Mathematisches Seminar

Matrizen

Leitung: Andreas Müller und Roy Seitz

Joshua Baer, Marius Baumann, Reto Fritsche, Alain Keller
Marc Kühne, Robine Luchsinger, Naoki Pross, Thomas Reichlin
Michael Schmid, Pascal Andreas Schmid, Adrian Schuler
Thierry Schwaller, Michael Steiner, Tim Tönz, Fabio Viecelli
Lukas Zogg

OST Ostschweizer Fachhochschule, Rapperswil, 2021

Inhaltsverzeichnis

Vorwort

Dieses Buch entstand im Rahmen des Mathematischen Seminars im Frühjahrssemester 2021 an der Ostschweizer Fachhochschule in Rapperswil. Die Teilnehmer, Studierende der Studiengänge für Elektrotechnik, Informatik und Bauingenieurwesen der OST, erarbeiteten nach einer Einführung in das Themengebiet jeweils einzelne Aspekte des Gebietes in Form einer Seminararbeit, über deren Resultate sie auch in einem Vortrag informierten.

Im Frühjahr 2021 war das Thema des Seminars die Matrizen. Ziel war, die Vielfalt der Anwendungsmöglichkeiten dieser einfachen Datenstruktur zu zeigen.

In einigen Arbeiten wurde auch Code zur Demonstration der besprochenen Methoden und Resultate geschrieben, soweit möglich und sinnvoll wurde dieser Code im Github-Repository dieses Kurses[1] [5] abgelegt. Im genannten Repository findet sich auch der Source-Code dieses Skriptes, es wird hier unter einer Creative Commons Lizenz zur Verfügung gestellt.

[1] `https://github.com/AndreasFMueller/SeminarMatrizen.git`

Teil I

Grundlagen

Einleitung

Die Mathematik befasst sich neben dem Rechnen mit Zahlen, der Arithmetik, mit einer Vielzahl von Abstraktionen, die oft überhaupt nichts mit Zahlen zu tun haben. Die Geometrie studiert zum Beispiel Objekte wie Punkte, Geraden, Kreise und deren Beziehungen untereinander, die man definieren kann ganz ohne das Wissen, was eine Zahl ist. Apollonius von Perga (262–190 BCE) hat in seinem Buch über Kegelschnitte als erster einen algebraischen Zusammenhang zwischen Zahlen festgestellt, die man also die Vorläufer heutiger Koordinaten eines Punktes ansehen könnte. Erst im 16. Jahrhundert entwickelte sich die Algebra allerdings weit genug, dass eine Algebraisierung der Geometrie möglich wurde. Pierre de Fermat und René Descartes schufen die sogenannte *analytische Geometrie*. Das rechtwinklige Koordinatensystem, nach Descartes auch karteisches Koordinatensystem genannt, beschreibt Punkte als Zahlenpaare (x, y) und Kurven in der Ebene durch ihre Gleichungen. Geraden können als Graphen der Funktion $f(x) = ax + b$ oder als Lösungsmenge linearer Gleichungen wie $ax + by = c$ verstanden werden. Eine Parabel kann als Graph einer quadratischen Funktion $f(x) = ax^2 + bx + c$ dargestellt werden. Die Punkte (x, y) eines Kreises lösen eine Gleichung der Form

$$(x - x_M)^2 + (y - y_M)^2 = r^2.$$

Mit dieser einfachen Idee konnte jedes geometrische Problem in der Ebene in ein algebraisches Problem übersetzt werden und umgekehrt.

Die Algebraisierung macht allerdings auch klar, dass dem Aufbau des Zahlensystems mehr Beachtung geschenkt werden muss. Zum Beispiel beschreibt die Gleichung

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

einen Kreis mit Radius 2 um den Punkt $(0, 1)$. Der Kreis hat natürlich zwei Schnittpunkte mit der x-Achse, wie mit jeder Gerade, deren Abstand vom Mittelpunkt des Kreises kleiner ist als der Radius. Schnittpunkte haben die Koordinaten $(x_S, 0)$ und x_S muss die Gleichung

$$x_S^2 + (0 - 1)^2 = x_S^2 + 1 = 4 \qquad \Rightarrow \qquad x_S^2 = 3$$

erfüllen. Eine solche Lösung ist nicht möglich, wenn man sich auf rationale Koordinaten $x_S \in \mathbb{Q}$ beschränkt, die Erweiterung auf reelle Zahlen ist notwendig.

Kapitel 1 übernimmt die Aufgabe, die Zahlensysteme klar zu definieren und ihre wichtigsten Eigenschaften zusammenzutragen. Sie bilden das Fundament aller folgenden Konstruktionen.

Die reellen Zahlen erweitern die rationalen Zahlen derart, dass damit zum Beispiel quaddratische Gleichungen gelöst werden können. Dies ist aber nicht die einzige mögliche Vorgehensweise. Die Zahl $\alpha = \sqrt{2}$ ist ja nur ein Objekt, mit dem gerechnet werden kann wie mit jeder anderen Zahl, welche aber die zusätzliche Rechenregel $\alpha^2 = 2$ erfüllt. Die Erweiterung von $\mathbb{R}$ zu den komplexen Zahl verlangt nur, dass man der Menge $\mathbb{R}$ ein neues algebraisches Objekt i hinzufügt, welches als spezielle

Eigenschaft die Gleichung $i^2 = -1$ hat. Bei $\sqrt{2}$ hat die geometrische Anschauung suggeriert, dass es eine solche Zahl "zwischen" den rationalen Zahlen gibt, aber für i gibt es keine solche Anschauung. Die imaginäre Einheit i erhielt daher auch diesen durchaus abwertend gemeinten Namen.

Die Zahlensysteme lassen sich also verstehen als einfachere Zahlensysteme, denen man zusätzliche Objekte mit besonderen algebraischen Eigenschaften hinzufügt. Doch was sind das für Objekte? Gibt es die überhaupt? Kann man deren Existenz einfach so postulieren, so wie man das mit i gemacht hat? Und was macht man, wenn man sich den nächsten "algebraischen Wunsch" erfüllen will, auch einfach wieder die Existenz des neuen Objektes postulieren?

Komplexen Zahlen und Matrizen zeigen, wie das gehen könnte. Indem man vier rationale Zahlen als 2×2-Matrix in der Form

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

gruppiert und die Rechenoperationen

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix} \\ AB &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} \end{aligned}$$

definiert, kann man neue Objekte mit zum Teil bekannten, zum Teil aber auch ungewohnten algebraischen Eigenschaften bekommen. Die Matrizen der Form

$$aI = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \quad a \in \mathbb{Q}$$

zum Beispiel erfüllen alle Regeln für das Rechnen mit rationalen Zahlen. $\mathbb{Q}$ kann man also als Teilmenge des neuen "Zahlensystems" ansehen. Aber die Matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

hat die Eigenschaft

$$J^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I.$$

Das neue Objekt J ist ein explizit konstruiertes Objekt, welches genau die rechnerischen Eigenschaften der imaginären Einheit i hat.

Die imaginäre Einheit ist nicht die einzige Grösse, die sich auf diese Weise konstruieren lässt. Zum Beispiel erfüllt die Matrix

$$W = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \qquad \text{die Gleichung} \qquad W^2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I,$$

die Menge der Matrizen

$$\mathbb{Q}(\sqrt{2}) = \left\{ \begin{pmatrix} a & 2b \\ b & a \end{pmatrix} \;\middle|\; a, b \in \mathbb{Q} \right\}$$

verhält sich daher genau so wie die Menge der rationalen Zahlen, denen man ein "imaginäres" neues Objekt $\sqrt{2}$ hinzugefügt hat.

Matrizen sind also ein Werkzeug, mit dem sich ein algebraisches Systeme mit fast beliebigen Eigenschaften konstruieren lässt. Dies führt zu einer Explosion der denkbaren algebraischen Strukturen. Kapitel 2 bringt etwas Ordnung in diese Vielfalt, indem die grundlegenden Strukturen charakterisiert und benannt werden.

In den folgenden Kapiteln sollen dann weitere algebraische Konstrukte studiert und mit Matrizen realisiert werden. Den Anfang machen in Kapitel 3 die Polynome. Polynome beschreiben grundlegende algebraische Eigenschaften eines einzelnen Objektes, sowohl $\sqrt{2}$ wie auch i sind Lösungen einer Polynomgleichung.

Eine besondere Rolle spielen in der Mathematik die Symmetrien. Eine der frühesten Anwendungen dieses Gedankens in der Algebra war die Überlegung, dass sich die Nullstellen einer Polynomgleichung permutieren lassen. Die Idee der Permutationsgruppe taucht auch in algebraischen Konstruktionen wie der Determinanten auf. Tatsächlich lassen sich Permutationen auch als Matrizen schreiben und die Rechenregeln für Determinanten sind ein direktes Abbild gewisser Eigenschaften von Transpositionen. Einmal mehr haben Matrizen ermöglicht, ein neues Konzept in einer bekannten Sprache auszudrücken.

Die Darstellungstheorie ist das Bestreben, nicht nur Permutationen, sondern beliebige Gruppen von Symmetrien als Mengen von Matrizen darzustellen. Die abstrakten Symmetriegruppen erhalten damit immer konkrete Realisierungen als Matrizenmengen. Auch kompliziertere Strukturen wie Ringe, Körper oder Algebren lassen sich mit Matrizen realisieren. Aber die Idee ist nicht auf die Geometrie beschränkt, auch analytische oder kombinatorische Eigenschaften lassen sich in Matrizenstrukturen abbilden und damit neuen rechnerischen Behandlungen zugänglich machen.

Das Kapitel 11 illustriert, wie weit dieser Plan führen kann. Die Konstruktion der Homologiegruppen zeigt, wie sich die Eigenschaften der Gestalt gewisser geometrischer Strukturen zunächst mit Matrizen, die kombinatorische Eigenschaften beschreiben, ausdrücken lassen. Anschliessend können daraus wieder algebraische Strukturen gewonnen werden. Gestalteigenschaften werden damit der rechnerischen Untersuchung zugänglich.

Die folgenden Kapitel sollen zeigen, wie Matrizen der Schlüssel dafür sein können, fast jede denkbare rechnerische Struktur zu verstehen und auch zum Beispiel für die Berechnung mit dem Computer zu realisieren.

Kapitel 1

Zahlen

Das Thema dieses Buches ist die Konstruktion interessanter mathematischer Objekte mit Hilfe von Matrizen. Die Einträge dieser Matrizen sind natürlich Zahlen. Wir wollen von diesen grundlegenden Bausteinen ausgehen. Dies schliesst natürlich nicht aus, dass man auch Zahlenmengen mit Hilfe von Matrizen beschreiben kann, wie wir es später für die komplexen Zahlen machen werden.

In diesem Kapitel sollen daher die Eigenschaften der bekannten Zahlensysteme der natürlichen, ganzen, rationalen, reellen und komplexen Zahlen nochmals in einer Übersicht zusammengetragen werden. Dabei wird besonderes Gewicht darauf gelegt, wie in jedem Fall einerseits neue Objekte postuliert, andererseits aber auch konkrete Objekte konstruiert werden können.

1.1 Natürliche Zahlen

Die natürlichen Zahlen sind die Zahlen, mit denen wir zählen. Sie abstrahieren das Konzept der Anzahl der Elemente einer endlichen Menge. Da die leere Menge keine Elemente hat, muss die Menge der natürlichen Zahlen auch die Zahl 0 enthalten. Wir schreiben

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}.$$

Peano-Axiome

Man kann den Zählprozess durch die folgenden Axiome von Peano beschreiben:

1. $0 \in \mathbb{N}$.
2. Jede Zahl $n \in \mathbb{N}$ hat einen *Nachfolger* $n' \in \mathbb{N}$.
3. 0 ist nicht Nachfolger einer Zahl.
4. Wenn zwei Zahlen $n, m \in \mathbb{N}$ den gleichen Nachfolger haben, $n' = m'$, dann sind sie gleich $n = m$.
5. Enthält eine Menge X die Zahl 0 und mit jeder Zahl auch ihren Nachfolger, dann ist $\mathbb{N} \subset X$.

Vollständige Induktion

Es letzte Axiom formuliert das Prinzip der vollständigen Induktion. Um eine Aussage $P(n)$ für alle natürlichen Zahlen n mit vollständiger Induktion zu beweisen, bezeichnet man mit X die Menge aller Zahlen, für die $P(n)$ wahr ist. Die Induktionsverankerung beweist, dass $P(0)$ wahr ist, dass also $0 \in X$. Der Induktionsschritt beweist, dass mit einer Zahl $n \in X$ auch der Nachfolger $n' \in X$ ist. Nach dem letzten Axiom ist $\mathbb{N} \subset X$, oder anders ausgedrückt, die Aussage $P(n)$ ist wahr für jede natürliche Zahl.

Addition

Aus der Nachfolgereigenschaft lässt sich durch wiederholte Anwendung die vertrautere Addition konstruieren. Um die Zahl $n \in \mathbb{N}$ um $m \in \mathbb{N}$ zu vermehren, also $n + m$ auszurechnen, kann man rekursive Regeln

$$\begin{aligned} n + 0 &= n \\ n + m' &= (n + m)' \end{aligned}$$

festlegen. Nach diesen Regeln ist

$$5 + 3 = 5 + 2' = (5 + 2)' = (5 + 1')' = ((5 + 1)')' = ((5 + 0')')' = (((5)')')'.$$

Dies ist genau die Art und Weise, wie kleine Kinder Rechnen lernen. Sie Zählen von 5 ausgehend um 3 weiter. Der dritte Nachfolger von 5 heisst üblicherweise 8.

Die algebraische Struktur, die hier konstruiert worden ist, heisst eine Halbgruppe. Allerdings kann man darin zum Beispiel nur selten Gleichungen lösen, zum Beispiel hat $3 + x = 1$ keine Lösung. Die Addition ist nicht immer umkehrbar.

Multiplikation

Es ist klar, dass auch die Multiplikation definiert werden kann, sobald die Addition definiert ist. Die Rekursionsformeln

$$\begin{aligned} n \cdot 0 &= 0 \\ n \cdot m' &= n \cdot m + n \end{aligned} \tag{1.1}$$

legen jedes Produkt von natürlichen Zahlen fest, zum Beispiel

$$5 \cdot 3 = 5 \cdot 2' = 5 \cdot 2 + 5 = 5 \cdot 1' + 5 = 5 \cdot 1 + 5 + 5 = 5 \cdot 0' + 5 + 5 = 5 \cdot 0 + 5 + 5 + 5 = 5 + 5 + 5.$$

Doch auch bezüglich der Multiplikation ist $\mathbb{N}$ unvollständig, die Beispielgleichung $3x = 1$ hat keine Lösung in $\mathbb{N}$.

Rechenregeln

Aus den Definitionen lassen sich auch die Rechenregeln ableiten, die man für die alltägliche Rechnung braucht. Zum Beispiel kommt es nicht auf die Reihenfolge der Summanden oder Faktoren an. Das *Kommutativgesetz* besagt

$$a + b = b + a \qquad \text{und} \qquad a \cdot b = b \cdot a.$$

Die Kommutativität der Addition werden wir auch in allen weiteren Konstruktionen voraussetzen. Die Kommutativität des Produktes ist allerdings weniger selbstverständlich und wird beim Matrizenprodukt nur noch für spezielle Faktoren zutreffen.

Eine Summe oder ein Produkt mit mehr als zwei Summanden bzw. Faktoren kann in jeder beliebigen Reihenfolge ausgewertet werden,

$$(a + b) + c = a + (b + c) \qquad \text{und} \qquad (a \cdot b) \cdot c = a \cdot (b \cdot c)$$

dies ist das Assoziativgesetz. Es gestattet auch eine solche Summe oder ein solches Produkt einfach als $a + b + c$ bzw. $a \cdot b \cdot c$ zu schreiben, da es ja keine Rolle spielt, in welcher Reihenfolge man die Teilprodukte berechnet.

