existieren, muss also das Minimalpolynom

$$\mu_A(\lambda) = 6-7\lambda+\lambda^3 = (\lambda-1)\cdot(\lambda-2)\cdot(\lambda+3)$$

verschwinden. Daher muss der Eigenwert λ einen der Werte aus dem Spektrum $\sigma_A=\{1,2,-3\}$ annehmen.

Um die zugehörigen Eigenvektoren zu finden, setzen wir diese Eigenwerte der Reihe nach in die gefundene obere Dreiecksmatrix $S(\lambda)$ ein und lösen das entstehende lineare, homogene Gleichungssystem $S(\lambda)\cdot\vec{x}=\vec{0}$.

1. Fall: $\lambda=1$. Das homogene System $S(\lambda)\cdot\vec{x}=\vec{0}$ hat die Koeffizientenmatrix

$$S(1) = \left(\begin{array}{ccc} -6 & 7 & -1 \\ 0 & 6 & -6 \\ 0 & 0 & 0 \end{array}\right), \qquad \left(\begin{array}{ccc} -6 & 7 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array}\right) \quad \left[\begin{array}{c} S_3(\frac{1}{6}) \\ \\ \end{array}\right]$$

Addition des (-7)-fachen der zweiten Zeile zur ersten ergibt die reduzierte Stufenform

$$\left(\begin{array}{ccc} -6 & 0 & 6 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array}\right) \quad \left[\begin{array}{c} Z_{21}(-7) \\ \\ \end{array}\right], \qquad \left(\begin{array}{ccc} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array}\right) \quad \left[\begin{array}{c} S_1(-\frac{1}{6}) \\ \\ \end{array}\right]$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir als Fixpunktraum von A den Eigenraum zum Eigenwert 1 der Matrix A.

$$V_1 = \left\{ t\left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right) \mid t\in\mathbb{R} \right\}$$

2. Fall: $\lambda = 2$. Das homogene System $S(\lambda) \cdot \vec{x} = \vec{0}$ hat die Koeffizientenmatrix

$$S(2) = \begin{pmatrix} -6 & 7 & -2 \\ 0 & 6 & -3 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \left(\begin{array}{ccc} -6 & 7 & -2 \\ 0 & 2 & -1 \\ 0 & 0 & 0 \end{array} \right) \quad \left[\begin{array}{c} S_2(\frac{1}{3}) \\ \\ \end{array} \right]$$

Addition des (-7)-fachen der zweiten Zeile zum 2-fachen der ersten ergibt die Stufenform

$$\left(\begin{array}{ccc} -12 & 0 & 3 \\ 0 & 2 & -1 \\ 0 & 0 & 0 \end{array} \right) \left[\begin{array}{c} ZS_{21}(-7,2) \\ \\ \end{array} \right], \qquad \left(\begin{array}{ccc} -4 & 0 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 0 \end{array} \right) \left[\begin{array}{c} S_1(-\frac{1}{3}) \\ \\ \end{array} \right]$$

Schreiben wir die Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert 2 der Matrix A.

$$V_2 = \left\{ t \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

3. Fall: $\lambda = -3$. Das homogene System $S(\lambda) \cdot \vec{x} = \vec{0}$ hat die Koeffizientenmatrix

$$S(-3) = \begin{pmatrix} -6 & 7 & 3 \\ 0 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \left(\begin{array}{ccc} -6 & 7 & 3 \\ 0 & 3 & 1 \\ 0 & 0 & 0 \end{array} \right) \left[\begin{array}{c} S_2(\frac{1}{2}) \\ \\ \end{array} \right]$$

Addition des (-7)-fachen der zweiten Zeile zum 3-fachen der ersten ergibt die Stufenform

$$\left(\begin{array}{ccc} -18 & 0 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 0 \end{array} \right) \left[\begin{array}{c} ZS_{21}(-7,3) \\ \\ \end{array} \right], \qquad \left(\begin{array}{ccc} -9 & 0 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 0 \end{array} \right) \left[\begin{array}{c} S_1(\frac{1}{2}) \\ \\ \end{array} \right]$$

Schreiben wir die Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert -3.

$$V_{-3} = \left\{ t \begin{pmatrix} 1 \\ -3 \\ 9 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

Aus den drei Basisvektoren $\vec{v}_j$ dieser drei Eigenräume lassen sich nun mit Hilfe des Exponentialansatzes

$$\vec{y}_j(k) = \lambda_j^k \vec{v}_j, \qquad (1 \leq j \leq 3)$$

die drei Basislösungen

$$\vec{y}_1(k) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{y}_2(k) = 2^k \cdot \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \qquad \vec{y}_3(k) = (-3)^k \cdot \begin{pmatrix} 1 \\ -3 \\ 9 \end{pmatrix}$$

unseres Systems bestimmen. Sie erfüllen also die Differenzengleichung

$$\vec{y}_j(k+1) = A \cdot \vec{y}_j(k), \qquad (1 \leq j \leq 3)$$

des Systems, wie man zur Kontrolle überprüfen kann. Man beachte, dass es sich bei der ersten Vektorfolge um einen Fixpunkt des betrachteten Systems handelt, während die zweite Vektorfolge mit den Zweierpotenzen anwächst und die dritte mit den Dreierpotenzen oszilliert.

Die Diagonalisierung von A und die zugehörige Matrizenpotenz A^k haben wir seinerzeit bei der Besprechung der Matrizenpotenzen bereits berechnet. Wir wollen nun sehen, wie sich das verwendete Verfahren aus der dynamischen Sichtweise ergibt. Diese alternative Sichtweise wird oft in der Analysis-Literatur beim Lösen von Systemen linearer Differenzailgleichungen mit konstanten Koeffizienten benutzt.

Fassen wir die soeben gefunden Basislösungen als Spalten einer Matrix auf, erhalten wir die *Fundamentalmatrix*

$$F(k) = \begin{pmatrix} 1 & 1 \cdot 2^k & 1 \cdot (-3)^k \\ 1 & 2 \cdot 2^k & -3 \cdot (-3)^k \\ 1 & 4 \cdot 2^k & 9 \cdot (-3)^k \end{pmatrix}$$

Weil ihre 3 unabhängigen Spalten Lösungen der gegebenen Differenzengleichung sind, erfüllt die Fundamentalmatrix, die Differenzengleichung

$$F(k+1) = A \cdot F(k)$$

Die Fundamentalmatrix erfüllt also die Differenzengleichung der Matrizenpotenz und man könnte vermuten, bereits die gesuchte Potenz A^k vor sich zu haben. Nun ist aber

$$F(0) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 1 & 4 & 9 \end{pmatrix}$$

die Matrix, deren Spalten aus den gefundenen Eigenvektoren besteht. Diese Matrix haben wir seinerzeit mit X bezeichnet. Offenbar gilt zur Zeit $k = 0$ die Bedingung $F(0) \neq E$, was wir von der Matrizenpotenz A^k zusätzlich verlangen.

Aus der Fundamentalmatrix lässt sich aber die gesuchte Matrizenpotenz nun leicht ermitteln. Weil bei einem linearen homogenen System jede Linearkombination von Lösungen wieder Lösung ist, so ist auch

$$\vec{y}(k) = c_1\vec{y}_1(k) + c_2\vec{y}_2(k) + c_3\vec{y}_3(k)$$

Lösung unserer Differenzengleichung. Unser Vorrat sollte aus Dimensionsgründen gross genug sein, um damit jede beliebige Lösung linear kombinieren zu können. Fassen wir die Konstanten c_j zum Vektor $\vec{c}$ zusammen, so lautet diese Vektorgleichung in Matrizenschreibweise

$$\vec{y}(k) = F(k) \cdot \vec{c}$$

Es bleiben noch die Koeffizienten c_j so zu bestimmen, dass die Anfangsbedingung $\vec{y}(0) = \vec{a}$ erfüllt ist. Setzen wir $k = 0$, erkennen wir, dass wir $\vec{c}$ so wählen müssen, dass die Gleichung

$$F(0) \cdot \vec{c} = \vec{a}$$

gilt. Das erreichen wir mit Hilfe der Lösung

$$\vec{c} = F^{-1}(0) \cdot \vec{a}$$

In unserem Beispiel ist

$$F^{-1}(0) = \frac{1}{20}\begin{pmatrix} 30 & -5 & -5 \\ -12 & 8 & 4 \\ 2 & -3 & 1 \end{pmatrix}$$

Setzen wir $\vec{c}$ in obige Matrizengleichung ein, erhalten wir die die gesuchte Lösung des ursprünglichen Anfangswertproblem in der Form

$$\vec{y}(k) = F(k) \cdot \vec{c} = F(k) \cdot F^{-1}(0) \cdot \vec{a}$$

Dieser Vektor erfüllt nämlich immer noch, wie $F(k)$, die Differenzengleichung und zusätzlich die gewünschte Anfangsbedingung. Daraus entnehmen wir, dass für die gesuchte Matrizenopotenz

$$A^k = F(k) \cdot F^{-1}(0)$$

gilt. Im Beispiel erhalten wir durch Multiplizieren der beiden Matrizen

$$F(k) \cdot F^{-1}(0) \quad = \quad \begin{pmatrix} 1 & 1 \cdot 2^k & 1 \cdot (-3)^k \\ 1 & 2 \cdot 2^k & -3 \cdot (-3)^k \\ 1 & 4 \cdot 2^k & 9 \cdot (-3)^k \end{pmatrix} \cdot \frac{1}{20}\begin{pmatrix} 30 & -5 & -5 \\ -12 & 8 & 4 \\ 2 & -3 & 1 \end{pmatrix}$$

tatsächlich die früher durch Diagonalisieren bestimmte Potenz A^k.

$$\frac{1}{20}\begin{pmatrix} 30 - 12\,2^k + \ 2\,(-3)^k & -5 + \ 8\,2^k - \ 3\,(-3)^k & -5 + \ 4\,2^k + \ (-3)^k \\ 30 - 24\,2^k - \ 6\,(-3)^k & -5 + 16\,2^k + \ 9\,(-3)^k & -5 + \ 8\,2^k - 3\,(-3)^k \\ 30 - 48\,2^k + 18\,(-3)^k & -5 + 32\,2^k - 27\,(-3)^k & -5 + 16\,2^k + 9\,(-3)^k \end{pmatrix}$$

Der Zusammenhang der beiden Verfahren zur Berechnung der Potenz A^k via die Fundamentalmatrix bzw. die Diagonalisierung ergibt sich, wenn man beachtet, dass die Fundamentalmatrix das Produkt $F(k) = F(0) \cdot \Lambda^k$ mit der Potenz Λ^k der früher benutzen Diagonalmatrix

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 2 & -3 \end{pmatrix}$$

ist, in deren Diagonalen die Eigenwerte stehen. Daher gilt die Beziehung

$$A^k = F(k) \cdot F^{-1}(0) = F(0) \cdot \Lambda^k \cdot F^{-1}(0)$$

und wir erhalten die benutzte Diagonalisierung von A, falls wir die Transformationsmatrix $X = F(0)$ verwenden. Nachdem wir den Zusammenhang zwischen Fundamentalmatrix und Diagonalisierung verstanden haben, werden wir uns in Zukunft vorwiegend um die algebraische direktere Diagonalisierung einer Matrix A kümmern und es dem Leser überlassen, falls erwünscht, den Übergang zu den Fundamentalmatrizen durchzuführen. ◯

Die Diagonalisierung und die allgemeinere Jordan'sche Normalform einer quadratischen Matrix A spielen auch beim Lösen von Systemen linearer Differentialgleichungen $\vec{y}\,'(t) = A \cdot \vec{y}(t)$ mit konstanten Koeffizienten eine fundamentale Rolle, weil der Propagator mit Hilfe der Diagonalisierung durch

$$e^{At} = X \cdot e^{\Lambda t} \cdot X^{-1}$$

berechnet werden kann. Damit ist die Differentialgleichung gelöst. Ferner können lineare Differenzen- bzw. Differentialgleichungen höherer Ordnung mit Hilfe der Begleitermatrix auf solche linearen Differenzen- bzw. Differentialgleichungssysteme reduziert werden. Die Begleitermatrix eines Polynoms hat die Eigenschaft, dass ihr charakteristisches Polynom gerade das vorgegebene Polynom ist. Weil das Spektrum der Begleitermatrix eines Polynoms die Menge der Nullstellen dieses Polynoms ist, gibt es ferner einen engen Zusammenhang zwischen den Nullstellen von Polynomen und den Eigenwerten von Matrizen.

4.2 Das Eigensystem

Wir schlagen nun also den zweiten Weg ein und untersuchen mit algebraischen Mitteln den einfachsten Fall, wo die Matrix $A \in \mathbb{R}^{n,n}$ eine Diagonalmatrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ als Normalform hat. Dann gibt es nach Definition eine gewisse invertierbare Matrix $X = (\vec{x}_1, \ldots, \vec{x}_n)$ und eine gewisse Diagonalmatrix Λ mit der Eigenschaft, dass $A = X \cdot \Lambda \cdot X^{-1}$ gilt. Die Diagonalelemente von Λ seien $\lambda_1, \ldots, \lambda_n$. Multiplikation dieser Gleichung mit X von rechts zeigt, dass $A \cdot X = X \cdot \Lambda$ gelten muss. Für die Spaltenvektoren der Matrix $X = (\vec{x}_1, \ldots, \vec{x}_n)$ gilt nach Definition des Matrizenproduktes

$$X \cdot \Lambda = (\vec{x}_1, \ldots, \vec{x}_1) \cdot \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = (\lambda_1 \vec{x}_1, \ldots, \lambda_n \vec{x}_n)$$

Aus dem selben Grund ist

$$A \cdot X = (A \cdot \vec{x}_1, \ldots, A \cdot \vec{x}_n)$$

Weil nun $A \cdot X = X \cdot \Lambda$ gelten soll, müssen also die Gleichungen

$$A \cdot \vec{x}_i = \lambda_i \vec{x}_i \quad \text{für} \quad 1 \leq i \leq n$$

erfüllt sein. Daraus folgt, dass die i-te Spalte von X ein Eigenvektor zum Eigenvektor λ_i sein muss.

In vielen Anwendungen der Matrizenrechnung im Zusammenhang mit dynamischen Systemen begegnet man also Systemen von n Gleichungen mit n Unbekannten, die sich als $A \cdot \vec{x} = \lambda \vec{x}$ für einen unbekannten Skalar λ schreiben lassen. Man interessiert sich in erster Linie dafür, für welche λ dieses Gleichungssystem nichttriviale Lösungen hat. Gleichwertig ausgedrückt geht es darum, λ so zu wählen, dass das homogene lineare Gleichungssystem

$$(A - \lambda E_n) \cdot \vec{x} = \vec{0}, \qquad (\vec{x} \neq \vec{0})$$

nichttriviale Lösungen hat und diese Lösungen zu bestimmen. Es handelt sich also um ein spezielles lineares Gleichungssystem in $\vec{x}$ mit Parameter λ, das prinzipiell nach der früher besprochenen Methode lösen lässt.

Definition. Sei A eine $n \times n$-Matrix und λ irgend ein Skalar. Den Lösungsraum des homogenen linearen Gleichungssystems

$$(A - \lambda E_n) \cdot \vec{x} = \vec{0}, \qquad (\vec{x} \neq \vec{0})$$

mit der *charakteristischen Matrix* als Koeffizientenmatrix bezeichnet man als *Eigenraum*, der zu λ gehört. Wir schreiben V_λ für diesen Eigenraum. Seine

Dimension heisst *geometrische Vielfachheit* von λ und seine Elemente heissen *Eigenvektoren*. Insbesondere bezeichnet V_0 den Kern und V_1 den *Fixpunktraum* der Matrix A.

Im Normalfall wird V_λ aus dem Nullvektor allein bestehen. Seine Dimension, d.h. die geometrische Vielfachheit von λ ist dann 0. Wir interessieren uns für die Sonderfälle d.h. für diejenigen λ, für die der Raum V_λ nichttriviale Elemente enthält und daher seine Dimension mindestens 1 ist.

Definition. Ein Skalar λ heisst *Eigenwert* der Matrix $A \in \mathbb{R}^{n.n}$, falls ein von Null verschiedener Vektor $\vec{x} \in \mathbb{R}^n$ existiert, für den die Eigenwertgleichung

$$A \cdot \vec{x} = \lambda \vec{x}, \qquad (\vec{x} \neq \vec{0})$$

gilt. Die Menge aller Eigenwerte heisst *Spektrum* σ_A der Matrix A.

Für einen Eigenwert λ ist die geometrische Vielfachheit grösser gleich 1.

Um ein effizientes Verfahren zur Berechnung des Minimalpolynoms einer Matrix A zu demonstrieren, führen wir es an geeigneten Beispielen vor.

Beispiel. Zur Berechnung des Eigensystems der Matrix

$$A = \begin{pmatrix} 2 & -3 & 1 \\ 3 & 1 & 3 \\ -5 & 2 & -4 \end{pmatrix}$$

gehen wir von der Koeffizientenmatrix der zugehörigen homogenen Eigenwertgleichung $(A - \lambda E_3) \cdot \vec{x} = \vec{0}$ aus. Für die charakteristische Matrix erhalten wir hier

$$A - \lambda E_3 = \begin{pmatrix} 2-\lambda & -3 & 1 \\ 3 & 1-\lambda & 3 \\ -5 & 2 & -4-\lambda \end{pmatrix}$$

Man beachte, dass ihre Elemente keine Zahlen, sondern Polynome aus dem Polynomring $\mathbb{Q}[\lambda]$ sind, mit denen wir nun rechnen müssen. Fassen wir λ als Parameter auf, müssen wir beim Rechnen daran denken, dass die interessanten Werte für λ, d.h. die gesuchten Eigenwerte, gerade Nullstellen gewisser dieser Polynome sein könnten. Wir müssen also beim Durchführen der Elementaroperationen aufpassen, dass wir nur solche benutzen, die für alle λ umkehrbar sind. Kritisch in dieser Hinsicht sind die Operationen vom Typ III und die damit zusammengebauten Operationen vom Typ IV.

Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die erste und die dritte Zeile und erhalten die Matrix

$$\begin{pmatrix} -5 & 2 & -4-\lambda \\ 3 & 1-\lambda & 3 \\ 2-\lambda & -3 & 1 \end{pmatrix} \qquad \begin{bmatrix} T_{13} \\ \\ T_{31} \end{bmatrix}$$

Dadurch ist das erste führende Elemente eine feste Zahl -5 geworden. Addition des 3-fachen der ersten Zeile zum 5-fachen der zweiten ist unbedenklich. Aber auch Addition des $(2-\lambda)$-fachen der ersten Zeile zum 5-fachen der dritten ist harmlos und wir erhalten schliesslich die Matrix

$$\begin{pmatrix} -5 & 2 & -4-\lambda \\ 0 & 11-5\lambda & 3-3\lambda \\ 0 & -11-2\lambda & -3+2\lambda+\lambda^2 \end{pmatrix} \qquad \begin{bmatrix} \\ Z_{12}(3,5) \\ Z_{13}(2-\lambda,5) \end{bmatrix}$$

Hier ist das zweite führende Element $11-5\lambda$ vom Parameter λ abhängig. Aber auch das Element darunter hängt von λ ab, so dass diesmal Vertauschen der beiden Zeilen nicht hilft. Um nach Möglichkeit an dieser Stelle zu einer festen Zahl zu kommen, schaffen wir das λ weg, indem wir das 2-fache der zweiten Zeile zum (-5)-fachen der dritten addieren. Diese Operation ist unbedenklich und liefert die Matrix

$$\left(\begin{array}{ccc} -5 & 2 & -4-\lambda \\ 0 & 11-5\lambda & 3-3\lambda \\ 0 & 77 & 21-16\lambda-5\lambda^2 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(2,-5) \end{array}\right]$$

In ihr ist die gewünschte feste Zahl 77 nun in der dritten Zeile entstanden und deshalb vertauschen wir nun die zweite und dritte Zeile.

$$\left(\begin{array}{ccc} -5 & 2 & -4-\lambda \\ 0 & 77 & 21-16\lambda-5\lambda^2 \\ 0 & 11-5\lambda & 3-3\lambda \end{array}\right) \qquad \left[\begin{array}{c} \\ T_{23} \\ T_{32} \end{array}\right]$$

In ihr ist auch das zweite führende Elemente 77 nun eine feste Zahl. Addition des $(5\lambda-11)$-fachen der zweiten Zeile zum 77-fachen der dritten Zeile ist unbedenklich und liefert die Stufenform.

$$\left(\begin{array}{ccc} -5 & 2 & -4-\lambda \\ 0 & 77 & 21-16\lambda-5\lambda^2 \\ 0 & 0 & 50\lambda-25\lambda^2-25\lambda^3 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ Z_{23}(5\lambda-11,77) \end{array}\right]$$

Deren dritte Zeile wird nun noch durch 25 dividiert haben, um betragsmässig möglichst kleine Koeffizienten zu erhalten.

$$S(\lambda)=\left(\begin{array}{ccc} -5 & 2 & -4-\lambda \\ 0 & 77 & 21-16\lambda-5\lambda^2 \\ 0 & 0 & 2\lambda-\lambda^2-\lambda^3 \end{array}\right) \qquad \left[\begin{array}{c} \\ \\ S_3(\frac{1}{25}) \end{array}\right]$$

Die entstanden Matrix $S(\lambda)$ hat einerseits Stufenform und ist aus der charakteristischen Matrix $A-\lambda E$ durch lauter umkehrbare Operationen hervorgegengen. Deshalb können wir statt der ursprünglichen homogenen Eigenwertgleichung das äquivalente System

$$S(\lambda)\cdot\vec{x}=\lambda\vec{x}, \qquad (\vec{x}\neq\vec{0})$$

lösen. Dieses System kann nur dann nichttriviale Lösungen haben, wenn eines der Diagonalelemente verschwindet. In diesem Beispiel ist das nur dann der Fall, wenn die letzte Zeile verschwindet, d.h. es muss die *charakteristische Gleichung*

$$2\lambda-\lambda^2-\lambda^3=0$$

gelten. Nach so viel fehleranfälliger Rechnerei ist man froh, das Ergebnis mit Hilfe des Satzes von Cayley-Hamilton kontrollieren zu können! Tatsächlich erfüllt die Matrix A ihre charakteristische Gleichung

$$2A-A^2-A^3=0$$

Weil die Matrix A den Eigenwert $\lambda=0$ hat, ist sie nicht invertierbar und wir dürfen die charakteristische Gleichung nicht einfach durch λ dividieren und

annehmen, dass das Minimalpolynom einen echt kleineren Grad hat. Um das Minimalpolynom von A tatsächlich zu finden, machen wir den Ansatz $\mu_A(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 - \lambda^3$ eines beliebigen normierten kubischen Polynoms. Einsetzen der Matrix A liefert die Linearkombination

$$\begin{aligned} & a_0E + a_1A + a_2A^2 \\ = \; & a_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 2 & -3 & 1 \\ 3 & 1 & 3 \\ -5 & 2 & -4 \end{pmatrix} + a_2 \begin{pmatrix} -10 & -7 & -11 \\ -6 & -2 & -6 \\ 16 & 9 & 17 \end{pmatrix} \\ = \; & \begin{pmatrix} a_0 + 2a_1 - 10a_2 & -3a_1 - 7a_2 & a_1 - 11a_2 \\ 3a_1 - 6a_2 & a_0 + a_1 - 2a_2 & 3a_1 - 6a_2 \\ -5a_1 + 16a_2 & 2a_1 + 9a_2 & a_0 - 4a_1 + 17a_2 \end{pmatrix} \\ = \; & \begin{pmatrix} 14 & 1 & 13 \\ 12 & 4 & 12 \\ -26 & -5 & -25 \end{pmatrix} = A^3 \end{aligned}$$

Der Koeffizientenvergleich liefert ein lineares Gleichungssystem, das die eindeutige Lösung $a_0 = 0, a_1 = 2, a_2 = -1$ hat, was unsere Vermutung

$$\mu_A(\lambda) = 2\lambda - \lambda^2 - \lambda^3$$

für das Minimalpolynom, das also hier mit dem charakteristischen Polynom übereinstimmt, bestätigt.

Um die gesuchten Eigenwerte zu bestimmen, benötigen wir die numerischen Lösungen der charakteristischen Gleichung. Dazu faktorisieren[3] wir das Minimalpolynom von A

$$\mu_A(\lambda) = 2\lambda - \lambda^2 - \lambda^3 = -\lambda(2 + \lambda)(\lambda - 1)$$

Für das Spektrum erhalten wir $\sigma_A = \{-2, 1, 0\}$.

Um nun die zugehörigen Eigenvektoren von A zu finden, setzen wir die jeweiligen Eigenwerte in die Matrix $S(\lambda)$ ein und lösen das entstehende lineare Gleichungssystem $S(\lambda) \cdot \vec{x} = \vec{0}$ durch Elimination.

1. Fall: $\lambda = -2$. Wir erhalten die numerische Matrix

$$S(-2) = \begin{pmatrix} -5 & 2 & -2 \\ 0 & 77 & 33 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -5 & 2 & -2 \\ 0 & 7 & 3 \\ 0 & 0 & 0 \end{pmatrix} \qquad \begin{bmatrix} \; S_2(\frac{1}{11}) \; \end{bmatrix}$$

deren zweite Zeile wir durch 11 dividiert haben. Addition des (-2)-fachen der zweiten Zeile zum 7-fachen der ersten liefert die reduzierte Stufenform

$$\begin{pmatrix} -35 & 0 & -20 \\ 0 & 7 & 3 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} \; ZS_{21}(-2, 7) \; \end{bmatrix}, \qquad \begin{pmatrix} -7 & 0 & -4 \\ 0 & 7 & 3 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} \; S_1(\frac{1}{5}) \; \end{bmatrix}$$

Ihre erste Zeile haben wir noch durch denn grössten gemeinsamen Teiler 5 dividiert. Ihr Lösungsraum ist der Eigenraum

$$V_{-2} = \left\{ t \begin{pmatrix} 4 \\ 3 \\ -7 \end{pmatrix} | t \in \mathbb{R} \right\}$$

[3] Was für Polynome in $\mathbb{Q}[x]$ mit einem Algorithmus von Kronecker immer möglich ist, aber für die Handrechnung für Polynome von grösserem Grad schnell unpraktikabel wird.

Er wird vom Eigenvektor

$$\vec{v}_1 = \begin{pmatrix} 4 \\ 3 \\ -7 \end{pmatrix}$$

der zum Eigenwert $\lambda = -2$ gehört, aufgespannt, wie man leicht überprüft.

2. Fall: $\lambda = 0$. Wir erhalten die numerische Matrix

$$S(0) = \begin{pmatrix} -5 & 2 & -4 \\ 0 & 77 & 21 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -5 & 2 & -4 \\ 0 & 11 & 3 \\ 0 & 0 & 0 \end{pmatrix} \qquad \begin{bmatrix} S_2(\frac{1}{7}) \\ \\ \end{bmatrix}$$

deren zweite Zeile wir durch 7 dividiert haben. Addition des (-2)-fachen der zweiten Zeile zum 11-fachen der ersten liefert die reduzierte Stufenform

$$\begin{pmatrix} -55 & 0 & -50 \\ 0 & 7 & 3 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} ZS_{21}(-2,11) \\ \\ \end{bmatrix}, \quad \begin{pmatrix} -11 & 0 & -10 \\ 0 & 7 & 3 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} S_1(\frac{1}{5}) \\ \\ \end{bmatrix}$$

Ihre erste Zeile haben wir noch durch denn grössten gemeinsamen Teiler 5 dividiert. Ihr Lösungsraum ist der Kern

$$V_0 = \left\{ t \begin{pmatrix} 10 \\ 3 \\ -11 \end{pmatrix} | t \in \mathbb{R} \right\}$$

Er wird vom Eigenvektor

$$\vec{v}_2 = \begin{pmatrix} 10 \\ 3 \\ -11 \end{pmatrix}$$

der zum Eigenwert $\lambda = 0$ gehört, aufgespannt, wie man leicht überprüft.

3. Fall: $\lambda = 1$. Wir erhalten die numerische Matrix

$$S(-2) = \begin{pmatrix} -5 & 2 & -5 \\ 0 & 77 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -5 & 2 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \begin{bmatrix} S_2(\frac{1}{77}) \\ \\ \end{bmatrix}$$

deren zweite Zeile wir durch 77 dividiert haben. Addition des (-2)-fachen der zweiten Zeile zur ersten liefert die reduzierte Stufenform

$$\begin{pmatrix} -5 & 0 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} Z_{21}(-2) \\ \\ \end{bmatrix}, \quad \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} S_1(-\frac{1}{5}) \\ \\ \end{bmatrix}$$

Ihre erste Zeile haben wir noch durch denn grössten gemeinsamen Teiler -5 dividiert. Ihr Lösungsraum ist der Fixpunktraum

$$V_1 = \left\{ t \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} | t \in \mathbb{R} \right\}$$

Er wird vom Eigenvektor

$$\vec{v}_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

der zum Eigenwert $\lambda = 1$ gehört, aufgespannt, wie man leicht überprüft.

Zur Berechnung der Diagonalisierung und einer expliziten Formel für die Potenzen A^k benutzen wir nun die drei gefundenen Eigenvektoren.

$$\vec{v}_1 = \begin{pmatrix} 4 \\ 3 \\ -7 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 10 \\ 3 \\ -11 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

Man beachte, dass wir für Einmal die Reihenfolge der gefundenen Eigenvektoren umgestellt haben, um den Leser zu überzeugen, dass die Reihenfolge keine wesentliche Rolle spielt, falls man sie konsequent beibehält. Mit Hilfe der drei gefundenen Eigenvektoren in der gewählten Reihenfolge bilden wir nun die invertierbare Transformationsmatrix

$$X = \begin{pmatrix} 4 & -1 & 10 \\ 3 & 0 & 3 \\ -7 & 1 & -11 \end{pmatrix}, \qquad X^{-1} = \frac{1}{6} \begin{pmatrix} -3 & -1 & -3 \\ 12 & 26 & 18 \\ 3 & 3 & 3 \end{pmatrix}$$

Ihre Spalten sind die gefundenen Eigenvektoren in der gewählten — unorthodoxen — Reihenfolge. Durch Multiplikation erhält man die *Diagonalisierung*

$$\Lambda = X^{-1} \cdot A \cdot X = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Auf der Diagonalen der entstanden Diagonalmatrix stehen die Eigenwerte von A, allerdings ebenfalls in der unorthodoxen Reihenfolge. Wir können also mit dieser Diagonalisierung weiterrechnen und erhalten für ihre Potenzen

$$\Lambda^k = X^{-1} \cdot A \cdot X = \begin{pmatrix} (-2)^k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0^k \end{pmatrix}$$

Offenbar müssen wir also in Zukunft $0^0 = 1$ setzen und daher

$$0^k = \begin{cases} 1 & k = 0 \\ 0 & k \geq 1 \end{cases}$$

definieren, damit $\Lambda^0 = E$ gilt.

Aus der gefundenen Faktorisierung $A = X \cdot \Lambda \cdot X^{-1}$ ergibt sich schliesslich mit der Gleichung $A^k = X \cdot \Lambda^k \cdot X^{-1}$ die Potenz

$$A^k = \begin{pmatrix} 4 & -1 & 10 \\ 3 & 0 & 3 \\ -7 & 1 & -11 \end{pmatrix} \cdot \begin{pmatrix} (-2)^k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0^k \end{pmatrix} \cdot \frac{1}{6} \begin{pmatrix} -3 & -1 & -3 \\ 12 & 26 & 18 \\ 3 & 3 & 3 \end{pmatrix} =$$

$$\frac{1}{6} \begin{pmatrix} -12 + 30\,0^k - 12\,(-2)^k & -26 + 30\,0^k - 4\,(-2)^k & -18 + 30\,0^k - 12\,(-2)^k \\ 90^k - 9\,(-2)^k & 90^k - 3\,(-2)^k & 90^k - 9\,(-2)^k \\ 12 - 33\,0^k + 21\,(-2)^k & 26 - 33\,0^k + 7\,(-2)^k & 18 - 33\,0^k + 21\,(-2)^k \end{pmatrix}$$

wie der Leser durch Einsetzen in die rekursive Definition der Matrizenpotenz überprüfen möge. ◯

Die nächsten Beispiele dienen zum Üben und Vertiefen und zeigen mögliche Komplikationen.

Beispiel. Wir berechnen die Eigenwerte und die zugehörigen Eigenräume der symmetrischen Matrix

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

Wir müssen durch geschickte Wahl des Parameters λ dafür sorgen, dass das homogene System $(A - \lambda E) \cdot \vec{x} = \vec{0}$ mit der Koeffizientenmatrix

$$A - \lambda E = \begin{pmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{pmatrix}$$

nichttriviale Lösungen hat. Dazu lösen wir das Gleichungssystem mit dem Eliminationsverfahren. Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die ersten beiden Zeilen und erhalten

$$\begin{pmatrix} -1 & 2-\lambda & -1 \\ 1-\lambda & -1 & 0 \\ 0 & -1 & 1-\lambda \end{pmatrix} \qquad \begin{bmatrix} T_{12} \\ T_{21} \\ \end{bmatrix}$$

Addition des $(1-\lambda)$-Vielfachen der ersten Zeile zur zweiten Zeile ist unbedenklich und ergibt die Matrix

$$\begin{pmatrix} -1 & 2-\lambda & -1 \\ 0 & \lambda^2-3\lambda+1 & \lambda-1 \\ 0 & -1 & 1-\lambda \end{pmatrix} \qquad \begin{bmatrix} Z_{12}(1-\lambda) \\ \end{bmatrix}$$

Um nicht unnötige Fälle untersuchen zu müssen, vertauschen wir nun die zweite und die dritte Zeile und erhalten die Matrix

$$\begin{pmatrix} -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \\ 0 & \lambda^2-3\lambda+1 & \lambda-1 \end{pmatrix} \qquad \begin{bmatrix} \\ T_{23} \\ T_{32} \end{bmatrix}$$

Man überzeuge sich, dass Addition des $(\lambda^2-3\lambda+1)$-fachen der zweiten Zeile zur dritten Zeile ebenfalls unbedenklich ist. Wir erhalten eine äquivalente Matrix, die Stufenform hat.

$$S(\lambda) = \begin{pmatrix} -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \\ 0 & 0 & -\lambda^3+4\lambda^2-3\lambda \end{pmatrix} \qquad \begin{bmatrix} \\ \\ Z_{23}(\lambda^2-3\lambda+1) \end{bmatrix}$$

Damit nichttriviale Lösungen existieren, muss eines der Diagonalelemente verschwinden d.h. es muss die *charakteristische Gleichung* $-\lambda^3 + 4\lambda^2 - 3\lambda = 0$ erfüllt sein. Einsetzen der Matrix A in die charakteristische Gleichung liefert

$$-A^3 + 4A^2 - 3A = 0$$

die Nullmatrix, wie es der Satz von Cayley-Hamilton verlangt. Auch diesmal dürfen wir das gefundene charakteristische Polynom nicht einfach durch λ dividieren, weil die Matrix A nicht invertierbar ist. Um das Minimalpolynom von

A tatsächlich zu finden, machen wir den Ansatz $\mu_A(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 - \lambda^3$ eines beliebigen normierten kubischen Polynoms. Einsetzen der Matrix A liefert die Linearkombination

$$\begin{aligned} & a_0E + a_1A + a_2A^2 \\ = \; & a_0\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + a_1\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} + a_2\begin{pmatrix} 2 & -3 & 1 \\ -3 & 6 & -3 \\ 1 & -3 & 2 \end{pmatrix} \\ = \; & \begin{pmatrix} a_0 + a_1 + 2a_2 & -a_1 - 3a_2 & a_2 \\ -a_1 - 3a_2 & a_0 + 2a_1 + 6a_2 & -a_1 - 3a_2 \\ a_2 & -a_1 - 3a_2 & a_0 + a_1 + 2a_2 \end{pmatrix} \\ = \; & \begin{pmatrix} 5 & -9 & 4 \\ -9 & 18 & -9 \\ 4 & -9 & 5 \end{pmatrix} = A^3 \end{aligned}$$

Der Koeffizientenvergleich liefert ein lineares Gleichungssystem, das die eindeutige Lösung $a_0 = 0, a_1 = -3, a_2 = 4$ hat, was unsere Vermutung

$$\mu_A(\lambda) = 2\lambda - \lambda^2 - \lambda^3$$

für das Minimalpolynom, das also hier mit dem charakteristischen Polynom übereinstimmt, bestätigt.

Die numerischen Lösungen der charakteristischen Gleichung sind die gesuchten Eigenwerte. Um sie zu finden, faktorisieren wir das *Minimalpolynom*

$$\mu_A(\lambda) = -\lambda^3 + 4\lambda^2 - 3\lambda = -\lambda(\lambda - 1)(\lambda - 3)$$

Daher muss λ einen der Werte aus dem Spektrum $\sigma_A = \{0, 1, 3\}$ annehmen.

Um nun die zugehörigen Eigenvektoren zu finden, setzen wir die jeweiligen Eigenwerte in die Matrix $S(\lambda)$ ein und lösen das entstehende lineare Gleichungssystem mt dem Eliminationsverfahren fertig auf.

1. Fall: $\lambda = 0$. Die obere Dreiecksmatrix lautet

$$S(0) = \begin{pmatrix} -1 & 2 & -1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Kern, d.h. den Eigenraum zum Eigenwert 0 der Matrix A.

$$V_0 = \left\{ t\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

2. Fall: $\lambda = 1$. Die obere Dreiecksmatrix lautet

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

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert 1. Seine Lösung lautet in vektorieller Form:

$$V_1 = \left\{ t \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

3. Fall: $\lambda = 3$. Die obere Dreiecksmatrix lautet

$$S(3) = \begin{pmatrix} -1 & -1 & -1 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert 3.

$$V_3 = \left\{ t \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

Der Leser möge überprüfen, dass wir tatsächlich Eigenvektoren zu den jeweiligen Eigenwerten gefunden haben. Dabei fällt ihm vielleicht auf, dass in diesem Beispiel die drei gefundenen Eigenvektoren paarweise zueinander orthogonal sind. Das ist eine Konsequenz der Symmetrie der Matrix A. ◯

Um eine gewisse Ahnung zu haben, warum der Satz von Cayley-Hamilton gilt, nehmen wir an, die Matrix A habe die 3 Eigenwerte λ_1, λ_2 und λ_3 mit den zugehörigen linear unabhängigen Eigenvektoren $\vec{v}_1, \vec{v}_2, \vec{v}_3$. Nach Voraussetzung gilt also

$$A \cdot \vec{v}_k = \lambda_k \vec{v}_k, \qquad (1 \leq k \leq 3)$$

Das Minimalpolynom hat dann die Faktorisierung

$$\mu_A(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)$$

und daher gilt auch

$$\mu_A(A) = (A - \lambda_1 E) \cdot (A - \lambda_2 E) \cdot (A - \lambda_3 E)$$

wobei dieses Produkt von der Reihenfolge der Faktoren unabhängig ist. Wir müssen überprüfen, dass für alle Vektoren $\vec{x}$ die Gleichung $\mu_A(A) \cdot \vec{x} = \vec{0}$ gilt. Weil die Eigenvektoren nach Voraussetzung linear unabhängig sind, lässt sich der Vektor $\vec{x}$ als Linearkombination

$$\vec{x} = x_1 \vec{v}_1 + x_2 \vec{v}_2 + x_3 \vec{v}_3$$

darstellen. Daher gilt wegen der Linearität von μ_A

$$\mu_A(A) \cdot \vec{x} = x_1 \mu_A(A) \cdot \vec{v}_1 + x_2 \mu_A(A) \cdot \vec{v}_2 + x_3 \mu_A(A) \cdot \vec{v}_3$$

Nun ist aber beispielsweise auf Grund der Definition von μ_A

$$\begin{aligned} \mu_A(A) \cdot \vec{v}_1 &= (A - \lambda_2 E) \cdot (A - \lambda_3 E) \cdot (A - \lambda_1 E) \cdot \vec{v}_1 \\ &= (A - \lambda_2 E) \cdot (A - \lambda_3 E) \cdot (A\vec{v}_1 - \lambda_1 \vec{v}_1) \\ &= (A - \lambda_2 E) \cdot (A - \lambda_3 E) \cdot (\lambda_1 \vec{v}_1 - \lambda_1 \vec{v}_1) = \vec{0} \end{aligned}$$

Analog zeigt man die Beziehungen $\mu_A(A) \cdot \vec{v}_2 = \vec{0}$ und $\mu_A(A) \cdot \vec{v}_3 = \vec{0}$. Daraus folgt dann wegen der Linearität der Matrizenmultiplikation die Aussage $\mu_A(A) \cdot \vec{x} = \vec{0}$ des Satzes von Cayley-Hamilton. Dieses Argument hat einige Lücken und Tücken, die man allerdings durch einen verfeinerten Beweis schliessen und überbrücken kann. Auf die Finessen wollen wir nicht eingehen und behaupten einfach, wie das seinerzeit auch Cayley gemacht hat, ohne einen wasserdichten Beweis, dass die Aussage des Satzes von Cayley-Hamilton für beliebige quadratische Matrizen gilt.

In den folgenden Beispielen zur Berechnung des Eigensystems wird auf weitere Komplikationen aufmerksam gemacht.

Beispiel. Zur Berechnung des Eigensystems der Matrix

$$A = \begin{pmatrix} 0 & 1 & -2 \\ 0 & -1 & 2 \\ -1 & -1 & 1 \end{pmatrix}$$

muss untersucht werden, wann das zugehörige homogene Gleichungssystem $(A - \lambda E_3) \cdot \vec{x} = \vec{0}$ mit der charakteristischen Koeffizientenmatrix

$$A - \lambda E = \begin{pmatrix} -\lambda & 1 & -2 \\ 0 & -1-\lambda & 2 \\ -1 & -1 & 1-\lambda \end{pmatrix}$$

nichttriviale Lösungen hat. Um dies zu entscheiden, lösen wir das betreffende Gleichungssystem mit Hilfe des Eliminantionsverfahren. Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die erste mit der dritten Zeile und erhalten

$$\begin{pmatrix} -1 & -1 & 1-\lambda \\ 0 & -1-\lambda & 2 \\ -\lambda & 1 & -2 \end{pmatrix} \qquad \begin{bmatrix} T_{13} \\ \\ T_{31} \end{bmatrix}$$

Addition des $(-\lambda)$-Vielfachen der ersten Zeile zur dritten Zeile ist unbedenklich und liefert die Matrix

$$\begin{pmatrix} -1 & -1 & 1-\lambda \\ 0 & -(1+\lambda) & 2 \\ 0 & 1+\lambda & \lambda^2-\lambda-2 \end{pmatrix}$$

Addition der zweiten zur dritten Zeile liefert die Stufenform

$$S(\lambda) = \begin{pmatrix} -1 & -1 & 1-\lambda \\ 0 & -(1+\lambda) & 2 \\ 0 & 0 & \lambda^2-\lambda \end{pmatrix} \qquad \begin{bmatrix} \\ \\ Z_{23}(1) \end{bmatrix}$$

Damit nichttriviale Lösungen dieses homogenen Gleichungssystems existieren, darf seine Koeffizientenmatrix nicht invertierbar sein. Das ist genau dann der Fall, wenn mindestens einer der führenden Koeffizienten verschwindet, d.h. es muss $-(1+\lambda) = 0$ oder $\lambda^2 - \lambda = 0$ sein. Weil diese beiden Polynome teilerfremd sind, fassen wir sie zusammen und verlangen, dass λ eine Nullstelle des Minimalpolynoms

$$\mu_A(\lambda) = -(1+\lambda)(\lambda^2-\lambda) = -\lambda^3 + \lambda$$

sein muss. Auch an diesem Beispiel lässt sich durch Nachrechnen bestätigen, dass die Matrix A Nullstelle ihres charakteristischen Polynoms ist. Es ist nämlich tatsächlich

$$\mu_A(A) = -A^3 + A = 0$$

die Nullmatrix.

Um zu bestätigen, dass wir tatsächlich das Minimalpolnom von A gefunden haben, machen wir den Ansatz $\mu_A(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 - \lambda^3$ eines beliebigen normierten kubischen Polynoms. Einsetzen der Matrix A liefert die Linearkombination

$$\begin{aligned}
& a_0E + a_1A + a_2A^2 \\
&= a_0\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + a_1\begin{pmatrix} 0 & 1 & -2 \\ 0 & -1 & 2 \\ -1 & -1 & 1 \end{pmatrix} + a_2\begin{pmatrix} 2 & 1 & 0 \\ -2 & -1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} a_0 + 2a_2 & a_1 + a_2 & -2a_1 \\ -2a_2 & a_0 - a_1 - a_2 & 2a_1 \\ -a1 - a2 & -a_1 - a_2 & a_0 + a_1 + 2a_2 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 & -2 \\ 0 & -1 & 2 \\ -1 & -1 & 1 \end{pmatrix} = A^3
\end{aligned}$$

Der Koeffizientenvergleich liefert ein lineares Gleichungssystem, das die eindeutige Lösung $a_0 = 0, a_1 = 1, a_2 = 0$ hat, was unsere Vermutung

$$\mu_A(\lambda) = \lambda - \lambda^3$$

für das Minimalpolynom, das also hier mit dem charakteristischen Polynom übereinstimmt, bestätigt.

Man beachte, dass es sich beim Minimalpolynom $\mu_A(\lambda)$ in diesem Beispiel um die Determinante der charakteristischen Matrix $A - \lambda E$ handelt. In den meisten Lehrbüchern wird dieses Invertierbarkeitskriterium zur Definition der Eigenwerte benutzt und das *charakteristische Polynom* durch

$$\chi_A(\lambda) = \det(A - \lambda E)$$

definiert. Weil die effiziente Berechnung der Determinante aber über die Stufenform läuft und die Determinante einer oberen Dreiecksmatrix das Produkt der Diagonalelemente ist, ist dadurch nichts gewonnen. Tatsächlich interessiert man sich gar nicht für das charakteristische Polynom, sondern für das Minimalpolynom $\mu_A(\lambda)$, dessen Grad im allgemeinen echt kleiner ist, als jener des charakteristischen Polynoms und das trotzdem die Matrix A als Lösung und die gewünschten Eigenwerte als Nullstellen hat. Das Minimalpolynom ist in dieser Hinsicht optimal und der Umweg über das charakteristische Polynom naiv. Ein weiterer rechnerischer Vorteil des benutzten Verfahrens gegenüber der Verwendung der Determinante ist der Umstand, dass zur Berechnung der Eigenvektoren die soeben erhaltene Matrix $S(\lambda)$ verwendet werden kann und die Vorwärtsphase des Eliminationsalgorithmus nicht für jeden Eigenvektor erneut durchgeführt werden muss. Ferner hat obige Rechnung bereits eine teilweise Faktorisierung des Minimalpolynoms ergeben.

Um die Lösungen der charakteristischen Gleichung $\mu_A(\lambda) = 0$ zu finden, faktorisieren wir das Minimalpolynom nun vollständig und erhalten

$$\mu_A(\lambda) = -\lambda^3 + \lambda = -(1+\lambda)(\lambda^2 - \lambda) = -(1+\lambda)\lambda(\lambda - 1)$$

Offenbar muss λ einen der Werte aus dem Spektrum $\sigma_A = \{0, 1, -1\}$ von A annehmen.

Um schliesslich die zugehörigen Eigenvektoren zu finden, setzen wir die jeweiligen Eigenwerte in die Matrix $S(\lambda)$ ein und lösen das entstehende lineare Gleichungssystem mit Hilfe der Rückwärtsphase der Elimination.

1. Fall: $\lambda = 0$. Die obere Dreiecksmatrix liefert

$$S(0) = \begin{pmatrix} -1 & -1 & 1 \\ 0 & -1 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -1 & 0 & -1 \\ 0 & -1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} Z_{21}(-1) \\ \\ \end{bmatrix}$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Kern, d.h. den Eigenraum zum Eigenwert 0 der Matrix A

$$V_0 = \left\{ t \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

2. Fall: $\lambda = 1$. Die obere Dreiecksmatrix lautet

$$S(1) = \begin{pmatrix} -1 & -1 & 0 \\ 0 & -2 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 0 & 2 \\ 0 & -2 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} Z_{21}(1,-2) \\ \\ \end{bmatrix}$$

Schreiben wir die Lösung in vektorieller Form, erhalten wir den Fixpunktraum, d.h. den Eigenraum zum Eigenwert 1 der Matrix A

$$V_1 = \left\{ t \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

3. Fall: $\lambda = -1$. Die obere Dreiecksmatrix lautet

$$S(-1) = \begin{pmatrix} -1 & -1 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -1 & -1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{bmatrix} Z_{21}(-1) \\ \\ \end{bmatrix}$$

Schreiben wir die Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert -1:

$$V_{-1} = \left\{ t \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

Wiederum kontorolliere man das gefundene Eigensystem an Hand der ursprünglichen Matrix. ◯

Man beachte, dass bei einer $(n \times n)$-Matrix A das entstehende Minimalpolynom, dessen Nullstellen die gesuchten Eigenwerte sind, höchstens vom Grad n

sein wird, weil die charakteristische Matrix $A - \lambda E_n$ den Parameter λ nur in der Diagonalen in der ersten Potenz enthält. Wendet man auf sie das Eliminationsverfahren an, wird dieses Polynom als Summe von Produkten von solchen Linearfaktoren entstehen. Da diese Produkte aus jeder Spalte höchstens einen Faktor enthalten, kommt also λ höchstens in der n-ten Potenz vor. Da jedes nicht verschwindende Polynom n-ten Grades höchstens n Nullstellen hat, gilt:

Korollar. Eine quadratische $(n \times n)$-Matrix A hat höchstens n Eigenwerte.

Um eine gewisse geometrische Vorstellung von Eigenwerten und ihren zugehörigen Eigenvektoren zu entwickeln, definieren wir mit Hilfe Matrix A die lineare Abbildung $\vec{x} \mapsto A \cdot \vec{x}$ und interpretieren damit die neuen Begriffe.

Beispiel. Die quadratische Matrix

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

liefert die lineare Abbildung $f_A \colon \mathbb{R}^2 \to \mathbb{R}^2$, die durch die Zuordnungsvorschrift

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3x \\ x + 2y \end{pmatrix}$$

gegeben ist. Sie lässt sich an Hand folgender Figur geometrisch interpretieren. Dabei gehen wir von einigen einfachen ebenen Figuren aus und bilden sie unter der gegebenen Abbildung ab. Der Vergleich von Urbild und dem zugehörigen Bild gibt einen gemetrischen Eindruck von der zugrundeliegenden Abbildung.

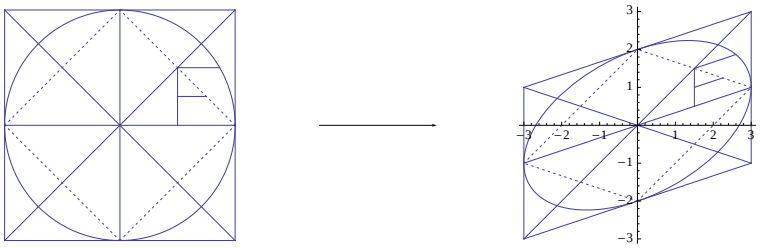


Abbildung 4.1: Erste geometrische Interpretation von Eigenvektoren.

Ein Eigenvektor dieser linearen Abbildung ist laut Definition ein Vektor $\vec{x}$ mit der Eigenschaft, dass er unter der Abbildung in ein Vielfaches übergeht. Einen solchen Vektor erkennt man in der Figur daran, dass er und sein Bild $A \cdot \vec{x}$ in die selbe Richtung zeigen. Ohne grosse Schwierigkeiten stellt man an Hand der Figur fest, dass dies für die beiden Vektoren

$$\vec{x}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{x}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

der Fall ist. Tatsächlich gilt

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

und

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Daher ist also $\vec{x}_1$ ein Eigenvektor zum Eigenwert 2 und $\vec{x}_2$ ist ein Eigenvektor zum Eigenwert 3.

Berechnet man das Eigensystem wie im letzten Beispiel vorgeführt, erhält man für unsere Matrix A das Minimalpolynom

$$\mu_A(\lambda) = \lambda^2 - 5\lambda + 6 = (3 - \lambda)(2 - \lambda)$$

Offenbar sind 2 und 3 die einzigen Eigenwerte von A. Für die zugehörigen Eigenräume ergibt sich:

$$V_2 = \{t \begin{pmatrix} 0 \\ 1 \end{pmatrix} | \in \mathbb{R}\} \qquad \text{und} \qquad V_3 = \{t \begin{pmatrix} 1 \\ 1 \end{pmatrix} | \in \mathbb{R}\}$$

Sie entsprechen den beiden Geraden, die durch die Vektoren $\vec{x}_1$ und $\vec{x}_2$ aufgespannt werden d.h. der vertikalen Achse und der ersten Winkelhalbierenden. Man beachte, dass etwa der Vektor

$$\vec{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

kein Eigenvektor unserer Abbildung ist, weil gilt:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Tatsächlich ist die horizontale Achse nicht parallel zu ihrem Bild.

Die soeben geschilderten Sachverhalte können noch mit Hilfe einer zweiten Graphik veranschaulicht werden. Dazu zeichnen wir in jedem Punkt des Einheitskreises, der durch den Vektor $\vec{x} \in S^1$ beschrieben wird, seinen Bildvektor $A \cdot \vec{x}$ ein. In unserem Fall erhalten wir dann das folgende Vektorfeld auf dem Einheitskreis:

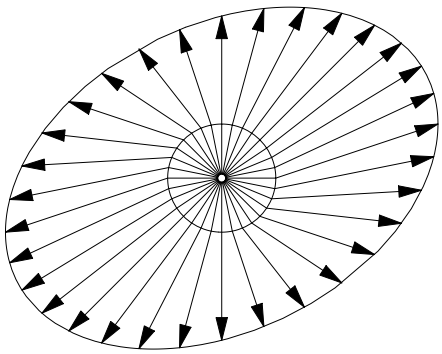


Abbildung 4.2: Zweite geometrische Interpretation von Eigenvektoren.

Die Eigenräume der Abbildung $\vec{x} \mapsto A \cdot \vec{x}$ werden nun durch diejenigen Vektoren aufgespannt, für die der Radiusvektor $\vec{x}$ und sein Bildvektor $A \cdot \vec{x}$ auf einer Geraden liegen, d.h. keinen Knick aufweisen. Aus der Figur wird klar, dass dies nur für die vertikale Achse und die erste Winkelhalbierende der Fall ist, die durch die beiden Vektoren $\vec{x}_1$ und $\vec{x}_2$ aufgespannt werden. Diese beiden Geraden gehen beide unter der linearen Abbildung in sich über und es sind die einzigen beiden Geraden mit dieser Eigenschaft. Die Eigenräume sind also *invariante Teilräume* der linearen Abbildung f_A. ◯

Weil bei einer Drehung die Drehachse und bei einer Ebenenspiegelung die Spiegelebene fix bleibt, kann man offenbar mit Hilfe des Eigenraumes V_1 insbesondere solche Fixelemente bestimmen.

4.3 Diagonalisierung

Neben diesen geometrischen Interpretationen hat das Eigensystem einer Matrix A auch wichtige algebraische Anwendungen. Falls eine quadratische Matrix A „genügend viele“ Eigenvektoren hat, lassen sich damit hohe Potenzen der Matrix leicht berechnen. Das Vorgehen soll an den bereits untersuchten Matrizen demonstriert werden.

Beispiel. Für die symmetrische Matrix

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

haben wir oben das Eigensystem bestimmt. Zunächst bilden wir mit Hilfe der verschiedenen Eigenräume nach Möglichkeit eine $n\times$n-Matrix X, deren Spalten unabhängige Basisvektoren der jeweiligen Eigenräume sind. In unserem Beispiel erhalten wir aus den drei bereits gefundenen Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

die Matrix

$$X = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix}$$

Der Eigenraum V_0 ist eindimensional und liefert die erste Spalte von X. Der Eigenraum V_1 ist ebenfalls eindimensional und liefert die zweite Spalte von X. Auch der Eigenraum von V_3 ist eindimensional und liefert die dritte Spalte von X. In unserem Beispiel haben wir eine Matrix X erhalten, die den selben Typ wie die ursprüngliche Matrix A hat. In ihren Spalten stehen lauter Eigenvektoren von A.

Diese Matrix X wird nun invertiert. Wir erhalten

$$X^{-1} = \frac{1}{6} \begin{pmatrix} 2 & 2 & 2 \\ -3 & 0 & 3 \\ 1 & -2 & 1 \end{pmatrix}$$

Das Produkt $X^{-1} \cdot A \cdot X$ liefert eine Diagonalmatrix $\Lambda(A)$, in deren Diagonale gerade die Eigenwerte von A stehen. Man beachte die korrekte Reihenfolge der drei Matrizen! In unserem Fall erhalten wir durch Ausmultiplizieren tatsächlich die Diagonalmatrix $X^{-1} \cdot A \cdot X = D(A)$.

$$\frac{1}{6}\begin{pmatrix} 2 & 2 & 2 \\ -3 & 0 & 3 \\ 1 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Wir haben die gekoppelte Information der Matrix A in die einzelnen Achsenrichtungen komplett entkoppeln können, da die Mischterme ausserhalb der Diagonalen verschwunden sind. Lösen wir die Gleichung $X^{-1} \cdot A \cdot X = \Lambda(A)$ nach der gegebenen Matrix auf, erhalten wir die Faktorisierung der gegebenen Matrix $A = X \cdot \Lambda(A) \cdot X^{-1}$, wobei $\Lambda(A)$ die Diagonalmatrix mit den Eigenwerten in der Diagonalen ist. Ihre Reihenfolge stimmt mit jener der Eigenvektoren in der Matrix X überein.

Damit ergibt sich schliesslich mit der Gleichung $A^k = X \cdot \Lambda^k \cdot X^{-1}$ die Potenz

$$\begin{aligned} A^k &= \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0^k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3^k \end{pmatrix} \cdot \frac{1}{6}\begin{pmatrix} 2 & 2 & 2 \\ -3 & 0 & 3 \\ 1 & -2 & 1 \end{pmatrix} \\ &= \frac{1}{6}\begin{pmatrix} 3+2\cdot 0^k+ & 3^k & 2\cdot 0^k-2\cdot 3^k & -3+2\cdot 0^k+ & 3^k \\ 2\cdot 0^k-2\cdot 3^k & & 2\cdot 0^k+4\cdot 3^k & 2\cdot 0^k-2\cdot 3^k & \\ -3+2\cdot 0^k+ & 3^k & 2\cdot 0^k-2\cdot 3^k & 3+2\cdot 0^k+ & 3^k \end{pmatrix} \end{aligned}$$

wie der Leser durch Einsetzen in die rekursive Definition der Matrizenpotenz überprüfen möge. ◯

Auch das andere der oben bereits behandelten Beispiele lässt sich nach dem selben Muster diagonalisieren.

Beispiel. Für die Matrix

$$A = \begin{pmatrix} 0 & 1 & -2 \\ 0 & -1 & 2 \\ -1 & -1 & 1 \end{pmatrix}$$

haben wir oben das Eigensystem bestimmt. Ihre Eigenräume zu den Eigenwerten aus dem Spektrum $\sigma_A = \{0, 1, -1\}$ werden durch die drei Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

aufgespannt. Mit diesen drei Vektoren bilden wir die Transformationsmatrix

$$X = \begin{pmatrix} -1 & -1 & -1 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Diese Matrix X wird nun invertiert und liefert die Inverse

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

Das Produkt $X^{-1} \cdot A \cdot X = \Lambda(A)$ liefert eine Diagonalmatrix in deren Diagonale gerade die Eigenwerte von A in der gewählten Reihenfolge stehen.

$$\begin{pmatrix} 1 & 1 & 0 \\ -1 & -1 & 1 \\ -1 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & -2 \\ 0 & -1 & 2 \\ -1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} -1 & -1 & -1 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

wie man leicht durch Nachrechnen überprüfen. Auf der Diagonalen der entstehenden Diagonalmatrix stehen die Eigenwerte von A.

Damit ergibt sich schliesslich mit der Gleichung $A^k = X \cdot \Lambda^k \cdot X^{-1}$ die Potenz

$$\begin{aligned} A^k &= \begin{pmatrix} -1 & -1 & -1 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0^k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & (-1)^k \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 0 \\ -1 & -1 & 1 \\ -1 & 0 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 1+(-1)^k-0^k & 1-0^k & -1+(-1)^k+0^{1+k} \\ -1-(-1)^k+2\cdot 0^k & -1+2\cdot 0^k & 1-(-1)^k+0^{1+k} \\ -1+0^k & -1+0^k & 1+0^{1+k} \end{pmatrix} \end{aligned}$$

wie der Leser durch Einsetzen in die rekursive Definition der Matrizenpotenz überprüfen möge. ◯

Die an den Beispielen gefundenen Sachverhalte lassen sich verallgemeinern. Zunächst beschreiben wir jene Matrizen, deren Information entkoppelt werden kann.

Definition. Man sagt, eine quadratische Matrix A sei *diagonalisierbar*, falls eine invertierbare Matrix X und eine Diagonalmatrix Λ existieren mit der Eigenschaft, dass die Faktorisierung

$$A = X \cdot \Lambda \cdot X^{-1}$$

gilt.

Aus dem Beweis des folgenden Satzes wird klar, dass diese Reihenfolge tatsächlich die richtige ist.

Satz. Eine quadratische Matrix $A \in \mathbb{R}^{n,n}$ ist genau dann diagonalisierbar, falls eine invertierbare Matrix X vom Typ $n \times n$ existiert, deren Spaltenvektoren die Eigenvektoren von A sind. Für eine solche Matrix X ist

$$\Lambda = X^{-1} \cdot A \cdot X$$

eine Diagonalmatrix mit den Eigenwerten von A in der Diagonalen.

Beweis. Die eine Richtung dieses Satzes haben wir bereits gezeigt und am Anfang des letzten Abschnittes zur Motivation für die Suche nach dem Eigensystem benutzt.

Wir nehmen nun umgekehrt an, die Vektoren $\vec{x}_1, \ldots, \vec{x}_n$ seien die Spaltenvektoren einer invertierbaren Matrix $X = (\vec{x}_1, \ldots, \vec{x}_n)$ und dass jeder dieser Vektoren $\vec{x}_i$ ein Eigenvektor zum Eigenwert λ_i ist. Dann gilt also nach Definition die Eigenwertgleichung

$$A \cdot \vec{x}_i = \lambda_i \vec{x}_i, \qquad (1 \leq i \leq n)$$

Nach Definition des Matrizenproduktes gilt damit

$$A \cdot X = A \cdot (\vec{x}_1, \ldots, \vec{x}_n) = (A \cdot \vec{x}_1, \ldots, A \cdot \vec{x}_n) = (\lambda_1 \vec{x}_1, \ldots, \lambda_n \vec{x}_n)$$

Die letzte Matrix ist aber gerade die Matrix $X \cdot \Lambda$, wobei $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ die Diagonalmatrix mit den Eigenwerten in der Diagonalen ist. Der Grund dafür liegt darin, dass der Effekt des Matrizenproduktes einer beliebigen Matrix X mit einer Diagonalmatrix Λ darin besteht, dass die i-te Spalte von X mit dem i-ten Diagonalelement von D multipliziert wird. Wir haben also aus der Eigenwertgleichung die Gleichung $A \cdot X = X \cdot \Lambda$ gefunden. Da wir X invertierbar vorausgesetzt haben, können wir diese Gleichung von rechts mit der Matrix X^{-1} multiplizieren und erhalten die Gleichung $A = X \cdot \Lambda \cdot X^{-1}$, die zeigt, dass die Matrix A diagonalisierbar ist und $\Lambda(A)$ die behauptete Eigenschaft hat. An Hand dieses Beweise wird auch klar, warum die gewählte Reihenfolge der drei Matrizen in der Diagonalisierung die richtige ist. Die Diagonalisierung kommt von der Gleichung $A \cdot X = X \cdot \lambda$ her, die ihrerseits den Eigenwertgleichungen für die Spalten von X entspricht. □

Unsere bisherigen Beispiele sind insofern irreführend, als es durchaus möglich ist, dass zu einem Eigenwert λ mehrere Eigenvektoren gehören können, die voneinander unabhängig sind. Solche Eigenwerte heissen *ausgeartet*. Die Dimension des zugehörigen Eigenraumes V_λ d.h. die *geometrische Vielfachheit* des entsprechenden Eigenwertes ist also grösser als 2. In die Transformationsmatrix X nehme man dann so viele verschiedene unabhängige Eigenvektoren auf, wie diese geometrische Vielfachheit angibt. Aus unseren weiteren Überlegungen wird folgen, dass die Matrix X genau dann invertierbar ist, wenn sie quadratisch ist, da ihre Spaltenvektoren unabhängig sind.

Beispiel. Wir bestimmen das Eigensystem der symmetrischen Matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Dazu gehen wir von der zugehörigen charakteristischen Matrix

$$A - \lambda E = \begin{pmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{pmatrix}$$

aus. Wir müssen dafür sorgen, dass das System der homogenen Eigenwertgleichungen eine nicht triviale Lösung hat. Dazu lösen wir das Gleichungssystem mit dem Eliminationsverfahren. Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die ersten beiden Zeilen und erhalten:

$$\begin{pmatrix} 1 & 2-\lambda & 1 \\ 2-\lambda & 1 & 1 \\ 1 & 1 & 2-\lambda \end{pmatrix} \qquad \begin{bmatrix} T_{12} \\ T_{21} \\ \end{bmatrix}$$

Addition des $(\lambda - 2)$-Vielfachen der ersten Zeile zur zweiten Zeile und Subtraktion der ersten Zeile von der dritten Zeile sind unbedenklich. Man erhält die Matrix

$$\begin{pmatrix} 1 & 2-\lambda & 1 \\ 0 & (1-\lambda)(\lambda-3) & \lambda-1 \\ 0 & \lambda-1 & 1-\lambda \end{pmatrix} \qquad \begin{bmatrix} \\ Z_{12}(\lambda-2) \\ Z_{13}(-1) \end{bmatrix}$$

Um nicht unnötige Fälle untersuchen zu müssen, vertauschen wir nun die zweite und die dritte Zeile und erhalten die Matrix

$$\begin{pmatrix} 1 & 2-\lambda & 1 \\ 0 & \lambda-1 & 1-\lambda \\ 0 & (1-\lambda)(\lambda-3) & \lambda-1 \end{pmatrix} \qquad \begin{bmatrix} \\ T_{23} \\ T_{32} \end{bmatrix}$$

Addition des $(\lambda-3)$-fachen der zweiten Zeile zur dritten Zeile ist ebenfalls unbedenklich. Wir erhalten die obere Dreiecksmatrix:

$$S(\lambda) = \begin{pmatrix} 1 & 2-\lambda & 1 \\ 0 & \lambda-1 & 1-\lambda \\ 0 & 0 & (\lambda-1)(\lambda-4) \end{pmatrix} \qquad \begin{bmatrix} \\ \\ Z_{23}(\lambda-3) \end{bmatrix}$$

Damit nichttriviale Lösungen existieren, muss die charakteristische Gleichung $(\lambda-1)^2(\lambda-4)=0$ erfüllt sein. Auch hier hat das charakteristische Polynom

$$\chi_A(\lambda) = (\lambda-1)^2(\lambda-4) = \lambda^3 - 6\lambda^2 + 9\lambda - 4$$

die Matrix A als Nullstelle, weil tatsächlich

$$\chi_A(A) = A^3 - 6A^2 + 9A - 4E_3 = 0$$

gilt, wie man leicht durch Nachrechnen überprüft.

Die Nullstellen des charakteristischen Polynoms χ_A sind die gesuchten Eigenwerte. Daher muss λ einen der Werte aus dem Spektrum $\sigma_A = \{1, 4\}$ annehmen. Man beachte, dass diese Matrix nur zwei verschiedene Eigenwerte hat. Das kommt daher, dass das charakteristische Polynom mehrfache Nullstellen hat.

Ein zweiter Blick auf die in unserem Beispiel durch das Eliminationsverfahren erhaltene Stufenform lässt vermuten, dass es diesmal sogar eine algebraische Gleichung von echt kleinerem Grad mit der Eigenschaft geben könnte, die von der Matrix A gelöst wird. Benutzen wir statt des charakteristischen nämlich das *Minimalpolynom*

$$\mu_A(\lambda) = (\lambda-1)(\lambda-4) = \lambda^2 - 5\lambda + 4$$

das ein Faktor des charakteristischen Polynoms ist, stellen wir durch Nachrechnen fest, dass die Matrix A sogar Nullstelle dieses Minimalpolynoms kleineren Grades ist, weil tatsächlich

$$\mu_A(A) = A^2 - 5A + 4E_3 = 0$$

gilt. Man beachte, dass das Minimalpolynom die selben Nullstellen wie das charakteristische Polynom hat und daher die selben Eigenwerte liefert.

Um diesen Befund zu bestätigen, dass wir tatsächlich das Minimalpolnom von A gefunden haben, machen wir den Ansatz $\mu_A(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 - \lambda^3$ eines beliebigen normierten kubischen Polynoms. Einsetzen der Matrix A liefert die Linearkombination

$$\begin{aligned} & a_0E + a_1A + a_2A^2 \\ = \; & a_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + a_1 \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} + a_2 \begin{pmatrix} 6 & 5 & 5 \\ 5 & 6 & 5 \\ 5 & 5 & 6 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} a_0+2a_1+6a_2 & a_1+5a_2 & a_1+5a_2 \\ a_1+5a_2 & a_0+2a_1+6a_2 & a_1+5a_2 \\ a_1+5a_2 & a_1+5a_2 & a_0+2a_1+6a_2 \end{pmatrix} \\
&= \begin{pmatrix} 22 & 21 & 21 \\ 21 & 22 & 21 \\ 21 & 21 & 22 \end{pmatrix} = A^3
\end{aligned}
$$

Der Koeffizientenvergleich liefert ein lineares Gleichungssystem.

$$
\begin{cases} a_0 + 2a_1 + 6a_2 &= 22 \\ a_1 + 5a_2 &= 21 \end{cases}
$$

Es hat die allgemeine Lösung $a_0 = -20 + 4t, a_1 = 21 - 5t, a_2 = t$. Weil wir für jeden Parameter t ein normiertes kubisches Polynom erhalten, das die gegebene Matrix als Nullstelle hat, wählen wir t so, dass möglichst viele aufeinanderfolgende Koeffizienten des entstehenden Polynoms verschwinden. Im vorliegenden Fall wählen wir t so, dass $a_0 = 0$ wird, was $t = 5$ und damit $a_1 = -4$ und $a_2 = 5$ zur Folge hat. Das entstehende Polynom hat also die Form $-4\lambda + 5\lambda^2 - \lambda^3$ und lässt sich diesmal durch λ dividieren, da $\lambda = 0$ kein Eigenwert der ursprünglichen Matrix war, was zum erwarteten Minimalpolynom

$$
\mu_A(\lambda) = -4 + 5\lambda - \lambda^2
$$

zweiten Grades der Matrix A führt. In diesem Fall ist das Minimalpolynom ein echter Teiler des charakteristischen Poloynoms $\chi_A(\lambda) = (\lambda - 1) \cdot \mu_A(\lambda)$, das zum Parameter $t = 6$ gehört.

Um nun die zugehörigen Eigenvektoren zu finden, setzen wir die jeweiligen Eigenwerte in die Matrix $S(\lambda)$ ein und lösen das entstehende lineare Gleichungssystem, indem wir das Eliminationsverfahren fortsetzen.

1. Fall: $\lambda = 1$. Die obere Dreiecksmatrix $S(\lambda)$ lautet dann

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

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Fixpunktraum, d.h. den Eigenraum zum Eigenwert 1 der Matrix A

$$
V_1 = \left\{ t \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + s \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \mid t, s \in \mathbb{R} \right\}
$$

Weil der Raum V_1 die Dimension 2 hat, ist der Eigenwert $\lambda = 1$ ausgeartet und hat die geometrische Vielfachheit 2.

2. Fall: $\lambda = 4$. Die obere Dreiecksmatrix $S(\lambda)$ lautet dann

$$
S(4) = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 3 & -3 \\ 0 & 0 & 0 \end{pmatrix}
$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert 4

$$V_4 = \left\{ t \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

Dieser Raum ist 1-dimensional. Der Eigenwert $\lambda = 4$ hat also die geometrische Vielfachheit 1. Da die Summe der geometrischen Vielfachheiten 3 beträgt, ist die Matrix A diagonalisierbar.

Fassen wir die gefundenen Eigenvektoren in gewohnter Weise zu einer Matrix zusammen, erhalten wir die Matrix

$$X = \begin{pmatrix} -1 & -1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Der Eigenraum V_1 ist zweidimensional und liefert die ersten beiden Spalten von X. Der Eigenraum V_4 ist eindimensional und liefert die dritte Spalte.

Diese Matrix X wird nun invertiert und wir erhalten die Matrix

$$X^{-1} = -\frac{1}{3} \begin{pmatrix} 1 & 1 & -2 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{pmatrix}$$

Das Produkt $X^{-1} \cdot A \cdot X$ liefert eine Diagonalmatrix $\Lambda(A)$, in deren Diagonale gerade die Eigenwerte von A stehen. In unserem Fall erhalten wir also

$$-\frac{1}{3} \begin{pmatrix} 1 & 1 & -2 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} -1 & -1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

wie man leicht durch Nachrechnen überprüfen kann. Auch in diesem Beispiel haben wir die gekoppelte Information der Matrix A in die einzelnen Achsenrichtungen komplett entflechten können. Lösen wir die Gleichung $X^{-1} \cdot A \cdot X = \Lambda(A)$ nach der gegebenen Matrix auf, erhalten wir auch hier die Matrizengleichung $A = X \cdot \Lambda(A) \cdot X^{-1}$, wobei $\Lambda(A)$ die Diagonalmatrix mit den Eigenwerten in der Diagonalen ist.

Zum Bestimmen der k-ten Potenz A^k der Matrix A gehen wir also in der gewohnten Art vor. Mit der Gleichung $A^k = X \cdot \Lambda^k \cdot X^{-1}$ erhalten wir auch hier die gesuchte Potenz

$$\begin{aligned} A^k &= \begin{pmatrix} -1 & -1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4^k \end{pmatrix} \cdot (-\frac{1}{3}) \begin{pmatrix} 1 & 1 & -2 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 2+4^k & 4^k-1 & 4^k-1 \\ 4^k-1 & 4^k+2 & 4^k-1 \\ 4^k-1 & 4^k-1 & 4^k+2 \end{pmatrix} \end{aligned}$$

wie der Leser durch Nachrechnen und Einsetzen in die rekursive Definition der Matrizenpotenz überprüfen möge. ◯

Es ist nicht etwa so, dass zu einer quadratischen Matrix $A \in \mathbb{R}^{n \cdot n}$ immer genau n unabhängige Eigenvektoren existieren und man damit eine invertierbare $(n \times n)$-Matrix X bilden kann in deren Spalten lauter Eigenvektoren stehen! Dies ist nur dann der Fall, wenn die Summe der geometrischen Vielfachheiten gleich n ist. Obiger Satz sagt, dass dies genau dann der Fall ist, wenn die Matrix A diagonalisierbar ist. Es gibt also nichtdiagonalisierbare Matrizen.