Die Konstruktion der Multiplikation als iterierte Addition mit Hilfe der Rekursionsformel (1.1) hat auch zur Folge, dass die *Distributivgesetze*

$$a \cdot (b + c) = ab + ac \qquad \text{und} \qquad (a + b)c = ac + bc$$

gelten. Bei einem nicht-kommutativen Produkt ist es hierbei notwendig, zwischen Links- und Rechts-Distributivgesetz zu unterscheiden.

Die Distributivgesetze drücken die wohlbekannte Regel des Ausmultiplizierens aus. Ein Distributivgesetz ist also grundlegend dafür, dass man mit den Objekten so rechnen kann, wie man das in der elementaren Algebra gelernt hat. Auch die Distributivgesetze sind daher Rechenregeln, die wir in Zukunft immer dann fordern werden, wenn Addition und Multiplikation definiert sind. Sie gelten immer für Matrizen.

Teilbarkeit

Die Lösbarkeit von Gleichungen der Form $ax = b$ mit $a, b \in \mathbb{N}$ gibt Anlass zum sehr nützlichen Konzept der Teilbarkeit. Die Zahl b heisst teilbar durch a, wenn die Gleichung $ax = b$ eine Lösung in $\mathbb{N}$ hat. Jede natürlich Zahl n ist durch 1 und durch sich selbst teilbar, denn $n \cdot 1 = n$. Andere Teiler sind dagegen nicht selbstverständlich. Die Zahlen

$$\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \dots\}$$

haben keine weiteren Teiler. Sie heissen *Primzahlen*. Die Menge der natürlichen Zahlen ist die naheliegende Arena für die Zahlentheorie.

Konstruktion der natürlichen Zahlen aus der Mengenlehre

Die Peano-Axiome postulieren, dass es natürliche Zahlen gibt. Es werden keine Anstrengungen unternommen, die natürlichen Zahlen aus noch grundlegenderen mathematischen Objekten zu konstruieren. Die Mengenlehre bietet eine solche Möglichkeit.

Da die natürlichen Zahlen das Konzept der Anzahl der Elemente einer Menge abstrahieren, gehört die leere Menge zur Zahl 0. Die Zahl 0 kann also durch die leere Menge $\emptyset = \{\}$ wiedergegeben werden.

Der Nachfolger muss jetzt als eine Menge mit einem Element konstruiert werden. Das einzige mit Sicherheit existierende Objekt, das für diese Menge zur Verfügung steht, ist $\emptyset$. Zur Zahl 1 gehört daher die Menge $\{\emptyset\}$, eine Menge mit genau einem Element. Stellt die Menge N die Zahl n dar, dann können wir die zu $n + 1$ gehörige Menge N' dadurch konstruieren, dass wir zu den Elemente von N ein zusätzliches Element hinzufügen, das noch nicht in N ist, zum Beispiel $\{N\}$:

$$N' = N \cup \{N\}.$$

Die natürlichen Zahlen existieren also, wenn wir akzeptieren, dass es Mengen gibt. Die natürlichen Zahlen sind dann nacheinander die Mengen

$$\begin{aligned}
0 &= \emptyset \\
1 &= 0 \cup \{0\} = \emptyset \cup \{0\} = \{0\} \\
2 &= 1 \cup \{1\} = \{0\} \cup \{1\} = \{0, 1\} \\
3 &= 2 \cup \{2\} = \{0, 1\} \cup \{2\} = \{0, 1, 2\} \\
&\vdots \\
n + 1 &= n \cup \{n\} = \{0, \ldots, n - 1\} \cup \{n\} = \{0, 1, \ldots, n\} \\
&\vdots
\end{aligned}$$

Natürliche Zahlen als Äquivalenzklassen

Im vorangegangenen Abschnitt haben wir die natürlichen Zahlen aus der leeren Menge schrittweise sozusagen "von unten" aufgebaut. Wir können aber auch eine Sicht "von oben" einnehmen. Dazu definieren wir, was eine endliche Menge ist und was es heisst, dass endliche Mengen gleiche Mächtigkeit haben.

Definition 1.1. *Eine Menge A heisst* endlich*, wenn es jede injektive Abbildung $A \to A$ auch surjektiv ist. Zwei endliche Mengen A und B heissen* gleich mächtig*, in Zeichen $A \sim B$, wenn es eine Bijektion $A \to B$ gibt.*

Der Vorteil dieser Definition ist, dass sie die früher definierten natürlichen Zahlen nicht braucht, diese werden jetzt erst konstruiert. Dazu fassen wir in der Menge aller endlichen Mengen die gleich mächtigen Mengen zusammen, bilden also die Äquivalenzklassen der Relation $\sim$.

Der Vorteil dieser Sichtweise ist, dass die natürlichen Zahlen ganz explizit als die Anzahlen von Elementen einer endlichen Menge entstehen. Eine natürlich Zahl ist also eine Äquivalenzklasse $[\![A]\!]$, die alle endlichen Mengen enthält, die die gleiche Mächtigkeit wie A haben. Zum Beispiel gehört dazu auch die Menge, die im vorangegangenen Abschnitt aus der leeren Menge aufgebaut wurde.

Die Mächtigkeit einer endlichen Menge A ist die Äquivalenzklasse, in der die Menge drin ist: $|A| = [\![A]\!] \in \mathbb{N}$ nach Konstruktion von $\mathbb{N}$. Aus logischer Sicht etwas problematisch ist allerdings, dass wir von der "Menge aller endlichen Mengen" sprechen ohne uns zu versichern, dass dies tatsächlich eine zulässige Konstruktion ist.

1.2 Ganze Zahlen

Die Menge der ganzen Zahlen löst das Problem, dass nicht jede Gleichung der Form $x + a = b$ mit $a, b \in \mathbb{N}$ eine Lösung $x \in \mathbb{N}$ hat. Dazu ist erforderlich, den natürlichen Zahlen die negativen Zahlen hinzuzufügen, also wieder die Existenz neuer Objekte zu postulieren, die die Rechenregeln weiterhin erfüllen.

Paare von natürlichen Zahlen

Die ganzen Zahlen können konstruiert werden als Paare (u, v) von natürlichen Zahlen $u, v \in \mathbb{N}$. Die Paare der Form $(u, 0)$ entsprechen den natürlichen Zahlen, die Paare $(0, v)$ sind die negativen Zahlen.

Die Rechenoperationen sind wie folgt definiert:

$$\begin{aligned}(a,b)+(u,v) &= (a+u, b+v)\\ (a,b)\cdot(u,v) &= (au+bv, av+bu)\end{aligned} \tag{1.2}$$

Die Darstellung ganzer Zahlen als Paare von natürlichen Zahlen findet man auch in der Buchhaltung, wo man statt eines Vorzeichen *Soll* und *Haben* verwendet. Dabei kommt es nur auf die Differenz der beiden Positionen an. Fügt man beiden Positionen den gleichen Betrag hinzu, ändert sich nichts. Viele der Paare (a,b) müssen also als äquivalent angesehen werden.

Äquivalenzrelation

Die Definition (1.2) erzeugt neue Paare, die wir noch nicht interpretieren können. Zum Beispiel ist $0 = 1 + (-1) = (1,0) + (0,1) = (1,1)$. Die Paare (u,u) müssen daher alle mit 0 identifiziert werden. Es folgt dann auch, dass alle Paare von natürlichen Zahlen mit "gleicher Differenz" den gleichen ganzzahligen Wert darstellen, allerdings können wir das nicht so formulieren, da ja die Differenz noch gar nicht definiert ist. Stattdessen gelten zwei Paare als äquivalent, wenn

$$(a,b) \sim (c,d) \qquad \Leftrightarrow \qquad a+d = c+b \tag{1.3}$$

gilt. Diese Bedingung erhält man, indem man zu $a - b = c - d$ die Summe $b + d$ hinzuaddiert. Ein ganzen Zahl z ist daher eine Menge von Paaren von natürlichen Zahlen mit der Eigenschaft

$$(a,b) \in z \ \wedge (a',b') \in z \qquad \Leftrightarrow \qquad (a,b) \sim (a',b') \qquad \Leftrightarrow \qquad a+b' = a'+b.$$

Man nennt eine solche Menge eine *Äquivalenzklasse* der Relation $\sim$.

Die Menge $\mathbb{Z}$ der *ganzen Zahlen* ist die Menge aller solchen Äquivalenzklassen. Die Menge der natürlichen Zahlen $\mathbb{N}$ ist in evidenter Weise darin eingebettet als die Menge der Äquivalenzklassen von Paaren der Form $(n,0)$.

Entgegengesetzter Wert

Zu jeder ganzen Zahl z dargestellt durch das Paar (a,b) stellt das Paar (b,a) eine ganze Zahl dar mit der Eigenschaft

$$z + (b,a) = (a,b) + (b+a) = (a+b, a+b) \sim (0,0) = 0. \tag{1.4}$$

Die von (b,a) dargestellte ganze Zahl wird mit $-z$ bezeichnet, die Rechnung (1.4) lässt sich damit abgekürzt als $z + (-z) = 0$ schreiben.

Lösung von Gleichungen

Gleichungen der Form $a = x + b$ können jetzt für beliebige ganze Zahlen immer gelöst werden. Dazu schreibt man $a, b \in \mathbb{N}$ als Paare und sucht die Lösung in der Form $x = (u,v)$. Man erhält

$$\begin{aligned}(a,0) &= (u,v) + (b,0)\\ (a+b,b) &= (u+b,v)\end{aligned}$$

Das Paar $(u,v) = (a,b)$ ist eine Lösung, die man normalerweise als $a - b = (a,0) + (-(b,0)) = (a,0) + (0,b) = (a,b)$ schreibt.

Ring

Die ganzen Zahlen sind ein Beispiel für einen sogenannten Ring, eine algebraische Struktur in der Addition, Subtraktion und Multiplikation definiert sind. Weitere Beispiel werden später vorgestellt, der Ring der Polynome $\mathbb{Z}[X]$ in Kapitel 3 und der Ring der $n \times n$-Matrizen in Kapitel 2. In einem Ring wird nicht verlangt, dass die Multiplikation kommutativ ist, Matrizenringe sind nicht kommutativ. $\mathbb{Z}$ ist ein kommutativer Ring ebenso sind die Polynomringe kommutativ. Die Theorie der nicht kommutativen Ringe ist sehr viel reichhaltiger und leider auch komplizierter als die kommutative Theorie.

1.3 Rationale Zahlen

In den ganzen Zahlen sind immer noch nicht alle linearen Gleichungen lösbar, es gibt keine ganze Zahl x mit $3x = 1$. Die nötige Erweiterung der ganzen Zahlen lernen Kinder noch bevor sie die negativen Zahlen kennenlernen.

Wir können hierbei denselben Trick anwenden, wie schon beim Übergang von den natürlichen zu den ganzen Zahlen. Wir kreieren wieder Paare (z, n), deren Elemente nennen wir *Zähler* und *Nenner*, wobei $z, n \in \mathbb{Z}$ und zudem $n \neq 0$. Die Rechenregeln für Addition und Multiplikation lauten

$$(a, b) + (c, d) = (ad + bc, bd) \qquad \text{und} \qquad (a, b) \cdot (c, d) = (ac, bd).$$

Die ganzen Zahlen lassen sich als in dieser Darstellung als $z \mapsto (z, 1)$ einbetten.

Ähnlich wie schon bei den ganzen Zahlen ist diese Darstellung aber nicht eindeutig. Zwei Paare sind äquivalent, wenn sich deren beide Elemente um denselben Faktor unterscheiden,

$$(a, b) \sim (c, d) \quad \Leftrightarrow \quad \exists \lambda \in \mathbb{Z} \colon \lambda a = c \wedge \lambda b = d.$$

Dass es sich hierbei wieder um eine Äquivalenzrelation handelt, lässt sich einfach nachprüfen.

Durch die neuen Regen gibt es nun zu jedem Paar (a, b) mit $a \neq 0$ ein Inverses (b, a) bezüglich der Multiplikation, wie man anhand der folgenden Rechnung sieht,

$$(a, b) \cdot (b, a) = (a \cdot b, b \cdot a) = (a \cdot b, a \cdot b) \sim (1, 1).$$

Brüche

Rationale Zahlen sind genau die Äquivalenzklassen dieser Paare (a, b) von ganzen Zahlen a und $b \neq 0$. Da diese Schreibweise recht unhandlich ist, wird normalerweise die Notation als Bruch $\frac{a}{b}$ verwendet. Die Rechenregeln werden dadurch zu den wohlvertrauten

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \qquad \text{und} \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

und die speziellen Brüche $\frac{0}{b}$ und $\frac{1}{1}$ erfüllen die Regeln

$$\frac{a}{b} + \frac{0}{d} = \frac{ad}{bd} \sim \frac{a}{b}, \qquad \frac{a}{b} \cdot \frac{0}{c} = \frac{0}{bc} \qquad \text{und} \qquad \frac{a}{b} \cdot \frac{1}{1} = \frac{a}{b}.$$

Wir sind uns gewohnt, die Brüche $\frac{0}{b}$ mit der Zahl 0 und $\frac{1}{1}$ mit der Zahl 1 zu identifizieren.

Kürzen

Wie bei den ganzen Zahlen entstehen durch die Rechenregeln viele Brüche, denen wir den gleichen Wert zuordnen möchten. Zum Beispiel folgt

$$\frac{ac}{bc} - \frac{a}{b} = \frac{abc - abc}{b^2c} = \frac{0}{b^2c},$$

wir müssen also die beiden Brüche als gleichwertig betrachten. Allgemein gelten die zwei Brüche $\frac{a}{b}$ und $\frac{c}{d}$ als äquivalent, wenn $ad - bc = 0$ gilt. Dies ist gleichbedeutend mit der früher definierten Äquivalenzrelation und bestätigt, dass die beiden Brüche

$$\frac{ac}{bc} \qquad \text{und} \qquad \frac{a}{b}$$

als gleichwertig zu betrachten sind. Der Übergang von links nach rechts heisst *Kürzen*, der Übergang von rechts nach links heisst *Erweitern*. Eine rationale Zahl ist also eine Menge von Brüchen, die durch Kürzen und Erweitern ineinander übergeführt werden können.

Die Menge der Äquivalenzklassen von Brüchen ist die Menge $\mathbb{Q}$ der rationalen Zahlen. In $\mathbb{Q}$ sind Addition, Subtraktion und Multiplikation mit den gewohnten Rechenregeln, die bereits in $\mathbb{Z}$ gegolten haben, uneingeschränkt möglich.

Kehrwert

Zu jedem Bruch $\frac{a}{b}$ lässt sich der Bruch $\frac{b}{a}$, der sogenannte *Kehrwert* konstruieren. Er hat die Eigenschaft, dass

$$\frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ba} = 1$$

gilt. Der Kehrwert ist also das multiplikative Inverse, jede von 0 verschiedene rationale Zahl hat eine Inverse.

Lösung von linearen Gleichungen

Mit dem Kehrwert lässt sich jetzt jede lineare Gleichung lösen. Die Gleichung $ax = b$ hat die Lösung

$$ax = \frac{a}{1}\frac{u}{v} = \frac{b}{1} \qquad \Rightarrow \qquad \frac{1}{a}\frac{a}{1}\frac{u}{v} = \frac{1}{a}\frac{b}{1} \qquad \Rightarrow \qquad \frac{u}{v} = \frac{b}{a}.$$

Dasselbe gilt auch für rationale Koeffizienten a und b. In der Menge $\mathbb{Q}$ kann man also beliebige lineare Gleichungen lösen.

Körper

$\mathbb{Q}$ ist ein Beispiel für einen sogenannten *Körper*, in dem die arithmetischen Operationen Addition, Subtraktion, Multiplikation und Division möglich sind mit der einzigen Einschränkung, dass nicht durch 0 dividiert werden kann. Körper sind die natürliche Bühne für die lineare Algebra, da sich lineare Gleichungssysteme ausschliesslich mit den Grundoperation lösen lassen.

Wir werden im Folgenden für verschiedene Anwendungszwecke weitere Körper konstruieren, zum Beispiel die reellen Zahlen $\mathbb{R}$ und die rationalen Zahlen $\mathbb{C}$. Wann immer die Wahl des Körpers keine Rolle spielt, werden wir den Körper mit $\mathbb{k}$ bezeichnen.

1.4 Reelle Zahlen

In den rationalen Zahlen lassen sich algebraische Gleichungen höheren Grades immer noch nicht lösen. Dass die Gleichung $x^2 = 2$ keine rationale Lösung hat, ist schon den Pythagoräern aufgefallen. Die geometrische Intuition der Zahlengeraden führt uns dazu, nach Zahlen zu suchen, die gute Approximationen für $\sqrt{2}$ sind. Wir können zwar keinen Bruch angeben, dessen Quadrat 2 ist, aber wenn es eine Zahl $\sqrt{2}$ mit dieser Eigenschaft gibt, dann können wir dank der Ordnungsrelation feststellen, dass sie in all den folgenden, kleiner werdenden Intervallen

$$\left[1, \frac{3}{2}\right], \left[\frac{7}{5}, \frac{17}{12}\right], \left[\frac{41}{29}, \frac{99}{70}\right], \left[\frac{239}{169}, \frac{577}{408}\right], \dots$$

enthalten sein muss[1]. Jedes der Intervalle enthält auch das nachfolgende Intervall, und die intervalllänge konvergiert gegen 0. Eine solche *Intervallschachtelung* beschreibt also genau eine Zahl, aber möglicherweise keine, die sich als Bruch schreiben lässt.

Die Menge $\mathbb{R}$ der reellen Zahlen kann man auch als Menge aller Cauchy-Folgen $(a_n)_{n\in\mathbb{N}}$ betrachten. Eine Folge ist eine Cauchy-Folge, wenn es für jedes $\varepsilon > 0$ eine Zahl $N(\varepsilon)$ gibt derart, dass $|a_n - a_m| < \varepsilon$ für $n, m > N(\varepsilon)$. Ab einer geeigneten Stelle $N(\varepsilon)$ sind die Folgenglieder also mit Genauigkeit ε nicht mehr unterscheidbar.

Nicht jede Cauchy-Folge hat eine rationale Zahl als Grenzwert. Da wir für solche Folgen noch keine Zahlen als Grenzwerte haben, nehmen wir die Folge als eine mögliche Darstellung der Zahl. Die Folge kann man ja auch verstehen als eine Vorschrift, wie man Approximationen der Zahl berechnen kann.

Zwei verschiedene Cauchy-Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ können den gleichen Grenzwert haben. So sind

$$\begin{aligned} a_n\colon &\quad 1, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \frac{41}{29}, \frac{99}{70}, \frac{239}{169}, \frac{577}{408}, \dots \\ b_n\colon &\quad 1, 1.4, 1.41, 1.412, 1.4142, 1.41421, 1.414213, 1.4142135, \dots \end{aligned}$$

beide Folgen, die die Zahl $\sqrt{2}$ approximieren. Im Allgemeinen tritt dieser Fall ein, wenn $|a_n - b_n|$ eine Folge mit Grenzwert 0 oder Nullfolge ist. Eine reelle Zahl ist also die Menge aller rationalen Cauchy-Folgen, deren Differenzen Nullfolgen sind.

Die Menge $\mathbb{R}$ der reellen Zahlen kann man also ansehen als bestehend aus Mengen von Folgen, die alle den gleichen Grenzwert haben. Die Rechenregeln der Analysis

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n \qquad \text{und} \qquad \lim_{n\to\infty} a_n \cdot b_n = \lim_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n$$

stellen sicher, dass sich die Rechenoperationen von den rationalen Zahlen auf die reellen Zahlen übertragen lassen.

1.5 Komplexe Zahlen

In den reellen Zahlen lassen sich viele algebraische Gleichungen lösen. Andere, z. B. die Gleichung

$$x^2 + 1 = 0, \tag{1.5}$$

[1] Die Näherungsbrüche konvergieren sehr schnell, sie sind mit der sogenannten Kettenbruchentwicklung der Zahl $\sqrt{2}$ gewonnen worden.

haben weiterhin keine Lösung. Der Grund dafür ist das Bestreben bei der Konstruktion der reellen Zahlen, die Ordnungsrelation zu erhalten. Diese ermöglicht, Näherungsintervall und Intervallschachtelungen zu definieren.

Die Ordnungsrelation sagt aber auch, dass $x^2 \geq 0$ ist für jedes $x \in \mathbb{R}$, so dass $x^2+1 > 0$ sein muss. Dies ist der Grund, warum die Gleichung 1.5 keine Lösung in $\mathbb{R}$ haben kann. Im Umkehrschluss folgt auch, dass eine Erweiterung der reellen Zahlen, in der die Gleichung (1.5) lösbar ist, ohne die Ordnungsrelation auskommen muss. Es muss darin Zahlen geben, deren Quadrat negativ ist und der Grössenvergleich dieser Zahlen untereinander ist nur eingeschränkt möglich.

Imaginäre und komplexe Zahlen

Den reellen Zahlen fehlen also Zahlen, deren Quadrat negativ ist. Nach inzwischen bewährtem Muster konstruieren wird die neuen Zahlen daher als Paare (a, b). Die erste Komponente soll die bekannten reellen Zahlen darstellen, deren Quadrat positiv ist. Die zweite Komponente soll für die Zahlen verwendet werden, deren Quadrat negativ ist. Die Zahl, deren Quadrat -1 sein soll, bezeichnen wir auch mit dem Paar $(0, 1)$ und schreiben dafür auch $i = (0, 1)$ mit $i^2 = -1$.

Die Rechenregeln sollen weiterhin erhalten bleiben, sie müssen daher wie folgt definiert werden:

$$\begin{aligned} (a, b) + (c, d) &= (a + c, b + d) \qquad & (a + bi) + (c + di) &= (a + c) + (b + d)i \\ (a, b) \cdot (c, d) &(ad - bd, ad + bc) \qquad & (a + bi) \cdot (c + di) &= ac - bd + (ad + bc)i. \end{aligned} \tag{1.6}$$

Diese Regeln ergeben sich ganz natürlich aus den Rechenregeln in $\mathbb{R}$ unter Berücksichtigung der Regel $i^2 = -1$.

Eine komplexe Zahl ist ein solches Paar, die Menge der komplexen Zahlen ist

$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}$$

mit den Rechenoperationen (1.6). Die Menge $\mathbb{C}$ verhält sich daher wie eine zweidimensionaler reeller Vektorraum.

Real- und Imaginärteil

Ist $z = a + bi$ eine komplexe Zahl, dann heisst a der Realteil $a = \Re z$ und b heisst der Imaginärteil $\Im z$. Real- und Imaginärteil sind lineare Abbildungen $\mathbb{C} \to \mathbb{R}$, sie projizieren einen Punkt auf die Koordinatenachsen, die entsprechend auch die reelle und die imaginäre Achse heissen.

Die Multiplikation mit i vertauscht Real- und Imaginärteil:

$$\Re(iz) = -b = -\Im z \qquad \text{und} \qquad \Im(iz) = a = \Re z.$$

Zusätzlich kehrt das Vorzeichen der einen Komponente. Wir kommen auf diese Eigenschaft zurück, wenn wir später in Abschnitt XXX komplexe Zahlen als Matrizen beschreiben.

Komplexe Konjugation

Der komplexen Zahl $u = a + bi$ ordnen wir die sogenannte *komplex konjugierte* Zahl $\overline{z} = a - bi$. Mit Hilfe der komplexen Konjugation kann man den Real- und Imaginärteil algebraisch ausdrücken:

$$\Re z = \frac{z + \overline{z}}{2} = \frac{a + bi + a - bi}{2} = \frac{2a}{2} = a \qquad \text{und} \qquad \Im z = \frac{z - \overline{z}}{2i} = \frac{a + bi - a + bi}{2i} = \frac{2bi}{2i} = b.$$

In der Gaussschen Zahlenebene ist die komplexe Konjugation eine Spiegelung an der reellen Achse.

Betrag

In $\mathbb{R}$ kann man die Ordnungsrelation dazu verwenden zu entscheiden, ob eine Zahl 0 ist. Wenn $x \geq 0$ ist und $x \leq 0$, dann ist $x = 0$. In $\mathbb{C}$ steht diese Ordnungsrelation nicht mehr zur Verfügung. Eine komplexe Zahl ist von 0 verschieden, wenn die Länge des Vektors in der Zahlenebene verschieden von 0 ist. Wir definieren daher den Betrag einer komplexen Zahl $z = a + bi$ als

$$|z|^2 = a^2 + b^2 = (\Re z)^2 + (\Im z)^2 \qquad \Rightarrow \qquad |z| = \sqrt{a^2 + b^2} = \sqrt{(\Re z)^2 + (\Im z)^2}.$$

Der Betrag lässt sich auch mit Hilfe der komplexen Konjugation ausdrücken, es ist $z\overline{z} = (a + bi)(a - bi) = a^2 + abi - abi + b^2 = |z|^2$. Der Betrag ist immer eine reelle Zahl.

Division

Die Erweiterung zu den komplexen Zahlen muss auch die Division erhalten. Dies ist durchaus nicht selbstverständlich. Man kann zeigen, dass ein Produkt von Vektoren eines Vektorraums nur für einige wenige, niedrige Dimensionen überhaupt möglich ist. Für die Division sind die Einschränkungen noch gravierender, die einzigen Dimensionen > 1, in denen ein Produkt mit einer Division definiert werden kann[2], sind 2, 4 und 8. Nur in Dimension 2 ist ein kommutatives Produkt möglich, dies muss das Produkt der komplexen Zahlen sein.

Wie berechnet man den Quotienten $\frac{z}{w}$ für zwei beliebige komplexe Zahlen $z = a + bi$ und $w = c + di$ mit $w \neq 0$? Dazu erweitert man den Bruch mit der komplex konjugierten des Nenners:

$$\frac{z}{w} = \frac{z\overline{w}}{w\overline{w}} = \frac{z\overline{w}}{|w|^2}$$

Da der Nenner $|w|^2 > 0$ eine reelle Zahl ist, ist die Division einfach, es ist die Multiplikation mit der reellen Zahl $1/|w|^2$.

Wir können den Quotienten auch in Komponenten ausdrücken:

$$\frac{z}{w} = \frac{a + bi}{c + di} = \frac{(a + bi)(c + di)}{(c + di)(c - di)} = \frac{ac - bd + (ad + bc)i}{c^2 + d^2}.$$

Gausssche Zahlenebene

Beschränkt man die Multiplikation auf einen reellen Faktor, wird $\mathbb{C}$ zu einem zweidimensionalen reellen Vektorraum. Man kann die komplexe Zahl $a + bi$ daher auch als Punkt (a, b) in der sogenannten Gaussschen Ebene betrachten. Die Addition von komplexen Zahlen ist in diesem Bild die vektorielle Addition, die Multiplikation mit reellen Zahlen werden wir weiter unten genauer untersuchen müssen.

Die Zahlenebene führt auf eine weitere Parametrisierung einer komplexen Zahl. Ein Punkt z der Ebene kann in Polarkoordinaten auch durch den Betrag und den Winkel zwischen der reellen Achse und dem Radiusvektor zum Punkt beschrieben werden.

[2] Der Beweis dieser Aussage ist ziemlich schwierig und wurde erst im 20. Jahrhundert mit Hilfe der Methoden der algebraischen Topologie erbracht. Eine Übersicht über den Beweis kann in Kapitel 10 von [1] gefunden werden.

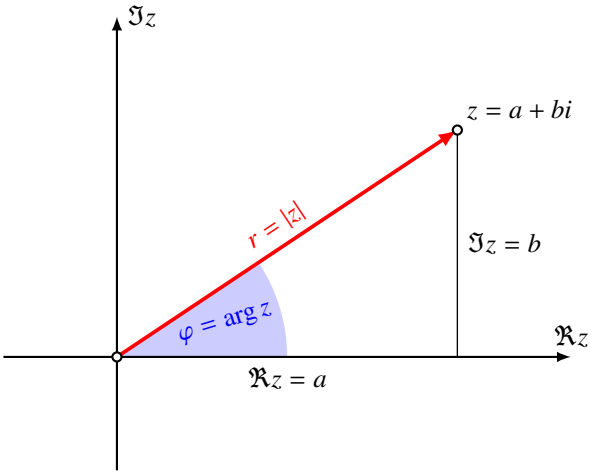


Abbildung 1.1: Argument und Betrag einer komplexen Zahl $z = a + ib$ in der Gaussschen Zahlenebene

Geometrische Interpretation der Rechenoperationen

Die Addition kompelxer Zahlen wurde bereits als Vektoraddition in der Gausschen Zahlenebene. Die Multiplikation ist etwas komplizierter, wir berechnen Betrag und Argument von zw separat. Für den Betrag erhalten wir

$$|zw|^2 = z\bar{z}w\overline{w} = |z|^2|w|^2$$

Der Betrag des Produktes ist also das Produkt der Beträge.

Für das Argument verwenden wir, dass

$$\tan\arg z = \frac{\Im z}{\Re z} = \frac{b}{a} \qquad \Rightarrow \qquad b = a\tan\arg z$$

und analog für w. Bei der Berechnung des Produktes behandeln wir nur den Fall $a \neq 0$ und $c \neq 0$, was uns ermöglicht, den Bruch durch ac zu kürzen:

$$\tan\arg wz = \frac{\Im wz}{\Re wz} = \frac{ad + bc}{ac - bd} = \frac{\frac{d}{c} + \frac{b}{a}}{1 - \frac{b}{a}\frac{d}{c}} = \frac{\tan\arg z + \tan\arg w}{1 + \tan\arg z \cdot \tan\arg w} = \tan(\arg z + \arg w).$$

Im letzten Schritt haben wir die Additionsformel für den Tangens verwendet. Daraus liest man ab, dass das Argument eines Produkts die Summe der Argumente ist. Die Multiplikation mit einer festen komplexen Zahl führt also mit der ganzen komplexen Ebene eine Drehstreckung durch. Auf diese geometrische Beschreibung der Multiplikation werden wir zurückkommen, wenn wir die komplexen Zahlen als Matrizen beschreiben wollen.

Algebraische Vollständigkeit

Die komplexen Zahlen $\mathbb{C}$ sind als Erweiterung von $\mathbb{R}$ so konstruiert worden, dass die Gleichung $x^2 + 1 = 0$ eine Lösung hat. Etwas überraschend ist dagegen, dass in dieser Erweiterung jetzt jede

beliebige algebraische Gleichung lösbar geworden. Dies ist der Inhalt des Fundamentalsatzes der Algebra.

Satz 1.2 (Fundamentalsatz der Algebra)**.** *Jede algebraische Gleichung der Form*

$$p(x) = x^n + a_{n-1}x^{n-1} + a_1 x + a_0 = 0, \qquad a_k \in \mathbb{C}$$

mit komplexen Koeffizienten hat n möglicherweise mit Vielfachheit gezähle Nullstellen $\alpha_1, \ldots, \alpha_m$*, d. h. das Polynom* $p(x)$ *lässt sich in Linearfaktoren*

$$p(x) = (x - \alpha_1)^{k_1}(x - \alpha_2)^{k_2} \cdot \ldots \cdot (x - \alpha_m)^{k_m}$$

zerlegen, wobei $k_1 + k_2 + \cdots + k_m = n$*. Die Zahlen* k_j *heisst die* Vielfachheit *der Nullstelle* α_j*.*

Der Fundamentalsatz der Algebra wurde erstmals von Carl Friedrich Gauss bewiesen. Seither sind viele alternative Beweise mit Methoden aus den verschiedensten Gebieten der Mathematik gegeben worden. Etwas salopp könnten man sagen, dass der Fundamentalsatz ausdrückt, dass die Konstruktion der Zahlensysteme mit $\mathbb{C}$ abgeschlossen ist, soweit damit die Lösbarkeit beliebiger Gleichungen angestrebt ist.