Beispiel. Wir gehen von der Drehung der Ebene um $\frac{\pi}{2}$ im Gegenuhrezeigersinn aus. Diese lineare Abbildung wird durch die Matrix

$$D_{\frac{\pi}{2}} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

beschrieben und betrachten das zugehörige tangentiale Vektorfeld auf dem Einheitskreis.

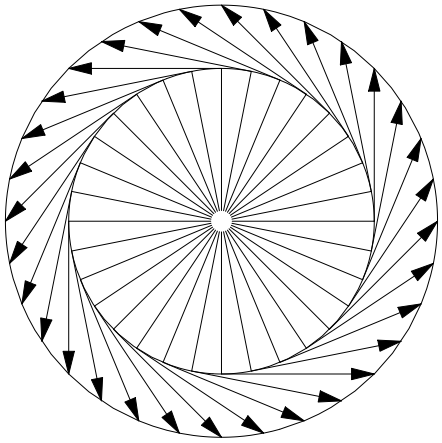


Abbildung 4.3: Vektorfeld einer Drehung in der Ebene.

Aus der Figur dieses Vektorfeldes wird klar, dass hier kein reeller Eigenvektor existieren kann, weil jeder Radiusvektor $\vec{x}$ und sein Bildvektor $A \cdot \vec{x}$ senkrecht aufeinander stehen. ◯

Bei Verwendung der komplexen Zahlen als Skalare haben Polynome in der Regel mehr Nullstellen, als wenn bloss reelle Lösungen zugelassen werden. Daher erwarten wir dann in der Regel auch mehr Eigenwerte als Nullstellen des charakteristischen Polynoms. Diese Eigenwerte liefern dann zugehörige Eigenvektoren, deren Komponenten allerdings in der Regel komplexe Zahlen sind.

Beispiel. Die im letzten Beispiel verwendete Drehmatrix

$$D_{\frac{\pi}{2}} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

hat das charakteristische Polynom $\chi(\lambda) = \lambda^2 + 1$ und damit die beiden komplexen Eigenwerte i und $-i$. Das komplexe Spektrum der Drehmatrix ist also

$\sigma = \{i, -i\}$. Die zugehörigen Eigenvektoren sind

$$\vec{v}_1 = \begin{pmatrix} i \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{v}_2 = \begin{pmatrix} -i \\ 1 \end{pmatrix}$$

Tatsächlich zeigt eine einfache Rechnung

$$\begin{pmatrix} i \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i \begin{pmatrix} i \\ 1 \end{pmatrix}$$

und

$$\begin{pmatrix} -i \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} -i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -i \end{pmatrix} = -i \begin{pmatrix} -i \\ 1 \end{pmatrix}$$

Daher ist also $\vec{v}_1$ ein Eigenvektor zum Eigenwert i und $\vec{v}_2$ ist ein Eigenvektor zum Eigenwert $-i$, die man allerdings in der reellen Geometrie nicht sieht.

Fassen wir diese Eigenvektoren zu einer Matrix zusammen, erhalten wir die komplexe Matrix

$$X = \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix}$$

Diese Transformationsmatrix X wird nun invertiert. Für die Inverse erhalten wir

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

Das Produkt $X^{-1} \cdot D_{\frac{\pi}{2}} \cdot X$ liefert auch hier eine Diagonalmatrix Λ, in deren Diagonale gerade die Eigenwerte von A stehen. Es ist nämlich

$$X^{-1} \cdot D_{\frac{\pi}{2}} \cdot X = \frac{1}{2} \begin{pmatrix} -i & 1 \\ i & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \Lambda$$

wie man leicht durch Nachrechnen überprüfen kann.

Diese komplexe Diagonalmatrix kann selbstverständlich, wie im reellen Fall, zur Berechnung der Matrizenpotenz verwendet werden. In unserem Fall ergibt sich

$$\begin{aligned} D_{\frac{\pi}{2}}^k = X \cdot \Lambda^k \cdot X^{-1} &= \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} i^k & 0 \\ 0 & (-i)^k \end{pmatrix} \cdot \frac{1}{2} \begin{pmatrix} -i & 1 \\ i & 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} i^k + (-i)^k & i\big(i^k - (-i)^k\big) \\ -i\big(i^k - (-i)^k\big) & i^k + (-i)^k \end{pmatrix} \end{aligned}$$

Diese komplexe Matrix kann mit Hilfe der Euler'schen Formeln

$$\cos(\varphi) = \frac{e^{i\varphi} + e^{-i\varphi}}{2}, \qquad \sin(\varphi) = \frac{-i(e^{i\varphi} - e^{-i\varphi})}{2}$$

in eine reelle Form umgerechnet werden. Aus den Polardarstellungen der Eigenwerte $i = e^{i\frac{\pi}{2}}$ und $-i = e^{-i\frac{\pi}{2}}$ erhalten wir

$$\begin{aligned} D_{\frac{\pi}{2}}^k &= \frac{1}{2} \begin{pmatrix} i^k + (-i)^k & i\big(i^k - (-i)^k\big) \\ -i\big(i^k - (-i)^k\big) & i^k + (-i)^k \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} e^{i\frac{k\pi}{2}} + e^{-i\frac{k\pi}{2}} & i\big(e^{i\frac{k\pi}{2}} - e^{-i\frac{k\pi}{2}}\big) \\ -i\big(e^{i\frac{k\pi}{2}} - e^{-i\frac{k\pi}{2}}\big) & e^{i\frac{k\pi}{2}} + e^{-i\frac{k\pi}{2}} \end{pmatrix} \\ &= \begin{pmatrix} \cos\left(\frac{k\pi}{2}\right) & -\sin\left(\frac{k\pi}{2}\right) \\ \sin\left(\frac{k\pi}{2}\right) & \cos\left(\frac{k\pi}{2}\right) \end{pmatrix} = D_{\frac{k\pi}{2}} \end{aligned}$$

wie man aus geometrischen Gründen sofort erkennt. Das zugehörige diskrete dynamische System durchläuft periodisch die Vektorfolge

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, D_{\frac{\pi}{2}} \cdot \vec{a} = \begin{pmatrix} -a_2 \\ a_1 \end{pmatrix}, D^2_{\frac{\pi}{2}} \cdot \vec{a} = \begin{pmatrix} -a_1 \\ -a_2 \end{pmatrix}, D^3_{\frac{\pi}{2}} \cdot \vec{a} = \begin{pmatrix} a_2 \\ -a_1 \end{pmatrix}$$

Sie bilden die Ecken eines Quadrates mit dem Symmetriezentrum im Ursprung. Im Unterschied zu den bisher bestimmten Potenzen von Matrizen, wo nur reelle Exponentialfunktionen eine Rolle gespielt haben, tauchen bei Matrizen mit komplexen Eigenwerten nun auch Kreisfunktionen auf, die dafür sorgen, dass das betreffende System nicht bloss wachsen, zerfallen oder oszillieren, sondern auch mit beliebigen Frequenzen schwingen kann. ◯

Schwingungen und damit komplexe Eigenwerte spielen selbstverständlich in vielen Anwendungen eine zentrale Rolle. Das folgende Beispiel soll ein Verfahren motivieren, das für jeden Techniker zum Standardrepertoire gehört.

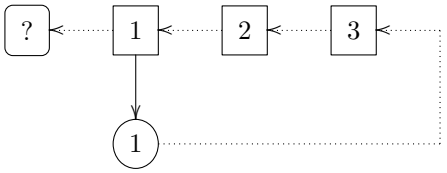
Beispiel. Wie jede Tanz-Banause weiss, wird Walzer im 3/4-Takt zur Taktfolge

k	0	1	2	3	4	5	$\cdots$
x_k	1	2	3	1	2	3	$\cdots$

getanzt. Weil sich diese Folge periodisch mit der Periode 3 wiederholt, wird sie durch die lineare Differenzengleichung

$$x_{k+3} = x_k$$

dritter Ordnung beschrieben. Sie kann durch $x_k = (k \bmod 3) + 1$ explizit beschrieben und durch das einfache Schieberegister



realisiert werden. Die Frage stellt sich, wie man diese Folge durch eine elementare Formel beschreiben kann.

Ohne zunächst den Umweg über die Vektorfolgen zu wählen, machen wir für die gesuchte Folge den *skalaren Exponentialansatz*

$$x_k = \lambda^k$$

Einsetzen in die Differenzengleichung zeigt, dass die gesuchte Konstante λ die charakteristische Gleichung

$$\lambda^3 = 1$$

erfüllen muss. Für λ kommt nur eine der drei komplexen dritten Einheitswurzeln

$$\begin{array}{lllll} \lambda_1 & = & \zeta_3^0 & = & 1 \\ \lambda_2 & = & \zeta_3^1 & = & e^{i\frac{2\pi}{3}} = \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right) = -\frac{1}{2} + i\frac{\sqrt{3}}{2} \\ \lambda_3 & = & \zeta_3^2 & = & e^{i\frac{4\pi}{3}} = \cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right) = -\frac{1}{2} - i\frac{\sqrt{3}}{2} \end{array}$$

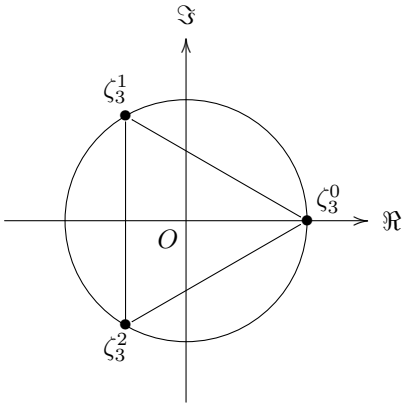


Abbildung 4.4: Lage der drei Eigenwerte in der komplexen Ebene $\mathbb{C}$.

in Frage, die in den Ecken eines gleichseitigen Dreiecks liegen, das seinerseits dem Einheitskreis einbeschrieben ist.

Man beachte, dass die drei Folgen λ_j^k die verlangte Anfangsbedingung der Walzerfolge nicht erfüllen, sondern sich auf den Ecken des gleichseitigen Dreiecks bewegen. Die erste Basislösung λ_1^k bleibt in der Ecke 1 der gleichseitigen Dreiecks fix. Die zweite Basislösung λ_2^k durchläuft die Ecken des gleichseitigen Dreiecks periodisch, indem sie in jedem Takt um den Winkel $\frac{2\pi}{3}$ im Gegenuhrzeigersinn rotiert. Auch die dritte Basislösung λ_3^k rotiert im Gegenuhrzeigersinn, besucht aber dabei immer die übernächste Ecke des gleichseitigen Dreiecks, da ihr Drehwinkel $\frac{4\pi}{3}$ beträgt. Alternativ kann man sie als im Uhrzeigersinn um den Winkel $-\frac{2\pi}{3}$ rotierend auffassen.

Aus den gefunden drei Basislösungen lassen sich nun sämtliche Lösungen der Differenzengleichungen linear kombinieren d.h. in der Form

$$x_k = c_1\,\lambda_1^k + c_2\,\lambda_2^k + c_3\,\lambda_3^k = c_1\,\zeta_3^0 + c_2\,\zeta_3^k + c_3\,\zeta_3^{2k}$$

zusammenbauen, wobei die drei komplexen Konstanten $c_1, c_2, c_3 \in \mathbb{C}$ von der Anfangsbedingung abhängen. Im Fall der Walzerzahlen müssen die drei linearen Bedingungen des linearen Gleichunssystems

$$\left\{\begin{array}{rcl} x_0 &=& c_1 + c_2 + c_3 = 1 \\ x_1 &=& c_1 + \zeta_3\, c_2 + \zeta_3^2\, c_3 = 2 \\ x_2 &=& c_1 + \zeta_3^2\, c_2 + \zeta_3\, c_3 = 3 \end{array}\right. \qquad \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & \zeta_3 & \zeta_3^2 & 2 \\ 1 & \zeta_3^2 & \zeta_3 & 3 \end{array}\right)$$

gelten. Die eindeutige Lösung dieses linearen Gleichungssystems sind die drei komplexen Zahlen

$$\begin{array}{rcl} c_1 &=& 2 \\ c_2 &=& -\frac{1}{2} + i\frac{\sqrt{3}}{6} = \frac{\sqrt{3}}{3} e^{i\frac{5\pi}{6}} \\ c_3 &=& -\frac{1}{2} - i\frac{\sqrt{3}}{6} = \frac{\sqrt{3}}{3} e^{i\frac{7\pi}{6}} \end{array}$$

wie man durch Einsetzen leicht überprüft. Sie lassen sich sehr schnell berechnen, weil das zu lösende lineare Gleichungssystem die Fouriermatrix F_3 als

Koeffizientenmatrix hat. Ihre Inverse F_3^{-1} lässt sich daher dank der Orthogonalitätsrelation sofort bestimmen.

Damit erhält man für die Walzerzahlen die gesuchte explizite Beschreibung

$$\begin{aligned} x_k &= 2 + c_2\,\zeta_3^k + c_3\,\zeta_3^{2k} = 2 + \Big(-\frac{1}{2} + \frac{\sqrt{3}}{6}i\Big)e^{ik\frac{2\pi}{3}} + \Big(\frac{1}{2} - \frac{\sqrt{3}}{6}i\Big)e^{ik\frac{4\pi}{3}} \\ &= 2 + \frac{\sqrt{3}}{3}e^{i(k\frac{2\pi}{3}+\frac{5\pi}{6})} + \frac{\sqrt{3}}{3}e^{i(k\frac{4\pi}{3}+\frac{7\pi}{6})} = 2 + \frac{\sqrt{3}}{3}\Big(e^{i(k\frac{2\pi}{3}+\frac{5\pi}{6})} + e^{i(k\frac{4\pi}{3}+\frac{7\pi}{6})}\Big) \\ &= 2 + \frac{\sqrt{3}}{3}\Big(\cos\big(k\frac{2\pi}{3} + \frac{5\pi}{6}\big) + \cos\big(k\frac{4\pi}{3} + \frac{7\pi}{6}\big)\Big) \end{aligned}$$

Die letzte Gleichheit ergibt sich aus dem Umstand, dass die Folge reellwertig sein muss! Sie lässt sich unter Verwendung des Additionstheorems auf die einfachere reelle Form

$$x_k = 2 - \cos\Big(k \cdot \frac{2\pi}{3}\Big) - \frac{\sqrt{3}}{3}\sin\Big(k \cdot \frac{2\pi}{3}\Big)$$

bringen, die man selbstverständlich auch geometrisch hätte finden können.

Mit den Polardarstellungen der beiden konjugiert komplexen Eigenwerte

$$\lambda_2 = r \cdot e^{i\varphi}, \qquad \lambda_3 = r \cdot e^{-i\varphi}$$

mit dem Betrag

$$r = |\lambda_2| = |\lambda_3| = 1$$

und dem Argument

$$\varphi = \frac{2\pi}{3}$$

hätte man für die Folge der Walzerzahlen auch direkt den reellen Ansatz

$$x_k = c_1\lambda_1^k + d_2 r^k \cos(k \cdot \varphi) + d_3 r^k \sin(k \cdot \varphi)$$

machen können. Die drei rellen Konstanten $c_1, d_2, d_3 \in \mathbb{R}$ müssen auch diesmal aus den Anfangsbedingungen bestimmt werden und erfüllen das lineare Gleichungssystem

$$\left\{\begin{array}{rcl} x_0 &=& \lambda_1^0\; c_1 + r^0\cos(0\cdot\varphi)d_2 + r^0\sin(0\cdot\varphi)d_3 = \;\; 1 \\ x_1 &=& \lambda_1\; c_1 + r^1\cos(1\cdot\varphi)d_2 + r^1\cos(1\cdot\varphi)d_3 = \;\; 2 \\ x_2 &=& \lambda_1^2\; c_1 + r^2\cos(2\cdot\varphi)d_2 + r^2\cos(2\cdot\varphi)d_3 = \;\; 3 \end{array}\right.$$

das in unserem numerischen Spezialfall die Gestalt

$$\left\{\begin{array}{rcl} x_0 &=& c_1 + \;\; d_2 \qquad\qquad = \;\; 1 \\ x_1 &=& c_1 - \frac{1}{2}d_2 + \frac{\sqrt{3}}{3}d_3 = \;\; 2 \\ x_2 &=& c_1 - \frac{1}{2}d_2 - \frac{\sqrt{3}}{3}d_3 = \;\; 3 \end{array}\right.$$

annimmt und die eindeutige Lösung

$$c_1 = 2, d_2 = -1, d_3 = -\frac{\sqrt{3}}{3}$$

hat. Damit ergibt sich für die Walzerzahlen die selbe einfache Form wie oben.

$$x_k = 2 - \cos(k \cdot \varphi) - \frac{\sqrt{3}}{3}\sin(k \cdot \varphi) = 2 - \cos\left(k\frac{2\pi}{3}\right) - \frac{\sqrt{3}}{3}\sin\left(k\frac{2\pi}{3}\right)$$

In gewohnter Manier ersetzen wir nun noch die lineare Rekursion dritter Ordnung $x_{k+3} = x_k$ der Walzerzahlen durch ein dreidimensionales lineares System erster Ordnung. Als Systemmatrix dieses System benutzen wir die Begleitermatrix des charakteristischen Polynoms

$$\chi(\lambda) = \lambda^3 - 1$$

und definieren die zyklische Matrix

$$T_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Diese Matrix hat also als charakteristische Polynom $\chi(\lambda) = \lambda^3 - 1$ mit den drei dritten komplexen Einheitswurzeln $\zeta_3^0, \zeta_3^1, \zeta_3^2$ als Nullstellen. Diese drei Punkte der komplexen Ebene bilden ein gleichseitiges Dreieck auf dem Einheitskreis, dessen eine Ecke selbstverständich 1 ist, da $1^3 = 1$ ist. Für die Normalformen der 3-ten Einheitswurzeln gilt also

k	ζ_3^k	$x_k + iy_k$	$x_k + iy_k$	$N(x_k + iy_k)$
0	$e^{i0\cdot\frac{2\pi}{3}}$	$\cos(0) + i\sin(0)$	1	1
1	$e^{i1\cdot\frac{2\pi}{3}}$	$\cos\left(1\cdot\frac{2\pi}{3}\right) + i\sin\left(1\cdot\frac{2\pi}{3}\right)$	$-\frac{1}{2} + i\frac{\sqrt{3}}{2}$	$0.5 + 0.866025i$
2	$e^{i2\cdot\frac{2\pi}{3}}$	$\cos\left(2\cdot\frac{2\pi}{3}\right) + i\sin\left(2\cdot\frac{2\pi}{3}\right)$	$-\frac{1}{2} - i\frac{\sqrt{3}}{2}$	$0.5 - 0.866025i$

Diese Koordinaten sind dann zyklisch modulo 3 zu nehmen.

Nach dem Satz von Cayley-Hamilton gilt

$$T_3^3 = E.$$

Die Matrix T_3 spielt eine wichtige Rolle, wie wir im Zusammenhang mit der Fouriertransformation gesehen haben. Zunächst beachten wir, dass sie algebraisch als Operator auf den Komponenten eines beliebigen Vektors $\vec{x} \in \mathbb{R}^3$ einen *zyklischen Shift* bewirkt, weil

$$T_3 \colon \mathbb{R}^3 \to \mathbb{R}^3, \qquad \vec{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \mapsto T_3 \cdot \vec{x} = \begin{pmatrix} b \\ c \\ a \end{pmatrix}$$

gilt. Um ihre geometrische Bedeutung zu erkennen, interpretieren wir T_3 graphentheoretisch als Adjazenzmatrix des zyklischen Graphen

Offensichtlich enthält dieser dreieckige Graph die kombinatorische Bedeutung der Walzerzahlen. Um nun T_3 geometrisch zu interpretieren, müssen wir nur daran denken, dass eine lineare Abbildung durch die Operation auf den Standardbasisvektoren vollständig bestimmt ist. Die durch T_3 beschriebene lineare Abbildung permutiert also die Standardbasisvektoren zyklisch, wie man aus Einbettung des gleichseitiges Dreieck in den Würfel erkennt.

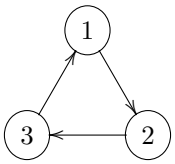


Abbildung 4.5: Zyklischer Graph mit drei Knoten.

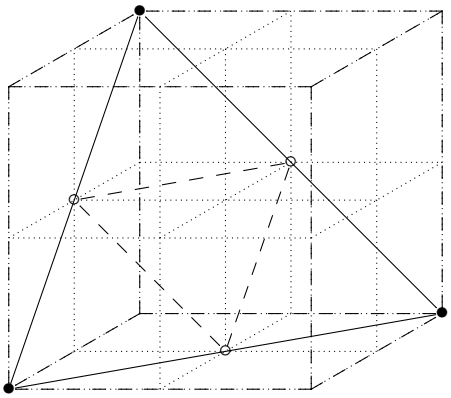


Abbildung 4.6: Das gleichseitige Dreieck im Würfel.

Daraus wird klar, dass T_3 eine Drehung um die Drehachse

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

mit dem Drehwinkel $\varphi = \frac{2\pi}{3}$ beschreibt. Tatsächlich ist T_3 als Permutationsmatrix orthogonal, weil ihre Spaltenvektoren das orthonormierte Dreibein der Standardbasisvektoren in zyklisch geshifteter Reihenfolge beschreiben. Es ist auch deutlich erkennbar, dass für die Determinante $\det(T_3) = 1$ gilt und daher T_3 die Orientierung erhält und deshalb eine Drehung sein muss. Daher ist ihre Inverse ebenfalls eine Drehung mit dem Winkel $-\varphi$ um die selbe Achse und wird durch die transponierte Matrix

$$T_3^T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

beschrieben. Sie bewirkt algebraisch einen *zyklischen Rechtsshift*

$$T_3^T \colon \mathbb{R}^3 \to \mathbb{R}^3, \qquad \vec{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \mapsto T_3^T \cdot \vec{x} = \begin{pmatrix} c \\ a \\ b \end{pmatrix},$$

während T_3 einen *zyklischen Linksshift* bewirkt.

Um eine an T_3 angepasste Basis zu finden, benötigen wir die zu den drei gefundenen Eigenwerten $\lambda_1 = \zeta_3^0, \lambda_2 = \zeta_3, \lambda_3 = \zeta_3^2$ gehörenden Eigenvektoren. Weil

die Drehachse beim Drehen fix bleibt, ist $\vec{v}_1$ ein Eigenvektor von T_3, der zum Eigenwert $\lambda_1 = \zeta_3^0 = 1$ gehört. Weil T_3 die zur Drehachse orthogonale Ebene

$$\vec{v}_1^{\perp} = \{\vec{x} \in \mathbb{R}^3 | \langle \vec{x}, \vec{v}_1 \rangle = 0\}, \qquad x + y + z = 0$$

ebenfalls fix lässt und in dieser Ebene eine Drehung um $\varphi = 120°$ bewirkt, können wir aus geometrischen Gründen keine weiteren reellen Eigenvektoren erwarten können und müssen mit Eigenvektoren rechnen, die zu den komplexen Eigenwerten λ_2 und λ_3 gehören.

Statt zu rechnen, lassen sich die beiden gesuchten Eigenvektoren, wie bei jeder Begleitermatrix, für einmal einfach hinschreiben. Wir behaupten, dass die beiden Vektoren

$$\vec{v}_2 = \begin{pmatrix} 1 \\ \zeta_3 \\ \zeta_3^2 \end{pmatrix}, \qquad \vec{v}_3 = \begin{pmatrix} 1 \\ \zeta_3^2 \\ \zeta_3^4 \end{pmatrix} = \begin{pmatrix} 1 \\ \zeta_3^2 \\ \zeta_3 \end{pmatrix}$$

das Gewünschte leisten. Tatsächlich rechnen wir mit $\zeta_3^3 = 1$ leicht nach, dass der Shift T_3 die Bedingungen

$$T_3 \cdot \vec{v}_2 = \begin{pmatrix} \zeta_3 \\ \zeta_3^2 \\ 1 \end{pmatrix} = \zeta_3 \begin{pmatrix} 1 \\ \zeta_3 \\ \zeta_3^2 \end{pmatrix} = \zeta_3 \vec{v}_2, \quad T_3 \cdot \vec{v}_3 = \begin{pmatrix} \zeta_3^2 \\ \zeta_3 \\ 1 \end{pmatrix} = \zeta_3^2 \begin{pmatrix} 1 \\ \zeta_3^2 \\ \zeta_3 \end{pmatrix} = \zeta_3^2 \vec{v}_3$$

erfüllt. Die drei gefundenen Eigenvektoren $\vec{v}_1, \vec{v}_2, \vec{v}_3$ bilden die Spalten der symmetrischen Transformationsmatrix, der wir wegen ihrer grossen Bedeutung für die Anwendungen ein eigenes Symbol geben.

$$F_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta_3 & \zeta_3^2 \\ 1 & \zeta_3^2 & \zeta_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \frac{1}{2}(-1+\sqrt{3}i) & \frac{1}{2}(-1-\sqrt{3}i) \\ 1 & \frac{1}{2}(-1-\sqrt{3}i) & \frac{1}{2}(-1+\sqrt{3}i) \end{pmatrix}$$

Multiplikation mit dieser Matrix liefert eine lineare Abbildung

$$\mathcal{F}_3 \colon \mathbb{C}^3 \to \mathbb{C}^3, \qquad \vec{x} \mapsto F_3 \cdot \vec{x}$$

die in der Technik als *diskrete Fourier-Transformation* (DFT) bezeichnet wird. Um diese Transformation umzukehren, benötigen wir die Inverse von F_3, die wir wiederum, statt blind zu rechnen, konzeptioneller bestimmen. Wegen der Orthogonalität der Ausgangsmatrix T_3 erwarten wir, dass auch die drei Eigenvektoren in einem geeigneten Sinn orthogonal sind. Die Rechnung zeigt allerdings, dass das Produkt $F_3 \cdot F_3^T$ nicht Vielfaches der Einheitsmatrix und daher die Erwartung falsch ist. Ein genaueres Hinsehen zeigt aber, dass die Vermutung nicht einfach grundfalsch ist, sondern leicht richtig gestellt werden kann, wenn man statt mit F_3^T mit der konjugiert komplexen Matrix $\overline{F}_3^T$ multipliziert. Beachtet man, dass die Einheitswurzeln durch Spiegeln an der reellen Achse die einfachen Beziehungen $\overline{\zeta_3} = \zeta_3^2$ und $\overline{\zeta_3^2} = \zeta_3$ erfüllen, so gilt tatsächlich

$$F_3 \cdot \overline{F}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta_3 & \zeta_3^2 \\ 1 & \zeta_3^2 & \zeta_3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta_3^2 & \zeta_3 \\ 1 & \zeta_3 & \zeta_3^2 \end{pmatrix} = 3 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Diese Beziehung erkennt man geometrisch leicht an Hand der Lage der dritten Einheitswurzeln auf dem Einheitskreis. Die Transformationsmatrix F_3 erfüllt also die Beziehung

$$F_3 \cdot \overline{F}_3^T = 3E_3 = \overline{F}_3^T \cdot F_3$$

Diese Bedingung zusammen mit der Symmetrie von F_3 bewirkt, dass die Matrix F_3 bis auf einen Faktor *unitär*[4] ist. Dabei handelt es sich um eine Bedingung, die im Umgang mit komplexen Matrizen die analoge Rolle spielt wie die Orthogonalität bei den reellen Matrizen, wobei wir allerdings auch das Skalarprodukt mit Hilfe der Konjugation definieren müssen. Aus dieser Bedingung lässt sich die gesuchte Inverse leicht bestimmen. Es ist

$$F_3^{-1} = \frac{1}{3} \cdot \overline{F}_3 = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta_3^2 & \zeta_3 \\ 1 & \zeta_3 & \zeta_3^2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \frac{1}{2}(-1-\sqrt{3}i) & \frac{1}{2}(-1+\sqrt{3}i) \\ 1 & \frac{1}{2}(-1+\sqrt{3}i) & \frac{1}{2}(-1-\sqrt{3}i) \end{pmatrix}$$

Multiplikation mit dieser Matrix liefert eine lineare Abbildung

$$\mathcal{F}_3^{-1}\colon \mathbb{C}^3 \to \mathbb{C}^3, \qquad \vec{x} \mapsto F_3^{-1} \cdot \vec{x}$$

die in der Technik[5] als *inverse diskrete Fourier-Transformation* (IDFT) bezeichnet wird. ◯

Beispiel. Um einen Vektor $\vec{x}$ nach Fourier zu transformieren, genügt es, ein Matrizenprodukt zu bilden.

$$\vec{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \mapsto \mathcal{F}_3(\vec{x}) = F_3 \cdot \vec{x} = \begin{pmatrix} 4 \\ \frac{1}{2}(-1+\sqrt{3}i) \\ \frac{1}{2}(-1-\sqrt{3}i) \end{pmatrix}$$

Für die inverse Transformation gehen wir entsprechend vor.

$$\vec{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \mapsto \mathcal{F}_3^{-1}(\vec{x}) = F_3^{-1} \cdot \vec{x} = \begin{pmatrix} 2 \\ -\frac{1}{2} + \frac{\sqrt{3}}{6}i \\ -\frac{1}{2} - \frac{\sqrt{3}}{6}i \end{pmatrix}$$

Seine Komponenten sind die für die explizite Beschreibung der Walzerzahlen erforderlichen Koeffizienten.

---

[4] Eine komplexe quadratische Matrix A heisst unitär, falls $A \cdot \overline{A}^T = E = \overline{A}^T \cdot A$ gilt.

[5] Die Bezeichnung unter Technikern, Physikern und Mathematikern in der Literatur ist nicht einheitlich. Die Abhängigkeit von der Dimension 3 des Raumes wird oft unterschlagen, die Rolle von $\mathcal{F}$ und $\mathcal{F}^{-1}$ wird oft vertauscht und der Faktor 3 wird anders verteilt. Physiker sagen, dass unter der Fourier-Transformation Vektoren aus dem Zeitbereich in den Spektralbereich übergehen und dass die inverse Fourier-Transformation einem Vektor aus dem Spektralbereich einen Vektor aus dem Zeitbereich zuordnet. Bei den Technikern wird zwischen dem physikalischen und dem Frequenzraum transformiert. Der Mathematiker sieht nicht ein, warum der selbe Raum $\mathbb{C}^3$ zwei verschiedene Namen erhalten soll und fasst $\mathcal{F}$ und $\mathcal{F}^{-1}$ als Symmetrie (Automorphismus) dieses Raumes auf, die interessante Eigenschaften haben, die alle aus den speziellen Beziehungen zwischern den komplexen Einheitswurzeln folgen.

Aus den Orthogonalitätsrelation der Fouriertransformation folgt, dass ein Paar entsprechender Vektoren die selbe Norm[6] hat, d.h. dass für jeden Vektor $\vec{x}$ die Beziehung

$$|F_3 \cdot \vec{x}| = |\vec{x}|$$

gilt. Diese für die Anwendungen wichtige Eigenschaft der Fourier-Transformation lässt sich in diesem Beispiel leicht explizit nachrechnen. ◯

Die Matrizen F_3 und F_3^{-1} wurden als Transformationsmatrizen so gewählt, dass sie die zyklische Matrix T_3 diagonalisieren. Das Produkt

$$F_3^{-1} \cdot T_3 \cdot F_3 = \operatorname{diag}(\zeta_3^0, \zeta_3^1, \zeta_3^2)$$

liefert also eine Diagonalmatrix mit den drei Eigenwerten auf der Diagonalen. Weil *jede* beliebige zyklische Matrix

$$Z_n = a_0 E + a_1 T_n + a_2 T_n^2 + \cdots + a_{n-1} T_n^{n-1} \in \mathbb{R}^{n,n}$$

die selben Eigenvektoren wie der Spezialfall T_n hat, erwarten wir, dass sie auch durch die Fouriertransformation F_n diagonalisiert wird. Zyklische Prozesse, die durch zyklische Matrizen beschrieben werden, führen früher oder später zur Fouriertransformation!

Was wir hier für die Periodenlänge 3 vorgeführt haben, lässt sich selbstverständlich auf beliebige Periodenlängen verallgemeinern. Dazu ist es zweckmässig, zuerst die Matrizenrechnung auf komplexe Skalare zu erweitern.

Nicht immer ist von Weitem klar, dass bei einem Problem komplexe Eigenvektoren eine Rolle spielen. An Hand des nächsten Beispiels aus der Kombinatorik wollen wir die neu erworbenen Mittel illustrieren.

Beispiel. Auf wieviele Arten können k Kinder in einer Reihe so aufgestellt werden, dass neben jedem Mädchen mindestens ein weiteres Mädchen als Anstandswauwau steht. Für die gesuchte Anzahl Muster schreiben wir x_k. Wer es lieber weniger frivol hat, kann auch nach der Anzahl k-stelliger Binärzahlen fragen, in denen immer mindestens zwei 0 nebeneinander stehen[7].

Zunächst brauchen wir verlässliches Anschauungsmaterial und untersuchen dazu die Verhältnisse für kleine Anzahlen. Die Mädchen bezeichnen wir mit ∘ und die Jungs mit •. Sie entsprechen den Binärziffern 0 und 1.

$k = 1$: Dann gibt es ein einziges erlaubtes Muster •. Daher ist $x_1 = 1$.

$k = 2$: Dann gibt es die beiden erlaubten Muster ••, ∘∘. Daher ist $x_2 = 2$.

$k = 3$: Dann gibt es die vier erlaubten Muster

••• •∘∘ ∘∘∘
∘∘•

[6] Damit Normen reell werden, ist die Norm des komplexen Vektors $\vec{x} \in \mathbb{C}^n$ als $|\vec{x}| := \sqrt{\overline{\vec{x}}^T \cdot \vec{x}}$ erklärt!

[7] Noch seriöser tönt die äquivalente Frage nach der Anzahl positiver Zöpfe mit 3 Strängen. Dabei versteht man unter einem Zopf mit n Strängen ein Element aus der Zopfgruppe B_n, die bekanntlich durch die Generatoren σ_i für $i = 1, \ldots, n-1$ erzeugt wird und die Zopf-Relationen

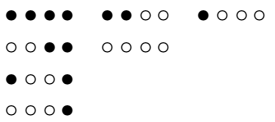
$$\begin{array}{ll} \sigma_i\sigma_j = \sigma_j\sigma_i & \text{für } |i-j| \geq 2 \\ \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} & \text{für } 1 \leq i \leq n-2 \end{array}$$

erfüllt. Ein positiver Zopf hat die zusätzliche Eigenschaft, dass in den Wörtern nur positive Potenzen der Erzeugenden σ_i benötigt werden. Die Inversen σ_i^{-1} sind also überflüssig.

Sie entsprechen den 3-stelligen Binärzahlen $111, 001, 100, 000$ und nicht erlaubt sind $110, 101, 011, 010$. Daher ist $x_3 = 4$.

Selbstverständlich wäre es jetzt vermessen, aus diesen wenigen Daten schliessen zu wollen, dass es sich bei der Folge x_k der Kinderzahlen um die Folge der 2-er Potenzen handelt! Tatsächlich gilt:

$k = 4$: Dann gibt es die folgenden sieben Muster

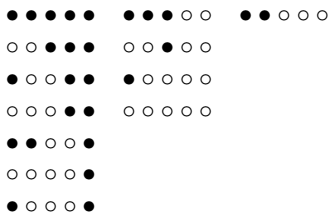


Sie entsprechen den sieben 4-stelligen Binärzahlen

$$1111, 0011, 1001, 0001, 1100, 0000, 1000$$

Damit uns das Gesetz der kleinen Zahlen nicht noch einmal zum Narren hält, berechnen wir weitere Beispiele.

$k = 5$: Dann gibt es die folgenden zwölf Muster



Zwar sind die nun vorhandenen Daten noch nicht ausreichend, um die vorliegende Folge in der "Encocyplopedia of Integer Sequences" bzw. in der On-Line-Datenbank `http://www.research.att.com/~njas/sequences` eindeutig zu indentifizieren, aber immerhin sollten wir genügend Information haben, um eine Chance haben, ihr Muster zu erkennen. Dazu fassen wir die gefundenen Werte in einer Tabelle zusammen, die wir zur Not durch das Auflisten weiterer Beispiele fortsetzen.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$
x_k	1	1	2	4	7	12	21	37	65	114	200	351	616	$\cdots$

Besser als weitere öde Knochenarbeit ist es nun, sich um eine rekursive Beschreibung der zugrundeliegenden Folge zu bemühen. Dabei versuchen wir, die Muster auf systematische Art vollständig und eindeutig aus den bereits konstruierten Muster mit echt weniger Elementen zu erzeugen.

Wir stellen uns vor, k Kinder seien regelgerecht aufgestellt. Dann gibt es drei disjunkte Fälle, die in obigen Beispielen den drei Spalten entsprechen.

1. Das letzte Kind ist ein Junge. Dann dürfen an den Plätzen davor die $k-1$ Kinder in einer der x_{k-1} erlaubten Arten aufgestellt werden und dabei werden keine Muster wiederholt. Die x_{k-1} Muster dieses Falles haben die Form $\cdots\bullet$.