Quaternionen und Octonionen

Die komplexen Zahlen ermöglichen eine sehr effiziente Beschreibung geometrischer Abbildungen wie Translationen, Spiegelungen und Drehstreckungen in der Ebene. Es drängt sich damit die Frage auf, ob sich $\mathbb{C}$ so erweitern lässt, dass man damit auch Drehungen im dreidimensionalen Raum beschreiben könnte. Da Drehungen um verschiedene Achsen nicht vertauschen, kann eine solche Erweiterung nicht mehr kommutativ sein.

William Rowan Hamilton propagierte ab 1843 eine Erweiterung von $\mathbb{C}$ mit zwei zusätzlichen Einheiten j und k mit den nichtkommutativen Relationen

$$i^2 = j^2 = k^2 = ijk = -1. \tag{1.7}$$

Er nannte die Menge aller Linearkombinationen

$$\mathbb{H} = \{a_0 + a_1 i + a_2 j + a_3 k \mid a_l \in \mathbb{R}\}$$

die *Quaternionen*, die Einheiten i, j und k heissen auch Einheitsquaternionen. Konjugation, Betrag und Division können ganz ähnlich wie bei den komplexen Zahlen definiert werden und machen $\mathbb{H}$ zu einer sogenannten *Divisionsalgebra*. Alle Rechenregeln mit Ausnahme der Kommutativität der Multiplikation sind weiterhin gültig und durch jede von 0 verschiedene Quaternion kann auch dividiert werden.

Aus den Regeln für die Quadrate der Einheiten in (1.7) folgt zum Beispiel $i^{-1} = -i$, $j^{-1} = -j$ und $k^{-1} = -k$. Die letzte Bedingung liefert daraus

$$ijk = -1 \qquad \Rightarrow \qquad \begin{cases} ij = ijkk^{-1} = -1k^{-1} = k \\ i^2 jk = -i = -jk \\ -j^2 k = -ji = k \end{cases}$$

Aus den Relationen (1.7) folgt also insbesondere auch, dass $ij = -ji$. Ebenso kann abgeleitet werden, dass $jk = -kj$ und $ik = -ki$. Man sagt, die Einheiten sind *antikommutativ*.

Die Beschreibung von Drehungen mit Quaternionen ist in der Computergraphik sehr beliebt, weil eine Quaternion mit nur vier Komponenten $a_0, \ldots, a_3$ vollständig beschrieben ist. Eine Transformationsmatrix des dreidimensionalen Raumes enthält dagegen neun Koeffizienten, die vergleichsweise komplizierte Abhängigkeiten erfüllen müssen. Quaternionen haben auch in weiteren Gebieten interessante Anwendungen, zum Beispiel in der Quantenmechanik, wo antikommutierende Operatoren bei der Beschreibung von Fermionen eine zentrale Rolle spielen.

Aus rein algebraischer Sicht kann man die Frage stellen, ob es eventuell auch noch grössere Divisionsalgebren gibt, die $\mathbb{H}$ erweitern. Tatsächlich hat Arthur Cayley 1845 eine achtdimensionale Algebra, die Oktonionen $\mathbb{O}$, mit vier weiteren Einheiten beschrieben. Allerdings sind die Oktonionen nur beschränkt praktisch anwendbar. Grund dafür ist die Tatsache, dass die Multiplikation in $\mathbb{O}$ nicht mehr assoziativ ist. Das Produkt von mehr als zwei Faktoren aus $\mathbb{O}$ ist von der Reihenfolge der Ausführung der Multiplikationen abhängig.

Kapitel 2

Vektoren und Matrizen

2.1 Lineare Algebra

In diesem Abschnitt tragen wir die bekannten Resultate der linearen Algebra zusammen. Meistens lernt man diese zuerst für Vektoren und Gleichungssyteme mit reellen Variablen. In der linearen Algebra werden aber nur die arithmetischen Grundoperationen verwendet, es gibt also keinen Grund, warum sich die Theorie nicht über einem beliebigen Zahlenkörper entwickeln lassen sollte. Die in Kapitel 4 untersuchten endlichen Körper sind zum Beispiel besser geeignet für Anwendungen in der Kryptographie oder für die diskrete schnelle Fourier-Transformation. Daher geht es in diesem Abschnitt weniger darum alles herzuleiten, sondern vor allem darum, die Konzepte in Erinnerung zu rufen und so zu formulieren, dass offensichtlich wird, dass alles mit einem beliebigen Zahlkörper $\mathbb{k}$ funktioniert.

2.1.1 Vektoren

Koordinatensysteme haben ermöglicht, Punkte als Zahlenpaare zu beschreiben. Dies ermöglicht, geometrische Eigenschaften als Gleichungen auszudrücken, aber mit Punkten kann man trotzdem noch nicht rechnen. Ein Vektor fasst die Koordinaten eines Punktes in einem Objekt zusammen, mit dem man auch rechnen und zum Beispiel Parallelverschiebungen algebraisieren kann. Um auch Streckungen ausdrücken zu können, wird auch eine Menge von Streckungsfaktoren benötigt, mit denen alle Komponenten eines Vektors multipliziert werden können. Sie heissen auch *Skalare* und liegen in $\mathbb{k}$.

Zeilen- und Spaltenvektoren

Vektoren sind Tupel von Elementen aus $\mathbb{k}$.

Definition 2.1. *Ein n-dimensionaler* Spaltenvektor *ist ein n-Tupel von Zahlen aus* $\mathbb{k}$ *geschrieben als*

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{k}^n.$$

Ein m-dimensionaler Zeilenvektor *wird geschrieben als*

$$u = \begin{pmatrix} u_1 & u_2 & \dots & u_m \end{pmatrix} \in \mathbb{k}^m.$$

Für Vektoren gleicher Dimension sind zwei Rechenoperationen definiert. Die *Addition von Vektoren* $a, a \in \mathbb{k}^n$ und die Multiplikation eines Vektors mit einem Skalar $\lambda \in \mathbb{k}$ erfolgt elementweise:

$$a + b = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ \vdots \\ a_n + b_n \end{pmatrix}, \qquad \lambda a = \lambda \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \vdots \\ \lambda a_n \end{pmatrix}.$$

Die üblichen Rechenregeln sind erfüllt, nämlich

Kommutativität:	$a + b = b + a$		$\forall a, b \in V$
Assoziativgesetze:	$(a + b) + c = a + (b + c)$	$(\lambda\mu)a = \lambda(\mu a)$	$\forall a, b, c \in V,\ \lambda, \mu \in \mathbb{k}$ (2.1)
Distributivgesetze:	$\lambda(a + b) = \lambda a + \lambda b$	$(\lambda + \mu)a = \lambda a + \mu a$	$\forall a, b \in V,\ \lambda, \mu \in \mathbb{k}.$

Diese Gesetze drücken aus, dass man mit Vektoren so rechnen kann, wie man das in der Algebra gelernt hat, mit der einzigen Einschränkung, dass man Skalare immer links von Vektoren schreiben muss. Die Distributivgesetze zum Beispiel sagen, dass man Ausmultipilizieren oder Ausklammern kann genauso wie in Ausdrücken, die nur Zahlen enthalten.

Man beachte, dass es im allgemeinen kein Produkt von Vektoren gibt. Das aus der Vektorgeometrie bekannte Vektorprodukt ist eine Spezialität des dreidimensionalen Raumes, es gibt keine Entsprechung dafür in anderen Dimensionen.

Standardbasisvektoren

In $\mathbb{k}^n$ findet man eine Menge von speziellen Vektoren, durch die man alle anderen Vektoren ausdrücken kann. Mit den sogenannten *Standardbasisvektoren*

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

kann der Vektor $a \in \mathbb{k}^n$ als

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \dots + a_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = a_1 e_1 + a_2 e_2 + \dots + a_n e_n$$

ausgedrückt werden.

Vektorraum

Die Rechnungen, die man gemäss der Rechengesetze (2.1) anstellen kann, verlangen nicht, dass Elemente a und b, mit denen man da rechnet, Zeilen- oder Spaltenvektoren sind. Jede Art von mathematischem Objekt, mit dem man so rechen kann, kann als (abstrakter) Vektor betrachtet werden.

Definition 2.2. *Eine Menge V von Objekten, auf der zwei Operationen definiert, nämlich die Addition, geschrieben $a + b$ für $a, b \in V$ und die Multiplikation mit Skalaren, geschrieben λa für $a \in V$ und $\lambda \in \mathbb{k}$, heisst ein* $\mathbb{k}$-Vektorraum *oder* Vektorraum über $\mathbb{k}$ *(oder einfach nur* Vektorraum*, wenn $\mathbb{k}$ aus dem Kontext klar sind), wenn die Rechenregeln* (2.1) *gelten*

Die Mengen von Spaltenvektoren $\mathbb{k}^n$ sind ganz offensichtlich Vektorräume. Die in Kapitel 3 studierten Mengen von Polynomen mit Koeffizienten in $\mathbb{k}$ sind ebenfalls Vektorräume.

Beispiel. Die Zahlenmenge $\mathbb{C}$ ist ein $\mathbb{R}$-Vektorraum. Elemente von $\mathbb{C}$ können addiert und mit reellen Zahlen multipliziert werden. Die Rechenregeln für die komplexen Zahlen umfassen auch alle Regeln (2.1), also ist $\mathbb{C}$ ein Vektorraum über $\mathbb{R}$. ○

Beispiel. Die Menge $C([a, b])$ der stetigen Funktionen $[a, b] \to \mathbb{R}$ bildet ein Vektorraum. Funktionen können addiert und mit reellen Zahlen multipliziert werden:

$$(f + g)(x) = f(x) + g(x) \qquad \text{und} \qquad (\lambda f)(x) = \lambda f(x).$$

Dies reicht aber noch nicht ganz, denn $f + g$ und λf müssen ausserdem auch *stetige* Funktionen sein. Das dem so ist, lernt man in der Analysis. Die Vektorraum-Rechenregeln (2.1) sind ebenfalls erfüllt. ○

Die Beispiele zeigen, dass der Begriff des Vektorraums die algebraischen Eigenschaften eine grosse Zahl sehr verschiedenartiger mathematischer Objekte beschreiben kann. Alle Erkenntnisse, die man ausschliesslich aus Vekotorraumeigenschaften gewonnen hat, sind auf alle diese Objekte übertragbar. Im folgenden werden wir alle Aussagen für einen Vektorraum V formulieren, wenn wir die Darstellung als Tupel $\mathbb{k}^n$ nicht brauchen.

Gleichungssysteme in Vektorform

Die Vektorraum-Operationen erlauben nun auch, lineare Gleichungssysteme in *Vektorform* zu schreiben:

$$\left.\begin{array}{rcl} a_{11}x_1 + & \dots + & a_{1n}x_n = \quad b_1 \\ \vdots & \ddots & \vdots \quad\ \vdots \\ a_{m1}x_1 + & \dots + & a_{1n}x_n = \quad b_m \end{array}\right\} \quad \Rightarrow \quad x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + \dots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \tag{2.2}$$

Die rechte Seite von (2.2) ist eine Linearkombination der Spaltenvektoren.

Definition 2.3. *Eine Linearkombination der Vektoren $v_1, \dots, v_n \in V$ ist ein Ausdruck der Form*

$$v = \lambda_1 v_1 + \dots + \lambda_n v_n$$

mit $\lambda_1, \dots, \lambda_n \in \mathbb{k}$.

Die Menge aller Vektoren, die sich als Linearkombinationen einer gegebenen Menge ausdrücken lässt, heisst der aufgespannte Raum.

Definition 2.4. *Sind $a_1, \dots, a_n \in V$ Vektoren, dann heisst die Menge*

$$\langle a_1, \dots, a_n \rangle = \{x_1 a_1 + \dots + x_n a_n \mid x_1, \dots, x_n \in \mathbb{k}\}$$

aller Vektoren, die sich durch Linearkombination aus den Vektoren $a_1, \dots, a_n$ gewinnen lassen, der von $a_1, \dots, a_n$ aufgespannte Raum.

Lineare Abhängigkeit

Die Gleichung (2.2) drückt aus, dass sich der Vektor b auf der rechten Seite als Linearkombination der Spaltenvektoren ausdrücken lässt. Oft ist eine solche Darstellung auf nur eine Art und Weise möglich. Betrachten wir daher jetzt den Fall, dass es zwei verschiedene Linearkombinationen der Vektoren $a_1, \dots, a_n$ gibt, die beide den Vektor b ergeben. Deren Differenz ist

$$\left.\begin{array}{llll} x_1a_1 + & \dots + & x_na_n = & b \\ x'_1a_1 + & \dots + & x'_na_n = & b \end{array}\right\} \quad \Rightarrow \quad \underbrace{(x_1 - x'_1)}_{\lambda_1}a_1 + \dots + \underbrace{(x_n - x'_n)}_{\lambda_n}a_n = 0. \tag{2.3}$$

Die Frage, ob ein Gleichungssystem genau eine Lösung hat, hängt also damit zusammen, ob es Zahlen $\lambda_1, \dots, \lambda_n$ gibt, für die die Gleichung erfüllt ist.

Definition 2.5. *Die Vektoren $a_1, \dots, a_n$ heissen linear abhängig, wenn es Zahlen $\lambda_1, \dots, \lambda_n \in \mathbb{k}$ gibt, die nicht alle* 0 *sind, so dass*

$$\lambda_1a_1 + \dots + \lambda_na_n = 0. \tag{2.4}$$

Die Vektoren heissen linear abhängig, wenn aus (2.4) *folgt, dass alle $\lambda_1, \dots, \lambda_n = 0$ sind.*

Lineare Abhängigkeit der Vektoren $a_1, \dots, a_n$ bedeutet auch, dass man einzelne der Vektoren durch andere ausdrücken kann. Hat man nämlich eine Linearkombination (2.4) und ist der Koeffizient $\lambda_k \neq 0$, dann kann man nach a_k auflösen:

$$a_k = -\frac{1}{\lambda_k}(\lambda_1a_1 + \dots + \widehat{\lambda_ka_k} + \dots + \lambda_na_n).$$

Darin bedeutet der Hut, dass der entsprechende Term weggelassen werden muss. Da dies für jeden von 0 verschiedenen Koeffizienten möglich ist, sagt man eben nicht, a_k ist linear abhängig von den anderen, sondern man sagt $a_1, \dots, a_n$ sind (untereinander) linear abhängig.

Basis

Ein lineares Gleichungssystem fragt danach, ob und wie ein Vektor b als Linearkombination der Vektoren $a_1, \dots, a_n$ ausgedrückt werden kann. Wenn dies eindeutig möglich ist, dann haben die Vektoren $a_1, \dots, a_n$ offenbar eine besondere Bedeutung.

Definition 2.6. *Eine linear unabhängig Menge von Vektoren $\mathcal{B} = \{a_1, \dots, a_n\} \subset V$ heisst* Basis *von V. Die maximale Anzahl linear unabhängiger Vektoren in V heisst* Dimension *von V.*

Die Standardbasisvektoren bilden eine Basis von $V = \mathbb{k}^n$.

Unterräume

Die Mengen $\langle a_1, \dots, a_n \rangle$ sind Teilmengen von V, in denen die Addition von Vektoren und die Multiplikation mit Skalaren immer noch möglich ist.

Definition 2.7. *Eine Teilmenge $U \subset V$ heisst ein* Unterraum *von V, wenn U selbst ein $\mathbb{k}$-Vektorraum ist, also*

$$\begin{array}{rcl} a, b \in U & \Rightarrow & a + b \in U \\ a \in U,\ \lambda \in \mathbb{k} & \Rightarrow & \lambda a \in U \end{array}$$

gilt.

2.1.2 Matrizen

Die Koeffizienten eines linearen Gleichungssystems finden in einem Zeilen- oder Spaltenvektor nicht Platz. Wir erweitern das Konzept daher in einer Art, dass Zeilen- und Spaltenvektoren Spezialfälle sind.