2. Die letzten beiden Kinder sind zwei Mädchen. Dann dürfen an den Plätzen davor die $k-2$ Kinder in einer der x_{k-2} erlaubten Arten aufgestellt werden und dabei werden keine Muster wiederholt und keines der im ersten Fall bereits aufgelisteten Muster kann nochmals vorkommen. Die x_{k-2} Muster dieses Falles haben die Form $\cdots \circ \circ$.

3. Die letzten drei Kinder sind drei Mädchen. Dabei müssen wir aber aufpassen, dass am Schluss nicht etwa vier Mädchen stehen, weil wir diese Muster bereits im zweiten Fall gezählt haben. Daher muss vor den drei Mädchen ein Junge stehen und an den Plätzen davor dürfen die $k-4$ Kinder in einer der x_{k-4} erlaubten Arten aufgestellt werden. Dabei werden keine Muster wiederholt und keines der im ersten und zweiten Fall bereits aufgelisteten Muster kann nochmals vorkommen. Die x_{k-4} Muster dieses Falls haben die Form $\cdots \bullet \circ \circ \circ$.

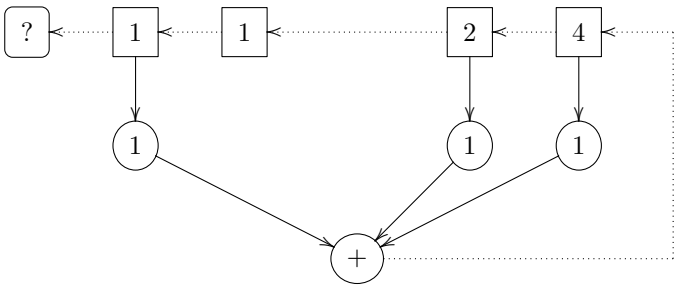
Weil man auf diese Art alle Muster der fraglichen Art genau einmal erhält, gilt die Rekursionsgleichung

$$x_k = x_{k-1} + x_{k-2} + x_{k-4}$$

die man nach einer Substituition von k durch $k+4$ in der Form

$$x_{k+4} = x_{k+3} + x_{k+2} + x_k$$

schreiben kann und mit jener der Fibonacci-Folge f_k vergleiche. Mit Hilfe dieser Rekursion lassen sich nun leicht weitere Werte aus der Tabelle bestimmen. Nachträglich stellt es sich als vernünftig heraus, $x_0 = 1$ zu setzen, weil dann die Rekursion auf für $k = 0$ gültig bleibt. Die Rekursionsgleichung erlaubt es auch sofort, ein Schieberegister zu entwerfen, das die Folge x_k produziert.



Die gefundene Rekursionsgleichung hat das charakteristische Polynom

$$\chi_A(\lambda) = \lambda^4 - \lambda^3 - \lambda^2 - 1$$

und die Folge x_k tritt als erste Komponente im Zustandsvektor $\vec{y}(k)$ der Vektorfolge $\vec{y}(k+1) = A \cdot \vec{y}(k)$ auf, deren Systemmatrix die Begleitermatrix A von $\chi(\lambda)$ ist und deren Anfangszustand sich aus den ersten vier Werten aus obiger Tabelle ergeben.

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}, \qquad \vec{y}(0) = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \end{pmatrix}$$

Die zugehörige Lösung $\vec{y}(k) = A^k \cdot \vec{y}(0)$ erlaubt es, die Folge mit Hilfe des Verdoppelungsverfahrens effizient zu berechnen. Das typische explosive Verhalten in der linken Teilfigur mit dem Graphen von x_k lässt vermuten, dass diese Folge exponentiell wächst. Diese Vermutung lässt sich erhärten, wenn man in der rechten Teilfigur den Graphen der logarithmierten Folge $\log(x_k)$ betrachtet, der praktisch geradlinig verläuft.

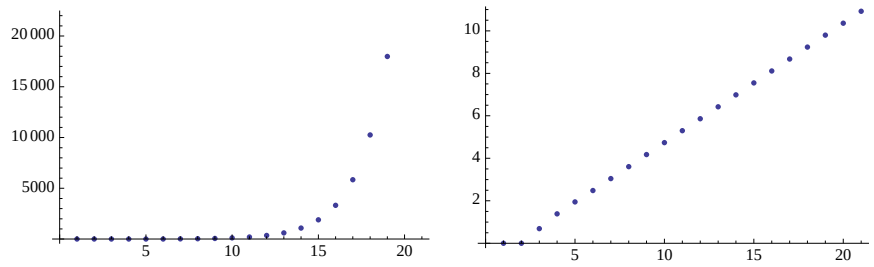
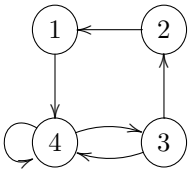


Abbildung 4.7: Graphen der beiden Folgen x_k und $\log(x_k)$.

Interpretieren wir die Matrix A als Adjazenzmatrix des Graphen



so beschreiben ihre Potenzen A^k die Anzahl Weg der Länge k in diesem Graphen. Aus den Matrizenpotenzen

$$A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 3 & 3 \end{pmatrix}, \quad A^4 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 3 & 3 \\ 3 & 2 & 4 & 6 \end{pmatrix}$$

lesen wir ab, dass es keinen Weg der Länge 4 vom Knoten 2 zum Knoten 1 gibt und dass man auf den 4 Wegen der Länge 4

$$\begin{array}{ll} 3 \to 2 \to 1 \to 4 \to 4, & 3 \to 4 \to 3 \to 4 \to 4 \\ 3 \to 4 \to 4 \to 3 \to 4, & 3 \to 4 \to 4 \to 4 \to 4 \end{array}$$

vom Knoten 3 zum Knoten 4 gelangen kann.

Die Summe

$$S_3 = A + A^2 + A^3 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 2 & 1 & 2 & 4 \\ 4 & 2 & 5 & 6 \end{pmatrix}$$

zeigt, dass jeder Weg von den Knoten 1 und 2 zu sich selber mindestens die Länge 4 haben muss. Aus der entsprechenden Summe

$$S_4 = A + A^2 + A^3 + A^4 = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 2 & 1 & 3 & 4 \\ 4 & 2 & 5 & 7 \\ 7 & 4 & 9 & 12 \end{pmatrix}$$

lesen wir ab, dass es vom Knoten 3 zum Knoten 4 neben den oben aufgelisteten Wegen der exakten Länge 4 noch 5 weitere Wege der Längen höchstens 4 gibt. Insgesamt existieren also 9 solche Wege.

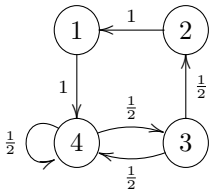
Länge	1	2	3	4
	$3 \to 4$	$3 \to 4 \to 4$	$3 \to 2 \to 1 \to 4$	$3 \to 2 \to 1 \to 4 \to 4$
			$3 \to 4 \to 3 \to 4$	$3 \to 4 \to 3 \to 4 \to 4$
			$3 \to 4 \to 4 \to 4$	$3 \to 4 \to 4 \to 3 \to 4$
				$3 \to 4 \to 4 \to 4 \to 4$

Insbesondere ist auf Grund von S_4 jeder Knoten mit jedem durch mindestens einen Weg der Länge höchstens 4 verbindbar ist. Der Graph ist also zusammenhängend.

Gehen wir zur zugehörigen stochastischen Matrix über, indem wir jede Spalte der positiven Matrix A durch ihre Spaltensumme dividieren, erhalten wir das ergodische stochastische System mit der Übergangsmatrix

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

die zum Übergangsgraphen



gehört. Wählen wir als Anfangszustand $\vec{p}(0)$ die Gleichverteilung, so liefert die Vektorfolge $\vec{p}(k) = P^k \cdot \vec{p}(0)$ eine Folge von Verteilungen $\vec{p}(k)$ mit den ersten vier Gliedern

$$\vec{p}(0) = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}, \quad \vec{p}(1) = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{2} \end{pmatrix}, \quad \vec{p}(2) = \begin{pmatrix} \frac{1}{8} \\ \frac{1}{16} \\ \frac{1}{4} \\ \frac{1}{16} \end{pmatrix}, \quad \vec{p}(3) = \begin{pmatrix} \frac{1}{16} \\ \frac{1}{8} \\ \frac{9}{32} \\ \frac{17}{32} \end{pmatrix}, \cdots$$

Die Graphen der vier zugehörigen Komponentenfolgen $p_1(k), p_2(k), p_3(k), p_4(k)$ findet man in den folgenden vier Teilfiguren.

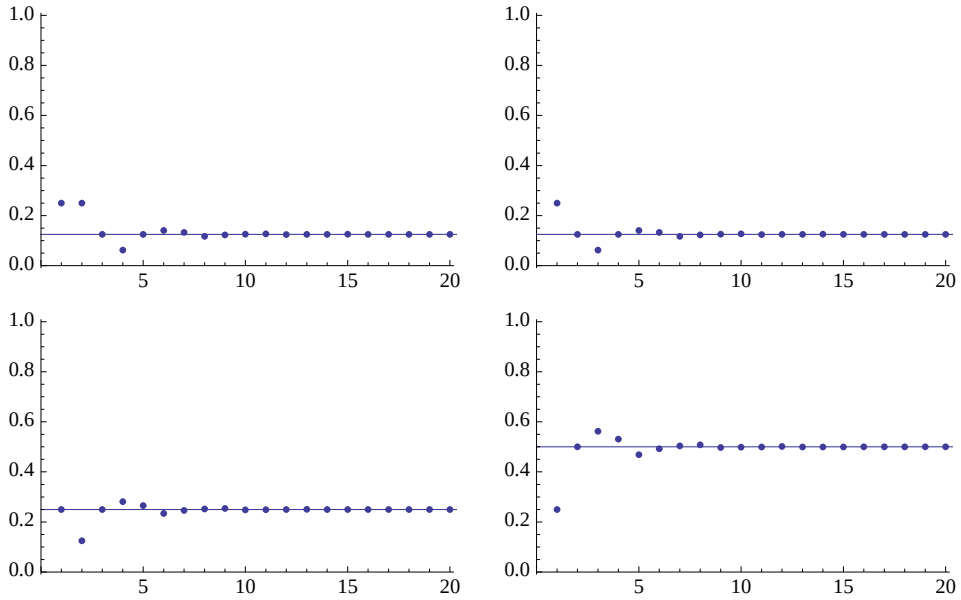


Abbildung 4.8: Graphen der Komponentenfolgen $p_1(k), p_2(k), p_3(k), p_4(k)$.

Man beachte, dass diese Folgen ein eigentümliches Schwingungsverhalten zeigen aber sich asymptotisch je einem Grenzwert zu näheren scheinen, nachdem sie genügend eingeschwungen sind. Um diese Grenzwerte zu bestimmen, untersuchen wir die Fixpunkte von P und müssen dazu die Fixpunktgleichung $P \cdot \vec{v} = \vec{v}$ bzw. das lineare, homogene Gleichungssystem $(P - E) \cdot \vec{v} = \vec{0}$ lösen. Selbstverständlich interessieren wir uns nur für nichttriviale Lösungen, die wir dann durch Normieren als Grenzverteilung interpretieren können. Um den Kern von $P - E$ zu finden, gehen wir von der zugehörigen Koeffizientenmatrix

$$P - E = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & \frac{1}{2} & 0 \\ 0 & 0 & -1 & \frac{1}{2} \\ 1 & 0 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}, \qquad \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 2 & 0 & 1 & -1 \end{pmatrix}$$

aus und führen nun geeignete Elementaroperationen durch. Um die Rechnung zu vereinfachen, haben wir die letzten drei Zeilen mit 2 multipliziert, um ganzzahlig weiterrechnen zu können. Addition des 2-fachen der ersten Zeile zur vierten Zeile liefert

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 2 & 1 & -1 \end{pmatrix}$$

Addition der zweiten Zeile zur vierten Zeile liefert

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 2 & -1 \end{pmatrix}$$

Addition der dritten Zeile zur vierten liefert die Stufenform, aus der klar wird, dass ein nicht trivialer Fixpunkt existiert.

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Um ihn zu finden, addieren wir die dritte Zeile zum 2-fachen der zweiten.

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -4 & 0 & 1 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Addition der zweiten Zeile zum 4-fachen der ersten liefert die reduzierte Stufenform

$$\begin{pmatrix} -4 & 0 & 0 & 1 \\ 0 & -4 & 0 & 1 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Die Lösungsmenge, d.h. der Fixpunktraum V_1 besitzt also die Parameterdarstellung

$$V_1 = t \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \\ 1 \end{pmatrix}, \qquad \vec{p}(\infty) = \begin{pmatrix} \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}$$

Weil der gesuchte Fixpunkt zusätzlich eine Wahrscheinlichkeitsverteilung liefern soll, muss seine Spaltensumme 1 ergeben und für den gesuchten Fixpunkt kommt nur der Vektor $\vec{p}(\infty)$ in Frage. Die Matrizenpotenzen von P werden sich also langfristig dem Grenzwert

$$P^\infty = \lim_{k\to\infty} P^k = \begin{pmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

nähern, deren Spalten jeweils der bestimmte Fixpunkt $\vec{p}(\infty)$ sind. Ein Blick auf die Graphen der Komponenetenfolge bestätigt diesen Befund.

Wegen $\vec{p}(\infty) = P^\infty \cdot \vec{p}(0)$ nähert sich die Vektorfolge $\vec{p}(k) = P^k \cdot \vec{p}(0)$ dem selben Fixpunkt $\vec{p}(\infty)$ völlig unabhängig von der Anfangsverteilung $\vec{p}(0)$. Daher kann dieser Fixpunkt dadurch approximativ bestimmt werden, dass man diese Folge für irgend eine beliebige Anfangsverteilung eine Weile lang verfolgt. Ist man also mit einem numerischen Näherungswert für den Fixpunkt zufrieden, muss die Fixpunktgleichung nicht exakt gelöst werden. Dieses Verfahren zum näherungsweisen Lösen ist dann speziell praktisch, wenn die übergansmatrix sehr gross ist, weil dann die direkte Lösung mit Hilfe des Eliminationsverfahrens sehr aufwändig ist. (Internet)

Der gefundene Fixpunkt als Grenzverteilung lässt sich wahrscheinichkeitstheoretisch so interpretieren, dass die Wahrscheinlichkeit, dass das System langfristig im Zustand j ist, durch die Komponente $\vec{p}_j(\infty)$ gegeben ist. Auf Grund dieses Gesetzes der grossen Zahlen lässt sich das dynamische System einfach simulieren. Man startet in einem Zustand und ändert ihn gemäss den gegebenen Übergangswahrscheinlichkeiten. Im vorliegenden Beispiel erhält man durch Werfen einer Münze folgenden typischen Zufallsweg der Länge $k = 10$, der im Knoten 3 beginnt:

$$3 \to 2 \to 1 \to 4 \to 3 \to 2 \to 1 \to 4 \to 3 \to 4 \to 4$$

Er besucht die einzelnen Zustände j mit den absoluten Häufigkeiten $H_j(k)$ und den relativen Häufigkeiten $h_j(k) = \frac{H_j(k)}{k}$

j	$H_j(10)$	$h_j(10)$	$H_j(100'000)$	$h_j(100'000)$	$\vec{p}_j(\infty)$	r_j
1	2	0.2	12′377	0.12377	0.125	8
2	2	0.2	12′378	0.12378	0.125	8
3	2	0.2	25′054	0.25054	0.25	4
4	4	0.4	50′191	0.50191	0.5	2

Die grobe Näherung kann als Anfangsverteilung für die weitere Approximation an den Fixpunkt benutzt werden. Sie ist schon recht ordentlich, weil das System tatsächlich $h_3(10) = 0.2 = 20\%$ der Zeit im Zustand 3 war und nach dem Gesetz der grossen Zahlen nach langer Zeit im Mittel $\vec{p}_3(\infty) = 0.25 = 25\%$ der Zeit in diesem Zustand sein wird. Führt man stattdessen die Simulation $k = 100'000$ Mal durch, erkennt man aus den zugehörigen Werten in der Tabelle, dass sich die relativen Häufigkeiten der einzelnen Zustände den entsprechenden Komponenten von $\vec{p}(\infty)$ nähern. Die Konvergenz ist also recht langsam.

Aus dem Fixpunkt $\vec{p}(\infty)$ lassen sich auch die mittleren Rückkehrzeiten des Systems bestimmen. Es stellt sich nämlich heraus, dass der reziproke Wert

$$r_j = \frac{1}{\vec{p}_j(\infty)}$$

angibt, wie lange es im Mittel dauern wird, bis das System das erste Mal vom Zustand j in den Zustand j zurückkehrt. Im Beispiel wird es also rund $r_3 = 4$ Zeitschritte dauern, bis das System das erste Mal in den Zustand 3 zurückkehrt, wenn es im Zustand 3 startet. Genau dieses Verhalten lässt sich an obgem Beispiel zweimal beobachten.

Die mittlere Übergangszeit m_{ji}, die das System braucht, um das erste Mal vom Zustand i in den Zustand j überzugehen, lässt sich mit Hilfe der Matrix

$$Z = (E - P + P^\infty)^{-1}$$

bestimmen. Es ist nämlich

$$m_{ji} = \frac{z_{jj} - z_{ji}}{\vec{p}_j(\infty)}$$

In unserem Beispiel ist

$$E - P + P^\infty = \begin{pmatrix} \frac{9}{8} & -\frac{7}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{9}{8} & -\frac{3}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{5}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{2} & 0 & 1 \end{pmatrix}$$

und ihre Inverse ist

$$Z = (E - P + P^\infty)^{-1} = \begin{pmatrix} \frac{47}{64} & \frac{39}{64} & \frac{7}{64} & -\frac{9}{64} \\ -\frac{9}{64} & \frac{47}{64} & \frac{15}{64} & -\frac{1}{64} \\ -\frac{1}{32} & -\frac{9}{32} & \frac{23}{32} & \frac{7}{32} \\ \frac{7}{16} & -\frac{1}{16} & -\frac{1}{16} & \frac{15}{16} \end{pmatrix}$$

Die mittleren Übergangszeiten m_{ji} für den Übergang $i \to j$ findet man in der Matrix

$$M = \begin{pmatrix} 0 & 1 & 5 & 7 \\ 7 & 0 & 4 & 6 \\ 3 & 4 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{pmatrix}$$

In obger Simulation hat das System 2 Zeitschritte gebraucht, um das erste Mal vom Zustand 3 in den Zustand 1 überzugehen. Im Mittel sind dazu

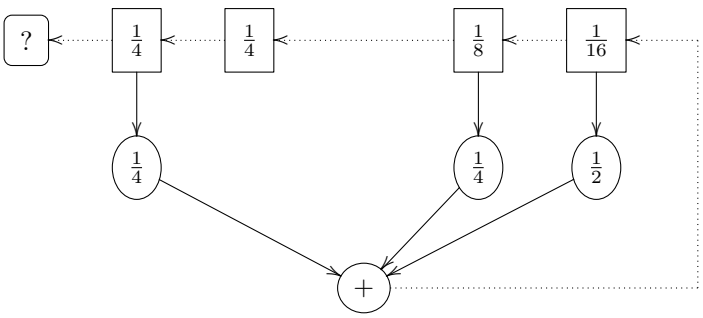
$$m_{13} = \frac{z_{11} - z_{13}}{\vec{p}_1(\infty)} = 5$$

Schritte notwendig.

Schliesslich wollen wir noch die unterwegs angetroffenen Folgen $p_j(k)$ explizit beschreiben. Dazu benötigen wir das charakteristische Polynom der stochastischen Matrix P und erhalten

$$\chi_P(\lambda) = -\frac{1}{4} - \frac{1}{4}\lambda^2 - \frac{1}{2}\lambda^3 + \lambda^4 = \frac{1}{4}(\lambda - 1)(1 + \lambda + 2\lambda^2 + 4\lambda^3)$$

Das zur Folge $\vec{p}_1(k)$ gehörende Schieberegister hat die Form



Für die restlichen Komponentenfolgen müssen die Anfangswerte entsprechend geändert werden.

Weil wir schon wissen, dass $\lambda_1 = 1$ ein Eigenwert von P ist, hat das charakteristische Polynom die angegebene Faktorisierung. Daher benötigen wir nur noch die Nullstellen des kubischen Polynoms $1 + \lambda + 2\lambda^2 + 4\lambda^3$.

Kürzen wir die beiden Radikale durch

$$w_1 = \sqrt[3]{\frac{47 + 3\sqrt{249}}{2}} \approx 3.61316\cdots, w_2 = \sqrt[3]{\frac{-47 + 3\sqrt{249}}{2}} \approx 0.553532\cdots$$

ab, so hat die reelle Lösung die Form

$$\lambda_2 = \frac{1}{6}\Big(-1 - w_1 + w_2\Big) \approx -0.676605\ldots$$

Die beiden anderen Nullstellen sind konjugiert komplex und haben die Form

$$\begin{aligned} \lambda_{3/4} &= \frac{1}{6}\Big(-1 + (\frac{w_1}{2} - \frac{w_2}{2}) \pm \sqrt{3}(\frac{w_1}{2} + \frac{w_2}{2})i\Big) \\ &= u \pm vi \approx 0.0883025\ldots \pm 0.60141\ldots i \end{aligned}$$

Die Folgen hat also eine explizite komplexe Beschreibung der Art

$$p_j(k) = c_1 + c_2\lambda_2^k + c_3\lambda_3^k + c_4\lambda_4^k$$

für geeignete Konstanten $c_1, c_2, c_3, c_4 \in \mathbb{C}$. Um eine reelle Beschreibung anzugeben, schreiben wir die beiden konjugiert komplexen Eigenwerte in Polarform

$$\lambda_3 = r \cdot e^{i\varphi}, \qquad \lambda_4 = r \cdot e^{-i\varphi}$$

mit dem Betrag

$$r = |\lambda_3| = |\lambda_4| \approx 0.676605\ldots$$

Für das Argument gilt

$$r \cdot \cos(\varphi) = u, \quad r \cdot \sin(\varphi) = v, \quad \tan(\varphi) = \frac{v}{u}, \quad \varphi \approx -1.42501\cdots$$

Dann gilt für die vier Komponentenfolgen

$$p_j(k) = c_{1j} + c_{2j}\lambda_2^k + d_{3j}r^k\cos(k\varphi) + d_{4j}r^k\sin(k\varphi)$$

und die reellen Konstanten müssen aus den Anfangsbedingungen durch Lösen eines linear Gleichungssysteme mit gemeinsamer Koeffizientenmatrix bestimmt werden. Für die Beträge der vier Eigenwerte gilt $|\lambda_1| = \lambda_1 = 1$ und

$$|\lambda_2| \approx 0.676605\ldots, \qquad |\lambda_3| = |\lambda_4| \approx 0.607858\ldots$$

Wegen

$$1 = |\lambda_1| > |\lambda_2| > |\lambda_3| = |\lambda_4|$$

hat $\lambda_1 = 1$ den grössten Betrag und da er er als einziger 1 ist, gilt asymptotisch $p_j(k) \sim c_{1j}$. Die Folgen nähern sich tatsächlich einem festen Wert und die hier aufgetauchten Kreisfunktionen erklären das beobachtete Schwingungsverhalten.

Nach diesem numerischen Abstecher in die Graphen- und Wahrscheinlichkeitstheorie kehren wir nun zur ursprünglichen Folge x_k der Kinderzahlen zurück und untersuchen auch diese Folge etwas genauer. Aus der gefundenen Rekursionsgleichung folgt insbesondere, dass die Folge x_k linear ist und ihre lineare Komplexität höchstens 4 ist. Um zu entscheiden, ob ihre lineare Komplexität nicht vielleicht kleiner ist, machen wir für die Rekursion den linearen Ansatz

$$x_k = ax_{k-1} + bx_{k-2} + cx_{k-3}$$

und versuchen nun die Koeffizienten zu berechnen. Setzt man die ersten Werte aus der Tabelle ein, erhält man das lineare Gleichungssystem

$$\left\{\begin{array}{lll} k=3: & 2a+1b+1c= & 4 \\ k=4: & 4a+2b+1c= & 7 \\ k=5: & 7a+4b+2c= & 12 \end{array}\right. \qquad \left(\begin{array}{ccc|c} 2 & 1 & 1 & 4 \\ 4 & 2 & 1 & 7 \\ 7 & 4 & 2 & 12 \end{array}\right) \qquad \left[\begin{array}{c} \\ Z_{12}(-2) \\ ZS_{13}(-7,2) \end{array}\right]$$

mit der danebenstehenden erweiterten Matrix. Zur Bestimmung seiner Lösungsmenge addieren wir das (-2)-fache der ersten Zeile zur zweiten Zeile und das (-7)-fache der ersten Zeile zum 2-fachen der dritten Zeile.

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 4 \\ 0 & 0 & -1 & -1 \\ 0 & 1 & -3 & -4 \end{array}\right) \quad \left[\begin{array}{c} \\ T_{23} \\ T_{32} \end{array}\right], \qquad \left(\begin{array}{ccc|c} 2 & 1 & 1 & 4 \\ 0 & 1 & -3 & -4 \\ 0 & 0 & -1 & -1 \end{array}\right) \quad \left[\begin{array}{c} Z_{31}(1) \\ Z_{32}(-3) \\ \end{array}\right]$$

Vertauschen der letzten beiden Zeilen liefert die Stufenform. Addition des (-3)-fachen der dritten Zeile zur zweiten und der dritten Zeile zur ersten liefert

$$\left(\begin{array}{ccc|c} 2 & 1 & 0 & 3 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & -1 \end{array}\right) \quad \left[\begin{array}{c} Z_{21}(-1) \\ \\ \end{array}\right]$$

Addition des (-1)-fachen der zweiten Zeile zur ersten liefert die reduzierte Stufenform

$$\left(\begin{array}{ccc|c} 2 & 0 & 0 & 4 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & -1 \end{array}\right)$$

aus der wir die eindeutig bestimmte Lösung $a = 2, b = -1, c = 1$ ablesen. Falls die Folge x_n also durch eine lineare Rekursionsgleichung kleinerer Ordnung beschrieben werden kann, so muss diese die Form

$$x_k = 2x_{k-1} - x_{k-2} + x_{k-3}$$

haben. Auf Grund der drei verwendeten Einzelfälle folgt noch nicht, dass diese neue Rekursionsgleichung genau die selbe Folge beschreibt, wie die alte. Tatsächlich gilt das folgende Resultat.

Satz. Die alte Rekursion

$$x_k = x_{k-1} + x_{k-2} + x_{k-4}, \quad x_0 = x_1 = 1, x_2 = 2, x_3 = 4$$

vierter Ordnung und die neue Rekursion

$$x_k = 2x_{k-1} - x_{k-2} + x_{k-3}, \quad x_0 = x_1 = 1, x_2 = 2$$

dritter Ordnung liefern für alle $k \geq 3$ die selben Zahlen.

Wir beweisen diesen Satz durch vollständige Induktion.

Beweis. Zunächst überprüfen wir, dass die Behauptung für irgend eine natürliche Zahl gilt. In unserem Fall ist es leicht nachzurechnen, dass die neue Rekursionsgleichug für $k = 3$ den selben Wert $x_3 = 4$ liefert, der zur alten Rekursion gehört. Das ist in unserem Fall banal, denn schliesslich haben wir die neue Rekursionsgleichung so gewählt, dass sie sogar für $k = 3, 4, 5$ die selben Werte wie die alte Rekursion liefert.

Wir können nun also annehmen, dass die beiden Rekrusionen an irgend einer Stelle k die selben Werte liefern und müssen nun zeigen, dass sie dann auch an der nächsten Stelle $k + 1$ den selben Wert liefern müssen. Nach Voraussetzung liefert die neue Rekursion den Wert

$$x_k = 2x_{k-1} - x_{k-2} + x_{k-3}$$

und wir müssen zeigen, dass der nächste Wert x_{n+1} ebenfalls die neue Rekusionsgleichung erfüllt d.h. dass

$$x_{k+1} = 2x_k - x_{k-1} + x_{k-2}$$

gilt. Die alte Rekursion liefert für diesen Nachfolgerwert

$$x_{k+1} = x_k + x_{k-1} + x_{k-3}$$

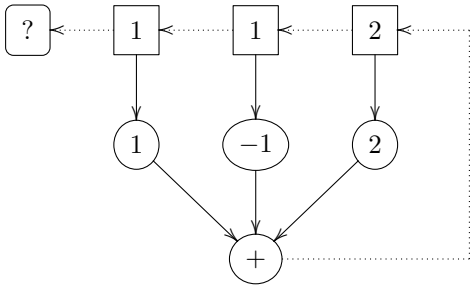
Damit gilt nach Voraussetzung also

$$\begin{aligned} x_{k+1} &= x_k + x_{k-1} + x_{k-3} \\ &= (2x_{k-1} - x_{k-2} + x_{k-3}) + x_{k-1} + x_{k-3} \\ &= 3x_{k-1} - x_{k-2} + 2x_{k-3} \\ &= 2x_k - x_{k-1} + x_{k-2} \end{aligned}$$

Die letzte Gleichheit gilt nämlich tatsächlich, weil sie nach dem Vereinfachen der Voraussetzung entspricht.

Wir haben also gesehen, dass die beiden Rekursionen mit k auch für $k + 1$ den selben Wert liefern. Daher müssen sie für alle natürlichen Zahlen $k \geq 3$ den selben Wert liefern, was behauptet wurde. □

In Zukunft können wir uns zur Untersuchung der Folge x_k also auf die einfachere neue Rekursion beschränken, da die lineare Komplexität dieser Folge bloss 3 beträgt. Aus der Rekursionsgleichung lesen wir direkt das Schema



für das zugehörige Schieberegister ab. Das charakteristische Polynom lautet

$$\chi(\lambda) = \lambda^3 - 2\lambda^2 + \lambda - 1$$

und hat die zugehörige Begleitermatrix

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 2 \end{pmatrix}, \qquad \vec{y}(0) = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

die nun aber negative Einträge hat und deshalb nicht ohne weiteres kombinatorisch interpretiert werden kann.

Um die Folge x_k mit Hilfe des Verdoppelungsverfahrens speziell effizient zu berechnen, machen wir aus der Rekursion dritter Ordnung eine Rekursion erster Ordnung in Dimension 3 indem wir den Zustandsvektor

$$\vec{y}(k) = \begin{pmatrix} x_k \\ x_{k+1} \\ x_{k+2} \end{pmatrix}$$

definieren. Der gegebenen Rekursionsgleichung entspricht die vektorielle Rekursion

$$\vec{y}(k+1) = B \cdot \vec{y}(k)$$

mit der Lösung $\vec{y}(k) = B^k \cdot \vec{y}(0)$. Um etwa x_{18} zu bestimmen, benötigen wir $\vec{y}(18-2) = \vec{y}(16)$ und dazu die Matrizenpotenz B^{16}, die wir numerisch durch iteriertes Quadrieren berechnen. Es ist

$$B^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & -1 & 2 \\ 2 & -1 & 3 \end{pmatrix}, \quad B^4 = \begin{pmatrix} 2 & -1 & 3 \\ 3 & -1 & 5 \\ 5 & -2 & 9 \end{pmatrix}$$

$$B^8 = \begin{pmatrix} 16 & -7 & 28 \\ 28 & -12 & 49 \\ 49 & -21 & 86 \end{pmatrix}, \quad B^{16} = \begin{pmatrix} 1432 & -616 & 2'513 \\ 2'513 & -1'081 & 4'410 \\ 4'410 & -1'897 & 7'739 \end{pmatrix}$$

Damit erhalten wir

$$\vec{y}(16) = B^{16} \cdot \vec{y}(0) = \begin{pmatrix} x_{16} \\ x_{17} \\ x_{18} \end{pmatrix} = \begin{pmatrix} 5'842 \\ 10'252 \\ 17'991 \end{pmatrix}$$

und können daraus schliesslich die gewünschte Anzahl $x_{18} = 17'991$ ablesen.

Um die Folge x_k explizit zu beschreiben, benötigen wir die Nullstellen des charakteristischen Polynoms

$$\chi_B(\lambda) = \lambda^3 - 2\lambda^2 + \lambda - 1$$

Es hat als einzige reelle Nullstelle

$$\lambda_1 = \frac{1}{3}\Big(2 + \sqrt[3]{25 - 3\sqrt{69}} + \sqrt[3]{25 + 3\sqrt{69}}\Big) \approx 1.75488\ldots$$

Kürzen wir die beiden Radikale durch

$$w_1 = \sqrt[3]{\frac{25-3\sqrt{69}}{2}} \approx 0.342178\cdots, \quad w_2 = \sqrt[3]{\frac{25+3\sqrt{69}}{2}} \approx 2.92245\cdots$$

ab, so hat die reelle Lösung die Form

$$\lambda_1 = \frac{1}{3}\Big(2 + w_1 + w_2\Big) \approx 1.75488\ldots$$

Die beiden anderen Nullstellen sind konjugiert komplex und haben die Form

$$\begin{aligned} \lambda_{2/3} &= \frac{1}{6}\Big(4 - (w_1+w_2) \pm \sqrt{3}(w_2 - w_1)i\Big) \\ &= u \pm vi \approx 0.122561\ldots \pm 0.744862\ldots i \end{aligned}$$

Die Folge hat also eine explizite komplexe Beschreibung der Art

$$x_k = c_1\lambda_1^k + c_2\lambda_2^k + c_3\lambda_3^k$$

für geeignete Konstanten $c_1, c_2, c_3 \in \mathbb{C}$. Zur Bestimmung setzen wir die ersten drei Folgenwerte ein und erhalten das lineare Gleichungssystem

$$\left\{\begin{array}{ll} k=0: & c_1 + \quad c_2 + \quad c_3 = 1 \\ k=1: & \lambda_1 c_1 + \lambda_2 c_2 + \lambda_3 c_3 = 1 \\ k=2: & \lambda_1^2 c_1 + \lambda_2^2 c_2 + \lambda_3^2 c_3 = 2 \end{array}\right. \qquad \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & 1 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & 2 \end{array}\right)$$

mit der Vandermonde-Matrix als Koeffizientenmatrix. Es hat die eindeutig bestimmte Lösung

$$\begin{aligned} c_1 &= \frac{2-\lambda_2-\lambda_3+\lambda_2\lambda_3}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)} \approx 0.722124\ldots, \\ c_2 &= \frac{2-\lambda_2-\lambda_3+\lambda_2\lambda_3}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)} \approx 0.138938\ldots + 0.20225\ldots i, \\ c_3 &= \frac{2-\lambda_1-\lambda_2+\lambda_1\lambda_2}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)} \approx 0.138938\ldots - 0.20225\ldots i, \end{aligned}$$

Um eine reelle Beschreibung der Folge x_k anzugeben, schreiben wir die beiden konjugiert komplexen Eigenwerte in Polarform

$$\lambda_2 = r\cdot e^{i\varphi}, \qquad \lambda_3 = r\cdot e^{-i\varphi}$$

mit dem Betrag

$$r = |\lambda_2| = |\lambda_3| \approx 0.754878\ldots$$

Für das Argument von λ_2 gilt

$$r\cdot\cos(\varphi) = u, \quad r\cdot\sin(\varphi) = v, \quad \tan(\varphi) = \frac{v}{u}, \quad \varphi \approx 1.40772\cdots$$

Dann ist

$$x_k = c_1\lambda_1^k + d_2 r^k\cos(k\varphi) + d_3 r^k \sin(k\varphi)$$

und für die Konstanten erhalten wir mit den Anfangsbedingungen

$$c_1 \approx 0.722124\ldots, d_2 \approx 0.277876\ldots, d_3 \approx -0.4045\ldots$$

Für die Beträge der drei Eigenwerte gilt

$$|\lambda_1| = \lambda_1 \approx 1.75488\ldots; \qquad |\lambda_2| = |\lambda_3| \approx 0.754878\ldots$$

Daher hat λ_1 den grössten Betrag. Weil er als einziger grösser als 1 ist, gilt asymptotisch

$$x_k \sim c_1\lambda_1^k$$

und die Folge wächst tatsächlich exponentiell und ihr Logarithmus linear, da gilt: $\log(x_k) \sim \log(c_1) + \log(\lambda_1) \cdot k$. ◯

4.4 Die Hauptachsen

Eine andere wichtige geometrische Interpretation und Anwendung des Eigensystems einer quadratischen Matrix ergibt sich wie folgt. Zunächst erinnern wir uns daran, dass ein Ellipsoid mit Hilfe einer sogn. *quadratischen Form* durch die Gleichung

$$\vec{v}^T \cdot M \cdot \vec{v} = 1$$

beschrieben werden kann, wobei M eine symmetrische Matrix bezeichnet. Wir interessieren uns für Lage und Länge der Symmetrie-Achsen dieses Ellipsoides.

Beispiel. Wir gehen von der Ellipse mit der Koordinatengleichung

$$\frac{5}{36}x^2 - \frac{1}{6}xy + \frac{1}{4}y^2 = 1, \qquad 5x^2 - 6xy + 9y^2 = 36$$

aus. Sie kann mit Hilfe der symmetrische Matrix

$$M = \begin{pmatrix} \frac{5}{36} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{4} \end{pmatrix} = \frac{1}{36}\begin{pmatrix} 5 & -3 \\ -3 & 9 \end{pmatrix}$$

geschrieben werden, weil dann die Ellipsengleichung $\vec{v}^T \cdot M \cdot \vec{v} = 1$ gilt.
Diese Ellipse kann als Bild des Einheitskreises unter der linearen Abbildung

$$f_A\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 3x \\ y + 2y \end{pmatrix}$$

erhalten werden, die wir bereits früher zur ersten geometrischen Interpretation des Eigensystems benutzt haben und für deren Abbildungsmatrix

$$A = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}, \qquad A^{-1} = B = \frac{1}{6}\begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix}$$

gilt. Damit erhalten wir nämlich die symmetrische Gram'sche Matrix

$$B^T \cdot B = \frac{1}{6 \cdot 6}\begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix} = \frac{1}{36}\begin{pmatrix} 5 & -3 \\ -3 & 9 \end{pmatrix} = M$$

mit der die Ellipse beschrieben wird.