Definition einer Matrix

Definition 2.8. *Eine $m \times n$-Matrix A (über $\mathbb{k}$) ist rechteckiges Schema*

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

mit $a_{ij} \in \mathbb{k}$. Die Menge aller $m \times n$-Matrizen wird mit

$$M_{m\times n}(\mathbb{k}) = \{A \mid A \text{ ist eine } m \times n\text{-Matrix}\}.$$

Falls $m = n$ gilt, heisst die Matrix A auch quadratisch *Man kürzt die Menge der quadratischen Matrizen als $M_n(\mathbb{k}) = M_{n\times n}(\mathbb{k})$ ab.*

Die m-dimensionalen Spaltenvektoren $v \in \mathbb{k}^m$ sind $m\times 1$-Matrizen $v \in M_{n\times 1}(\mathbb{k})$, die n-dimensionalen Zeilenvetoren $u \in \mathbb{k}^n$ sind $1 \times n$-Matrizen $v \in M_{1\times n}(\mathbb{k})$. Eine $m \times n$-Matrix A mit den Koeffizienten a_{ij} besteht aus den n Spaltenvektoren

$$a_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \quad a_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix}, \dots, a_n = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}.$$

Sie besteht auch aus den m Zeilenvektoren

$$\begin{pmatrix} a_{k1} & a_{k2} & \dots & a_{kn} \end{pmatrix}$$

mit $k = 1, \dots, m$.

Addition und Multiplikation mit Skalaren

Die $m \times n$-Matrizen $M_{m\times n}(\mathbb{k})$ bilden eine Vektorraum, die Addition von Matrizen und die Multiplikation wird wie folgt definiert.

Definition 2.9. *Sind $A, B \in M_{m\times n}(\mathbb{k})$ und $\lambda \in \mathbb{k}$, dann setzt man*

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix} \quad \text{und} \quad \lambda A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}.$$

Multiplikation

Will man ein lineares Gleichungssystem mit Hilfe der Matrix A der Koeffizienten schreiben, bekommt es die Form $Ax = b$, wobei der Vektor der rechten Seiten ist, und x ein Vektor von unbekannten Zahlen. Dies ist jedoch nur sinnvoll, wenn das Produkt Ax sinnvoll definiert werden kann.

Definition 2.10. *Eine $m \times n$-Matrix $A \in M_{m\times n}(\mathbb{k})$ und eine $n \times l$-Matrix $B \in M_{n\times l}(\mathbb{k})$ haben als Produkt eine $n \times l$-Matrix $C = AB \in M_{n\times l}(\mathbb{k})$ mit den Koeffizienten*

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{2.5}$$

Die Koeffizienten a_{ik} kommen aus der Zeile i von A, die Koeffizienten b_{kj} stehen in der Spalte j von B, die Multiplikationsregel (2.5) besagt also, dass das Element c_{ij} entsteht als das Produkt der Zeile i von A mit der Spalte j von C.

Einheitsmatrix

Welche $m \times m$-Matrix $I \in M_m(\mathbb{k})$ hat die Eigenschaft, dass $IA = A$ für jede beliebige Matrix $A \in M_{m\times n}(\mathbb{k})$. Wir bezeichnen die Einträge von I mit δ_{ij}. Die Bedingung $IA = A$ bedeutet

$$a_{ij} = \delta_{i1} a_{1j} + \cdots + \delta_{im} a_{mj},$$

Da auf der linken Seite nur a_{ij} vorkommt, müssen alle Terme auf der rechten Seite verschwinden ausser dem Term mit a_{ij}, dessen Koeffizient $\delta_{ii} = 1$ sein muss. Die Koeffizienten sind daher

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases}$$

Die Zahlen δ_{ij} heissen auch das *Kronecker-Symbol* oder *Kronecker-Delta*. Die Matrix I hat die Einträge δ_{ij} und heisst die *Einheitsmatrix*

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

2.1.3 Gleichungssysteme

Lineare Gleichungssysteme haben wir bereits in Vektorform geschrieben. Matrizen wurden eingeführt, um sie noch kompakter in der Matrixform $Ax = b$ zu schreiben. In diesem Abschnitt sollen die bekannten Resultate über die Lösung von linearen Gleichungssytemen zusammengetragen werden.

Eindeutige Lösung

Mit Hilfe der Vektorform eines linearen Gleichungssystems wurde gezeigt, dass die Lösung genau dann eindeutig ist, wenn die Spaltenvektoren der Koeffizientenmatrix linear unabhängig sind. Dies bedeutet, dass das Gleichungssystem

$$\begin{array}{ccc} a_{11}x_1 + & \dots + & a_{1n}x_n = 0 \\ \vdots & \ddots & \vdots \quad \vdots \\ a_{m1}x_1 + & \dots + & a_{mn}x_n = 0 \end{array} \tag{2.6}$$

eine nichttriviale Lösung haben muss. Das Gleichungssystem $Ax = b$ ist also genau dann eindeutig lösbar, wenn das homogene Gleichungssystem $Ax = 0$ nur die Nulllösung hat.

Inhomogene und homogene Gleichungssysteme

Ein Gleichungssystem mit 0 auf der rechten Seite ist also bereits ausreichend um zu entscheiden, ob die Lösung eindeutig ist. Ein Gleichungssystem mit rechter Seite 0 heisst *homogen*. Zu jedem *inhomogenen* Gleichungssystem $Ax = b$ mit $b \neq 0$ ist $Ax = 0$ das zugehörige homogene Gleichungssystem.

Ein homogenes Gleichungssytem $Ax = 0$ hat immer mindestens die Lösung $x = 0$, man nennt sie auch die *triviale* Lösung. Eine Lösung $x \neq 0$ heisst auch eine nichttriviale Lösung. Die Lösungen eines inhomgenen Gleichungssystem $Ax = b$ ist also nur dann eindeutig, wenn das zugehörige homogene Gleichungssystem eine nichttriviale Lösung hat.

Gauss-Algorithmus

Der Gauss-Algorithmus oder genauer Gausssche Eliminations-Algorithmus löst ein lineare Gleichungssystem der Form (2.2). Die Koeffizienten werden dazu in das Tableau

$$\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{array}\right]$$

geschrieben. Die vertikale Linie erinnert an die Position des Gleichheitszeichens. Es beinhaltet alle Informationen zur Durchführung des Algorithmus. Der Algorithmus is so gestaltet, dass er nicht mehr Speicher als das Tableau benötigt, alle Schritte operieren direkt auf den Daten des Tableaus.

In jedem Schritt des Algorithmus wird zunächst eine Zeile i und Spalte j ausgewählt, das Elemente a_{ij} heisst das Pivotelement. Die *Pivotdivision*

$$\left[\begin{array}{ccccc|c} a_{11} & \dots & a_{1j} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} & b_i \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} & b_m \end{array}\right] \rightarrow \left[\begin{array}{ccccc|c} a_{11} & \dots & a_{1j} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ \frac{a_{i1}}{a_{ij}} & \dots & 1 & \dots & \frac{a_{in}}{a_{ij}} & \frac{b_i}{a_{ij}} \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} & b_m \end{array}\right]$$

stellt sicher, dass das Pivot-Element zu 1 wird. Dies ist gleichbedeutend mit der Auflösung der Gleichung i noch der Variablen x_j. Mit der *Zeilensubtraktion* auf Zeile $k \neq i$ können die Einträge in der Spalte j zu Null gemacht werden. Dazu wird das a_{kj}-fache der Zeile i von Zeile k subtrahiert:

$$\left[\begin{array}{ccccc|c} \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{i1} & \dots & 1 & \dots & a_{in} & b_i \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{k1} & \dots & a_{kj} & \dots & a_{kn} & b_m \\ \vdots & & \vdots & \ddots & \vdots & \vdots \end{array}\right] \rightarrow \left[\begin{array}{ccccc|c} \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{i1} & \dots & 1 & \dots & a_{in} & b_i \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ a_{k1} - a_{kj}a_{i1} & \dots & 0 & \dots & a_{kn} - a_{kj}a_{in} & b_m - a_{kj}b_n \\ \vdots & & \vdots & \ddots & \vdots & \vdots \end{array}\right]$$

Typischerweise werden nach jeder Pivotdivision mehrer Zeilensubtraktionen durchgeführt um alle anderen Elemente der Pivotspalte ausser dem Pivotelement zu 0 zu machen. Beide Operationen können in einem Durchgang durchgeführt werden.

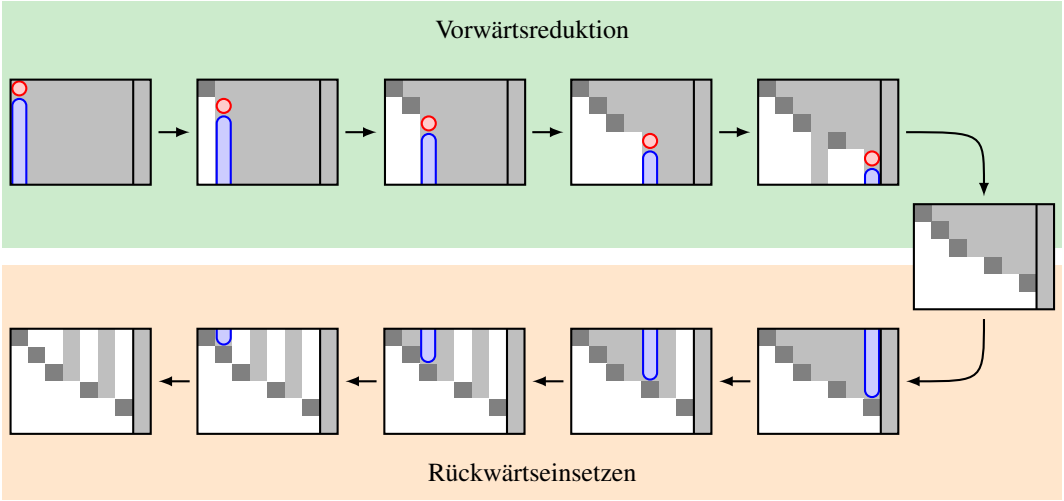


Abbildung 2.1: Zweckmässiger Ablauf der Berechnung des Gauss-Algorithmus. Falls in einer Spalte kein weiteres von 0 verschiedenes Pivotelement zur Verfügung steht, wird die Zeile übersprungen. Weisse Felder enthalten 0, dunkelgraue 1. Die roten Kreise bezeichnen Pivot-Elemente, die blauen Felder die mit einer Zeilensubtraktion zu 0 gemacht werden sollen.

Die beiden Operationen Pivotdivision und Zeilensubtraktion werden jetzt kombiniert um im linken Teil des Tableaus möglichst viele Nullen und Einsen zu erzeugen. Im Idealfall wird ein Tableau der Form

$$\left(\begin{array}{cccc|c} 1 & 0 & \dots & 0 & u_1 \\ 0 & 1 & \dots & 0 & u_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & u_m \end{array}\right)$$

erreicht, was natürlich nur $m = n$ möglich ist. Interpretiert man die Zeilen dieses Tableaus wieder als Gleichungen, dann liefert die Zeile i den Wert $x_i = u_i$ für die Variable i. Die Lösung kann also in der Spalte rechts abgelesen werden.

Die effizienteste Strategie für die Verwendung der beiden Operationen ist in Abbildung 2.1 dargestellt. In der Phase der *Vorwärtsreduktion* werden Pivotelemente von links nach rechts möglichst auf der Diagonale gewählt und mit Zeilensubtraktionen die darunterliegenden Spalten freigeräumt. Während des Rückwärtseinsetzens werden die gleichen Pivotelemente von rechts nach links genutzt, um mit Zeilensubtraktionen auch die Spalten über den Pivotelemnten frei zu räumen. Wenn in einer Spalte kein von 0 verschiedenes Element als Pivotelement zur Verfügung steht, wird diese Spalte übersprungen. Die so erzeuge Tableau-Form heisst auch die *reduzierte Zeilenstufenform* (*reduced row echelon form*, RREF).

Da der Ablauf des Gauss-Algorithmus vollständig von den Koeffizienten der Matrix A bestimmt ist, kann er gleichzeitig für mehrere Spalten auf der rechten Seite oder ganz ohne rechte Seite durchgeführt werden.

Lösungsmenge

Die Spalten, in denen im Laufe des Gauss-Algorithmus kein Pivotelement gefunden werden kann, gehören zu Variablen, nach denen sich das Gleichungssystem nicht auflösen lässt. Diese Variablen sind daher nicht bestimmt, sie können beliebig gewählt werden. Alle anderen Variablen sind durch diese frei wählbaren Variablen bestimmt.

Für ein Gleichungssystem $Ax = b$ mit Schlusstableau

$$\left(\begin{array}{ccccccccccc|c}
x_1 & x_2 & \dots & x_{j_i-1} & x_{j_1} & x_{j_1+1} & \dots & x_{j_2-1} & x_{j_2} & \dots & x_{j_k} & \\
1 & 0 & \dots & 0 & c_{1j_1} & 0 & \dots & 0 & c_{1j_2} & \dots & c_{1j_k} & d_1 \\
0 & 1 & \dots & 0 & c_{2j_1} & 0 & \dots & 0 & c_{2j_2} & \dots & c_{1j_k} & d_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 1 & c_{i_1,j_1} & 0 & \dots & 0 & c_{i_1,j_2} & \dots & c_{i_1j_k} & d_{i_1} \\
0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & c_{i_1+1,j_2} & \dots & c_{i_1+1,j_k} & d_{i_1+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & c_{i_2,j_2} & \dots & c_{i_2j_k} & d_{i_2} \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & c_{i_2+1,j_k} & d_{i_2+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & d_m
\end{array}\right) \tag{2.7}$$

mit den k frei wählbaren Variablen $x_{j_1}, x_{j_2}, \dots, x_{j_k}$ kann die Lösungsmenge als

$$\mathbb{L} = \left\{ \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{i_1} \\ d_{i_1+1} \\ \vdots \\ d_{i_2} \\ d_{i_2+1} \\ \vdots \\ d_m \end{pmatrix} + x_{j_1} \begin{pmatrix} -c_{1j_1} \\ -c_{2j_1} \\ \vdots \\ -c_{i_1,j_1} \\ 1 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + x_{j_1} \begin{pmatrix} -c_{1j_2} \\ -c_{2j_2} \\ \vdots \\ -c_{j_1,j_2} \\ -c_{j_1+1,j_2} \\ \vdots \\ -c_{i_2,j_2} \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \dots + x_{j_k} \begin{pmatrix} -c_{1j_k} \\ -c_{2j_k} \\ \vdots \\ -c_{j_1,j_k} \\ -c_{j_1+1,j_k} \\ \vdots \\ -c_{i_2,j_k} \\ -c_{i_2+1,j_k} \\ \vdots \\ 0 \end{pmatrix} \middle| \; x_{i_1}, x_{i_2}, \dots, x_{i_k} \in \mathbb{k} \right\}$$

geschrieben werden. Insbesondere ist die Lösungsmenge k-dimensional.

Inverse Matrix

Zu jeder quadratischen Matrix $A \in M_n(\mathbb{k})$ kann man versuchen, die Gleichungen

$$Ac_1 = e_1, \quad Ac_2 = e_2, \dots, Ac_n = e_n$$

mit den Standardbasisvektoren e_i als rechten Seiten zu lösen, wobei die c_i Vektoren in $\mathbb{k}^n$ sind. Diese Vektoren kann man mit Hilfe des Gauss-Algorithmus finden:

$$\left(\begin{array}{cccc|cccc}
a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\
a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1
\end{array}\right) \rightarrow \left(\begin{array}{cccc|cccc}
1 & 0 & \dots & 0 & c_{11} & c_{12} & \dots & c_{1n} \\
0 & 1 & \dots & 0 & c_{21} & c_{22} & \dots & c_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 1 & c_{n1} & c_{n2} & \dots & c_{nn}
\end{array}\right)$$

Die Vektoren c_k sind die Spaltenvektoren der Matrix C mit den Einträgen c_{ij}.

Mit den Vektoren c_k können jetzt beliebige inhomogene Gleichungssysteme $Ax = b$ gelöst werden. Da $b = b_1e_1 + b_2e_2 + \cdots + b_ne_n$, kann man die Lösung x als $x = b_1c_1 + b_2c_2 + \cdots + b_nc_n$ konstruieren. Tatsächlich gilt

$$\begin{aligned} Ax &= A(b_1c_1 + b_2c_2 + \cdots + b_nc_n) \\ &= b_1Ac_1 + b_2Cc_2 + \cdots + b_nAc_n \\ &= b_1e_1 + b_2e_2 + \cdots + b_ne_n = b. \end{aligned}$$

Die Linearkombination $x = b_1c_1 + \cdots + b_nc_n$ kann in Vektorform als $x = Cb$ geschrieben werden.

Die Konstruktion von C bedeutet auch, dass $AC = E$, daher heisst C auch die zu A *inverse Matrix*. Sie wird auch $C = A^{-1}$ geschrieben.

Die Definition der inversen Matrix stellt sicher, dass $AA^{-1} = I$ gilt, daraus folgt aber noch nicht, dass auch $A^{-1}A = I$ ist. Diese Eigenschaft kann man jedoch wie folgt erhalten. Sei C die inverse Matrix von A, also $AC = I$. Sei weiter D die inverse Matrix von C, also $CD = I$. Dann ist zunächst $A = AE = A(CD) = (AC)D = ID = D$ und weiter $CA = CD = I$. Mit der Bezeichnung $C = A^{-1}$ erhalten wir also auch $A^{-1}A = I$.

Die Eigenschaften der Matrizenmultiplikation stellen sicher, dass die Menge der invertierbaren Matrizen eine Struktur bilden, die man Gruppe nennt, die in Abschnitt 2.3.1 genauer untersucht wird. In diesem Zusammenhang wird dann auf Seite 47 die Eigenschaft $A^{-1}A = I$ ganz allgemein gezeigt.

Determinante

XXX TODO

2.1.4 Lineare Abbildungen

Der besondere Nutzen der Matrizen ist, dass sie auch lineare Abbildungen zwischen Vektorräumen beschreiben können. In diesem Abschnitt werden lineare Abbildungen abstrakt definiert und die Darstellung als Matrix mit Hilfe einer Basis eingeführt.

Definition

Eine lineare Abbildung zwischen Vektorräumen muss so gestaltet sein, dass die Operationen des Vektorraums erhalten bleiben. Dies wird von der folgenden Definition erreicht.

Definition 2.11. *Eine Abbildung $f\colon V \to U$ zwischen Vektorräumen V und U heisst linear, wenn*

$$\begin{aligned} f(v + w) &= f(v) + f(w) && \forall v, w \in V \\ f(\lambda v) &= \lambda f(v) && \forall v \in V, \lambda \in \mathbb{k} \end{aligned}$$

gilt.

Lineare Abbildungen sind in der Mathematik sehr verbreitet.

Beispiel. Sie $V = C^1([a, b])$ die Menge der stetig differenzierbaren Funktionen auf dem Intervall $[a, b]$ und $U = C([a, b])$ die Menge der stetigen Funktion aif $[a, b]$. Die Ableitung $\frac{d}{dx}$ macht aus einer Funktion $f(x)$ die Ableitung $f'(x)$. Die Rechenregeln für die Ableitung stellen sicher, dass

$$\frac{d}{dx}\colon C^1([a, b]) \to C([a, b]) : f \mapsto f'$$

eine lineare Abbildung ist. ○

Beispiel. Sei V die Menge der Riemann-integrierbaren Funktionen auf dem Intervall $[a, b]$ und $U = \mathbb{R}$. Das bestimmte Integral

$$\int_a^b : V \to U : f \mapsto \int_a^b f(x)\,dx$$

ist nach den bekannten Rechenregeln für bestimmte Integrale eine lineare Abbildung. ○

Matrix

Um mit linearen Abbildungen rechnen zu können, ist eine Darstellung mit Hilfe von Matrizen nötig. Sei also $\mathcal{B} = \{b_1, \ldots, b_n\}$ eine Basis von V und $C = \{c_1, \ldots, c_m\}$ eine Basis von U. Das Bild des Basisvektors b_i kann als Linearkombination der Vektoren $c_1, \ldots, c_m$ dargestellt werden. Wir verwenden die Bezeichnung

$$f(b_i) = a_{1i}c_1 + \cdots + a_{mi}c_m.$$

Die lineare Abbildung f bildet den Vektor x mit Koordinaten $x_1, \ldots, x_n$ ab auf

$$\begin{aligned} f(x) &= f(x_1b_1 + \ldots x_nb_n) \\ &= x_1f(b_1) + \ldots x_nf(b_n) \\ &= x_1(a_{11}c_1 + \cdots + a_{m1}c_m) + \cdots + x_n(a_{1n}c_1 + \cdots + a_{mn}c_m) \\ &= (a_{11}x_1 + \cdots + a_{1n}x_n)c_1 + \cdots + (a_{m1}x_1 + \cdots + a_{mn}x_n)c_m \end{aligned}$$

Die Koordinaten von $f(x)$ in der Basis C in U sind also gegeben durch das Matrizenprodukt Ax, wenn x der Spaltenvektor aus den Koordinaten in der Basis $\mathcal{B}$ in V ist.

Die Matrix einer linearen Abbildung macht Aussagen über eine lineare Abbilung der Rechnung zugänglich. Allerdings hängt die Matrix einer linearen Abbildung von der Wahl der Basis ab. Gleichzeitig ist dies eine Chance, durch Wahl einer geeigneten Basis kann man eine Matrix in eine Form bringen, die zur Lösung eines Problems optimal geeignet ist.

Basiswechsel

In einem Vektorraum V seien zwei Basen $\mathcal{B} = \{b_1, \ldots, b_n\}$ und $\mathcal{B}' = \{b'_1, \ldots, b'_n\}$ gegeben. Ein Vektor $v \in V$ kann in beiden beiden Basen dargestellt werden. Wir bezeichnen mit dem Spaltenvektor x die Koordinaten von v in der Basis $\mathcal{B}$ und mit dem Spaltenvektor x' die Koordinaten in der Basisi $\mathcal{B}'$. Um die Koordinaten umzurechnen, muss man die Gleichung

$$x_1b_1 + \cdots + x_nb_n = x'_1b'_1 + \cdots + x'_nb'_n \tag{2.8}$$

lösen.

Stellt man sich die Vektoren b_i und b'_j als m-dimensionale Spaltenvektoren vor mit $m \geq n$, dann bekommt (2.8) die Form eines Gleichungssystems

$$\begin{array}{ccccccc} b_{11}x_1 + & \ldots + & b_{1n}x_n = & b'_{11}x'_1 + & \ldots + & b'_{1n}x'_n \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{m1}x_1 + & \ldots + & b_{mn}x_n = & b'_{m1}x'_1 + & \ldots + & b'_{mn}x'_n \end{array}$$

Dieses Gleichungssystem kann man mit Hilfe eines Gauss-Tableaus lösen. Wir schreiben die zugehörigen Variablen

x_1	$\dots$	x_n	x'_1	$\dots$	x'_n
b_{11}	$\dots$	b_{1n}	b'_{11}	$\dots$	v'_{1n}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
b_{n1}	$\dots$	b_{nn}	b'_{n1}	$\dots$	v'_{nn}
$b_{n+1,1}$	$\dots$	$b_{n+1,n}$	$b'_{n+1,1}$	$\dots$	$v'_{n+1,n}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
b_{m1}	$\dots$	b_{mn}	b'_{m1}	$\dots$	v'_{mn}

$\rightarrow$

x_1	$\dots$	x_n	x'_1	$\dots$	x'_n
1	$\dots$	0	t_{11}	$\dots$	t_{1n}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	$\dots$	1	t_{n1}	$\dots$	t_{nn}
0	$\dots$	0	0	$\dots$	0
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	$\dots$	0	0	$\dots$	0

Das rechte untere Teiltableau enthält lauter Nullen genau dann, wenn jeder Vektor in V sich in beiden Mengen $\mathcal{B}$ und $\mathcal{B}'$ ausdrücken lässt. Dies folgt aber aus der Tatsache, dass $\mathcal{B}$ und $\mathcal{B}'$ beide Basen sind, also insbesondere den gleichen Raum aufspannen. Die $n \times n$-Matrix T mit Komponenten t_{ij} rechnet Koordinaten in der Basis $\mathcal{B}'$ um in Koordinaten in der Basis $\mathcal{B}$.

Umkehrabbbildung

Sei f eine umkehrbare lineare Abbildung $U \to V$ und $g\colon V \to U$. die zugehörige Umkehrabbildung. Für zwei Vektoren u und w in U gibt es daher Vektoren $a = g(u)$ und $b = g(w)$ in V derart, dass $f(a) = u$ und $f(b) = w$. Weil f linear ist, folgt daraus $f(a + b) = u + w$ und $f(\lambda a) = \lambda a$ für jedes $\lambda \in \mathbb{k}$. Damit kann man jetzt

$$\begin{aligned} g(u + w) &= g(f(a) + f(b)) = g(f(a + b)) = a + b = g(u) + g(w) \\ g(\lambda u) &= g(\lambda f(a)) = g(f(\lambda a)) = \lambda a = \lambda g(u) \end{aligned}$$

berechnen, was zeigt, dass auch g eine lineare Abbildung ist. Hat f in geeignet gewählten Basen die Matrix F, dann hat die Umkehrabbildung $g = f^{-1}$ die Matrix $G = F^{-1}$. Da auch $f(g(y)) = y$ gilt für jeden Vektor $y \in V$ folgt, dass $FF^{-1} = E$ und $F^{-1}F = E$.

Kern und Bild

Für die Eindeutigkeit der Lösung eines linearen Gleichungssytems ist entscheidend, ob das zugehörige homogene Gleichungssystem $Ax = 0$ eine nichttriviale Lösung hat. Seine Lösungmenge spielt also eine besondere Rolle, was rechtfertigt, ihr einen Namen zu geben.

Definition 2.12. *Ist f eine lineare Abbildung $U \to V$, dann heisst die Menge*

$$\ker f = \{x \in U \mid f(x) = 0\}$$

der Kern *oder* Nullraum *der linearen Abbildung f. Ist $A \in M_{m\times n}(\mathbb{k})$ Matrix, dann gehört dazu eine lineare Abbildung $f\colon \mathbb{k}^n \to \mathbb{k}^m$. Der Kern oder Nullraum der Matrix A ist die Menge*

$$\ker A = \{x \in \mathbb{k}^m \mid Ax = 0\}.$$

Der Kern ist ein Unterraum, denn für zwei Vektoren $u, w \in \ker f$

$$\begin{aligned} f(u + v) = f(u) + f(v) = 0 + 0 = 0 \quad &\Rightarrow \quad u + v \in \ker f \\ f(\lambda u) = \lambda f(u) = \lambda \cdot 0 = 0 \quad &\Rightarrow \quad \lambda u \in \ker f \end{aligned}$$

gilt.

Ob ein Gleichungssystem $Ax = b$ überhaupt eine Lösung hat, hängt davon, ob der Vektor b als Bild der durch A beschriebenen linearen Abbildung $\mathbb{k}^n \to \mathbb{k}^m$ enthalten ist. Wir definieren daher das Bild einer linearen Abbildung oder Matrix.

Definition 2.13. *Ist $f\colon V \to U$ eine lineare Abbildung dann ist das Bild von f der Unterraum*

$$\operatorname{im} f = \{f(v) \mid v \in V\} \subset U$$

von U. Das Bild einer $m \times n$-Matrix A ist die Menge

$$\operatorname{im} A = \{Av \mid v \in \mathbb{k}^n\} \subset \mathbb{k}^m.$$

Zwei Vektoren $a, b \in \operatorname{im} f$ haben Urbilder $u, w \in V$ mit $f(u) = a$ und $f(w) = b$. Für Summe und Multiplikation mit Skalaren folgt

$$\begin{aligned} a + b = f(u) + f(v) = f(u + v) \quad &\Rightarrow a + b \in \operatorname{im} f \\ \lambda a = \lambda f(u) = f(\lambda u) \quad &\Rightarrow \lambda a \in \operatorname{im} f, \end{aligned}$$

also ist auch das Bild $\operatorname{im} f$ ein Unterraum von U. Das Bild der Matrix A ist der Unterraum

$$\{x_1 f(b_1) + \dots x_n f(b_n) | x_i \in \mathbb{k}\} = \langle f(b_1), \dots, f(b_n)\rangle = \langle a_1, \dots, a_n\rangle$$

von $\mathbb{k}^m$, aufgespannt von den Spaltenvektoren a_i von A.

Rang und Defekt

Die Dimensionen von Bild und Kern sind wichtige Kennzahlen einer Matrix.

Definition 2.14. *Sei A eine Matrix $A \in M_{m\times n}(\mathbb{k})$. Der* Rang *der Matrix A ist die Dimension des Bildraumes von A:* $\operatorname{rank} A = \dim \operatorname{im} A$. *Der* Defekt *der Matrix A ist die Dimension des Kernes von A:* $\operatorname{def} A = \dim \ker A$.

Da der Kern mit Hilfe des Gauss-Algorithmus bestimmt werden kann, können Rang und Defekt aus dem Schlusstableau eines homogenen Gleichungssystems mit A als Koeffizientenmatrix abgelesen werden.

Satz 2.15. *Ist $A \in M_{m\times n}(\mathbb{k})$ eine $m \times n$-Matrix, dann gilt*

$$\operatorname{rank} A = n - \operatorname{def} A.$$

Quotient

TODO: $\operatorname{im} A \simeq \mathbb{k}^m / \ker A$

2.2 Skalarprodukt

In der bisher dargestellten Form ist die lineare Algebra nicht in der Lage, unsere vom Abstandsbegriff dominierte Geometrie adäquat darzustellen. Als zusätzliches Hilfsmittel wird eine Methode benötigt, Längen und Winkel auszudrücken. Das Skalarprodukt passt in den algebraischen Rahmen der linearen Algebra, bringt aber auch einen Abstandsbegriff hervor, der genau der geometrischen Intuition entspricht.

2.2.1 Bilinearformen und Skalarprodukte

Damit man mit einem Skalarprodukt rechnen kann wie mit jedem anderen Produkt, müssen man auf beiden Seiten des Zeichesn ausmultiplizieren können:

$$(\lambda x_1 + \mu x_2) \cdot y = \lambda x_1 \cdot y + \mu x_2 \cdot y$$
$$x \cdot (\lambda y_1 + \mu y_2) = \lambda x \cdot y_1 + \mu x \cdot y_2.$$

Man kann dies interpretieren als Linearität der Abbildungen $x \mapsto x \cdot y$ und $y \mapsto x \cdot y$. Dies wird Bilinearität genannt und wie folgt definiert.

Definition 2.16. *Seien U, V, W $\mathbb{k}$-Vektorräume. Eine Abbildung $f\colon U \times V \to W$ heisst* bilinear*, wenn die partiellen Abbildungen $U \to W : x \mapsto f(x, y_0)$ und $V \to W : y \mapsto f(x_0, y)$ linear sind für alle $x_0 \in U$ und $y_0 \in V$, d. h.*

$$f(\lambda x_1 + \mu x_2, y) = \lambda f(x_1, y) + \mu f(x_2, y)$$
$$f(x, \lambda y_1 + \mu y_2) = \lambda f(x, y_1) + \mu f(x, y_2)$$

Eine bilineare Funktion mit Werten in $\mathbb{k}$ heisst auch Bilinearform*.*

Symmetrische bilineare Funktionen

Das Skalarprodukt hängt nicht von der Reihenfolge der Faktoren ab. In Frage dafür kommen daher nur Bilnearformen $f\colon V \times V \to \mathbb{k}$, die zusätzlich $f(x, y) = f(y, x)$ erfüllen. Solche Bilinearformen heissen symmetrisch. Für eine symmetrische Bilinearform gilt die binomische Formel

$$\begin{aligned} f(x + y, x + y) &= f(x, x + y) + f(y, x + y) = f(x, x) + f(x, y) + f(y, x) + f(y, y) \\ &= f(x, x) + 2f(x, y) + f(y, y) \end{aligned}$$

wegen $f(x, y) = f(y, x)$.