Wir berechnen nun die Eigenwerte und die Eigenvektoren der Matrix M nach dem uns geläufigen Verfahren. Für das charakteristische Polynom der Matrix M erhält man $\chi_M(\lambda) = \lambda^2 - \frac{7}{18}\lambda + \frac{1}{36}$. Die Eigenwerte der Matrix M erhält man als Nullstellen dieses Polynoms, d.h. als Lösung einer quadratischen Gleichung. Es gilt:

$$\lambda_1 = \frac{1}{36}(7 - \sqrt{13}) \quad \text{und} \quad \lambda_2 = \frac{1}{36}(7 + \sqrt{13})$$

mit den zugehörigen Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 2+\sqrt{13} \\ 3 \end{pmatrix} \qquad \text{und} \qquad \vec{v}_2 = \begin{pmatrix} 2-\sqrt{13} \\ 3 \end{pmatrix}$$

Für die zugehörigen Eigenräume ergibt sich daher

$$V_{\lambda_1} = \left\{ t \begin{pmatrix} 2+\sqrt{13} \\ 3 \end{pmatrix} | \in \mathbb{R} \right\} \qquad \text{und} \qquad V_{\lambda_2} = \left\{ t \begin{pmatrix} 2-\sqrt{13} \\ 3 \end{pmatrix} | \in \mathbb{R} \right\}$$

Die Frage stellt sich, ob wir dieses Eigensystem von M an Hand der Ellipse geometrisch interpretieren können. Zeichnen wir die beiden Eigenvektoren $\vec{v}_1$ und $\vec{v}_2$ ein, erhalten wir in unserem Fall folgende, uns bekannte, Figur.

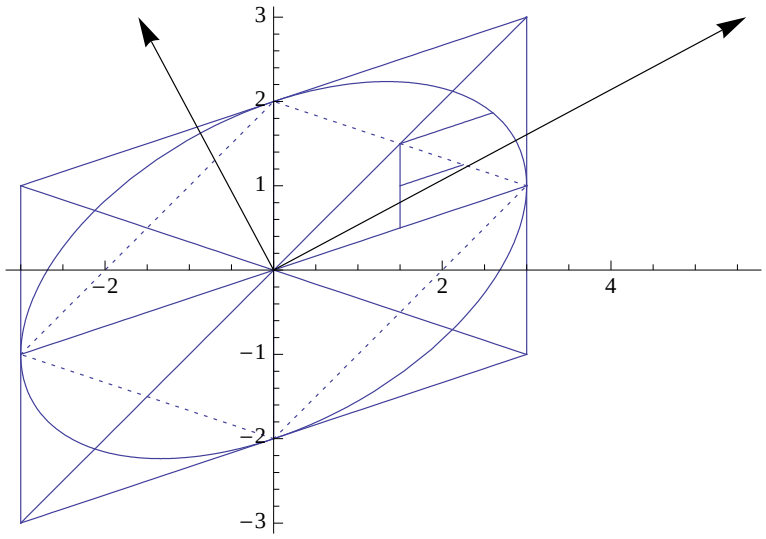


Abbildung 4.9: Dritte geometrische Interpretation von Eigenvektoren.

Wir stellen zunächst einmal fest, dass die beiden gefundenen Eigenvektoren senkrecht aufeinander stehen. Diese Eigenschaft lässt sich algebraisch mit Hilfe des Skalarproduktes durch die Gleichung $\langle \vec{v}_1, \vec{v}_2 \rangle = 0$ bestätigen. Untersucht man die Figur genauer, stellt man weiter fest, dass die Eigenvektoren genau in Richtung der *Symmetrieachsen* der Ellipse zeigen. Mit Hilfe von Eigenvektoren der Matrtix M kann man also jene Richtungen bestimmen, in die der Abstand des jeweiligen Ellipsenpunkunkte vom Ursprung minimal bzw. maximal ist. Die

vier Punkte, die als Schnitte der Ellipse mit den Symmetrieachsen entstehen, heissen *Scheitelpunkte*. Sie bestimmen den kleinsten und den grössten Radius der Ellipse. Der entsprechende minimale bzw. maximale Abstand der Scheitelpunkte vom Ursprung nennt man *kleine* bzw. *grosse Halbachse* der Ellipse. Es stellt sich heraus, dass die beiden Eigenwerte λ_1 und λ_2 mit diesen Halbachsen in Beziehung stehen. Ihre Längen sind nämlich $1/\sqrt{\lambda_1}$ und $1/\sqrt{\lambda_2}$. Das Eigensystem liefert also gewissermassen das natürliche Koordinatensystem für Ellipsen. ◯

Das beschriebene Verfahren lässt sich nicht nur für Ellipsen, sondern allgemeiner für beliebige *Kegelschnitte in Mittelpunktslage*, d.h. Gebilde mit der Koordinatengleichung

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 = 1$$

verwenden. Ein solcher Kegelschnitt kann mit Hilfe seiner darstellenden Matrix M beschrieben werden. Wählen wir für M folgende symmetrische Matrix

$$M = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}, \qquad \vec{v} = \begin{pmatrix} x \\ y \end{pmatrix}$$

so kann die Kegelschnittgleichung in Mittelpunktslage matriziell in der Form

$$\vec{v}^T \cdot M \cdot \vec{v} = 1$$

beschrieben werden.

Durch Lösen des Eigenwertproblems für die symmetrische Strukturmatrix M erhalten wir zwei orthogonale Vektoren $\vec{v}_1$ und $\vec{v}_2$ in Richtung der Hauptachsen. Aus den Eigenwerten lassen sich die Längen der Halbachsen berechnen.

Bei genauerem Hinsehen stellt man fest, dass nicht jeder Kegelschnitt das affine Bild eines Kreises d.h. eine Ellipse ist. Das hängt damit zusammen, dass nicht jede symmetrische Matrix M als Gram'sche Matrix d.h. in der Form $M = B^T \cdot B$ mit einer gewissen invertierbaren Matrix B faktorisiert werden kann. Der folgende zentrale Begriff sondert solche Matrizen aus.

Definition. Eine symmetrische Matrix M heisst *positiv semidefinit*, falls eine Matrix B existiert mit der Eigenschaft $M = B^T \cdot B$. Falls M zusätzlich invertierbar ist, heisst sie *positiv definit*.

Solche sogn. positv semidefinite symmetrische Matrizen, die als Gram'sche Matrizen in der Form $B^T \cdot B$ zerlegt werden können, spielen in vielen Anwendungen eine zentrale Rolle, weil sie mit der Energie dynamischer Systeme zusammenhängen.

Beispiel. Eine Matrix $M = (r)$ vom Typ 1×1 ist ein Skalar und daher symmetrisch. Sie ist genau dann positiv semidefinit, falls $r \geq 0$ ist und positiv definit, falls $r > 0$ gilt. Allgemeiner ist eine Diagonalmatrix symmetrisch und genau dann positiv semidefinit, wenn alle ihre Diagonalelemente positiv sind. Sie ist genau dann positiv definit, falls alle ihre Diagonalelemente strikt positiv sind. Insbesondere ist nicht jede Matrix positv semidefinit und schon gar nicht positiv definit. ◯

Für beliebige symmetrische Matrizen M lässt sich durch Transformation in das Eigensystem (Hauptachsentransformation) der zugehörige Kegelschnitt klassifizieren, d.h. die Gleichung des Kegelschnittes auf eine möglichst einfache Normalform bringen. Es stellt sich heraus, dass eine solche Kegelschnittgleichung

im 2-dimensionalen Raum zu einem der 6 Typen *Ellipse*, *Parabel*, *Hyperbel*, *Geradenpaar*, *Doppelgerade*, *Einsiedlerpunkt* äquivalent ist. Die ersten drei Typen heissen *nicht entartet*. Genau bei ihnen ist die Matrix M invertierbar.

Ein analoges Vorgehen lässt sich auf höhere Dimensionen für Quadriken übetragen. In der Dimension 3 erhalten wir die *Quadriken in Mittelpunkslage* mit der Gleichung

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{13}xz + 2a_{23}yz = 1$$

Wählen wir für M die symmetrische Matrix

$$M = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

so kann die Gleichung der Quadrik matriziell in der Form $\vec{v}^T \cdot M \cdot \vec{v} = 1$ geschrieben werden. Durch Lösen des Eigenwertproblems für die Matrix M erhalten wir drei orthogonale Vektoren $\vec{v}_1, \vec{v}_2$ und $\vec{v}_3$ in Richtung der Hauptachen. Aus den Eigenwerten lassen sich auch hier die Längen der Halbachsen berechnen und die Quadriken klassifizieren. Es zeigt sich hier, dass jede Quadrik im 3-dimensionalen Raum zu einem der 8 Typen *Ellipsoid*, *Einschaliges Hyperboloid*, *Zweischaliges Hyperboloid*, *Elliptisches Paraboloid*, *Hyperbolisches Paraboloid*, *Kegel*, *Zylinder* oder *Einsiedlerpunkt* äquivalent ist.

4.5 Das Hauptsystem

Zwar können Matrizen bei Verwendung der komplexen Zahlen zusätzliche komplexe Eigenwerte und zugehörige Eigenvektoren besitzen. Es gibt aber Matrizen, die auch bei Verwendung der komplexen Zahlen zu wenig linear unabhängige Eigenvektoren haben und damit nicht diagonalisierbar sind.

Beispiel. Wir gehen von der linearen Abbildung $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ aus, die als Komposition der Orthogonalprojektion $P_y\colon \mathbb{R}^2 \to \mathbb{R}^2$ der Ebene auf die y-Achse mit einer Spiegelung $S_{\frac{\pi}{4}}\colon \mathbb{R}^2 \to \mathbb{R}^2$ an der ersten Winkelhalbierenden entsteht. Diese lineare Abbildung wird durch folgende Matrix beschrieben:

$$N_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Sie erfüllt die Gleichung $N_2^2 = 0$ und wegen dieser Nilpotenz vom Grad 2 stirbt der dynamische Prozess N_2 nach einem Schritt aus und ist daher sicher nicht umkehrbar. Die Matrix benimmt sich wie eine verkappte Nullmatrix.

Diese nilpotente Matrix N_2 kann als Begleitermatrix interpretiert werden. Daher ist ihr Minimalpolynom $\chi_{N_2}(\lambda) = \lambda^2$ und ihr Spektrum besteht aus dem einzigen komplexen Eigenwert $\lambda = 0$. Für den zugehörigen Kern gilt

$$V_0 = \left\{ t \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

Weil der einzige Eigenwert 0 einen eindimensionalen Eigenraum hat, ist die Summe der geometrischen Vielfachheiten 1 und daher N_2 nicht diagonalisierbar. Das zugehörige Vektorfeld auf dem Einheitskreis sieht wie folgt aus:

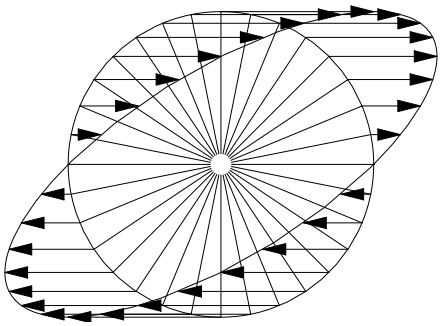


Abbildung 4.10: Vektorfeld einer nilpotenten Matrix.

Der Eigenraum V_0 beschreibt die x-Achse und wird durch den Vektor

$$\vec{v}^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

aufgespannt. Aus der Figur wird klar, dass die Vektoren, die die Schnittpunkte des Einheitskreises mit der x-Achse beschreiben, auf $\vec{0}$ abgebildet werden; das Vektorfeld hat in diesen Schnittpunkten eine Singularität. Alle andern Vektoren kommen als Eigenvektoren nicht in Frage, weil der Radiusvektor $\vec{v}$ und der Bildvektor $A \cdot \vec{v}$ einen Knick haben. ◯

Für nicht diagonalisierbare Matrizen, die also „zu wenig“ linear unabhängige Eigenvektoren haben, findet man genügend sogn. *Hauptvektoren*, mit deren Hilfe man zwar keine Diagonalform, sondern die allgemeinere Jordan'sche Normalform $J(A)$ herstellen kann.

Beispiel. In obigem Beispiel wird der zusätzliche Hauptvektor durch den Vektor

$$\vec{v}^{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

in Richtung der y-Achse beschrieben. Man beachte, dass es sich dabei *nicht* um einen Eigenvektor handeln kann, weil der Radiusvektor $\vec{v}^{(2)}$ und der Bildvektor $A \cdot \vec{v}^{(2)}$ einen Knick von 90° haben. Aus der Figur geht allerdings hervor, dass dieser Hauptvektor die Eigenschaft hat, dass sein Bild $A \cdot \vec{v}^{(2)}$ horizontal verläuft, d.h. im Eigenraum V_0 liegt. Tatsächlich ist

$$N_2 \cdot \vec{v}^{(2)} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \vec{v}^{(1)}$$

Dieser Hauptvektor kann als verallgemeinerter Eigenvektor betrachtet und an Stelle des fehlenden Eigenvektors verwendet werden. Dabei wird als Transformationsmatrix Xdie Einheitsmatrix E_2 entstehen und die nilpotente Ausgangsmatrix N_2 ist bereits in Normalform d.h. es ist $J(N_2) = N_2$. Man beachte, dass statt des gewählten Hauptvektors $\vec{v}^{(2)}$ auch der Vektor

$$\vec{v}^{(1)} + \vec{v}^{(2)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

benutzt werden könnte, da er die selben Eigenschaften hat. Diese Wahl würde dann zwar eine andere invertierbare Transformationsmatrix $X = E_2+N_2$ liefern, mit der aber die selbe Jordan'sche Normalform $J(N_2) = X^{-1} \cdot N_2 \cdot X$ entstanden wäre. ◯

Ein weiteres Beispiel ähnlichen Art ergibt sich aus der soeben benutzten Transformationsmatrix.

Beispiel. Die Scherung in Richtung der x-Achse mit dem Scherungsfaktor 1 ist umkehrbar und wird durch die Matrix

$$Z = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = E_2 + N_2$$

beschrieben. Ihre Potenzen Z^k lassen sich geometrisch leicht bestimmen, da es sich einfach um die Scherung in Richtung der x-Achse mit dem Scherungsfaktor k handelt, die durch die Matrix

$$Z^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} = E_2 + kN_2$$

beschrieben wird. Das folgt algebraisch aus der Umstand, dass Z als Summe der beiden Matrizen E_2 und N_2 zerlegen lässt, die kommutieren. Daher liefern die Binomische Formel und die Nilpotenz von N_2 tatsächlich die Beziehung

$$Z^k = (E_2 + N_2)^k = E_2 + kN_2$$

Aus geometrischen Gründen ist klar, dass der Vektor

$$\vec{v}^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

in Richtung der Scherungsachse unter dieser Scherung fix bleibt und daher ein Eigenvektor zum Eigenwert $\lambda = 1$ ist. Geometrisch ist ferner klar, dass die Vektoren in Richtung der Scherachse die einzigen Vektoren sind, die unter dieser Scherung fix bleibt, weil alle andern Vektoren beim Scheren ihre Richtung ändern. Da die Matrix Z bereits Stufenform hat, ist ihr Minimalpolynom

$$\mu_A(\lambda) = (1 - \lambda)^2$$

und als Eigenwerte kommt nur 1 in Frage. Weil der zugehörige Eigenraum

$$V_1 = \left\{ t \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

aus den Fixpunkten der Abbildung besteht und eindimensional ist, kann Z nicht diagonalisierbar. Das zugehörige Vektorfeld auf dem Einheitskreis sieht wie folgt aus:
Diesmal gilt für den zusätzlichen Hauptvektor in Richtung der y-Achse

$$\vec{v}^{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

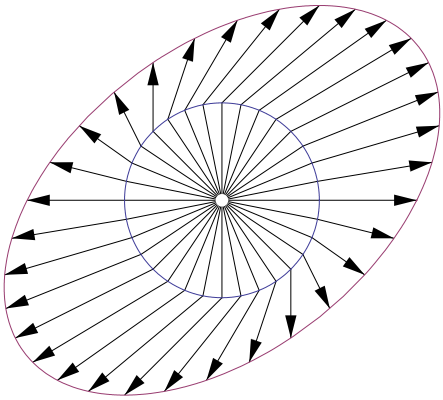


Abbildung 4.11: Vektorfeld einer Scherung in Richtung der x-Achse.

Er kann wiederum sicher kein Eigenvektor sein, da er mit dem zugehörigen Radiusvektor einen Knick von 45° aufweist. Bildet man ihn mit Z ab, erhält man den gescherten Vektor

$$Z \cdot \vec{v}^{(2)} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \vec{v}^{(2)} + \vec{v}^{(1)}$$

aus der folgt, dass der zweite Hauptvektor unter Z tatsächlich nicht in den durch sich selbst aufgespannten Teilraum abgebildet wird. Sie besagt jedoch, dass er in den Teilraum abgebildet wird, der durch $\vec{v}^{(2)}$ *und* durch $\vec{v}^{(1)}$ aufgespannt wird. Die gefundene Beziehung lässt sich durch die allgemeine Eigenwertgleichung

$$(Z - E) \cdot \vec{v}^{(2)} = \vec{v}^{(1)}$$

ausdrücken. Ihre allgemeinen Lösung liefert den Lösungsraum, der die Form

$$V_1^{(2)} = \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

hat und parallel zum Eigenraum V_1 ist. Verwendet man diesen Hauptvektor als verallgemeinerter Eigenvektor an Stelle des fehlenden Eigenvektors, so entsteht als Transformationsmatrix X die Einheitsmatrix E_2 und die Ausgangsmatrix Z ist bereits in Normalform d.h. es ist $J(Z) = Z$. ◯

Falls die Matrix A nicht diagonalisierbar ist, kann man ihre Jordan'sche Normalform $J(A)$ verwenden, um ihre Potenzen zu berechnen, weil nach endlich vielen Schritten der nilpotente Anteil ausstirbt und dann nur noch der diagonalisierbare Anteil übrigbleibt

Beispiel. Wir untersuchen das Eigensystem der Matrix

$$A = \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix}$$

Dazu gehen wir von der charakteristischen Matrix aus.

$$A - \lambda E = \begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 2-\lambda & -4 \\ 0 & 1 & 6-\lambda \end{pmatrix}$$

Wir müssen dafür sorgen, dass das System $(A - \lambda E) \cdot \vec{x} = \vec{0}$ eine nicht triviale Lösung hat. Dazu lösen wir das Gleichungssystem mit dem Eliminationsverfahren. Um keine unnötigen Fallunterscheidungen durchführen zu müssen, vertauschen wir die letzten beiden Zeilen und erhalten:

$$\begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 1 & 6-\lambda \\ 0 & 2-\lambda & -4 \end{pmatrix} \qquad \begin{bmatrix} \\ T_{23} \\ T_{32} \end{bmatrix}$$

Addition des $(\lambda-2)$-Vielfachen der zweiten Zeile zur dritten Zeile ist unbedenklich und liefert die obere Dreiecksmatrix.

$$S(\lambda) = \begin{pmatrix} 4-\lambda & 1 & 2 \\ 0 & 1 & 6-\lambda \\ 0 & 0 & -(\lambda-4)^2 \end{pmatrix} \qquad \begin{bmatrix} \\ \\ Z_{23}(\lambda-2) \end{bmatrix}$$

Das charakteristische Polynom $\chi_A(\lambda) = -(\lambda-4)^3 = 64-48\lambda+12\lambda^2-\lambda^3$ hat die gesuchten Eigenwerte als Nullstellen. Daher kann λ nur den Wert 4 annehmen. Das Spektrum unserer Matrix ist also die $\sigma_A = \{4\}$. Es ist

$$\chi_A(A) = 64E - 48A + 12A^2 - A^3 = 0$$

Ein zweiter Blick auf die erhaltene obere Dreiecksmatrix zeigt allerdings, dass auch hier ein Polynom kleineren Grades mit dieser Eigenschaft existiert. Statt des charakteristischen Polynoms benutzt man besser das quadratische Minimalpolynom

$$\mu_A(\lambda) = -(\lambda-4)^2 = -16 + 8\lambda - \lambda^2$$

Es gilt nämlich tatsächlich

$$\mu_A(A) = -16E + 8A - A^2 = 0$$

Das Minimalpolynom ist ein Teiler des charakteristischen Polynoms.

Um nun die zugehörigen Eigenvektoren zu finden, setzen wir den Eigenwert in die gefundene Dreiecksmatrix ein und lösen das entstehende lineare Gleichungssystem durch Elimination.

Für $\lambda = 4$ lautet die obere Dreiecksmatrix

$$S(4) = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

Schreiben wir seine Lösung in vektorieller Form, erhalten wir den Eigenraum zum Eigenwert 4:

$$V_4 = \left\{ t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix} \mid t, s \in \mathbb{R} \right\}$$

Der Eigenwert $\lambda = 4$ ist ausgeartet und hat die geometrische Vielfachheit 2, weil der Raum V_4 die Dimension 2 hat und durch die beiden Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}_2 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$$

aufgespannt ist. Weil keine weiteren unabhängigen Eigenvektoren mehr zur Verfügung stehen, ist die Matrix A nicht diagonalisierbar.

Um die fehlenden Eigenvektoren durch Hauptvektoren zu ergänzen, untersuchen wir nun statt der bereits diskutiereten Eigenwertgleichung

$$A \cdot \vec{v}^{(1)} = \lambda \vec{v}^{(1)}, \qquad \text{bzw.} \qquad (A - \lambda E) \cdot \vec{v}^{(1)} = \vec{0}$$

die *verallgemeinerte Eigenwertgleichung*

$$(A - \lambda E) \cdot \vec{v}^{(2)} = \vec{v}^{(1)}$$

Hier können wir für $\vec{v}^{(1)}$ weder den einen noch den anderen der beiden soeben gewählten Basisvektoren des Eigenraumes V_4 brauchen, weil das zugehörige lineare Gleichungssystem nach Konstruktion keine Lösung hat. Deshalb verwenden wir in Zukunft als Basis von V_4 die neuen[8] Eigenvektoren

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \vec{v}^{(1)} = \vec{v}_1 + \vec{v}_2 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

Damit und mit dem bereits bestimmten Eigenwert $\lambda = 4$ kann die verallgemeinerte Eigenwertgleichung $(A - 4E) \cdot \vec{v}^{(2)} = \vec{v}^{(1)}$ in matrizieller Form durch die erweiterte Matrix

$$\left(\begin{array}{ccc|c} 0 & 1 & 2 & 1 \\ 0 & -2 & -4 & -2 \\ 0 & 1 & 2 & 1 \end{array}\right), \qquad \left(\begin{array}{ccc|c} 0 & 1 & 2 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 2 & 1 \end{array}\right) \quad \left[\begin{array}{c} S_2(-\frac{1}{2}) \end{array}\right]$$

beschrieben wird. Um mit möglichst kleinen ganzen Zahlen weiterrechnen zu können, haben wir die zweite Zeile durch (-2) dividiert. Addition des (-1)-fachen der ersten Zeile zu den anderen beiden Zeilen liefert die Matrix

$$\left(\begin{array}{ccc|c} 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \quad \left[\begin{array}{c} \\ Z_{12}(-1) \\ Z_{13}(-1) \end{array}\right]$$

Ihre Lösung hat die vektorielle Form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$$

[8] Der Leser überzeuge sich, dass die Summe $\vec{v}^{(1)}$ der beiden gefundenen Eigenvektoren wirklich auch wieder ein Eigenvektor zum Eigenwert 4 ist. Das hängt konzeptionell damit zusammen, dass der Eigenraum V_4 unter Linearkombinationen abgeschlossen ist.

Weil wir nur irgend eine Lösung brauchen, wählen wir der Einfachheit halber $s = t = 0$ und benutzen für den gesuchten Hauptvektor der Stufe 2

$$\vec{v}^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Für ihn ist gilt also die verallgemeinerte Eigenwertgleichung

$$(A - \lambda E) \cdot \vec{v}^{(2)} = \vec{v}^{(1)}, \qquad A \cdot \vec{v}^{(2)} = \lambda \vec{v}^{(2)} + \vec{v}^{(1)}$$

wie man leicht kontrolliert.

Fassen wir die gefundenen verallgemeinerten Eigenvektoren zu einer Matrix zusammen, erhalten wir die Matrix

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Der Eigenraum V_4 ist zweidimensional und die gewählte Basis liefert die ersten beiden Spalten $\vec{v}_1$ und $\vec{v}^{(1)}$ von X. In der dritten Spalte steht der gefundenen Hauptvektor $\vec{v}^{(2)}$.

Diese Matrix X wird nun invertiert. In unserem Beispiel erhalten wir die Matrix

$$X^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

Das Produkt $X^{-1} \cdot A \cdot X$ liefert für die gesuchte Normalform von A die Matrix

$$J(A) = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}$$

Es in diesem entarteten Beispiel keine Diagonalmatrix, sondern die leicht gestörte Jordan-Matrix

$$J(A) = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \Lambda + N$$

Zusätzlich zu einer Diagonalmatrix Λ, in deren Diagonalen der Eigenwert steht, erscheint hier noch eine nilpotente Matrix N als Störung, deren Quadrat $N^2 = 0$ verschwindet und die aber mit der Diagonalmatrix Λ kommutiert. Es ist nämlich

$$\Lambda \cdot N = N \cdot \Lambda = 4N$$

wie man leicht durch Nachrechnen überprüfen kann. In diesem Beispiel haben wir die gekoppelte Information der Matrix A in die einzelnen Achsenrichtungen nicht komplett entkoppeln können. Lösen wir die Gleichung $X^{-1} \cdot A \cdot X = J(A)$ nach der gegebenen Matrix auf, erhalten wir die Faktorisierung (Jordanzerlegung)

$$A = X \cdot J(A) \cdot X^{-1},$$

wobei $J(A)$ die Jordan-Matrix ist.

Entscheidend für die beabsichtigte Anwendung zum Potenzieren der Matrix A, bzw. zum Berechnen des Propagators ist die Tatsache, dass die Matrix $J(A)$ immer noch einfach zum Potenzieren ist, obwohl sie nicht ganz diagonal ist. Auf Grund der Binomischen Formel ist wegen der Nilpotenz von N

$$J^k = (\Lambda + N)^k = \Lambda^k + k\Lambda^{k-1} \cdot N = \begin{pmatrix} 4^k & 0 & 0 \\ 0 & 4^k & k4^{k-1} \\ 0 & 0 & 4^k \end{pmatrix}$$

Damit ergibt sich mit der Gleichung $A^k = X \cdot J^k \cdot X^{-1}$ die gesuchte Potenz

$$\begin{aligned} A^k &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 4^k & 0 & 0 \\ 0 & 4^k & k4^{k-1} \\ 0 & 0 & 4^k \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 4^k & k4^{k-1} & 2k4^{k-1} \\ 0 & 4^k - 2k4^{k-1} & -k4^k \\ 0 & k4^{k-1} & 4^k + 2k4^{k-1} \end{pmatrix} \end{aligned}$$

wie man durch Einsetzen in die Rekursionsgleichung der Potenzen kontrollieren kann. Beim Potenzieren von Matrizen spielen neben den Exponentialfunktionen und den Kreisfunktionen auch Polynome eine zentrale Rolle ◯

Kapitel 5

Ungenaue Körper

Viele Anwendungen (Netzwerke, finite Elemente, Ausgleichsrechnung, Tomographie, Differentialgleichungen, Poisson-Problem, Bildverarbeitung etc.) liefern sehr grosse Matrizen bzw. lineare Gleichungssysteme (tausende von Variablen), die dann maschinell gelöst werden müssen. Die zur Lösung verwendeten Algorithmen basieren meistens auf dem Eliminationsverfahren oder sind iterativer Natur. Eine detailierte Diskussion der konkreten numerischen Lösung linearer Gleichungssystemen liegt ausserhalb unseres Weges. Einige Bemerkungen über praktische Aspekte, mögliche Schwierigkeiten und Auswege müssen also genügen. Wir wollen uns für den Rest dieses Abschnittes der besseren Übersicht wegen auf lineare Gleichungssysteme mit regulärer (d.h. quadratischer, invertierbarer) Koeffizientenmatrix A vom Typ $n \times n$ beschränken.

5.1 Reduktion des Speicherplatzes

Die in der Praxis vorkommenden Matrizen haben häufig eine grosse Anzahl Nullen. Solche Matrizen heissen *dünn besetzt*. Insbesondere treten im Zusammenhang mit der Diskretisierung kontinuierlicher Phänomene z.B. von Differentialgleichungen oder der Methode der finiten Elemente, meistens Bandmatrizen auf. Asymmetrische Prozesse werden bei geeigneter Nummereierung der Zustände durch Dreiecksmatrizen kodiert. Es ist natürlich ineffizient, in einem Computerprogramm alle diese Nullen zu speichern und zu manipulieren. Für die praktische Anwendung ist es also sehr wichtig, dass man die besondere Gestalt solcher Matrizen ausnutzt und beispielsweise Bandmatrizen oder symmetrische Matrizen nicht in einem zweidimensionalen Array, sondern in einer speziellen Datenstruktur, z.B. in einer verketteten Liste, welche neben den wesentlichen Elementen auch deren Position enthält, speichert. Bei dünn besetzten oder symmetrischen Matrizen sind spezielle Algorithmen für das Lösen von linearen Gleichungssystemen erforderlich, die darauf Rücksicht nehmen, dass die dünn besetzten Elemente nicht im Laufe der Rechnung aufgefüllt werden bzw. die Symmetrie zerstört wird. Solche Algorithmen existieren insbesondere für Bandmatrizen und für symmetrische Matrizen. Sie helfen, Speicherplatz und Rechenzeit zu sparen.

5.2 Komplexität der Matrizenmultiplikation

Wir interessieren uns für den Rechenaufwand für die bisher vorgestellten zentralen Algorithmen der linearen Algebra. Zur Vereinfachung zählen wir nur die Grundoperationen und fassen Additionen und Subtraktionen zur Grundoperation Addition zusammen. Entsprechend fassen wir Multiplikationen und Division zur Grundoperation Multiplikation zusammen. Heutige Rechner benötigen etwa $2 \cdot 10^{-6}$ [s] pro Multiplikation und $0.5 \cdot 10^{-6}$ [s] pro Addition.

Um eine Vorstellung davon zu haben, wie gross der Rechenaufwand zur Berechnung eines Matrizen-Produktes ist, benutzen wir das folgende Programm-Fragment zur Berechnung des Matrizen-Produktes $C = A \cdot B$ zweier $n \times n$-Matrizen A, B:

```
For i := 1 to n do
    begin
    For j := 1 to n do
        begin
        t := 0
        For k :=1 to n do
            begin
            t := t + a_ik · b_kj
            end
        c_ij := t
        end
    end
```

Da jedes der n^2 Elemente in der Produkt-Matrix C mit Hilfe von n Multiplikationen berechnet wird, sind auf diese Art n^3 Multiplikationen erforderlich, um zwei $n \times n$-Matrizen miteinander zu multiplizieren.

Jedermann glaubte, dass für die Multiplikation zweier $n \times n$-Matrizen unbedingt n^3 Multiplikationen erforderlich seien. Man war deshalb sehr überrascht, als 1968 V. Strassen einen Algorithmus publizierte, der mit weniger Multiplikationen auskommt.

Die Idee dieses Algorithmus besteht darin, den Umfang des Problems zu Halbieren; dies entspricht der Zerlegung jeder der Matrizen in Viertel, jedes vom Typ $\frac{n}{2} \times \frac{n}{2}$. (Teile und Herrsche) Das Problem ist damit zur Multiplikation von 2×2-Matrizen äquivalent. Nun gelang es Strassen, eine Möglichkeit zu finden, die Elemente so zu kombinieren, dass sich die Anzahl Multiplikationen, die für die Multiplikation von 2×2-Matrizen erforderlich sind, von 8 auf 7 reduziert. Die Anordnung und die Terme, die dafür benötigt werden, sind kompliziert. Um das Produkt

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

zu berechnen, bildet man zunächst folgende Ausdrücke:

$$\begin{aligned} m_1 &= (a_{12} - a_{22})(b_{21} + b_{22}) \\ m_2 &= (a_{11} + a_{22})(b_{11} + b_{22}) \end{aligned}$$

$$\begin{aligned} m_3 &= (a_{11} - a_{21})(b_{11} + b_{12}) \\ m_4 &= (a_{11} + a_{12})b_{22} \\ m_5 &= a_{11}(b_{12} - b_{22}) \\ m_6 &= a_{22}(b_{21} - b_{11}) \\ m_7 &= (a_{21} + a_{22})b_{11} \end{aligned}$$

Nun berechnet man die gesuchte Matrix mit Hilfe von:

$$\begin{aligned} c_{11} &= m_1 + m_2 - m_4 + m_6 \\ c_{12} &= m_4 + m_5 \\ c_{21} &= m_6 + m_7 \\ c_{22} &= m_2 - m_3 + m_5 - m_7 \end{aligned}$$

Weil eine $2n \times 2n$-Matrix in vier $n \times n$-Matrizen (Blockmatrizen) zerlegt werden kann, kann man die Idee rekursiv benutzen um das Produkt von zwei $2^k \times 2^k$-Matrizen mit 7^k statt mit den üblichen $(2^k)^3 = 8^k$ Multiplikationen zu berechnen. Mit Hilfe dieser Überlegung kann man zwei $n \times n$-Matrizen unter Verwendung von $n^{\log_2(7)} = n^{2.82}$ Multiplikationen multiplizieren.

Es ist wichtig festzustellen, dass wir nur die Anzahl Multiplikationen gezählt haben. Bevor der Algorithmus für eine praktische Anwendung in Frage kommt, müssen auch die Kosten der zusätzlichen Additionen (im Falle der 2×2-Matrizen sind es immerhin 18 statt der üblichen 4) und die Kosten der rekursiven Aufrufe berücksichtigt werden. Diese Buchhaltungskosten hängen stark von der speziellen Implementation oder vom verwendeten Computer ab. Dieser zusätzliche Ballast bewirkt, dass die Methode von Strassen für kleine Matrizen weniger effizient ist als die Standardmethode. Numerische Experimente zeigten, dass die Methode erst für $n \geq 40$ eine Verbesserung bringt. Selbst für grosse Matrizen ist die Verbesserung der Methode von Strassen jedoch kaum spürbar. Bei $n = 100$ lassen sich mit dem Algorithmus von Strassen etwa 7% der Rechenzeit einsparen.

Der „beste“ Algorithmus für die Multiplikation von Matrizen ist bis heute nicht gefunden worden. Es handelt sich hier um eines der bekanntesten noch immer offenen Probleme der Informatik. Inbesondere kennt man bisher keinen Grund, warum sich das Produkt zwei Matrizen aus $\mathbb{R}^{n,n}$ nicht mit n^2 Multiplikationen berechnen lässt.

5.3 Komplexität der Elimination

Um eine Vorstellung vom Rechenaufwand beim Lösen eines linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ zu erhalten, müssen wir die Anzahl der dabei jeweils ausgeführten Operationen bestimmen.

Man sieht leicht, dass die Berechnung der Lösung via Stufenform und dann Rückwärtseinsetzen (Gauss Elimination) oder via reduzierter Stufenform (Verfahren von Gauss-Jordan) die gleiche Anzahl Operationen benötigt. Beide Verfahren beginnen mit der Reduktion der erweiterten Matrix auf Stufenform. Da die zweite Phase (Rückwärtseinsetzen bzw. berechnen der reduzierten Stufenform) die gleiche Anzahl Operationen benötigen, sind beide Verfahren gleich

schnell. Den notwendigen Aufwand an Rechenoperationen entnimmt man folgender Tabelle.

Daneben haben wir noch das Lösungsverfahren mit Hilfe der inversen Matrix angetroffen: $\vec{x} = A^{-1} \cdot \vec{b}$. Mit diesem Verfahren werden in vielen Taschenrechner reguläre lineare Gleichungssystem gelöst. Dazu muss zunächst die Inverse berechnet werden. Sorgfältiges Zählen der erforderlichen Operationen zeigt, dass der vorgestellte Algorithmus zur Berechnung der Inversen einer invertierbaren $n \times n$-Matrix einen Aufwand von $n^3 - 2n^2 + n$ Additionen und von n^3 Multiplikationen erfordert. Die Multiplikation mit dem Konstantenvektor $\vec{b}$ erfordert noch zusätzliche $n(n-1)$ Additionen und n^2 Multiplikationen, so dass also insgesamt der in der folgenden Tabelle aufgeführte Aufwand erforderlich ist.

Ferner sind wir noch einem weiteren Verfahren zur Lösung von regulären linearen Gleichungssystemen — der sogn. Cramer'schen Regel — begegnet.

Der folgende Satz ermöglicht einen Vergleich des Rechenaufwandes für die drei Lösungsverfahren.

Satz. Die Anzahl Rechenoperationen für das Lösen eines linearen Gleichungssystems mit n Gleichungen und n Unbekannten nach dem Eliminationsverfahren, mit Hilfe der Inversen bzw. nach der Cramer'schen Methode beträgt:

Methode	Additionen	Multiplikationen
Gauss	$\frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n$	$\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$
Inverse	$n^3 - n^2$	$n^3 + n^2$
Cramer	$\frac{1}{3}n^4 - \frac{1}{6}n^3 - \frac{1}{3}n^2 + \frac{1}{6}n$	$\frac{1}{3}n^4 + \frac{1}{3}n^3 + \frac{2}{3}n^2 + \frac{2}{3}n - 1$

Dabei werden Subtraktionen als Additionen und Divisionen als Multiplikationen gezählt.

Man merke sich, dass das Eliminationsverfahren einen Aufwand von $\sim \frac{n^3}{3}$ Additionen und Multiplikationen erfordert. Die Berechnung mit Hilfe der Inversen erfordert etwa den dreifachen Aufwand und die Cramer'sche Regel ist praktisch unbrauchbar. Insbesondere liefert also das Eliminationsverfahren eine polynomiale Methode zur Lösung eines linearen Gleichungssystems. Ferner sind alle erforderlichen Operationen rationaler Art.

Es ist naheliegend, die Frage zu stellen, ob es Algorithmen gibt, die wesentlich schneller sind, als das Eliminationsverfahren. Tatsächlich wurden in den letzten Jahren Algorithmen entwickelt, die mit weniger Multiplikationen auskommen. Man kann zeigen, dass das Lösen linearer Gleichungssysteme zur Multiplikation von Matrizen äquivalent ist. Daher existieren also Algorithmen (z.B. der früher skizzierte Algorithmus von Strassen), mit denen Systeme von n Gleichungen in n Unbekannten in einer zu $n^{\log_2(7)}$ proportionalen Zeit gelöst werden können. Allerdings werden sie in der Praxis kaum eingesetzt, da sie schwierig zu programmieren sind und sehr viele Additionen benötigen. Man kann also zur Zeit davon ausgehen, dass keine wesentlich besseren Verfahren als die Gauss-Elimination existieren.

5.4 Iterative Verfahren

Oft benötigt man für ein lineares Gleichungssystem gar keine exakten Lösungen oder eine Einsicht in das betreffende System, weil z.B. die Komponenten des Konstantenvektors gemessene Grössen sind. Um approximative Lösungen zu finden, die man bei Bedarf verbessern kann, verwendet man eines der klassischen iterativen Näherungsverfahren: Statt direkt die exakte Lösung des Gleichungssystems zu bestimmen, macht man mit möglichst wenig Aufwand aus einer angenäherten Lösung des Problems eine besser angenäherte. Die Hoffnung besteht, dass die Folge dieser Vektoren gegen die exakte Lösung konvergiert. Bei iterativen Verfahren stellen sich also Konvergenzfragen und Fragen der Konvergenzgeschwindigkeit, die theoretisch ganz schön heikel sein können. Konvergente iterative Verfahren erweisen sich dafür als recht unempfindlich gegen Rundungsfehler.

Das folgende einfach zu beschreibende iterative Verfahren für das Lösen des linearen Gleichungssystems $A \cdot \vec{x} = \vec{b}$ stammt von Jacobi (1804 – 1896) und ist speziell geeignet, falls die Matrix A diagonal-dominant ist.

Definition. Die $n \times n$-Matrix A heisst *diagonal-dominant*, falls in jeder Zeile der Betrag des Diagonalelementes strikt grösser als die Summe der Beträge der anderen Elemente ist.

Gleichungssysteme, die von Anwendungen (Differentialgleichungen, finite Elemente) herrühren, haben oft diese Eigenschaft. Man kann zeigen, dass eine diagonal-dominante Matrix invertierbar ist.

Beispiel. Um das Verfahren an einem konkreten Beispiel zu sehen, gehen wir aus von folgendem linearen Gleichungssystem mit 3 Unbekannten:

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{array}$$

Durch Auflösen nach x_i erhalten wir die sogn. *Fixpunktgleichung*:

$$\begin{array}{rcrcrcl} x_1 & = & \frac{b_1}{a_{11}} & - & \frac{a_{12}}{a_{11}}x_2 & - & \frac{a_{13}}{a_{11}}x_3 \\ x_2 & = & \frac{b_2}{a_{22}} & - & \frac{a_{21}}{a_{22}}x_1 & - & \frac{a_{23}}{a_{22}}x_3 \\ x_3 & = & \frac{b_3}{a_{33}} & - & \frac{a_{31}}{a_{33}}x_1 & - & \frac{a_{32}}{a_{33}}x_2 \end{array}$$

Nun wendet man diese Fixpunktgleichung zunächst auf eine grobe Schätzung $\vec{x}(0)$ der Lösung an und erhält eine erste Näherung $\vec{x}(1)$ der Lösung; dann wendet man auf diese Näherung das Verfahren wieder an und erhält die zweite Näherung $\vec{x}(2)$. So weiterfahrend erhält man eine Vektorfolge $\vec{x}(k)$, die man so lange verfolgt, bis eine hinreichend genaue Näherung entstanden ist. Ganz im Sinne von Gauss, der einem seiner Schüler dazu schrieb:

> So fahre ich fort, bis nichts mehr zu corrigieren ist. Ich empfehle Ihnen diesen Modus zur Nachahmung. Schwerlich werden sie je wieder direct eliminiren, wenigstens nicht, wenn Sie mehr als 2 Unbekannte haben. Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken.

Offenbar ist der Programmieraufwand für das Jacobi-Verfahren sehr gering. Man beachte ferner, dass die einzelnen Rechenschritte von einander unabhängig sind. Damit ist dieses Verfahren für den Einsatz auf Parallelrechnern gut geeignet. ◯

Für theoretische Untersuchungen ist es zweckmässig, das Jacobi-Verfahren matriziell zu beschreiben. Dazu gehen wir vom Gleichungssystem $A \cdot \vec{x} = \vec{b}$ aus und zerlegen die Matrix A als Summe

$$A = U + D + O$$

wobei U untere und O obere Dreiecksmatrix sowie D Diagonalmatrix sind. Setzt man diese Zerlegung in das Gleichungssystem ein, ergibt sich:

$$A \cdot \vec{x} = (U + D + O) \cdot \vec{x} = U \cdot \vec{x} + D \cdot \vec{x} + O \cdot \vec{x} = \vec{b}$$

Durch Umformen erhält man:

$$D \cdot \vec{x} = \vec{b} - (U + O) \cdot \vec{x}$$

Falls die Matrix A digonal-dominant ist, sind ihre Diagonalelemente von Null verschieden und die Matrix D ist invertierbar. Durch Multiplikation mit D^{-1} erhält man schliesslich die *Fixpunktgleichung*:

$$\vec{x} = D^{-1} \cdot \vec{b} - D^{-1}(U + O) \cdot \vec{x}$$

Man sieht leicht, dass das gegebene lineare Gleichungssystem und diese Fixpunktgleichung die selben Lösungen haben. Wir wählen also für $\vec{x}(0)$ irgend einen Anfangsvektor, den wir mit Hilfe der Rekursion:

$$\vec{x}(k+1) = D^{-1} \cdot \vec{b} - D^{-1}(U + O) \cdot \vec{x}(k)$$

zu verbessern versuchen. Der ermittelte Vektor $\vec{x}(k)$ wird in diese Rekursiongleichung so lange eingesetzt, bis sich die Vektoren $\vec{x}(k+1)$ und $\vec{x}(k)$ hinreichend wenig unterscheiden.

Mit Hilfe des sogn. Banach'schen Fixpunktsatzes kann man zeigen, dass das Verfahren gegen die korrekte Lösung konvergiert, falls die Koeffizientenmatrix A diagonal-dominant ist.

Beispiel. Um die matrizielle Beschreibung am konkreten Beispiel zu sehen, gehen wir aus von der 3×3-Matrix und ihrer Zerlegung:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

Durch Nachrechnen findet man hier in der Tat die behauptete Fixpunktgleichung:

$$\begin{array}{lcccccc} x_1 & = & \frac{b_1}{a_{11}} & - & \frac{a_{12}}{a_{11}} x_2 & - & \frac{a_{13}}{a_{11}} x_3 \\ x_2 & = & \frac{b_2}{a_{22}} & - & \frac{a_{21}}{a_{22}} x_1 & - & \frac{a_{23}}{a_{22}} x_3 \\ x_3 & = & \frac{b_3}{a_{33}} & - & \frac{a_{31}}{a_{33}} x_1 & - & \frac{a_{32}}{a_{33}} x_2 \end{array}$$

◯

Im Laufe der Jahre sind weitere Iterationsverfahren entwickelt worden, die alle ihr Vor- und Nachteile haben. Insbesondere kann mit Hilfe von sogn. Mehrgitterverfahren die Konvergenz beträchtlich beschleunigt werden. Die Details findet man wiederum in der Fachliteratur zur numerischen linearen Algebra, etwa in [**?**].

5.5 Vermeiden von Rundungsfehlern

Ein Thema von ebenfalls grosser praktischer Bedeutung beim Rechnen in ungenauen Körpern ist der Einfluss von Überlauf und Rundungsfehlern.

Dem mit einem De-Luxe-Modell eines Taschenrechners ausgestatteten Leser, der müde lächelnd die Schulter zuckt und glaubt, numerische Lotterie-Ergebnisse könnten ihm mit seiner Maschine nicht passieren, sei dringend empfohlen, folgendes lineares Gleichungssystem mit seiner Maschine im Gleitkommazahlen-Modus zu lösen:

$$\left\{\begin{array}{rcrcrcrcl} -367'296x_1 & - & 43'199x_2 & + & 519'436x_3 & - & 954'302x_4 & = & 1 \\ 259'718x_1 & - & 477'151x_2 & - & 367'295x_3 & - & 1'043'199x_4 & = & 1 \\ 886'731x_1 & + & 88'897x_2 & - & 1'254'026x_3 & - & 1'132'096x_4 & = & 1 \\ 627'013x_1 & + & 566'048x_2 & - & 886'732x_3 & + & 911'103x_4 & = & 0 \end{array}\right.$$

Anschliessend löse man dieses Gleichungssystem von Hand, indem man möglichst lange ganzzahlig rechne. Für die Berechnung des grössten gemeinsamen Teilers der jeweiligen Zeilen, durch den man nach jeder Elementaroperation zur Vermeidung von sehr grossen Zahlen divieren sollte, leistet der Euklid'sche Algorithmus wertvolle Dienste.

Im Gegensatz zur symbolischen Behandlung oder zur Rechnung mit ganzen Zahlen, hat bei numerischer Rechnung die Reihenfolge der Eliminationsschritte einen Einfluss auf das Eliminationsverfahren. In der Praxis ist es beim Rechnen mit ungenauen Zahlen ratsam, mehr zu tun, als nur eine Zeile mit einem von Null verschiedenen Wert zu finden. Der Grund dafür ist, dass bei der Berechnung erhebliche Rundungsfehler entstehen können.

Beispiel. Ein einfaches Beispiel soll die Problematik demonstrieren. Gegeben sei das reguläre lineare Gleichungssystem mit dem Parameter $\varepsilon \neq 2$:

$$\left\{\begin{array}{rcrcl} \varepsilon x & + & 2y & = & 1 \\ x & + & y & = & 1 \end{array}\right.$$

mit der zugehörigen erweiterten Matrix

$$\left(\begin{array}{cc|c} \varepsilon & 2 & 1 \\ 1 & 1 & 1 \end{array}\right)$$

Addition des $-\frac{1}{\varepsilon}$-fachen der ersten Zeile zur zweiten Zeile liefert bereits die Stufenform:

$$\left(\begin{array}{cc|c} \varepsilon & 2 & 1 \\ 0 & (1-\frac{2}{\varepsilon}) & (1-\frac{1}{\varepsilon}) \end{array}\right)$$

und die symbolische Lösung $y = \frac{\varepsilon-1}{\varepsilon-2}$ sowie $x = -\frac{1}{\varepsilon-2}$. Stellen wir die beiden

Gleichungen um, erhalten wir das Gleichungssystem

$$\left\{\begin{array}{rcrcl} x & + & y & = & 1 \\ \varepsilon x & + & 2y & = & 1 \end{array}\right.$$

mit der zugehörigen erweiterten Matrix

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ \varepsilon & 2 & 1 \end{array}\right)$$

Addition des $-\varepsilon$-fachen der ersten Zeile zur zweiten Zeile liefert hier die Stufenform:

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 2-\varepsilon & 1-\varepsilon \end{array}\right)$$

Seine symbolische Lösung ist wiederum $y = \frac{1-\varepsilon}{2-\varepsilon}$ sowie $x = \frac{1}{2-\varepsilon}$.

Wählen wir konkret $\varepsilon = 10^{-2}$, so erhalten wir als Lösung die beiden Brüche $y = \frac{99}{199} = 0.497\ldots$ und $x = \frac{100}{199} = 0.502\ldots$. Vom Standpunkt der symbolischen Rechnung oder der rationalen Zahlen gibt es keine Probleme. ◯

Rationale Zahlen werden in einem Computer als Paare von ganzen Zahlen exakt gespeichert. Oft wird in der Ingenieur-Mathematik mit „reellen Zahlen“ gerechnet. Da in Maschinen nur endliche Information gespeichert werden kann, müssen die meisten Zahlen gerundet, bzw. abgeschnitten werden. Reelle Zahlen werden häufig in *normierter Gleitkommadarstellung* näherungsweise dargestellt. Bei n-stelliger Gleitkommadarstellung gilt für eine Zahl $x \neq 0$:

$$x = \pm \sum_{i=1}^{n} x_i 10^{-i} \cdot 10^k \quad \text{mit} \quad x_0 \neq 0$$

Der *Exponent* k ist eine ganze Zahl. Die *Mantisse*

$$\frac{1}{10} \leq \sum_{i=1}^{n} x_i 10^{-i} < 1$$

hat eine 0 vor dem Dezimalpunkt und n Stellen dahinter. Der Exponent wird immer so gewählt, dass die erste Nachkommaziffer von Null verschieden ist. Das ist offenbar, ausser bei 0, immer möglich.

Beispiel. Bei 4-stelliger Gleitkommadarstellungen entsprechen sich folgende Zahlen:

Normalform	Gleitkommadarstellung
57	$0.5700 \cdot 10^2$
$94'751$	$0.9475 \cdot 10^5$
$127'300'000$	$0.1273 \cdot 10^9$
$\frac{1}{3}$	$0.3333 \cdot 10^0$
3.14159	$0.3141 \cdot 10^1$
0.00825	$0.8250 \cdot 10^{-2}$

◯

Der Exponent gibt an, um wieviele Stellen der Dezimalpunkt der Mantisse beim Übergang von der Gleitkommadarstellung in die Normalform nach links bzw. nach rechts verschoben werden muss. Beim Übergang in Gleitkommadarstellung müssen die meisten Zahlen gerundet werden. Die Resultate der meisten Operationen sind mit Rundungsfehlern behaftet.

Beispiel. Nehmen wir jetzt an, dass wir eine (ganz primitive) Maschine zur Verfügung haben, die nur 2-stellige Gleitkommarechnungen ausführen kann. Die 3.te Nachkommastelle wird gerundet. Dann lautet die erweiterte Matrix von obigem System für $\varepsilon = 10^{-2}$ im ersten Fall

$$\left(\begin{array}{cc|c} 0.10 \cdot 10^{-1} & 0.20 \cdot 10^{1} & 0.10 \cdot 10^{1} \\ 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} \end{array}\right)$$

Addition des $-0.1 \cdot 10^{3}$-fachen der ersten Zeile zur zweiten Zeile liefert die Stufenform:

$$\left(\begin{array}{cc|c} 0.10 \cdot 10^{-1} & 0.20 \cdot 10^{1} & 0.10 \cdot 10^{1} \\ 0.00 \cdot 10^{0} & -0.20 \cdot 10^{3} & -0.99 \cdot 10^{2} \end{array}\right)$$

Für die gerundete Näherungslösungen ergibt sich $\hat{y} = 0.50 \cdot 10^{0}$ und daraus mit Hilfe von $\hat{x} = \frac{1}{\varepsilon}(1 - 2\hat{y})$ die sehr schlechte Näherungslösung $\hat{x} = 0.00 \cdot 10^{0}$.

Im zweiten Fall vertauschen wir beim ursprünglichen Problem die beiden Zeilen und erhalten die erweiterte Matrix

$$\left(\begin{array}{cc|c} 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} \\ 0.10 \cdot 10^{-3} & 0.20 \cdot 10^{1} & 0.10 \cdot 10^{1} \end{array}\right)$$

Nach Addition des $-0.10 \cdot 10^{1}$-fachen der ersten Zeile zur zweiten Zeile ergibt sich die Stufenform

$$\left(\begin{array}{cc|c} 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} & 0.10 \cdot 10^{1} \\ 0.00 \cdot 10^{0} & 0.20 \cdot 10^{1} & 0.99 \cdot 10^{0} \end{array}\right)$$

mit der Näherungslösung $\tilde{y} = 0.50 \cdot 10^{0}$ und daraus mit Hilfe von $\hat{x} = 1 - \tilde{y}$ die viel bessere Näherungslösung $\tilde{x} = 0.50 \cdot 10^{0}$.

Der geometrische Grund für dieses Verhalten liegt darin, dass bei der Scherung, die zur ersten Elementaroperation gehört, Längen und Winkel extrem verzerrt werden. Das führt dann zu zwei Geraden mit „schleifendem Schnitt“, wie das folgende Bild zeigt.

Das geschilderte Problem ist prinzipieller Art und wird bei jeder Genauigkeit auftreten. Bei n-stelliger Gleitkommaarithmetik (z.B. $n = 12$) rechne man das Beispiel mit dem Wert $\varepsilon = 10^{-n}$ auf beide Arten durch. ◯

Wir sehen also, dass beim Eliminationsverfahren die Wahl des führenden Elementes einen erheblichen Einfluss auf die Rundungsfehler hat. In der numerischen Praxis treten dann Schwierigkeiten mit dem Eliminationsalgorithmus auf, wenn die Grössenordnungen der Elemente der Matrix sehr verschieden sind. Im Beispiel sehen wir, dass es offenbar gefährlich ist, betragsmässig kleine Elemente zum Durchdividieren zu verwenden. Statt also nur zu verlangen, dass das

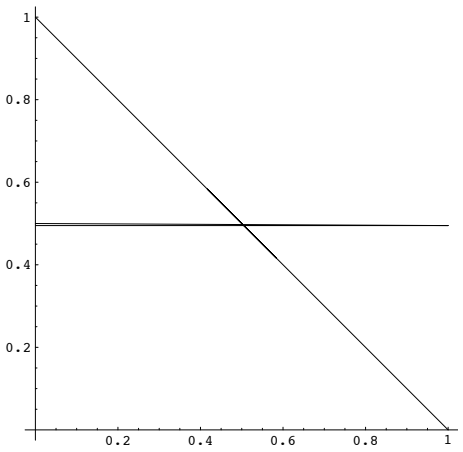


Abbildung 5.1: Schleifender Schnitt als Ursache für Rundungsfehler.

führende Element von Null verschieden ist, benutze man als Pivot-Element jene Zahl in der Pivot-Spalte, die beträgsmässig am grössten ist, indem man eine Zeilenvertauschung durchführe. Aber die Wahl der betragsmässig grösstmöglichen führenden Koeffizienten führt nicht immer zum Erfolg. Über die optimale Wahl gibt die Fachliteratur der numerische Mathematik unter dem Stichwort „Pivot-Strategie“ Auskunft. Dort werden in der Regel auch Spalten vertauscht, was einer Umnummerierung der Unbekannten gleichkommt. Statt einer Umspeicherung der Variablen nimmt man eine sorgfältige Buchhaltung über die durchgeführten Vertauschungen vor.

Eine weitere wichtige praktische Frage betrifft den Einfluss von Rundungsfehlern auf die Lösung eines linearen Gleichungssystems. Auch bei exakter Rechnung kann nämlich die Präzision verloren gehen!

Beispiel. Als Beispiel, das dieses Phänomen demonstrieren soll, gehen wir aus vom linearen Gleichungssystem:

$$\left\{\begin{array}{rcrcl} x & + & y & = & 2 \\ x & + & 1.0001y & = & 2.0001 \end{array}\right.$$

Seine exakte Lösung lautet $x = 1$ und $y = 1$. Nehmen wir nun an, durch Messungenauigkeiten würden wir dazu geführt, stattdessen das folgende, leicht gestörte, lineare Gleichungssystem lösen:

$$\left\{\begin{array}{rcrcl} x & + & y & = & 2 \\ x & + & 1.0001y & = & 2.0002 \end{array}\right.$$

Seine exakte Lösung ist $x = 0$ und $y = 2$. Wir stellen also fest, dass der Unterschied in der vierten Nachkommastelle einer Komponente die Lösung dramatisch — nämlich um den Faktor 10^4 — ändert! Unter Umständen können kleine Änderungen in den Koeffizienten des Systems grosse Änderungen in den Lösungen zur Folge haben. Gleichungssysteme, die auf Rundungsfehler sensibel reagieren,

heissen *schlecht konditioniert*. Das besprochene Beispiel ist geometrisch gesprochen deshalb schlecht konditioniert, weil die beiden zugehörigen Geraden fast parallel sind. Wackeln wir, wie hier, an einer der Geraden etwas, wird sich der Schnittpunkt in der Regel sehr stark verschieben. Rechnerisch lässt sich die schlechte Kondition dieses Systems dadurch feststellen, dass man beobachtet, dass die Koeffizientenmatrix „nahe bei“ einer singulären Matrix ist. Das drückt sich dadurch aus, dass ihre Determinante „fast“ Null ist. Von schlecht konditionierten Systemen lässt sich insbesondere die inverse Koeffizientenmatrix nur schwer berechnen. Schlechte Kondition ist also nicht auf ein fehlerhaftes Lösungsverfahren zurückzuführen, sondern stellt eine wesentliche Eigenschaft des Problems selbst dar. Bei schlecht konditionierten Problemen sind Fehler im Ergebnis durch typischerweise unvermeidbare Fehler der Daten bedingt. Dieser Effekt lässt sich durch kein noch so aufwändiges numerisches Verfahren vermeiden. Schlechte Kondition ist also eine Eigenschaft des Problems, die unabhängig von der verwendeten Lösungsmethode vorhanden ist. Umgekehrt kann aber die numerische Lösung eines gut konditionierten Problems, das also gegen Datenfehler unempfindlich ist, durch einen unsachgemässen Algorithmus vollkommen falsch sein. Zum Beispiel kann eine Elementaroperation ein gut konditioniertes System in ein schlecht konditioniertes überführen. Wie erwähnt, lässt sich eine solche numerische Instabilität durch geeignete Pivot-Strategie vermeiden. ◯

Für das Lösen von linearen Gleichungssystemen mit Gleitkommaarithmetik liegen ausgefeilte Bibliotheksprogramme vor. Den Kode des wohl wichtigsten solchen Programmes mit dem Namen LINPACK kann man im Internet unter der elektronischen Adresse ⟨netlib@research.att.com⟩ beziehen. Dieser Dienst ist beschrieben in [**?**]. Das Handbuch dazu ist erhältlich unter der Referenz [**?**]. Dieser Dienst enthält auch ein optimiertes Bibliotheksprogramm mit dem Namen EISPACK, mit dessen Hilfe man Eigenwerte numerisch berechnen kann. Die zugrundeliegenden Ideen gehören in eine Vorlesung über Numerik.

Noch heikler als das numerische Lösen linearer Gleichungssysteme ist das numerische Lösen des Eigenwertproblems bzw. die numerische Berechnung der Jordanschen Normalform einer Matrix. Falls die Matrix A mehrfache Eigenwerte hat oder “fast” eine Matrix mit degeneriertem Spektrum ist, ist ihre Jordansche Normalform sensitiv auf kleine Störungen.

Beispiel. Um auch diese Phänomen wieder an einem typischen Beispiel zu veranschaulichen, gehen wir von der Matrix

$$A(\varepsilon) = \begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}$$

aus. Die Matrix $A(0)$ hat das Minimalpolynom

$$\mu_{A(0)}(\lambda) = (\lambda - 1)^2 = \lambda^2 - 2\lambda + 1$$

mit dem doppelten Eigenwert $\lambda_{1,2} = 1$ und der Jordanschen Normalform

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

Für $\varepsilon \neq 0$ hat die Matrix $A(\varepsilon)$ das Minimalpolynom

$$\mu_{A(\varepsilon)}(\lambda) = (\lambda - 1)^2 = \lambda^2 - 2\lambda + (1 - \varepsilon)$$

mit den beiden Eigenwerten $\lambda_{1,2} = 1 \pm \sqrt{\varepsilon}$ und ist deshalb diagonalisierbar mit der Jordanschen Normalform

$$\begin{pmatrix} 1+\sqrt{\varepsilon} & 0 \\ 0 & 1-\sqrt{\varepsilon} \end{pmatrix}$$

Die schlechte Konditionierung des Problems erschwert die Suche nach einem robusten numerischen Algorithmus zur Berechnung der Jordanschen Normalform ungemein, weil das Ergebnis wesentlich davon abhängt, ob die beiden Eigenwerte gleich sind oder nicht. Daher vermeiden Numeriker die Jordansche Normalform wenn möglich und verwenden die stabile Schur-Zerlegung. ◯

Aber auch wer ganzzahlig oder symbolisch rechnen will, kann mit dem Eliminationsalgorithmus böse Überraschungen erleben. Die einzelnen Matrixelemente können nämlich im Laufe der Rechnung exponentiell anwachsen.

Beispiel. Dieses Phänomen sieht man deutlich an der Matrix, die lauter 2 auf der Diagonale und 1 unmittelbar unterhalb der Diagonale hat. Alle anderen Elemente seien 0. Als typisches Beispiel betrachte man die Matrix:

$$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

Der Eliminationsalgorithmus liefert dann der Reihe nach folgende Matrizen:

$$\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & -16 \end{pmatrix}$$

Das Anschwellen der Zahlen kann beim Berechnen von grossen linearen Gleichungssystemen Probleme durch Überlauf machen. ◯

Weil Determinanten quadratischer Matrizen als Abfallprodukt des Eliminationsverfahren entstehen, werden beim symbolischen Rechnen mit diesem Algorithmus die Terme mindestens so kompliziert wie die Determinante. Es gibt eine explizite Formel für die Determinante einer $n \times n$-Matrix A, die Leibniz zugeschrieben und in alten Schulbüchern häufig statt der rekursiven Entwicklung nach Zeilen oder Spalten verwendet wird. Es gilt

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sig}(\sigma) \prod_{j=1}^{n} a_{j\sigma(j)}$$

Daraus folgt insbesondere, dass die Determinante einer solchen Matrix insgesamt aus $n!$ Summanden besteht, die zur Permutationsgruppe S_n gehören. Insbesondere wächst also diese Formel für die Determinante exponentiell und damit auch die Grösse der symbolischen Terme im Eliminationsverfahren.

Beispiel. Für die Determinante einer beliebigen 4×4-Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{11} \\ a_{21} & a_{22} & a_{23} & a_{21} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

erhält man die Monster-Formel

$$\begin{aligned}\det(A) \;=\; & a_{14}a_{23}a_{32}a_{41} - a_{13}a_{24}a_{32}a_{41} - a_{14}a_{22}a_{33}a_{41} + a_{12}a_{24}a_{33}a_{41} + \\ & a_{13}a_{22}a_{34}a_{41} - a_{12}a_{23}a_{34}a_{41} - a_{14}a_{23}a_{31}a_{42} + a_{13}a_{24}a_{31}a_{42} + \\ & a_{14}a_{21}a_{33}a_{42} - a_{11}a_{24}a_{33}a_{42} - a_{13}a_{21}a_{34}a_{42} + a_{11}a_{23}a_{34}a_{42} + \\ & a_{14}a_{22}a_{31}a_{43} - a_{12}a_{24}a_{31}a_{43} - a_{14}a_{21}a_{32}a_{43} + a_{11}a_{24}a_{32}a_{43} + \\ & a_{12}a_{21}a_{34}a_{43} - a_{11}a_{22}a_{34}a_{43} - a_{13}a_{22}a_{31}a_{44} + a_{12}a_{23}a_{31}a_{44} + \\ & a_{13}a_{21}a_{32}a_{44} - a_{11}a_{23}a_{32}a_{44} - a_{12}a_{21}a_{33}a_{44} + a_{11}a_{22}a_{33}a_{44}\end{aligned}$$

die tatsächlich aus $4! = 24$ Summanden besteht, die sich ihrerseits je aus 4 Faktoren zusammensetzen. Würde man das Eliminationsverfahren auf die Matrix A zur Invertierung anwenden, würde diese Formel en passant entstehen. Sie lässt sich übrigens nicht mehr mit der in den Schweizer-Schulen leider weit verbreiteten Sarrus-Regel merken, weil sie eben 24 statt der fälschlicherweise erwarteten 8 Summanden enthält! ◯

Kapitel 6

Anwendungen

Der folgende Abschnitt soll die Frage behandeln, wie und wo „grosse“ lineare Gleichungssysteme überhaupt vorkommen.

6.1 Tomographie

Als erstes Beispiel verwenden wir die *Computertomographie*. Hier sieht man deutlich, wie trockene Theorie[1] eines Tages zur gefeierten Praxis werden kann.

Für die Diagnose eines Gehirnschlages oder von (Gehirn-) Tumoren und für andere medizinische Zwecke ist es von grossem Interesse, über ebene *Schnittbilder* des Gehirnes und anderer Körperteile zu verfügen. Solche Bilder sind auch für die zerstörungsfreie Qualitätskontrolle fester Materialien usw. von Bedeutung. Übliche Röntgenaufnahmen stellen, wie eine Photographie, nur Projektionen räumlicher Gebilde auf eine Ebene dar.

Die CT-Bilder stellen keine Röntgenaufnahmen im üblichen Sinne dar. Sie bestehen vielmehr aus tausenden von Pixeln in verschiedenen Graustufen bzw. Farben. Der Grauwert der einzelnen Pixeln wird dabei *errechnet* und danach auf einem Monitor wiedergegeben bzw. auf Film übertragen. Der Grauwert jedes Pixels im Bild entspricht genau der Massendichte im Original. Eine Graustufenskala am Rand des Bildes erlaubt die Ermittlung der konkreten numerischen Werte.

Die ersten praktisch verwendbaren Computertomographen wurden im Jahre 1973 gebaut. Heute existieren Scanner, die auf anderen physikalischen Grundlagen als den Röntgenstrahlen beruhen (z.B. Kernspin-Magnet-Tomographen). Die mathematischen Prinzipien sind allerdings die selben.

Die Idee der Tomographie besteht darin, dass man einen Teilchenstrahl durch das Gehirn sendet. Auf seinem Weg durch das Gehirn verliert der Strahl an Energie (Intensität) durch Absorption. Die Gesamtabsorption des Strahls ist die Differenz der Ausgangsintensität I und der Eingangsintensität I_0 und kann gemessen werden. Für den Energieverlust gilt also:

$$b = I - I_0$$

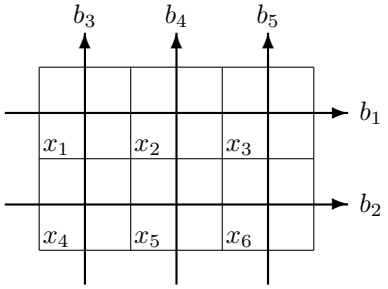
[1] Der Mathematiker J. Radón beschäftigte sich 1917 mit diesen Fragen.

Zur Diskretisierung des Problems denken wir uns auf das interessierende Gebiet genügend feines Häuschenpapier gelegt, d.h. wir zerlegen es in quadratische Zellen Z_i für $1 \leq i \leq n$. Der Energieverlust ist dann die Summe der Energieverluste E_i, die der Strahl beim Passieren der einzelnen Zellen erleidet, die auf seinem Weg liegen. In der Zelle Z_i ist die Absorption E_i proportional zur Dichte x_i der i-ten Zelle. Natürlich wollen wir die Zellen klein genug wählen, dass diese Dichte innerhalb einer Zelle als konstant betrachtet werden kann. Gerade an diesen unbekannten Werten x_i ist man aber interessiert. Bei der Bildproduktion — etwa auf einem Monitor mit mindestens so vielen Pixeln wie Zellen — soll ja dann später die Intensität der Färbung proportional zur Dichte x_i sein. Wegen der Kleinheit der Pixel kann man die Weglänge des Strahles in der i-ten Zelle 0 oder 1 setzen, je nachdem, ob der Strahl die i-te Zelle trifft oder nicht. Setzt man noch den Proportionalitätsfaktor zwischen Dichte und Absorption gleich 1, so erhalten wir für diesen Strahl die lineare Gleichung:

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b; \qquad a_i \text{ ist 0 oder 1}$$

Diese Gleichung enthält also so viele Unbekannte, wie es Zellen hat und das werden je nach gewünschter Auflösung tausende sein. Die Konstante b kann gemessen werden und die Koeffizienten a_i hängen von der gegenseitigen Lage der Zellen ab. Für jede andere Lage des Strahls wird man eine weitere derartige lineare Gleichung erhalten — insgesamt also ein Gleichungssystem mit tausenden von Gleichungen und Unbekannten, das man lösen muss.

Beispiel. Um ein konkretes Beispiel vor uns zu haben, das man von Hand bearbeiten kann, untersuchen wir das „Kleinhirn“, das aus folgenden $2 \times 3 = 6$ Zellen besteht:



und wählen die Strahlen in horizontaler und vertikaler Richtung. Es ergibt sich das folgende lineare Gleichungssystem:

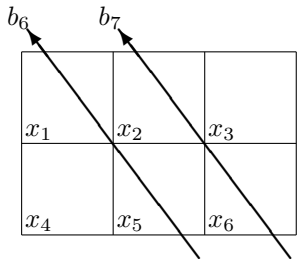
$$\begin{cases} x_1 + x_2 + x_3 = b_1 \\ x_4 + x_5 + x_6 = b_2 \\ x_1 + x_4 = b_3 \\ x_2 + x_5 = b_4 \\ x_3 + x_6 = b_5 \end{cases}$$

Es stellt sich heraus, dass nicht alle diese 5 Gleichungen voneinander unabhängig sind. Man stellt nämlich fest, dass der Rang dieses Gleichungssystems 4 ist. Also existieren nach dem Hauptsatz $6 - 4 = 2$ freie Variablen. Die Lösbarkeitsbedingung lautet:

$$b_1 + b_2 - b_3 - b_4 - b_5 = 0$$

Falls Messfehler vorliegen, könnte also das System keine Lösung haben. Um in dieser Richtung keine Komplikationen zu haben, lassen wir etwa die letzte

Gleichung weg. Um eine eindeutige Lösung zu erhalten, wählen wir noch zwei zusätzliche Strahlenrichtungen, die wir wie in der folgenden Figur legen:



Das vergrösserte lineare Gleichungssystem und die zugehörige erweiterte Matrix lauten nun:

$$\begin{cases} x_1 + x_2 + x_3 = b_1 \\ x_4 + x_5 + x_6 = b_2 \\ x_1 + x_4 = b_3 \\ x_2 + x_5 = b_4 \\ x_1 + x_5 = b_6 \\ x_2 + x_6 = b_7 \end{cases} \qquad \left(\begin{array}{cccccc|c} 1 & 1 & 1 & 0 & 0 & 0 & b_1 \\ 0 & 0 & 0 & 1 & 1 & 1 & b_2 \\ 1 & 0 & 0 & 1 & 0 & 0 & b_3 \\ 0 & 1 & 0 & 0 & 1 & 0 & b_4 \\ 1 & 0 & 0 & 0 & 1 & 0 & b_6 \\ 0 & 1 & 0 & 0 & 0 & 1 & b_7 \end{array}\right)$$

Man beachte, dass die Koeffizientenmatrix in diesem Fall nur 0 und 1 enthält. Solche sogn. BOOLE'sche Matrizen treten in den Anwendungen oft auf. Das zugehörige Gleichungssystem ist regulär. Die Inverse der Koeffizientenmatrix lautet:

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

○

Man beachte, dass es sich hier lohnt, das allgemeine Problem zu lösen bzw. die inverse Matrix zu bestimmen, da sich ja bei einer neuen Messung nur der Konstantenvektor $\vec{b}$ ändert. Dadurch wird der Rechenaufwand für die Einzelmessung erheblich reduziert. Statt jedesmal ein Gleichungssystem mit dem asymptotischen Aufwand $\sim \frac{1}{3}n^3$ zu lösen, braucht man nur noch mit dem Konstantenvektor zu multiplizieren. Der erforderliche Aufwand ist $\sim n^2$.

In der Praxis wird man noch weitere Richtungen wählen und damit ein überbestimmtes System erhalten. Nun lassen sich mit Hilfe der sogn. Ausgleichsrechnung allfällige Messfehler minimieren.

6.2 Die Poisson-Gleichung

Grosse lineare Gleichungssysteme treten auch bei der *Diskretisierung von partiellen Differentialgleichungen* auf. Differentialgleichugen sind Gleichungen, die

eine Beziehung zwischen einer gesuchten Funktion und ihren Ableitungen herstellt. Wenn es sich um Funktionen von mehreren Variablen handelt, treten die Ableitungen nach den einzelnen Variablen, die sogn. partiellen Ableitungen, auf. Die zugehörigen Differentialgleichungen heissen dann partielle Differentialgleichungen.