Positiv definite Bilinearformen und Skalarprodukt

Bilinearität alleine genügt nicht, um einen Vektorraum mit einem nützlichen Abstandsbegriff auszustatten. Dazu müssen die berechneten Abstände vergleichbar sein, es muss also eine Ordnungsrelation definiert sein, wie wir sie nur in $\mathbb{R}$ kennen. Wir sind daher gezwungen uns auf $\mathbb{R}$- oder $\mathbb{Q}$-Vektorräume zu beschränken.

Man lernt in der Vektorgeometrie, dass sich mit einer Bilinearform $f\colon V \times V \to \mathbb{R}$ die Länge eines Vektors x definieren lässt, indem man $\|x\|^2 = f(x, x)$ setzt. Ausserdem muss $f(x, x) \geq 0$ sein für alle x, was die Bilinearität allein nicht garantieren kann. Verschiedene Punkte in einem Vektorraum sollen in dem aus der Bilinearform abgeleiteten Abstandsbegriff immer unterscheidbar sein. Dazu muss jeder von 0 verschiedene Vektor positive Länge haben.

Definition 2.17. *Eine Bilinearform $f\colon V \times V \to \mathbb{R}$ heisst* positiv definit*, wenn*

$$f(x, x) > 0 \qquad \forall x \in V \setminus \{0\}.$$

Das zugehörige Skalarprodukt *wird $f(x, y) = \langle x, y \rangle$ geschrieben. Die l^2-*Norm *$\|x\|_2$ eines Vektors ist definiert durch $\|x\|_2^2 = \langle x, x \rangle$.*

Dreiecksungleichung

Damit man sinnvoll über Abstände sprechen kann, muss die Norm $\| \cdot \|_2$ der geometrischen Intuition folgen, die durch die Dreiecksungleichung ausgedrückt wird. In diesem Abschnitt soll gezeigt werden, dass die l^2-Norm diese immer erfüllt. Dazu sei V ein $\mathbb{R}$-Vektorraum mit Skalarprodukt $\langle\ ,\ \rangle$.

Satz 2.18 (Cauchy-Schwarz-Ungleichung)**.** *Für $x, y \in V$ gilt*

$$|\langle x, y\rangle| \leq \|x\|_2 \cdot \|y\|_2$$

mit Gleichheit genau dann, wenn x und y linear abhängig sind.

Beweis. Wir die Norm von $z = x - ty$:

$$\|x - ty\|_2^2 = \|x\|_2^2 - 2t\langle x, y\rangle + t^2\|y\|_2^2 \geq 0.$$

Sie nimmt den kleinsten Wert genau dann an, wenn es ein t gibt derart, dass $x = ty$. Die rechte Seite ist ein quadratischer Ausdruck in t, er hat sein Minimum bei

$$t = -\frac{-2\langle x, y\rangle}{2\|y\|_2^2} \quad \Rightarrow \quad \left\|x - \frac{\langle x, y\rangle}{\|y\|_2^2} y\right\|_2^2 = \|x\|_2^2 - 2\frac{(\langle x, y\rangle)^2}{\|y\|_2^2} + \frac{(\langle x, y\rangle)^2}{\|y\|_2^4}\|y\|_2^2$$
$$= \|x\|_2^2 - \frac{(\langle x, y\rangle)^2}{\|y\|_2^2} = \frac{\|x\|_2^2 \cdot \|y\|_2^2 - (\langle x, y\rangle)^2}{\|y\|_2^2} \geq 0$$

Es folgt

$$\Rightarrow \quad \|x\|_2^2 \cdot \|y\|_2^2 - (\langle x, y\rangle)^2 \geq 0$$
$$\Rightarrow \quad \|x\|_2 \cdot \|y\|_2 \geq |\langle x, y\rangle|$$

mit Gleichheit genau dann, wenn es ein t gibt mit $x = ty$. □

Satz 2.19 (Dreiecksungleichung)**.** *Für $x, y \in V$ ist*

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$$

mit Gleichheit genau dann, wenn $x = ty$ ist für ein $t \geq 0$.

Beweis.

$$\begin{aligned}\|x + y\|_2^2 &= \langle x + y, x + y\rangle = \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle \\ &= \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2 = \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2 \leq \|x\|_2^2 + 2\|x\|_2 \cdot \|y\|_2 + \|y\|_2^2 \\ &= (\|x\|_2 + \|y\|_2)^2 \\ \|x\|_2 + \|y\|_2 &\leq \|x\|_2 + \|y\|_2,\end{aligned}$$

Gleichheit tritt genau dann ein, wenn $\langle x, y\rangle = \|x\|_2 \cdot \|y\|_2$. Dies tritt genau dann ein, wenn die beiden Vektoren linear abhängig sind. □

Polarformel

Auf den ersten Blick scheint die Norm $\|x\|_2$ weniger Information zu beinhalten, als die symmetrische Bilinearform, aus der sie hervorgegangen ist. Dem ist aber nicht so, denn die Bilinearform lässt sich aus der Norm zurückgewinnen. Dies ist der Inhalt der sogenannte Polarformel.

Satz 2.20 (Polarformel)**.** *Ist $\|\cdot\|_2$ eine Norm, die aus einer symmetrischen Bilinearform $\langle\ ,\ \rangle$ hervorgegangen ist, dann kann die Bilinearform mit Hilfe der Formel*

$$\langle x, y\rangle = \frac{1}{2}(\|x+y\|_2^2 - \|x\|_2^2 - \|y\|_2^2) \tag{2.9}$$

für $x, y \in V$ wiedergewonnen werden.

Beweis. Die binomischen Formel

$$\|x+y\|_2^2 = \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2$$

kann nach $\langle x, y\rangle$ aufgelöst werden, was

$$\langle x, y\rangle = \frac{1}{2}(\|x+y\|_2^2 - \|x\|_2^2 - \|y\|_2^2)$$

ergibt. Damit ist die Polarformel (2.9) bewiesen. □

Komplexe Vektorräume und Sesquilinearformen

Eine Bilinearform auf einem komplexen Vektorraum führt nicht auf eine Grösse, die sich als Norm eignet. Selbst wenn $\langle x, x\rangle > 0$ ist,

$$\langle ix, iy\rangle = i^2\langle x, y\rangle = -\langle x, y\rangle < 0.$$

Dies kann verhindert werden, wenn verlangt wird, dass der Faktor i im ersten Faktor der Bilinearform als $-i$ aus der Bilinearform herausgenommen werden muss.

Definition 2.21. *Seien U, V, W komplexe Vektorräume. Eine Abbildung $f\colon U \times V \to W$ heisst* sesquilinear[1] *wenn gilt*

$$f(\lambda x_1 + \mu x_2, y) = \overline{\lambda} f(x_1, y) + \overline{\mu} f(x_2, y)$$
$$f(x, \lambda y_1 + \mu y_2) = \lambda f(x, y_1) + \mu f(x, y_2)$$

Für die Norm $\|x\|_2^2 = \langle x, x\rangle$ bedeutet dies jetzt

$$\|\lambda x\|_2^2 = \langle \lambda x, \lambda x\rangle = \overline{\lambda}\lambda\langle x, x\rangle = |\lambda|^2\|x\|_2^2 \qquad \Rightarrow \qquad \|\lambda x\|_2 = |\lambda|\,\|x\|_2.$$

[1] Das lateinische Wort *sesqui* bedeutet eineinhalb, eine Sesquilinearform ist also eine Form, die in einem Faktor (dem zweiten) linear ist, und im anderen nur halb linear.

2.2.2 Orthognormalbasis

Gram-Matrix

Sei V ein Vektorraum mit einem Skalarprodukt und $\{b_1, \ldots, b_n\}$ eine Basis von V. Wie kann man das Skalarprodukt aus den Koordinaten ξ_i und η_i der Vektoren

$$x = \sum_{i=1}^{n} \xi_i b_i, \quad \text{und} \quad y = \sum_{i=1}^{n} \eta_i b_i$$

berechnen? Setzt man x und y in das Skalarprodukt ein, erhält man

$$\langle x, y\rangle = \left\langle \sum_{i=1}^{n} \xi_i b_i, \sum_{j=1}^{n} \eta_j b_j \right\rangle = \sum_{i,j=1}^{n} \xi_i \eta_j \langle b_i, b_j\rangle.$$

Die Komponente $g_{ij} = \langle b_i, b_j\rangle$ bilden die sogenannte Gram-Matrix G. Mit ihr kann das Skalarprodukt auch in Vektorform geschrieben werden als $\langle x, y\rangle = \xi^t G \eta$.

Orthonormalbasis

Eine Basis $\{a_1, \ldots, a_n\}$ aus orthogonalen Einheitsvektoren, also mit $\langle a_i, a_j\rangle = \delta_{ij}$ heisst *Orthonormalbasis*. In einer Orthonormalbasis ist die Bestimmung der Koordinaten eines beliebigen Vektors besonders einfach, ist nämlich

$$v = \sum_{i=1}^{n} \langle v, a_i\rangle a_i. \tag{2.10}$$

Die Gram-Matrix einer Orthonormalbasis ist die Einheitsmatrix.

Gram-Schmidt-Orthonormalisierung

Mit Hilfe des Gram-Schmidtschen Orthonormalisierungsprozesses kann aus einer beliebige Basis $\{a_1, a_2, \ldots, a_n\} \subset V$ eines Vektorraums mit einem SKalarprodukt eine orthonormierte Basis $\{b_1, b_2, \ldots, b_n\}$ gefunden werden derart, dass für alle k $\langle b_1, \ldots, b_k\rangle = \langle a_1, \ldots, a_k\rangle$. Der Zusammenhang zwischen den Basisvektoren b_i und a_i ist gegeben durch

$$\begin{aligned}
b_1 &= \frac{a_1}{\|a_1\|_2} \\
b_2 &= \frac{a_2 - b_1\langle b_1, a_2\rangle}{\|a_2 - b_1\langle b_1, a_2\rangle\|_2} \\
b_3 &= \frac{a_3 - b_1\langle b_1, a_3\rangle - b_2\langle b_2, a_3\rangle}{\|a_3 - b_1\langle b_1, a_3\rangle - b_2\langle b_2, a_3\rangle\|_2} \\
&\vdots \\
b_n &= \frac{a_n - b_1\langle b_1, a_n\rangle - b_2\langle b_2, a_n\rangle - \cdots - b_{n-1}\langle b_{n-1}, a_n\rangle}{\|a_n - b_1\langle b_1, a_n\rangle - b_2\langle b_2, a_n\rangle - \cdots - b_{n-1}\langle b_{n-1}, a_n\rangle\|_2}.
\end{aligned}$$

Die Gram-Matrix der Matrix $\{b_1, \ldots, b_n\}$ ist die Einheitsmatrix.

Orthogonalisierung

Der Normalisierungsschritt im Gram-Schmidt-Orthonormalisierungsprozess ist nur möglich, wenn Quadratwurzeln unbeschränkt gezogen werden können. Das ist in $\mathbb{R}$ möglich, nicht jedoch in $\mathbb{Q}$. Es ist aber mit einer kleinen Anpassung auch über $\mathbb{Q}$ immer noch möglich, aus einer Basis $\{a_1, \dots, a_n\}$ eine orthogonale Basis zu konstruieren. Man verwendet dazu die Formeln

$$\begin{aligned} b_1 &= a_1 \\ b_2 &= a_2 - b_1\langle b_1, a_2\rangle \\ b_3 &= a_3 - b_1\langle b_1, a_3\rangle - b_2\langle b_2, a_3\rangle \\ &\vdots \\ b_n &= a_n - b_1\langle b_1, a_n\rangle - b_2\langle b_2, a_n\rangle - \cdots - b_{n-1}\langle b_{n-1}, a_n\rangle. \end{aligned}$$

Die Basisvektoren b_i sind orthogonal, aber $\|b_i\|_2$ kann auch von 1 abweichen. Damit ist es zwar nicht mehr so einfach wie in (2.10), einen Vektor in der Basis zu zerlegen. Ein Vektor v hat nämlich in der Basis $\{b_1, \dots, b_n\}$ die Zerlegung

$$v = \sum_{i=1}^{n} \frac{\langle b_i, v\rangle}{\|b_i\|_2^2} b_i, \tag{2.11}$$

Die Koordinaten bezüglich dieser Basis sind also $\langle b_i, v\rangle / \|b_i\|_2^2$.

Die Gram-Matrix einer Orthogonalen Basis ist immer noch diagonal, auf der Diagonalen stehen die Normen der Basisvektoren. Die Nenner in der Zerlegung (2.11) sind die Einträge der inverse Matrix der Gram-Matrix.

Orthonormalbasen in komplexen Vektorräumen

Die Gram-Matrix einer Basis $\{b_1, \dots, b_n\}$ in einem komplexen Vektorraum hat die Eigenschaft

$$g_{ij} = \langle b_i, b_j\rangle = \overline{\langle b_j, b_i\rangle}, = \overline{g}_{ji} \quad 1 \le i, j \le n.$$

Sie ist nicht mehr symmetrisch, aber selbstadjungiert, gemäss der folgenden Definition.

Definition 2.22. *Sei A eine komplexe Matrix mit Einträgen a_{ij}, dann ist $\overline{A}$ die Matrix mit komplex konjugierten Elementen $\overline{a}_{ij}$. Die* adjungierte *Matrix ist $A^* = \overline{A}^t$. Eine Matrix heisst selbstadjungiert, wenn $A^* = A$.*

2.2.3 Symmetrische und selbstadjungierte Abbilungen

In Definition 2.22 wurde der Begriff der selbstadjungierten Matrix basierend eingeführt. Als Eigenschaft einer Matrix ist diese Definition notwendigerweise abhängig von der Wahl der Basis. Es ist nicht unbedingt klar, dass derart definierte Eigenschaften als von der Basis unabhängige Eigenschaften betrachtet werden können. Ziel dieses Abschnitts ist, Eigenschaften wie Symmetrie oder Selbstadjungiertheit auf basisunabhängige Eigenschaften von linearen Abbildungen in einem Vektorraum V mit Skalarprodukt $\langle\ ,\ \rangle$ zu verstehen.

Symmetrische Abbildungen

Sei $f\colon V \to V$ eine lineare Abbildung. In einer Basis $\{b_1, \ldots, b_n\} \subset V$ wird f durch eine Matrix A beschrieben. Ist die Basis orthonormiert, dann kann man die Matrixelemente mit $a_{ij} = \langle b_i, Ab_j\rangle$ berechnen. Die Matrix ist symmetrisch, wenn

$$\langle b_i, Ab_j\rangle = a_{ij} = a_{ji} = \langle b_j, Ab_i\rangle = \langle Ab_i, b_j\rangle$$

ist. Daraus leitet sich jetzt die basisunabhängige Definition einer symmetrischen Abbildung ab.

Definition 2.23. *Eine lineare Abbildung $f\colon V \to V$ heisst* symmetrisch*, wenn $\langle x, Ay\rangle = \langle Ax, y\rangle$ gilt für beliebige Vektoren $x, y \in V$.*

Für $V = \mathbb{R}^n$ und das Skalarprodukt $\langle x, y\rangle = x^t y$ erfüllt eine symmetrische Abbildung mit der Matrix A die Gleichung

$$\left.\begin{aligned}\langle x, Ay\rangle &= x^t Ay\\ \langle Ax, y\rangle &= (Ax)^t y = x^t A^t y\end{aligned}\right\} \quad\Rightarrow\quad x^t A^t y = x^t Ay \quad \forall x, y \in \mathbb{R}^n,$$

was gleichbedeutend ist mit $A^t = A$. Der Begriff der symmetrischen Abbildung ist also eine natürliche Verallgemeinerung des Begriffs der symmetrischen Matrix.