Ein besonders wichtiges Beispiel für eine solche partielle Differentialgleichung ist die *Poisson-Gleichung*:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = g(x, y)$$

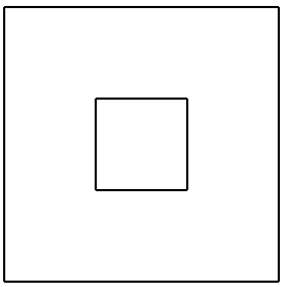
Dabei ist g eine gegebene Funktion im Inneren des Gebietes G der (x, y)-Ebene.

Im Spezialfall, wo die Funktion g überall 0 ist, redet man von der *Potentialgleichung*. Funktionen f, welche die Potentialgleichung

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0$$

erfüllen, heissen *harmonische Funktionen*.

Beispiel. Um etwas Konkretes vor Augen zu haben, wählen wir für G ein Quadrat mit einem quadratischen Loch:



○

Die gesuchte Funktion f soll auch auf dem Gebiet G definiert sein und dort die Poisson-Gleichung erfüllen. Durch die partielle Differentialgleichung ist die Funktion f nicht eindeutig bestimmt. Um f eindeutig festzulegen, muss man noch weitere Bedingungen vorschreiben, z.B. welche Werte die Funktion f auf dem Rand von G annehmen soll. (Randwertproblem) Die Berechnung der Lösung f ist im allgemeinen ein theoretisch sehr schwieriges, aber praktisch wichtiges Problem. Randwertaufgaben lassen sich nur für besonders einfache Gebiete (Rechteck, Kreis Kugel usw.) exakt bzw. mit Hilfe eines Reihenansatzes lösen. Im allgemeinen ist man auf numerische Methoden angewiesen.

Das Poisson-Problem spielt in der Elektrotechnik eine zentrale Rolle. Die Grundgleichung der Elektrostatik lautet:

$$\mathrm{div}(\vec{E}) = 4\pi\rho(\vec{x})$$

wobei $\vec{E}$ das elektrische Feld und $\rho(\vec{x})$ die Ladungsverteilung bezeichnet. Ist $\vec{E}$ ein Potentialfeld, d.h. gilt

$$\vec{E} = \mathrm{grad}(\Phi)$$

so genügt dieses Potential Φ der Poissson-Gleichung:

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 4\pi\rho(x, y)$$

Entsprechend erfüllt das Potential eines elektrischen Feldes im Vakuum die Potentialgleichung.

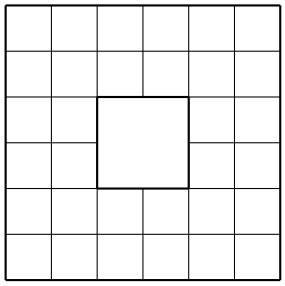
Die Potentialgleichung spielt auch in anderen Bereichen der Physik eine zentrale Rolle. Stellt sich in einem homogenen wärmeleitenden Medium infolge von Isolation der Oberfläche eine stationäre Temperaturverteilung $T(x, y)$ ein, so erfüllt diese Temperaturverteilung die Potentialgleichung:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$$

Die stationäre Temperaturverteilung ist unabhängig von den Materialkonstanten und hängt einzig von der Geometrie des betreffenden Gebietes G ab.

Bei der *Diskretisierung* verzichten wir darauf, die Werte der gesuchten Funktion f für alle Punkte von G zu kennen, sondern begnügen uns mit der Kenntnis der Werte $f(x_k)$ für eine diskrete Menge $x_1, x_2, \ldots, x_n$. Dazu wählt man die Punkte x_i hinreichend dicht beieinander, am einfachsten in gleichen Abständen d.h. man überzieht die Ebene mit einem quadratischen Gitter der Maschenweite h und approximieren den Bereich G durch ein Gebäude Q von Gitterquadraten:

Beispiel. Bei unserem Musterbeispiel erhalten wir:



○

Wir skalieren das Problem so, dass die Maschenweite gerade 1 Einheit ist. Die vorgeschriebenen Randwerte ersetzen wir durch plausible Werte in den Rand-Gitterpunkten dieses Gebäudes und die partielle Differentialgleichung durch eine gewisse *Differenzengleichung*, die die Werte benachbarter innerer Gitterpunkte miteinander verknüpft.

Um die Idee dieser Differenzengleichung zu erklären, betrachten wir zunächst eine Funktion $f\colon x \mapsto f(x)$ einer Variablen auf einem Intervall $[0, N]$. Wir teilen das Intervall in N gleiche Teile. Die Teilpunkte seien $x_i = i$ für $i = 0, \cdots, N$. Die *Änderung* der Funktion f im i-ten Teilintervall ist:

$$\Delta f(i) = f(i+1) - f(i)$$

Für die Änderung zweiter Stufe gilt entsprechend:

$$\begin{aligned} \Delta^2 f(i) &= \Delta\Delta f(i) = \Delta f(i+1) - \Delta f(i) \\ &= f(i+1) - 2f(i) + f(i-1) \end{aligned}$$

Für Leser mit Analysis-Kentnissen sei gesagt, dass wir hier einfach die Ableitung diskretisieren. Für die Ableitung $f'(x)$ gilt bekanntlich:

$$f'(x) = \frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Die Diskretisierung besteht nun grob gesprochen darin, dass wir diese Ableitung durch den Differenzenquotien

$$\frac{\Delta f}{\Delta x} = \frac{f(x+h) - f(x)}{h}$$

ersetzen. In unseren Einheiten ist nämlich $h = 1$. Entsprechend ergibt sich für die Diskretisierung der zweiten Ableitung:

$$\frac{\Delta^2 f}{\Delta x^2} = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

Analog gehen wir bei partiellen Differentialgleichungen vor. Bei unserer Poisson-Gleichung benutzen wir im Bereich G die Gitterpunkte P_{ij}, deren Koordinaten ganzzahlig sind. Die partiellen Ableitungen

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

ersetzen wir durch:

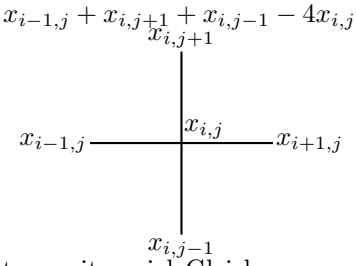
$$f(i+1,j) - 2f(i,j) + f(i-1,j) + f(i,j+1) - 2f(i,j) + f(i,j-1)$$

Daher erfüllen die unbekannten zu berechnenden Näherungen

$$x_{i,j} = f(i,j)$$

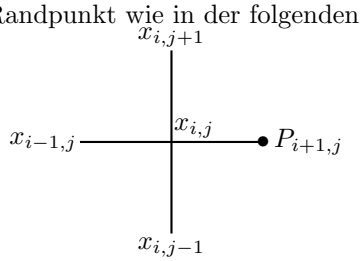
im Inneren des Bereiches folgende linearen Gleichungen:

$$x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1} - 4x_{i,j} = g(i,j)$$



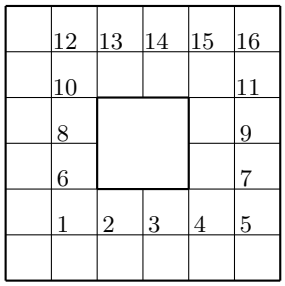
Wir haben also ein System mit soviel Gleichungen und Unbekannten, wie es innere Punkte gibt.

Die rechte Seite der letzten Gleichung ist als bekannt anzusehen. Man beachte aber, dass dies nicht die einzigen Konstanten dieses Gleichungssystems sind. Zusätzlich sind nämlich die Werte auf dem Rand durch die Randbedingung gegeben. Liegt also ein Randpunkt wie in der folgenden Figur:



so ist $x_{i+1,j}$ nicht als Unbekannte, sondern als gegeben zu betrachten. Alle derartigen Randpunkte liefern somit weitere Konstanten, die wir bei der Formulierung des linearen Gleichungssystems auf die rechte Seite nehmen müssen. Der Konstantenvektor hängt von den Werten der Funktion g *und* von den Randwerten ab.

Beispiel. Um ein konkretes Beispiel vor Augen zu haben, denken wir uns den Querschnitt eines quadratischen Kamins mit einem quadratischen Loch. Die Temperatur im Inneren des Kamins sei konstant bei T_i. Der obere Rand liege an einer Aussenwand mit der Temperatur T_a. Die restlichen drei Randstücke liegen im Gebäudeinnern mit einer Raumtemperatur T_r. Wir interessieren uns für die stationäre Temperaturverteilung und haben ein Dirichlet-Problem zu lösen. Die Diskretisierung und die Nummerierung der Unbekannten wählen wir wie in der folgenden Figur:



Die spärlich besetzte erweiterte Matrix des entstehenden linearen Gleichungssystems lautet:

$$\left(\begin{array}{cccccccccccccccc|c}
4 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2T_a \\
-1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & -1 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2T_r \\
-1 & 0 & 0 & 0 & 0 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & 0 & -1 & 0 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 4 & 0 & -1 & 0 & 0 & 0 & 0 & T_i + T_r \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 4 & 0 & 0 & 0 & 0 & -1 & T_i + T_r \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 4 & -1 & 0 & 0 & 0 & T_a + T_r \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & T_i + T_a \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & T_i + T_a \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & T_i + T_a \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 4 & T_a + T_r
\end{array}\right)$$

Für die Lösung dieses Gleichungssystems gilt:

$$\begin{aligned}
x_1 &= \frac{19\ T_a+1940\ T_i+7353\ T_r}{9312} && (32.4694)\\
x_2 &= \frac{19\ T_a+13580\ T_i+18993\ T_r}{32592} && (44.9913)\\
x_3 &= \frac{19\ T_a+29876\ T_i+35289\ T_r}{65184} && (47.4956)\\
x_4 &= \frac{19\ T_a+13580\ T_i+18993\ T_r}{32592} && (44.9913)\\
x_5 &= \frac{19\ T_a+1940\ T_i+7353\ T_r}{9312} && (32.4694)\\
x_6 &= \frac{247\ T_a+13580\ T_i+18765\ T_r}{32592} && (44.8863)\\
x_7 &= \frac{247\ T_a+13580\ T_i+18765\ T_r}{32592} && (44.8863)\\
x_8 &= \frac{19\ T_a+308\ T_i+345\ T_r}{672} && (47.0759)\\
x_9 &= \frac{19\ T_a+308\ T_i+345\ T_r}{672} && (47.0759)\\
x_{10} &= \frac{3439\ T_a+13580\ T_i+15573\ T_r}{32592} && (43.4172)\\
x_{11} &= \frac{3439\ T_a+13580\ T_i+15573\ T_r}{2592} && (43.4172)\\
x_{12} &= \frac{3667\ T_a+1940\ T_i+3705\ T_r}{9312} && (26.5931)\\
x_{13} &= \frac{15307\ T_a+13580\ T_i+3705\ T_r}{32592} && (37.9552)\\
x_{14} &= \frac{31603\ T_a+29876\ T_i+3705\ T_r}{65184} && (40.2276)\\
x_{15} &= \frac{15307\ T_a+13580\ T_i+3705\ T_r}{32592} && (37.9552)\\
x_{16} &= \frac{3667\ T_a+1940\ T_i+3705\ T_r}{9312} && (26.5931)
\end{aligned}$$

Man beachte, wie sich die Symmetrie der Fragestellung auf die Symmetrie der Lösung auswirkt. Bei einer angenommenen Innentemperatur $T_i = 80$, einer Aussentemperatur $T_a = 5$ und einer Raumtemperatur $T_r = 20$ erhalten wir die in Klammern angegebene numerische Lösung dieses Problems. ◯

Wie wir sehen, erhalten wir durch Diskretisierung der Poisson-Gleichung sehr schnell sehr grosse lineare Gleichungssysteme, die man in der Regel nicht exakt lösen muss. Hier ist also ein Näherungsverfahren zur Lösung angezeigt. Man beachte, dass die entstehende Matrix immer diagonal-dominant ist und daher das Verfahren von Jacobi konvergiert. Etwas praktischer ist hier eine Variante, die unter dem Namen Gauss-Seidel-Verfahren bekannt ist. Zunächst schreiben wir die definierende Gleichung des Gitterpunktes $P_{i,j}$ um, indem wir nach $x_{i,j}$ auflösen. Wir erhalten die Beziehung:

$$x_{i,j} = \frac{1}{4}(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1} - g(i,j))$$

Unser Ziel besteht darin, für jeden inneren Punkt P_{ij} eine Folge von Näherungswerten

$$x^0_{i,j}, x^1_{i,j}, x^2_{i,j}, \cdots$$

zu berechnen, die gegen die wahre Lösung des Gleichungssystems konvergiert. Die nullte Näherung wird irgendwie geschätzt. Ist die k-te Näherungslösung bereits bekannt, so erhält man die $(k+1)$-te Näherungslösung, indem man einen inneren Gitterpunkt $P_{i,j}$ auswählt und den dort vorhandenen Wert $x^k_{i,j}$ ersetzt durch den Wert, der sich aus der obigen Gleichung mit Hilfe der vorhandenen Nachbarwerte ergibt. Man setzt also:

$$x^{k+1}_{i,j} = \frac{1}{4}(x^k_{i+1,j} + x^k_{i-1,j} + x^k_{i,j+1} + x^k_{i,j-1} - g(i,j))$$

Alle anderen Werte bleiben unverändert. Nun durchläuft man alle inneren Gitterpunkte der Reihe nach immer wieder, bis man die gesuchte Genauigkeit erreicht hat.

Dem Gauss-Seidel-Verfahren liegt im Wesentlichen die selbe Idee wie dem Jacobi-Verfahren zu Grunde. Das Prinzip besteht darin, dass man für jede Komponente des Lösungsvektors den *jeweils zuletzt berechneten* Näherungswert einsetzt. Das Verfahren wird so theoretisch etwas unübersichtlicher, dafür konvergieren aber die Näherungswerte oft etwas rascher. Man zerlegt die Koeffizientenmatrix wie früher in $A = U + D + O$. Durch Umformen ergibt sich wie seinerzeit die Gleichung $U \cdot \vec{x} + D \cdot \vec{x} + O \cdot \vec{x} = \vec{b}$. Diese Gleichung lässt sich umformen zu

$$(U + D) \cdot \vec{x} = \vec{b} - O \cdot \vec{x}$$

Nach Multiplikation mit $(U+D)^{-1}$ ergibt sich die Fixpunktgleichung für diesen Fall:

$$\vec{x} = (U + D)^{-1} \cdot \vec{b} - (U + D)^{-1} \cdot O \cdot \vec{x}$$

Für die Konvergenz des Jacobi-Verfahrens ist die Matrix $D^{-1} \cdot (U+O)$ ausschlaggebend. Sie ist im Verfahren von Gauss-Seidel zu ersetzen durch die Matrix $(U+D)^{-1} \cdot O$. Man kann zeigen, dass beide Verfahren genau dann konvergieren, wenn der betragsgrösste Eigenwert dieser jeweiligen Matrix kleiner als 1 ist. Je kleiner dieser Betrag ist, um so schneller konvergiert das jeweilige Verfahren.

Wir wollen noch die Frage der Existenz und Eindeutigkeit der Lösung der linearen Gleichungssysteme untersuchen, die aus Poisson-Problemen entstehen. Beide Fragen lassen sich in einem Schlag erledigen, ohne dass man sich irgend welche Gedanken über die unter Umständen komplizierte Gestalt des Quadrat-Gebäudes machen muss. Wir verwenden den Äquivalenzsatz und zeigen, dass das zugehörige homogene Gleichungssystem nur die triviale Lösung hat. Weil der Konstantenvektor des Gleichungssystems von der Funktion g und von den Randdaten herrührt, wird das zu untersuchende System homogen, wenn $g = 0$ ist und alle Randwerte 0 sind. Um zu zeigen, dass zu jedem Poisson-Problem genau eine Lösung existiert müssen wir überprüfen:

Lemma. Verschwindet $x_{i,j}$ in allen Randpunkten, so besitzt das zugehörige Dirichlet-Problem nur die Nullösung.

Der Grund für dieses Lemma liegt in einer wichtigen Eigenschaft von harmonischen Funktionen — sie erfüllen die *Maximumseigenschaft*. Die diskrete Version davon lautet:

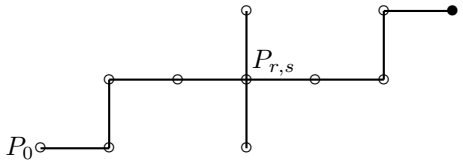
Satz. Die Funktion $x_{i,j}$, die auf einem Gebäude Q erklärt ist, und dort die Mittelwertseigenschaft

$$x_{i,j} = \frac{1}{4}(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1})$$

erfüllt, nimmt ihr Maximum M und ihr Minimum m auf dem Rand an.

Beweis. Es sei P_0 ein Punkt, in dem $x_{i,j}$ das Maximum M annimmt. Wir verbinden den Punkt P_0 durch einen Gitterweg mit dem Rand und markieren auf diesem Weg den letzten Gitterpunkt, in dem $x_{i,j} = M$ ist. Dies sei der

Punkt $P_{r,s}$.



Der zu diesem Gitterpunkt gehörige Wert $x_{r,s}$ ist aber wegen der Mittelwertseigenschaft der Mittelwert der vier Nachbarwerte. Diese vier Nachbarwerte sind nach Voraussetzung $\leq M$. Einer der Nachbarwerte ist aber nach Konstruktion von $P_{r,s}$ sogar strikt kleiner als M. Also ist der Mittelwert $x_{r,s}$ strikt kleiner als M, was der Konstruktion widerspricht. Daher muss das Maximum auf dem Rand liegen. Analog schliesst man für das Minimum. □

Aus diesem Satz ergibt sich nun aber auch unmittelbar das Lemma.

Beweis. Falls nämlich alle Randwerte 0 sind, so sind Maximum und Minimum beide 0. Daher müssen alle Zwischenwerte 0 sein. □

6.3 Netzwerke

In der Elektrotechnik studiert man Netzwerke, die aus vielen miteinander verbundenen Elementen bestehen. Kennt man in einem elektrischen Netzwerk die Spannungs- und Stromquellen sowie das Verhalten der Elemente, kann man die Spannungen über und die Stromstärken durch die einzelnen Elemente mit Hilfe der beiden KIRCHHOFF'schen Regeln berechnen.

1. **Knotenregel**: Die Summe der Stromstärken der ankommenden Ströme ist in jedem Knoten des Leitersystems zu jedem Zeitpunkt gleich der Summe der Stromstärken der abfliessenden Ströme. Für die gerichteten Ströme gilt also

   $$\sum_k I_k(t) = 0 \qquad (I_k(t) : \text{Strom im } k. \text{ Zweig des Knotens zur Zeit } t)$$

   Die Knotenregel drückt die Ladungserhaltung aus.

2. **Maschenregel**: In jeder Masche (geschlossener Weg) ist die Summe der Spannungen gleich der Summe der Spannungsabfälle an den einzelen Elemente. Für die gerichteten Spannungen gilt also

   $$\sum_k U_k(t) = 0 \qquad (U_k(t) : \text{Spannung im } k. \text{ Zweig der Masche zur Zeit } t)$$

   Die Maschenregel drückt die Energieerhaltung aus.

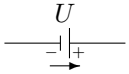
In diesem Abschnitt befassen wir uns ausschliesslich mit Gleichstromnetzen, deren Spannungen und Ströme sich also nicht mit der Zeit ändern. Im Laufe der Zeit wird klar werden, wie sich die hier entwickelten Methoden auch auf den allgemeineren zeitabhängigen Fall anwenden lassen.

Beispiel. Neben dem idealen elektrischen Leiter (Metall)

und dem idealen Isolator (Porzellan)

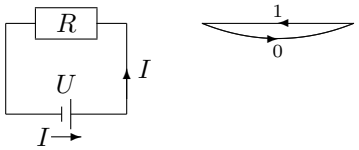
. .

die als degenerierte Elemente interpretiert werden können, spielt die ideale Spannungsquelle (Batterie mit der Quellenspannung U)



eine fundamentale Rolle als Energiequelle. Die negativ geladenen Elektronen bewegen sich in Pfeilrichtung. (Physikalische Stromrichtung) ◯

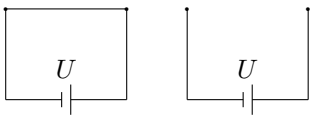
Beispiel. Die simpelste nicht degenerierte Schaltung besteht aus einem einzigen Element mit dem Symbol ⊣□⊢, das wir als *Widerstand* (Kohlenstoff) bezeichnen. In der folgenden Figur findet man das zugehörige Schema.



Im rechtsstehenden gerichteten, markierten Graphen findet man die topologische Information dieser Schaltung. Jeder Zweig der Schaltung wird durch eine gerichtete Kante dagestellt und mit der Nummer des bereffenden Elementes versehen.

Den Spannungabfall am Widerstand R berechnet man für lineare Netzwerke nach dem Ohm'schen Gesetz $U = R \cdot I$ als Produkt aus der gerichteten Stromstärke und dem Widerstandswert R. Solche Widerstände sind weder Strom- noch Stromrichtungsabhängig.

Als Grenzfall ist der Widerstand eines Leiters definitionsgemäss 0. Der Strom, der durch den Leiter fliesst, ist definitionsgemäss ∞. (Kurzschluss) Dual dazu fliesst durch einen Isolator der Strom 0 (Leerlauf) und sein Widerstand ist definitionsgemäss ∞. Die beiden degenerierten Betriebstzustände werden durch die folgenden beiden Schemata beschrieben.



Oft ist es bequemer, statt des Widerstandes R seinen reziproken Wert $G = 1/R$, den *Leitwert* zu benutzen. Er ist dann gross, wenn der Widerstand klein ist und umgekehrt. Insbesondere hat ein Leiter den Leitwert ∞ und durch ihne fliesst der Strom ∞. Ein Isolator hat den Leitwert 0 und durch ihn fliesst der Strom 0. Das Ohm'sche Gesetz nimmt dann die Form $I = G \cdot U$ an.

Weil ein Leiter keinen Widerstand hat, dürfen wir einen Leiter zu einem Knoten zusammenziehen.

•——• ∼ .

ohne das Verhalten eines Netzwerkes zu verändern.

Ein Widerstand kann die Werte $R \in \mathbb{R}^+ \cup \{\infty\}$ annehmen. Falls er nur die zwei degenerierten Zustände $0, \infty \in \mathbf{Z}_2$, d.h. die beiden Leitwerte 0 (isolierend) und ∞ (leitend) annehmen kann, wird er zu einem *Schalter* und die Elektrotechnik zur Digitaltechnik. Je nach ihrem Grundzustand kommen Schalter *schliessend* und *öffnend* vor und wir benutzen die beiden Symbole (die einzige Spannungsquelle unterdrücken wir im Schema)

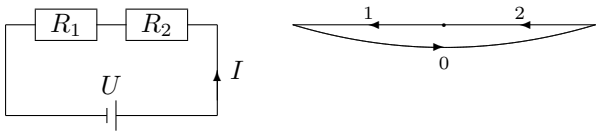
Der schliessende Schalter G ist im Grundzustand isolierend und hat dann den Leitwert 0. Im betätigten Zustand ist er leitend und dann ist sein Leitwert ∞. Umgekehrt ist der öffnende Schalter $\overline{G}$ im Grundzustand leitend und hat dann den Leitwert ∞. Im betätigten Zustand ist er isolierend und dann ist sein Leitwert 0. Bezeichnet man den Grundzustand eines Schalters mit 0 und den betätigten Zustand mit ∞, so findet man in den folgenden Tabellen die Leitwerte der beiden Schalter:

	G
0	0
∞	∞

	$\overline{G}$
0	∞
∞	0

In der Informatik wird statt des Symbols ∞ meistens das Symbol 1 benutzt. Es kann natürlich nicht als numerischer Leitwert interpretiert werden. ◯

Beispiel. Im nächsten Beispiel möchten wir wissen, wie gross der Gesamtwiderstand zweier Widerstände R_1 und R_2 ist, wenn man sie *seriell* schaltet. Dazu bauen wir sie in die folgende Schaltung ein und untersuchen das zugehörige Netzwerk:



Weil dieses einfache Netzwerk keine Knoten (Verzweigungspunkt von drei und mehr Leitungen) und eine einzige Masche enthält, die wir willkürlich im Gegenuhrzeigersinn durchlaufen, ergeben die Kirchhoff'schen Regel hier eine einzige Bedingungsgleichung:

$$R_1 I + R_2 I = U \quad \text{bzw.} \quad I = \frac{U}{R_1 + R_2}$$

Damit können wir den Gesamtwiderstand R berechnen, der an der Batterie liegt. Nach dem Ohm'schen Gesetz gilt für den Ersatzwiderstand

$$\frac{U}{I} = R_s = R_1 + R_2$$

Wir stellen also fest, dass sich die Widerstände bei Serieschaltung addieren. Insbesondere wird bei Serieschaltung der Gesamtwiderstand *grösser* als der grösste Teilwiderstand sein, d.h. es ist immer

$$R \geq \max(R_1, R_2)$$

was intuitiv einzuleuchten scheint.

Für den Leitwerte bei Serieschaltung ergibt sich dann

$$G_s = \frac{G_1 G_2}{G_1 + G_2}; \qquad G_s \leq \min(G_1, G_2)$$

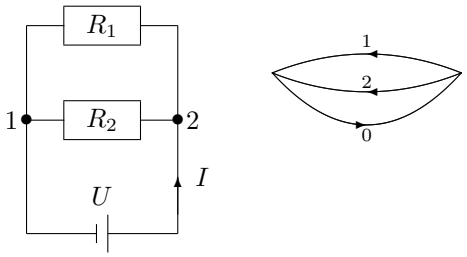
was weniger anschaulich erscheint.

Für die Teilspannungen an den beiden Widerständen ergibt sich damit

$$U_1 = R_1 I = \frac{R_1}{R_1 + R_2} U \qquad U_2 = R_2 I = \frac{R_2}{R_1 + R_2} U$$

Die Teilspannungen bei einer Serieschaltung verhalten sich wie die Teilwiderstände — eine Gesetzmässigkeit, die als *Spannungsteilerregel* bekannt ist. ◯

Beispiel. Als nächstes möchten wir wissen, wie gross der Gesamtwiderstand zweier gegebener Widerstände R_1 und R_2 ist, wenn man sie *parallel* schaltet. Dazu untersuchen wir folgendes Netzwerk:



Es hat zwei Knoten und zwei Maschen. Zur Beantwortung der Frage würde ein im Bruchrechnen sattelfester Mittelschüler etwa wie folgt argumentieren. Wir nehmen an, der Strom durch den Widerstand R_1 sei I_1. Für den Strom I_2 durch den Widerstand R_2 gilt natürlich $I_2 = I - I_1$. Da beide Widerstände an der selben Spannung liegen, ergibt sich nach dem Ohm'schen Gesetz die Gleichung

$$U = R_1 I_1 = R_2 (I - I_1)$$

Durch Auflösen dieser Gleichung nach I_1 erhält man

$$I_1 = \frac{R_2}{R_1 + R_2} I$$

Für die Spannung ergibt sich also durch Einsetzen:

$$U = R_1 I_1 = \frac{R_1 R_2}{R_1 + R_2} I \quad \text{bzw.} \quad I = \frac{R_1 + R_2}{R_1 R_2} U$$

Schliesslich liefert das Ohm'schen Gesetz $U = RI$ nach einer Division durch I die gesuchte Lösung:

$$R_p = \frac{R_1 R_2}{R_1 + R_2}$$

Diese Formel kann man sich etwas leichter einprägen, wenn man sie in der Form

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2}$$

auswendig lernt. Falls zum Beispiel die Widerstände $R_1 = 20$ [Ω] und $R_2 = 50$ [Ω] betragen, so erhält man für den Gesamtwiderstand $R = \frac{100}{7} \approx 14.285\ldots$ [Ω]. Dieses Beispiel belegt, dass bei Parallelschaltung der Gesamtwiderstand *kleiner* als der kleinste Teilwiderstand ist. Es ist tatsächlich immer

$$R_p \leq \min(R_1, R_2)$$

Noch einfacher zu merken wird die gefunden Formel, wenn man sie mit Hilfe von Leitwerten ausdrückt. Für den Leitwerte bei Parallelschaltung ergibt sich:

$$G_p = G_1 + G_2; \qquad G_p \geq \max(G_1, G_2)$$

Man beachte die Dualität zwischen den Formeln bei Serie- und Parallelschaltung unter dem Übergang von den Widerständen zu den Leitwerten.

Für die Teilströme an den beiden Widerständen ergibt sich damit

$$I_1 = \frac{R_2}{R_1 + R_2} I \qquad I_2 = \frac{R_1}{R_1 + R_2} I$$

Die Teilströme bei einer Parallelschaltung verhalten sich umgekehrt wie die Teilwiderstände — eine Gesetzmässigkeit, die als *Stromteilerregel* bekannt ist. ◯

Beispiel. Ein digitales Netzwerk (Schaltkreis) aus zwei seriell geschalteten öffnenden Schaltern hat nur noch die vier Leitwertzustände

$$(G_1, G_2) \in \{(0,0), (0,\infty), (\infty,0), (\infty,\infty)\}$$

Für den Gesamtleitwert bei Serieschaltung der beiden Schalter erhalten wir in diesen vier Fällen die Leitwerte:

G_1	G_2	G_s
0	0	0
0	∞	0
∞	0	0
∞	∞	∞

Diese Tabelle kann mit der Abmachung $0 < \infty$ durch die Formel

$$G_s = \min(G_1, G_2) = G_1 \wedge G_2$$

zusammengefasst werden. Die Serieschaltung zweier Schalter wird zur Konjunktion (UND) und die lineare Algebra zur Boole'schen Algebra. Dort wird statt des Leitwertes ∞ meistens das Symbol 1 benutzt um anzudeuteten, dass der betreffende Schalter offen ist. Der Leitwert 0 symbolisiert, dass der Schalter geschlossen ist.

Ein Vergleich der beiden Formeln für die Serieschaltung in der Elektrotechnik und in der Digitaltechnik zeigt, dass die arithmetische Formel beim Bilden der Grenzwerte in die entsprechende BOOLE'sche Formel übergeht. Den Unterschied zwischen kontinuierlicher Elektrotechnik und diskreter Digitaltechnik erkennt man, wenn man die Höhenlinien der beiden Funktionen G_s vergleicht. In der linken Figur

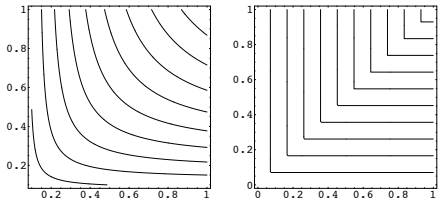
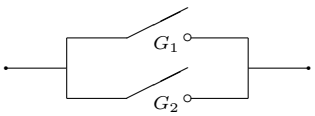


Abbildung 6.1: Niveaulinien von G_s bei Serieschaltung.

findet man den Fall $\frac{G_1 G_2}{G_1+G_2}$ der Widerstände und rechts jenen der Schalter $\min(G_1, G_2)$.

Werden die beiden Schalter parallel geschaltet, erhalten wir für den Gesamtleitwert der vier möglichen Schalterzustände folgende Leitwerte:

G_1	G_2	G_p
0	0	0
0	∞	∞
∞	0	∞
∞	∞	∞



Diese Tabelle kann durch die Vorschrift

$$G_p = \max(G_1, G_2) = G_1 \vee G_2$$

zusammengefasst werden. Die Parallelschaltung zweier Schalter wird zur Disjunktion (ODER).

Wiederum geht die die arithmetische Formel G_p für die Parallelschaltung beim Bilden der Grenzwerte in die entsprechende Boole'sche Formel über. In der linken Figur

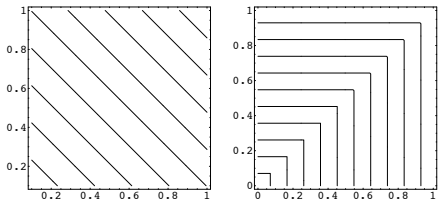


Abbildung 6.2: Niveaulinien von G_p bei Parallelschaltung.

findet man die Höhenlinien für den Fall $G_1 + G_2$ der Widerstände und rechts der Schalter $\max(G_1, G_2)$. ◯

Der Ingenieur beschäftigt sich vor allem deshalb mit Mathematik, um an vernetzte Information heranzukommen.

Beispiel. Um zwei unbekannte Widerstände R_1 und R_2 berechnen zu können, schalten wir sie einmal seriell und dann parallel zusammen und messen beide

Male den Gesamtstrom, der an einer Batterie mit der Spannung U liegt. Nach unseren Überlegungen gilt für den Strom I_S, der beim Serieschalten fliesst:

$$R_1 I_S + R_2 I_S = U$$

Entsprechend gilt für den Strom I_P, der in der Parallelschaltung fliesst, die Bedingung

$$U = \frac{R_1 R_2}{R_1 + R_2} I_P$$

Löst man die erste Gleichung nach R_2 auf, erhält man

$$R_2 = \frac{U}{I_S} - R_1$$

Setzt man das Resultat in die zweite Gleichung ein, erhält man nach einigen Umformungen die quadratische Gleichung für R_1:

$$I_S I_P R_1^2 - I_P U R_1 + U^2 = 0$$

Ihre Lösungen findet man durch quadratische Ergänzung. Die beiden positiven Lösungen lauten

$$R_1 = \frac{I_P U + \sqrt{I_P^2 U^2 - 4 I_S I_P}}{2 I_S I_P}; \qquad R_2 = \frac{I_P U - \sqrt{I_P^2 U^2 - 4 I_S I_P}}{2 I_S I_P}$$

Weil das Problem in den beiden Widerständen symmetrisch ist, erhalten wir durch Vertauschen von R_1 und R_2 die entsprechende quadratische Gleichung für den Widerstand R_2 mit den selben beiden Lösungen. Damit das Problem eine Lösung hat, muss der Radikand positiv sein. Daher muss also die Bedingung

$$U \geq 2\sqrt{\frac{I_S}{I_P}}$$

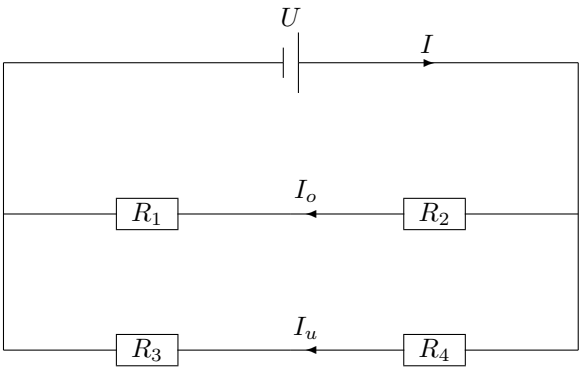
erfüllt sein. Falls die Spannung $U = 220\ [V]$ beträgt und die beiden Ströme $I_S = 0.9\ [A]$ bzw. $I_P = 6\ [A]$ gemessen wurden, so ergeben sich für die Widerstände die Werte $R_1 \approx 44.9\ldots\ [\Omega]$ und $R_2 \approx 199.5\ [\Omega]$. ◯

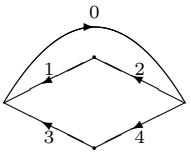
Mit Hilfe der beiden hergeleiteten Formeln für die Serie- und die Parallelschaltung lassen sich nun eine ganze Reihe von Gesamtwiderständen von Netzwerken berechnen, in dem man sie schrittweise in einzelne Seriell- und Parallelschaltungen zerlegt. Diese sogn. seriell-parallel-Schaltungen lassen sich also rekursiv aus den beiden besprochenen Sonderfällen zusammenbauen. Ihre präzise Definition lautet also.

Definition. Ein Netzwerk heisst *seriell-parallel*, falls es entweder eine Serie- oder eine Parallelschaltung von seriell-parallelen Netzwerken ist. Ein einziges Element ist seriell-parallel.

Es lässt sich zeigen, dass ein Netzwerk genau dann seriell-parallel bezüglich der beiden Knoten a und b ist, falls durch jedes seiner Elemente mindestens ein Pfad von a nach b existiert, der durch keinen Knoten mehr als einmal führt und so, dass keine zwei solchen Pfade in unterschiedlicher Richtung durch das selbe Element führen.

Beispiel. Der folgende Stromkreis enthält eine Batterie mit der Quellenspannung U und vier Widerstände.





Da es sich um eine seriell-parallel-Schaltung handelt, ist die Berechnung einfach. Durch Zerlegung in zwei serielle Teilschaltungen ergibt sich für die Widerstände im oberen bzw. im unteren Zweig

$$R_o = R_1 + R_2; \qquad R_u = R_3 + R_4$$

und daher gilt nach dem OHM'schen Gesetz für die entsprechenden Ströme

$$I_o = \frac{U}{R_1 + R_2}; \qquad I_u = \frac{U}{R_3 + R_4}$$

Für den Gesamtwiderstand erhalten wir durch Parallelschalten der beiden Zweige:

$$R = \frac{R_o R_u}{R_o + R_u} = \frac{(R_1 + R_2)(R_3 + R_4)}{R_1 + R_2 + R_3 + R_4} = \frac{R_1R_3 + R_1R_4 + R_2R_3 + R_2R_4}{R_1 + R_2 + R_3 + R_4}$$

und für den Batteriestrom I nach dem OHM'schen Gesetz:

$$I = \frac{(R_1 + R_2 + R_3 + R_4)U}{R_1R_3 + R_1R_4 + R_2R_3 + R_2R_4}$$

Für $U = \frac{3}{2}$ [V], $R_1 = \frac{1}{4}$ [Ω], $R_2 = \frac{5}{4}$ [Ω], $R_3 = \frac{5}{4}$ [Ω], $R_4 = \frac{7}{4}$ [Ω] ist $I_o = 1$ [A], $I_u = \frac{1}{2}$ [A] und für den Batteriestrom gilt $I = \frac{3}{2}$ [A]. Insbesondere sind beide Zweigströme kleiner als der Batteristrom. Man beachte die Symmetrien dieser Formeln und überlege sich, wie sie mit den Symmetrien des Netzwerkes zusammenhängen.

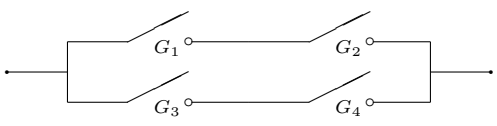
Die selbe Information lässt sich natürlich auch mit Hilfe der Leitwerte ausdrücken. Für den Leitwert des oberen Zweiges bzw. des unteren Zweiges gilt

$$G_o = \frac{G_1G_2}{G_1 + G_2}; \qquad G_u = \frac{G_3G_4}{G_3 + G_4}$$

Daraus erhalten wir für den Leitwert der gesamten Schaltung

$$G = G_o + G_u = \frac{G_1G_2}{G_1 + G_2} + \frac{G_3G_4}{G_3 + G_4} = \frac{G_1G_2G_3 + G_1G_2G_4 + G_1G_3G_4 + G_2G_3G_4}{(G_1 + G_2)(G_3 + G_4)}$$

Das zugehörige digitale Netzwerk mit der Beschreibung $G = (G_1 \wedge G_2) \vee (G_3 \wedge G_4)$ und dem Schema

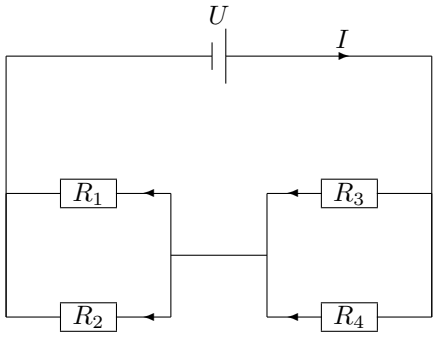


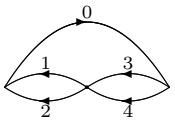
wird durch die 16 Werte der Tabelle

G_1	G_2	G_3	G_4	G	G_1	G_2	G_3	G_4	G
0	0	0	0	0	∞	0	0	0	0
0	0	0	∞	0	∞	0	0	∞	0
0	0	∞	0	0	∞	0	∞	0	0
0	0	∞	∞	∞	∞	0	∞	∞	∞
0	∞	0	0	0	∞	∞	0	0	∞
0	∞	0	∞	0	∞	∞	0	∞	∞
0	∞	∞	0	0	∞	∞	∞	0	∞
0	∞	∞	∞	∞	∞	∞	∞	∞	∞

beschrieben. ◯

Beispiel. Ersetzen wir in der soeben untersuchten Schaltung parallele durch serielle Verschaltungen, erhält man das folgende duale Netzwerk, das wiederum Seriell-Parallel ist.





Durch Zerlegung in zwei in Serie geschaltete Parallelschaltungen ergibt sich für die Widerstände im linken bzw. im rechten Zweig

$$R_l = \frac{R_1 R_2}{R_1 + R_2}; \qquad R_r = \frac{R_3 R_4}{R_3 + R_4}$$

Für den Gesamtwiderstand erhalten wir durch Serieschalten der beiden Zweige:

$$R = R_l + R_r = \frac{R_1 R_2}{R_1 + R_2} + \frac{R_3 R_4}{R_3 + R_4} = \frac{R_1 R_2 R_3 + R_1 R_2 R_4 + R_1 R_3 R_4 + R_2 R_3 R_4}{(R_1 + R_2)(R_3 + R_4)}$$

Die selbe Information lässt sich natürlich auch mit Hilfe der Leitwerte ausdrücken. Für den Leitwert des linken Zweiges bzw. des rechten Zweiges gilt

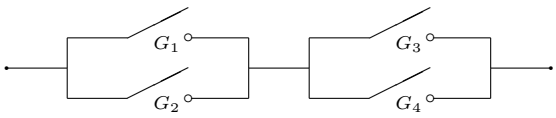
$$G_l = G_1 + G_2; \qquad G_r = G_3 + G_4$$

Daraus erhalten wir für den Leitwert der gesamten Schaltung

$$G = \frac{G_l G_r}{G_l + G_r} = \frac{(G_1 + G_2)(G_3 + G_4)}{G_1 + G_2 + G_3 + G_4} = \frac{G_1 G_3 + G_1 G_4 + G_2 G_3 + G_2 G_4}{G_1 + G_2 + G_3 + G_4}$$

Man beachte die Dualität zu den Formeln im letzten Beispiel.

Das zugehörige digitale Netzwerk mit der Beschreibung $G = (G_1 \vee G_2) \wedge (G_3 \vee G_4)$ und dem Schema

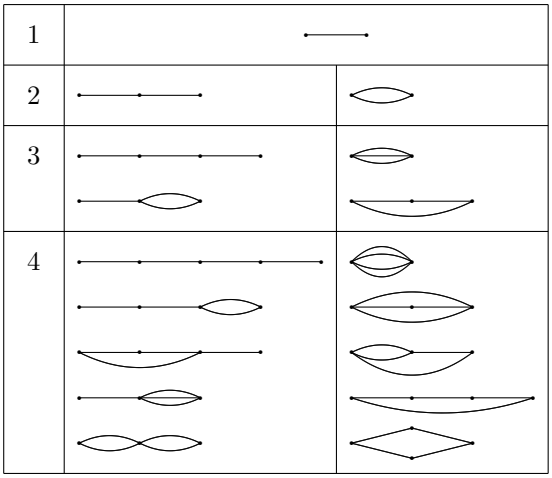


wird durch die 16 Werte der Tabelle

G_1	G_2	G_3	G_4	G	G_1	G_2	G_3	G_4	G
0	0	0	0	0	∞	0	0	0	0
0	0	0	∞	0	∞	0	0	∞	∞
0	0	∞	0	0	∞	0	∞	0	∞
0	0	∞	∞	0	∞	0	∞	∞	∞
0	∞	0	0	0	∞	∞	0	0	0
0	∞	0	∞	∞	∞	∞	0	∞	∞
0	∞	∞	0	∞	∞	∞	∞	0	∞
0	∞	∞	∞	∞	∞	∞	∞	∞	∞

beschrieben. ◯

Für $1 \leq n \leq 4$ findet man die durch folgende ungerichtete Graphen beschriebene verschiedene Schaltungen, wobei wir je ein duales Paar einander gegenüber gestellt haben.



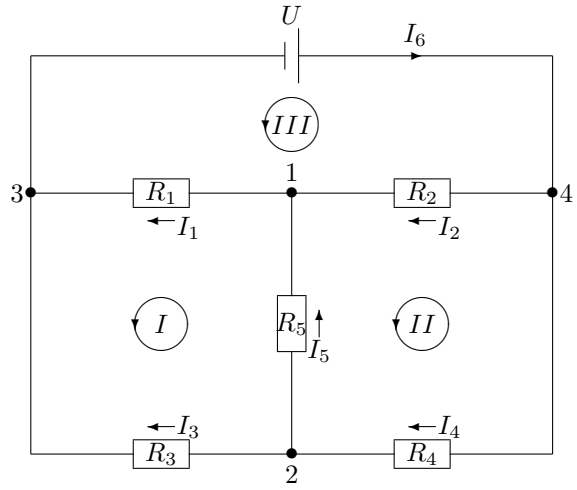
Man beachte, dass alle diese Schaltungen seriell-parallel sind.

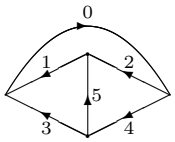
Die Sache beginnt dann komplex zu werden, wenn ein Netzwerk vorliegt, das nicht diese einfache Topologie hat. Um ein Netzwerk zu konstruieren, das nicht mehr seriell-parallel ist, benötigt man mindestens 5 Elemente. Dann wird das Mittelschüler-Vorgehen schnell versagen. Dem Schüler kommt der Umstand zu Gute, dass er es mit sehr wenig Unbekannten — in der Regel zwei oder drei — zu tun hat. Daher kann er (meist unbewusst) eine der Unbekannten sofort durch die anderen ausdrücken und den entstandenen Ausdruck in die restlichen Gleichungen einsetzen. Diese weit verbreitete Einsetzungsmethode, die unter dem

missverständlichen Namen Elimination bekannt ist, liefert schliesslich eine Gleichung mit nur einer Unbekannten, die der Schüler nun leicht mit Hilfe der neu erworbenen Bruchrechnung lösen kann. Falls aber mehrere, stark vermaschte unbekannte Ströme in einem Netzwerk vorkommen, wird sich diese Elimination schnell nicht mehr im Kopf bewältigen lassen. Insbesondere besteht die Gefahr, dass man den Überblick verliert und „im Kreis herum“ rechnet. Praktiker haben sich noch eine Reihe von Faustregeln, Tricks und vor allem Vereinfachungsmöglichkeiten[2] zugelegt, die bei wirklich komplexen Netzwerken, insbesondere solchen, deren Koeffizienten komplexe Zahlen sind, wie sie in der Wechselstromtechnik vorkommen, zwar die Grösse des linearen Gleichungssystems verkleinern; an der prinzipiellen Aufgabe, ein gewisses lineares Gleichungssystem lösen zu müssen, ändern sie aber nichts. Vor allem benötigen diese Methoden ziemlich subtile mathematische Methoden, wenn man genau verstehen will, warum sie eigentlich funktionieren.

Für das Verständnis von anspruchsvollen Netzwerken benötigt man also ein systematisches Verfahren, dass die gesuchte Information garantiert in möglichst kurzer Zeit liefert. Es basiert auf der unmittelbaren Anwendung der KIRCHHOFF'schen Regeln. Zur Illustration untersuchen wir das einfachste Netzwerk, das nicht seriell-parallel ist.

Beispiel. Der folgende Stromkreis[3] enthält eine Batterie mit der Quellenspannung U. Es besteht aus $z = 6$ Zweigen und $k = 4$ Knoten.





Wir wollen die Zweigströme $I_1, I_2, I_3, I_4, I_5, I_6$ in diesem Netzwerk bei gegebener Batteriespannung U und gegeben Widerständen R_1, R_2, R_3, R_4, R_5 berechnen. Nachdem wir die Stromrichtungen in den Zweigen mit dem Index des zugehörigen Zweigstromes numeriert haben und in den Maschen den Linksumlauf (Gegenuhrzeigersinn) willkürlich festgelegt haben, benutzen wir die KIRCH-

[2] Dreieck-Stern-Umwandlung; Maschenstromverfahren; Knotenpotentialverfahren; Überlagerung; Ersatzspannungsquellen.

[3] sogn. Wheatstone-Brücke; Wheatstone: britischer Physiker (1802 – 1875).

HOFF'schen Regeln und erhalten folgende Gleichungen:

$$\text{Knotengleichung:} \quad \begin{cases} I_2 + I_5 = I_1 & -I_1 + I_2 + I_5 = 0 \\ I_4 = I_3 + I_5 & -I_3 + I_4 - I_5 = 0 \\ I_1 + I_3 = I_6 & I_1 + I_3 - I_6 = 0 \\ I_6 = I_2 + I_4 & -I_2 - I_4 + I_6 = 0 \end{cases}$$

$$\text{Maschengleichungen:} \quad \begin{cases} -R_3 I_3 + R_1 I_1 + R_5 I_5 = 0 \\ -R_4 I_4 - R_5 I_5 + R_2 I_2 = 0 \\ -R_2 I_2 - R_1 I_1 = -U \end{cases}$$

Als nächstes schreiben wir diese Gleichungen so übersichtlich wie möglich, indem wir zunächst Variablen mit dem gleichen Namen untereinanderschreiben d.h. allenfalls Platz offenlassen. Die konstanten Terme bringen wir auf die rechte Seite. Die Tatsache, dass wir unsere Information ordnen können, spielt im Eliminationsverfahren eine zentrale Rolle. Die (geordneten) Gleichungen lauten:

$$\begin{cases} -I_1 + I_2 + I_5 = 0 \\ -I_3 + I_4 - I_5 = 0 \\ I_1 + I_3 - I_6 = 0 \\ -I_2 - I_4 + I_6 = 0 \\ R_1 I_1 - R_3 I_3 + R_5 I_5 = 0 \\ R_2 I_2 - R_4 I_4 - R_5 I_5 = 0 \\ R_1 I_1 + R_2 I_2 = U \end{cases}$$

Unsere Aufgabe besteht also darin, diese Gleichungen, d.h. dieses Gleichungssystem, das die uns interessierende Variablen miteinander verknüpft, zu lösen. Man beachte, dass sämtliche vorkommenden Variablen nur in der ersten Potenz vorkommen. Die Gleichungen enthalten keine Produkte oder Wurzeln ihrer Variablen. Solche Gleichungen heissen linear. Also handelt es sich um ein *lineares Gleichungssystem*. Die Kirchhoff'schen Regeln liefern immer lineare Gleichungssysteme.

Eine *Lösung* dieses linearen Gleichungssystems besteht aus 6 Zahlen, die *alle* vorkommenden Gleichungen des Systems erfüllen. Die Gesamtheit aller Lösungen heisst *Lösungsraum*. Für eine solche Lösung schreiben wir

$$\vec{I} = \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \end{pmatrix} \in \mathbb{R}^6$$

und nennen dies ein 6-*Tupel* von Zahlen oder einen 6-dimensionalen *Spaltenvektor*. Wir schreiben kurz $\vec{I}$ dafür. Ein weiteres solches Tupel erscheint auf der rechten Seite der Gleichheitszeichen. Es enthält die Konstanten des Systems und hängt nur von den Spannungsquellen ab. Wir definieren also den 7-dimensiona-

len *Konstantenvektor*:

$$\vec{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ U \end{pmatrix} \in \mathbb{R}^7$$

Da die Namen der Unbekannten nicht von Bedeutung sind, fassen wir nun noch die Koeffizienten des linearen Gleichungssystems zu einem eigenen Objekt zusammen, indem wir uns die Position der Unbekannten, der „+“-Zeichen und der „=“-Zeichen merken. So erhalten wir ein rechteckiges Zahlenschema mit 7 *Zeilen* und 6 *Spalten*. Diese *Matrix* vom Typ 7×6 füllen wir natürlich dort mit 0 auf, wo gewisse Variablen fehlen:

$$M = \begin{pmatrix} -1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & -1 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ R_1 & 0 & -R_3 & 0 & R_5 & 0 \\ 0 & R_2 & 0 & -R_4 & -R_5 & 0 \\ R_1 & R_2 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Unser lineares Gleichungssystem können wir nun symbolisch schreiben als:

$$M \cdot \vec{I} = \vec{b}$$

Wir wollen also annehmen, wir hätten eine Multiplikation der Matrix M mit dem Spaltenvektor $\vec{I}$ so definiert, dass diese Gleichung genau obiges lineares Gleichungssystem beschreibt. Interpretieren wir diese Matrix als Operator $M\colon \mathbb{R}^6 \to \mathbb{R}^7$, so geht es darum, jene Zustände $\vec{I} \in \mathbb{R}^6$ zu finden, die auf $\vec{b} \in \mathbb{R}^7$ abgebildet werden.

Als nächstes beachten wir, dass die Knotengleichungen voneinander abhängig sind. Addiert man nämlich die vier Knotengleichungen, so erhält man die Gleichung $0 = 0$. Der Grund für diese Abhängigkeit liegt darin, dass jeder Zweigstrom einmal aus einem Knoten heraus- und in einen anderen Knoten hineinführt und daher in zwei Knotengleichungen vorkommt und zwar mit unterschiedlichen Vorzeichen. Daher muss die Summe der Knotengleichungen immer verschwinden. Das hat die wichtige Konsequenz, dass mindestens eine der Knotengleichungen bis auf ein Vorzeichen als Summe der übrigen geschrieben und daher weggelassen werden kann. In unserem Beispiel ist etwa die dritte Gleichung das Negative der Summe der ersten, zweiten und vierten Gleichung. Daher dürfen wir die dritte Gleichung weglassen, ohne den Lösungsraum zu verändern. Haben wir nämlich eine Lösung für die restlichen drei Gleichungen vor uns, so erfüllt sie die dritte Gleichung automatisch! Unser verkleinertes Gleichungssystem lautet damit:

$$\begin{cases} -I_1 & +I_2 & & & +I_5 & & = & 0 \\ & & -I_3 & I_4 & -I_5 & & = & 0 \\ & -I_2 & & -I_4 & & +I_6 & = & 0 \\ R_1 I_1 & & -R_3 I_3 & & +R_5 I_5 & & = & 0 \\ & R_2 I_2 & & -R_4 I_4 & -R_5 I_5 & & = & 0 \\ R_1 I_1 & +R_2 I_2 & & & & & = & U \end{cases}$$

Wiederum können wir dieses verkleinerte Gleichungssystem matriziell schreiben, indem wir die 6×6-Matrix A wie folgt definieren:

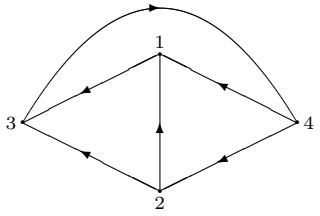
$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ R_1 & 0 & -R_3 & 0 & R_5 & 0 \\ 0 & R_2 & 0 & -R_4 & -R_5 & 0 \\ R_1 & R_2 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Erklären wir wie oben eine Multiplikation der Matrix A mit dem 6-Tupel $\vec{I}$ und den Konstantenvektor $\vec{b}$, so lautet das lineare Gleichungssystem nun kurz:

$$A \cdot \vec{I} = \vec{b}$$

Dieses lineare Gleichungssystem kann man lösen. Wir werden sehen, dass der Rechenaufwand für das Lösen eines linearen Gleichungssystems proportional zur 3. Potenz der Anzahl Gleichungen ist. In unserem Fall sind also rund $6^3 = 216$ Rechenoperationen nötig.

Auf Grund der speziellen Herkunft unseres Gleichungssystems aus einem gerichteten Graphen mit 4 Knoten



mit der *Inzidenzmatrix*

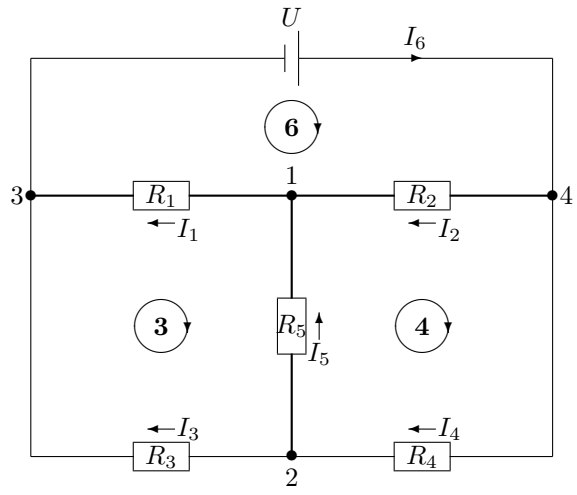
$$N = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

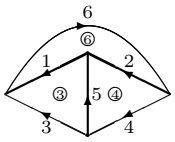
lässt sich das Problem mit Hilfe des *Maschenstromverfahrens* weiter vereinfachen. Dazu bestimmen wir im Graphen zunächst einen *maximalen Baum*. Darunter versteht man einen Teilgraphen mit folgenden beiden Eigenschaften:

1. Er enthält sämtliche Knoten des betreffenden Netzwerkes.

2. Er besitzt keine Maschen.

In einem Graphen gibt es in der Regel verschiedene Möglichkeiten, einen maximalen Baum zu wählen, von denen wir in der Figur eine fett eingezeichnet haben. Er besteht aus den *Baumzweigen* $1, 2, 5$. Die restlichen Zweige $3, 4, 6$ des Netzwerkes bilden die *Verbindungszweige*.





Wie oben geben wir uns die Richtungen der gesuchten Ströme beliebig vor. Als nächstens wählen wir die Umlaufsrichtungen der Maschen so, dass sie den in den zugehörigen Verbindungszweigen fliessenden Strömen entsprechen. Die einzelnen Maschen werden den zugehörigen Verbindungszweigen entsprechend numeriert. Man beachte den Unterschied zur früher benutzten Wahl der Orientierung, die damals willkürlich im Gegenuhrzeigersinn festgelegt wurde.

Entscheidend ist nun die Beobachtung, dass sich aus den drei Strömen I_3, I_4, I_6 in den Verbindungszweigen (sogn. Maschenströme) alle restlichen Ströme in den Baumzweigen durch Überlagern einfach bestimmen lassen. Aus der Knotenregel folgt nämlich

$$\begin{aligned} I_6 &= I_1 + I_3 & \qquad I_1 &= I_6 - I_3 \\ I_6 &= I_2 + I_4 & \qquad I_2 &= I_6 - I_4 \\ I_4 &= I_3 + I_5 & \qquad I_5 &= I_4 - I_3 \end{aligned}$$

Weil wir die Ströme der Baumzweige einfach auf die Maschenströme zurückführen können, genügt es also, die Maschenströme zu berechnen. Dadurch reduziert sich die zur Berechnung erforderliche Anzahl Gleichungen.

Zur Berechnung der Maschenströme I_3, I_4, I_6 benutzen wir die zugehörigen drei *Maschengleichungen*, die wir mit Hilfe der gewählten Maschenströme formulieren:

$$\left\{ \begin{array}{rcl} -R_1(I_6 - I_3) - R_5(I_4 - I_3) + R_3 I_3 &=& 0 \\ -R_2(I_6 - I_4) + R_4 I_4 + R_5(I_4 - I_3) &=& 0 \\ R_1(I_6 - I_3) + R_2(I_6 - I_4) &=& U \end{array} \right.$$

Ordnen wir diese Gleichungen nach den unbekannten Strömen um, nehmen sie die Form

$$\left\{ \begin{array}{rrrcl} (R_1 + R_3 + R_5)I_3 & -R_5 I_4 & -R_1 I_6 &=& 0 \\ -R_5 I_3 & +(R_2 + R_4 + R_5)I_4 & -R_2 I_6 &=& 0 \\ -R_1 I_3 & -R_2 I_4 & +(R_1 + R_2)I_6 &=& U \end{array} \right.$$

an. Weil dieses Gleichungssystem nur noch drei Unbekannte hat, wird sich der Lösungsaufwand um den Faktor 8 auf $3^3 = 27$ Operationen reduzieren.

Dieses lineare Gleichungssystem kann in gewohnter Weise $A \cdot \vec{I} = \vec{b}$ matriziell beschrieben werden, indem man die *Widerstandsmatrix* A und den Vektor $\vec{I}$ der Maschenströme definiert.

$$A = \begin{pmatrix} R_1 + R_3 + R_5 & -R_5 & -R_1 \\ -R_5 & R_2 + R_4 + R_5 & -R_2 \\ -R_1 & -R_2 & R_1 + R_2 \end{pmatrix}, \qquad \vec{I} = \begin{pmatrix} I_3 \\ I_4 \\ I_6 \end{pmatrix}$$

Der Konstantenvektor $\vec{b}$ ist definiert als

$$\vec{b} = \begin{pmatrix} 0 \\ 0 \\ U \end{pmatrix}$$

und enthält die Information der Energiequellen. Man beachte, dass die Koeffizientenmatrix symmetrisch ist und man deshalb Speicherplatz sparen kann. Dieses Gleichungssystem, bzw. die zugehörigen Matrizen können direkt aus der gegebenen Schaltung aufgestellt werden. Zunächst wählt man im betreffenden Netzwerk, das k Knoten und z Zweige haben mag, einen maximalen Baum. Die Umlaufsrichtungen der Maschen werden, wie beschrieben, den Verbindungszweigen entsprechend gewählt. Weil jeder Baum mit k Knoten aus topologischen Gründen $k-1$ Kanten hat, erhalten wir in unserem Fall $k-1$ Baumzweige und $z-k+1$ Verbindungszweige. Wir wollen ferner annehmen, dass zwischen zwei Knoten nur ein einziger Widerstand vorkommt. Sind mehrer Widerstände vorhanden, ersetzen wir sie durch seriell-parallel-Schaltungen. Zum Aufstellen der Matrizen geht man wie folgt vor.

1. Die Maschenströme entsprechen den $z-k+1$ Verbindungszweigen bzw. den Maschen. Sie bilden im aufzustellenden Gleichungssystem die Unbekannten.

2. Für jeden Maschenstrom bestimmt man den *Umlaufwiderstand* d.h. die Summe derjenigen Widerstände, der betreffenden Masche. Dieser Umlaufwiderstand tritt in der Widerstandsmatrix als Koeffizient für den zugehörigen Maschenstrom auf der Diagonalen auf und ist positiv.

3. Je zwei Maschen werden nun durch die *Kopplungswiderstände* miteinander in Beziehung gebracht. Es ist die Summe der in den beiden Maschen gemeinsam vorkommenden Widerstände. Ihr Vorzeichen ist positiv, wenn die beiden Maschen die gleiche Orientierung haben; anderenfalls ist es negativ.

4. Im Konstantenvektor steht die Summe der Quellenspannungen, die in der betreffenden Masche auftreten. Jede Quellenspannung wird negativ gewichtet, wenn ihr physikalischer Richtungssinn der Maschenorientierung entgegengesetzt ist. Sonst erhält sie ein positives Gewicht.

Statt der ursprünglich $(k-1)+(z-k+1) = z$ Gleichungen müssen wir nun also noch ein lineares Gleichungssystem mit $z-k+1$ Gleichungen in gleich vielen Unbekannten lösen.

Im Beispiel findet man für die Maschenströme nach einiger Rechnung die Werte

$$I_3 \quad = \quad \frac{U(R_{12} + R_{14} + R_{15} + R_{25})}{\Delta}$$

$$\begin{aligned} I_4 &= \frac{U(R_{12}+R_{15}+R_{23}+R_{25})}{\Delta} \\ I_6 &= \frac{U(R_{12}+R_{14}+R_{15}+R_{23}+R_{25}+R_{34}+R_{35}+R_{45})}{\Delta} \end{aligned}$$

Dabei haben wir zur besseren Übersicht die unübliche Abkürzung $R_{ij} = R_i \cdot R_j$ benutzt. Mit der entsprechenden Abkürzung $R_{ijk} = R_i \cdot R_j \cdot R_k$ können wir den gemeinsamen Nenner schreiben als:

$$\Delta = R_{123} + R_{124} + R_{134} + R_{135} + R_{145} + R_{234} + R_{235} + R_{245}$$

Die Ströme in den Verbindungszweigen ergeben sich daraus nun, wie oben erwähnt, durch Überlagerung. Insgesamt erhalten wir für die Lösung $\vec{I}$ unseres Problems

$$\vec{I} = \begin{pmatrix} \frac{U(R_{23}+R_{34}+R_{35}+R_{45})}{\Delta} \\ \frac{U(R_{14}+R_{34}+R_{35}+R_{45})}{\Delta} \\ \frac{U(R_{12}+R_{14}+R_{15}+R_{25})}{\Delta} \\ \frac{U(R_{12}+R_{15}+R_{23}+R_{25})}{\Delta} \\ \frac{U(R_{23}-R_{14})}{\Delta} \\ \frac{U(R_{12}+R_{14}+R_{15}+R_{23}+R_{25}+R_{34}+R_{35}+R_{45})}{\Delta} \end{pmatrix}$$

Nachdem wir eine Vorstellung davon haben, wie die Lösung aussieht, ist es höchste Zeit, dass wir uns fragen, welchen Nutzen wir nun aus dieser Lösung ziehen können.

Einerseits können wir natürlich beliebige Zahlen einsetzen und mit unseren Gleichungen die zughörigen Ströme ausrechnen.

Beispiel. Es sei $U = 20$ [V], $R_1 = 15$ [Ω], $R_2 = 8$ [Ω], $R_3 = 7$ [Ω], $R_4 = 9$ [Ω], $R_5 = 13$ [Ω]. Für die Ströme gilt: $I_1 = \frac{6540}{8153} \approx 0.802$ [A], $I_2 = \frac{8120}{8153} \approx 0.996$ [A], $I_3 = \frac{11080}{8153} \approx 1.359$ [A], $I_4 = \frac{9500}{8153} \approx 1.165$ [A], $I_5 = -\frac{1580}{8153} \approx -0.194$ [A], $I_6 = \frac{17620}{8153} \approx 2.161$ [A]. Man beachte, dass kleine natürliche Zahlen als Daten recht grosse Brüche als Ergebnis liefern können. Ob ein Gehirn ein solches Netzwerk realisiert, scheint mehr als fraglich. ◯

Offenbar gibt es immer genau eine Lösung für unser Problem, falls der gemeinsame Nenner $\Delta \neq 0$ ist. Da aber sämtliche Widerstände positive Zahlen sind, kann Δ nur dann 0 sein, wenn ein degeneriertes Netzwerk vorliegt.

Beispiel. Wir können aus den bisherigen Ergebnissen den Gesamtwiderstand berechnen, der an der Batterie liegt. Nach dem Ohm'schen Gesetz ist:

$$R = \frac{U}{I_6} = \frac{R_{123}+R_{124}+R_{134}+R_{135}+R_{145}+R_{234}+R_{235}+R_{245}}{R_{12}+R_{14}+R_{15}+R_{23}+R_{25}+R_{34}+R_{35}+R_{45}}$$

Im Spezialfall, wo alle Widerstände gleich 1 [Ω] sind, ergibt sich $I_1 = I_2 = I_3 = I_4 = \frac{1}{2}$, $I_5 = 0$ und $I_6 = 1$. Der Gesamtwiderstand R der Schaltung ist dann

auch 1 [Ω]. In diesem Fall fliesst also durch den Widerstand R_5 kein Strom. Man könnte ihn ohne Schaden aus der Schaltung herausschneiden. ◯

Viel wichtiger als weiterer Zahlensalat ist allerdings das tiefere Verständnis, das wir über das vorliegende Netzwerk erworben haben und das wir nun konstruktiv verwenden können. Es soll hier stellvertretend nur eine der vielen Anwendungen der Wheatstone-Brücke skizziert werden.

Da man in der Praxis den Wert 0 speziell gut messen kann, fragen wir uns, wann denn gewisse der Ströme 0 sein können. Dazu kommt nur die einfachste Komponente der Lösung, d.h. der Strom I_5 in Betracht. Die allgemeine Lösung zeigt, dass gilt:

$$I_5 = \frac{U(R_2R_3 - R_1R_4)}{\Delta}$$

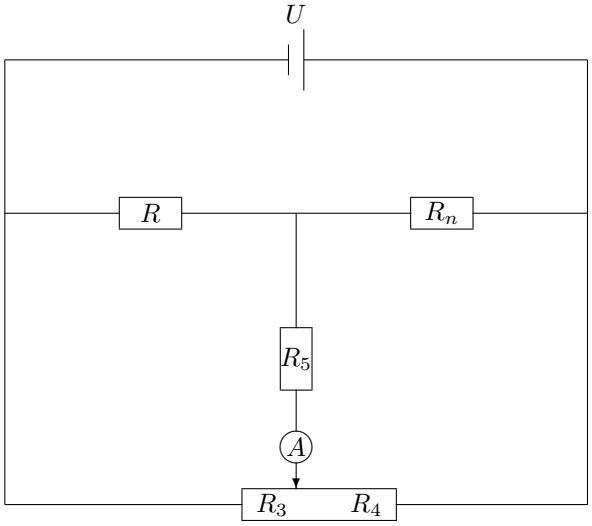
Der Strom I_5 ist bei $\Delta \neq 0$ genau dann 0, wenn der Zähler dieses Bruches verschwindet, d.h. bei $U \neq 0$ also genau dann, falls die sogn. *Abgleichbedingung*

$$R_2R_3 = R_1R_4$$

erfüllt ist. Man beachte, dass diese Gleichung unabhängig vom Widerstand R_5 ist. Diese Gleichung erinnert einen an den Strahlensatz oder an die fundamentalen Gleichungen der geometrischen Optik bzw. an das Hebelgesetz und dürfte deshalb ebenso vielfältige Anwendungen nach sich ziehen.

Die Wheatstone-Brücke lässt sich zur Widerstandsmessung mit Hilfe der Längenmessung verwenden. Das Vorgehen ist wie folgt:

1. Man baut sich mit dem zu bestimmenden Widerstand R und einem bekannten Eichwiderstand R_n die folgende Schaltung zusammen:



2. Man gleicht die Brücke ab, d.h sorgt für $I_5 = 0$. Das ist sehr genau möglich, dadurch dass man das Potentiometer geeignet verstellt und am Ampèremeter die Mittel-Nullstellung abliest.

3. Aus der Abgleichbedingung lässt sich der gesuchte Widerstand berechnen via die Gleichung:
$$R = R_n \cdot \frac{R_3}{R_4}$$
Das Widerstandsverhältnis $\frac{R_3}{R_4}$ am Potentiometer ist proportional zum Längenverhältnis $\frac{l_3}{l_4}$ der abgegriffenen Teile eines gespannten Widerstandsdrahtes. Dieses Längenverhältnis lässt sich mit grosser Genauigkeit messen. Für den gesuchten Widerstand ergibt sich schliesslich:
$$R = R_n \cdot \frac{l_3}{l_4}$$

Als Bonus ergibt uns unsere Lösung noch die Einsicht, dass die ganze Messung offenbar von der Spannung U unabhängig ist, d.h. die Spannung brauchen wir nicht auch noch zu messen. Der mathematische Grund dafür liegt darin, dass die rechten Seite unseres linearen Gleichungssystems sehr spärlich besetzt ist, nämlich nur einen einzigen von 0 verschiedenen Wert enthält.

Der tiefere Grund dafür, dass der Strom I_5 eine Sonderrolle spielt, liegt in der Symmetrie dieses speziellen Netzwerkes. Diese Symmetrie wirkt sich nämlich auch auf die Lösungen aus. ◯

Dieses Beispiel hat hoffentlich etwas Neugier geweckt. Zum Beispiel könnte man sich fragen:

1. Wie beschreibt man ein Netzwerk ohne es zeichnen zu müssen und wie findet man dann das zugehörige lineare Gleichungssystem auf systematische Art?

2. Wieviel Abhängigkeiten zwischen den Gleichungen gibt es und wovon hängen sie ab? Offenbar braucht man nicht immer alle Knoten mit einzubeziehen. Wie viele muss man mindestens? Wie steht es mit den Maschen. Kann es auch da Abhängigkeiten geben? Wie findet man eine Basismenge von Maschen? Hängen Knoten und Maschen gar irgend wie zusammen? Wie findet man alle möglichen Abhängigkeiten und wie reduziert man damit das lineare Gleichungssystem auf systematische Art?

3. Wie kann man von einem Graphen entscheiden, ob er seriell-parallel ist?

4. Was passiert mit dem linearen Gleichungssystem und seiner Lösung, wenn man die Batterie umpolt oder einen der frei gewählten Orientierungen der Zweige umkehrt?

5. Welche Operationen darf man eigentlich mit linearen Gleichungssystemen durchführen, ohne die Lösungsmenge zu verändern?

6. Wie löst man lineare Gleichungssysteme systematisch? Kann man das programmieren?

7. Wie gross ist der Rechenaufwand für alle diese gewünschten systematischen Operationen? Geht es eventuell mit kleinerem Aufwand?

8. Kann man spärliche Besetzung oder Symmetrie der Koeffizientenmartrix irgendwie systematisch ausnutzen?

9. Wie kann man sämtliche Lösungen auf systematische Art beschreiben?
   (a) Hat jedes lineare Gleichungssystem überhaupt eine Lösung?
   (b) Ist diese Lösung immer eindeutig bestimmt?
10. Ist in jedem linearen Gleichungssystem die Lösung immer von der Art, dass nur Brüche über einem gemeinsamen Nenner vorkommen? Welche Rolle spielt dieser gemeinsame Nenner? Wie gross können eigentlich die vorkommenden Zähler und Nenner im schlimmsten Fall werden?
11. Kann man immer ein Netzwerk finden, das ein gegebenes lineares Gleichungssystem realisiert? Wie findet man das kleinste solche Netzwerk?
12. Treten lineare Gleichungssysteme und die damit verbundenen Phänomene in anderen für uns relevanten Zusammenhängen auf? Können wir dadurch Modellvorstellungen (Analogien) für den elektrischen Strom und den Informationsfluss in einem Rechner gewinnen?

    In Cambridge erzählt man sich zu diesem Thema folgende Geschichte von Maxwell: Maxwell weckte in einer seiner Vorlesungen einen schlafenden Studenten mit den Worten auf: "Young man, what is electricity?". Der verdatterte Student gab zur Antwort: "I'm terribly sorry, Sir, I knew the answer but I have forgotten it". Maxwell's Reaktion zu seiner Klasse war: "Gentlemen, you have just witnessed the greatest tragedy in the history of science. The one person who knew what electricity is has forgotten it".
13. Können wir intuitiv besser verstehen, was hier eigentlich los ist, damit wir nicht einfach blind dem Zahlen- bzw. Formelsalat der Lösung ausgeliefert sind. Wir möchten uns ein Bild von der ganzen Sache machen — d.h. die Sache geometrisch interpretieren!
14. Wie wirken sich Symmetrien des Netzwerkes auf Symmetrien des Gleichungssystems und auf Symmetrien der Lösung aus? Wieviele Symmetrien hat unser Problem und wie finden wir sie auf systematische Art?
15. Kann man mit diesen Methoden auch Wechselstromkreise analysieren?