Selbstadjungierte Abbildungen

In einem komplexen Vektorraum ist das Skalarprodukt nicht mehr bilinear und symmetrisch, sondern sesquilinear und konjugiert symmetrisch.

Definition 2.24. *Eine lineare Abbildung $f\colon V \to V$ heisst* selbstadjungiert*, wenn $\langle x, fy\rangle = \langle fx, y\rangle$ für alle $x, y \in \mathbb{C}$.*

Im komplexen Vektorraum $\mathbb{C}^n$ ist das Standardskalarprodukt definiert durch $\langle x, y\rangle = \overline{x}^t y$.

Die Adjungierte

Die Werte der Skalarprodukte $\langle x, y\rangle$ für alle $x \in V$ legen den Vektor y fest. Gäbe es nämlich einen zweiten Vektor y' mit den gleichen Skalarprodukten, also $\langle x, y\rangle = \langle x, y'\rangle$ für alle $x \in V$, dann gilt wegen der Linearität $\langle x, y - y'\rangle = 0$. Wählt man $x = y - y'$, dann folgt $0 = \langle y - y', y - y'\rangle = \|y - y'\|_2$, also muss $y = y'$ sein.

Definition 2.25. *Sei $f\colon V \to V$ eine lineare Abbildung. Die lineare Abbildung $f^*\colon V \to V$ definiert durch*

$$\langle f^* x, y\rangle = \langle x, fy\rangle, \qquad x, y \in V$$

heisst die Adjungierte *von f.*

Eine selbstadjungierte Abbildung ist also eine lineare Abbildung, die mit ihrer Adjungierte übereinstimmt, als $f^* = f$. In einer orthonormierten Basis $\{b_1, \ldots, b_n\}$ hat die Abbildung f die Matrixelemente $a_{ij} = \langle b_i, fb_j\rangle$. Die adjungierte Abbildung hat dann die Matrixelemente

$$\langle b_i, f^* b_j\rangle = \overline{\langle f^* b_j, b_i\rangle} = \overline{\langle b_j, fb_i\rangle} = \overline{a_{ji}},$$

was mit der Definition von A^* übereinstimmt.

2.2.4 Orthogonale und unitäre Matrizen

Von besonderer geometrischer Bedeutung sind lineare Abbildung, die die Norm nicht verändern. Aus der Polarformel (2.9) folgt dann, dass auch das Skalarprodukt erhalten ist, aus dem Winkel berechnet werden können. Abbildungen, die die Norm erhalten, sind daher auch winkeltreu.

Definition 2.26. *Eine lineare Abbildung $f\colon V \to V$ in einem reellen Vektorraum mit heisst* orthogonal*, wenn $\langle fx, fy\rangle = \langle x, y\rangle$ für alle $x, y \in V$ gilt.*

Die adjungierte einer orthogonalen Abbildung erfüllt $\langle x, y\rangle = \langle fx, fy\rangle = \langle f^* fx, y\rangle$ für alle $x, y \in V$, also muss $f^* f$ die identische Abbildung sein, deren Matrix die Einheitsmatrix ist. Die Matrix O einer orthogonalen Abbildung erfüllt daher $O^t O = I$.

Für einen komplexen Vektorraum erwarten wir grundsätzlich dasselbe. Lineare Abbildungen, die die Norm erhalten, erhalten das komplexe Skalarprodukt. Auch in diesem Fall ist $f^* f$ die identische Abbildung, die zugehörigen Matrixen U erfüllen daher $U^* U = I$.

Definition 2.27. *Eine lineare Abbildung $f\colon V \to V$ eines komplexen Vektorraumes V mit Skalarprodukt heisst unitär, wenn $\langle x, y\rangle = \langle fx, fy\rangle$ für alle Vektoren $x, y \in V$. Eine Matrix heisst unitär, wenn $U^* U = I$.*

Die Matrix einer unitären Abbildung in einer orthonormierten Basis ist unitär.

2.2.5 Orthogonale Unterräume

2.2.6 Andere Normen auf Vektorräumen

Das Skalarprodukt ist nicht die einzige Möglichkeit, eine Norm auf einem Vektorraum zu definieren. In diesem Abschnitt stellen wir einige weitere mögliche Normdefinitionen zusammen.

l^1-Norm

Definition 2.28. *Die l^1-Norm in $V = \mathbb{R}^n$ oder $V = \mathbb{C}^n$ ist definiert durch*

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|$$

für $v \in V$.

Auch die l^1-Norm erfüllt die Dreiecksungleichung

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1} |x_i| + \sum_{i=1} |y_i| = \|x\|_1 + \|y\|_1 .$$

Die l^1-Norm kommt nicht von einem Skalarprodukt her. Wenn es ein Skalarprodukt gäbe, welches auf diese Norm führt, dann müsste

$$\langle x, y\rangle = \frac{1}{2}(\|x + y\|_1^2 - \|x\|_1^2 - \|y\|_1^2)$$

sein. Für die beiden Standardbasisvektoren $x = e_1$ und $y = e_2$ bedeutet dies

$$\left.\begin{aligned} \|e_1\|_1 &= 2 \\ \|e_2\|_1 &= 2 \\ \|e_1 \pm + e_2\|_1 &= 2 \end{aligned}\right\} \quad \Rightarrow \quad \langle e_1, \pm e_2\rangle = \frac{1}{2}(2^2 - 1^2 - 1^2) = 1$$

Die Linearität des Skalarproduktes verlangt aber, dass $1 = \langle e_1, -e_2\rangle = -\langle e_1, e_2\rangle = -1$, ein Widerspruch.

l^∞-Norm

Definition 2.29. *Die l^∞-Norm in $V = \mathbb{R}^n$ und $V = \mathbb{C}^n$ ist definiert*

$$\|v\|_\infty = \max_i |v_i|.$$

Sie heisst auch die Supremumnorm.

Auch diese Norm erfüllt die Dreiecksungleichung

$$\|x + y\|_\infty = \max_i |x_i + y_i| \leq \max_i (|x_i| + |y_i|) \leq \max_i |x_i| + \max_i |y_i| = \|x\|_\infty + \|y\|_\infty.$$

Auch diese Norm kann nicht von einem Skalarprodukt herkommen, ein Gegenbeispiel können wir wieder mit den ersten beiden Standardbasisvektoren konstruieren. Es ist

$$\left.\begin{aligned} \|e_1\|_\infty &= 1 \\ \|e_2\|_\infty &= 1 \\ \|e_1 \pm e_2\|_\infty &= 1 \end{aligned}\right\} \quad \Rightarrow \quad \langle e_1, \pm e_2\rangle = \frac{1}{2}(\|e_1 \pm e_2\|_\infty^2 - \|e_1\|_\infty^2 - \|e_2\|_\infty^2) = \frac{1}{2}(1 - 1 - 1) = -\frac{1}{2}.$$

Es folgt wieder $-\frac{1}{2} = \langle e_1, -e_2\rangle = -\langle e_1, e_2\rangle = \frac{1}{2}$, ein Widerspruch.

Operatornorm

Der Vektorraum der linearen Abbildungen $f\colon U \to V$ kann mit einer Norm ausgestattet werden, wenn U und V jeweils eine Norm haben.

Definition 2.30. *Seien U und V Vektorräume über $\mathbb{R}$ oder $\mathbb{C}$ und $f\colon U \to V$ eine lineare Abbildung. Die* Operatorname *der linearen Abbildung ist*

$$\|f\| = \sup_{x \in U \wedge \|x\| \leq 1} \|fx\|.$$

Nach Definition gilt $\|fx\| \leq \|f\| \cdot \|x\|$ für alle $x \in U$. Die in den Vektorräumen U und V verwendeten Normen haben einen grossen Einfluss auf die Operatornorm, wie die beiden folgenden Beispiele zeigen.

Beispiel. Sei V ein komplexer Vektorraum mit einem Skalarprodukt und $y \in V$ ein Vektor. y definiert die lineare Abbildung

$$l_y\colon V \to \mathbb{C} : x \mapsto \langle y, x\rangle.$$

Zur Berechnung der Operatorname von l_y

$$|l_y(x)|^2 = |\langle y, x\rangle|^2 \leq \|y\|_2^2 \cdot \|x\|_2^2$$

mit Gleichheit genau dann, wenn x und y linear abhängig sind. Dies bedeutet, dass $\|l_y\| = \|y\|$, die Operatorname von l_y stimmt mit der Norm von y überein. ◯

Beispiel. Sei $V = \mathbb{C}^n$. Dann definiert $y \in V$ eine Linearform

$$l_y \colon V \to \mathbb{C} : x \mapsto y^t x.$$

Wir suchen die Operatornorm von l_y, wenn V mit der l^1-Norm ausgestattet wird. Sei k der Index der betragsmässig grössten Komponente von y_k, also $\|y\|_\infty = |y_k|$. Dann gilt

$$|l_y(x)| = \left|\sum_{i=1}^n y_i x_i\right| \le \sum_{i=1}^n |y_i| \cdot |x_i| \le |y_k| \sum_{i=1}^n |x_i| = \|y\|_\infty \cdot \|x\|_1.$$

Gleichheit wird erreicht, wenn die Komponente k die einzige von 0 verschiedene Komponente des Vektors x ist. Somit ist $\|l_y\| = \|y\|_\infty$. ○

Normen auf Funktionenräumen

Alle auf $\mathbb{R}^n$ und $\mathbb{C}^n$ definierten Normen lassen sich auf den Raum der stetigen Funktionen $[a, b] \to \mathbb{R}$ oder $[a, b] \to \mathbb{C}$ verallgemeinern.

Die Supremumnorm auf dem Vektorraum der stetigen Funktionen ist

$$\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|$$

für $f \in C([a, b], \mathbb{R})$ oder $f \in C([a, b], \mathbb{C})$.

Für die anderen beiden Normen wird zusätzlich das bestimmte Integral von Funktionen auf $[a, b]$ benötigt. Die L^2-Norm wird erzeugt von dem Skalarprodukt

$$\langle f, g\rangle = \frac{1}{b-a}\int_a^b \overline{f}(x) g(x)\, dx \qquad \Rightarrow \qquad \|f\|_2^2 = \frac{1}{b-a}\int_a^b |f(x)|^2\, dx.$$

Die L^2-Norm ist dagegen

$$\|f\|_1 = \int_a^b |f(x)|\, dx.$$

2.3 Algebraische Strukturen

Im Laufe der Definition der Vektorräume $\mathbb{k}^n$ und der Operationen für die Matrizen in $M_{m\times n}(\mathbb{k})$ haben wir eine ganze Reihe von algebraischen Strukturen kennengelernt. Nicht immer sind alle Operationen verfügbar, in einem Vektorraum gibt es normalerweise kein Produkt. Und bei der Konstruktion des Zahlensystems wurde gezeigt, dass additive oder multiplikative Inverse nicht selbstverständlich sind. Sinnvolle Mathematik lässt sich aber erst betreiben, wenn zusammen mit den vorhandenen Operationen auch einige Regeln erfüllt sind. Die schränkt die Menge der sinnvollen Gruppierungen von Eigenschaften ein. In diesem Abschnitten sollen diesen sinnvollen Gruppierungen von Eigenschaften Namen gegeben werden.

2.3.1 Gruppen

Die kleinste sinnvolle Struktur ist die einer Gruppe. Eine solche besteht aus einer Menge G mit einer Verknüpfung, die additiv

$$G \times G \to G : (g, h) = gh$$

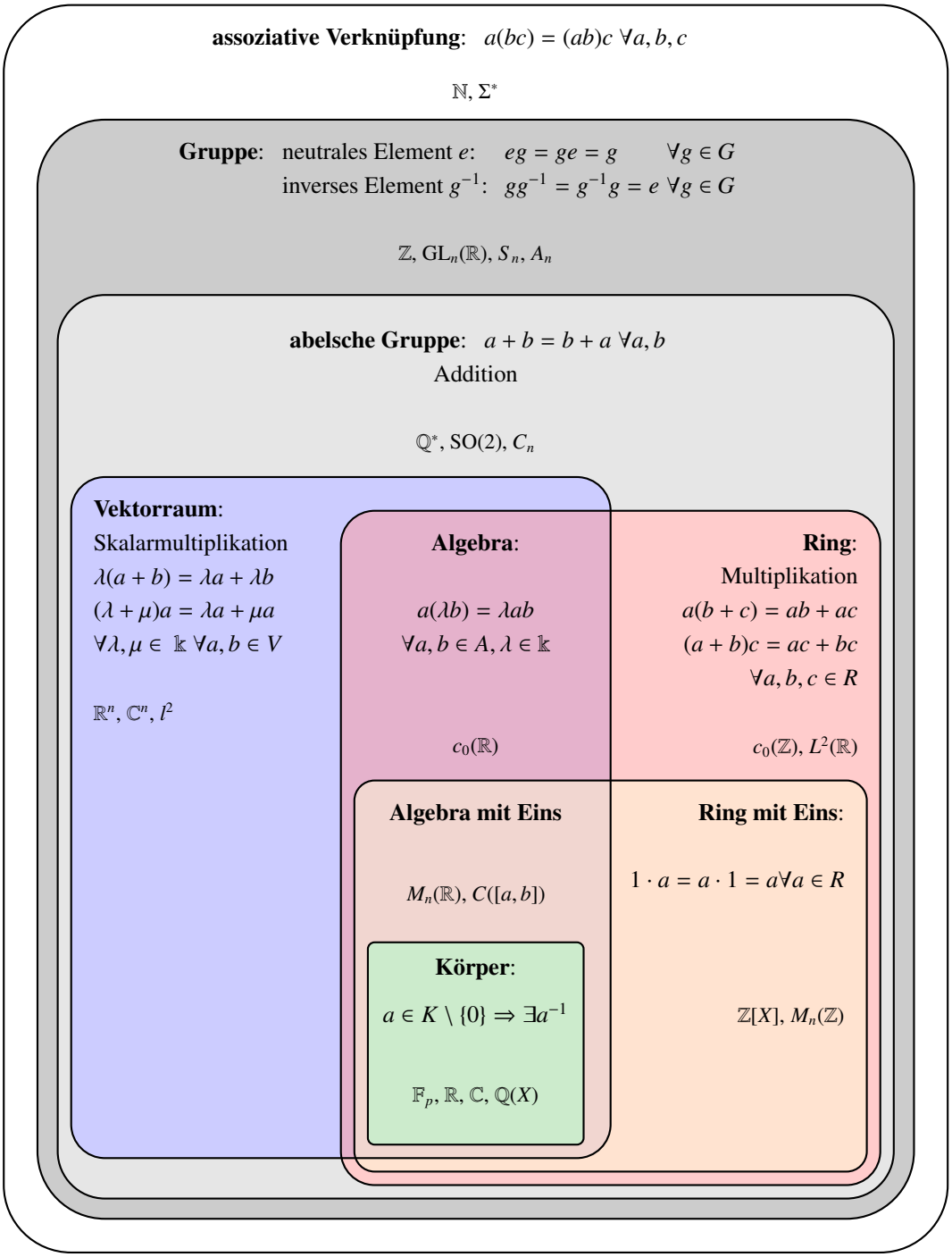


Abbildung 2.2: Übersicht über die verschiedenen algebraischen Strukturen, die in Abschnitt 2.3 zusammengestellt werden.

oder multiplikativ

$$G \times G \to G : (g, h) = g + h$$

geschrieben werden kann. Ein Element $0 \in G$ heisst *neutrales Element* bezüglich der additiv geschriebenen Verknüpfung falls $0 + x = x$ für alle $x \in G$. Ein Element $e \in G$ heisst neutrales Element bezüglich der multiplikativ geschriebneen Verknüpfung, wenn $ex = x$ für alle $x \in G$. In den folgenden Definitionen werden wir immer die multiplikative Schreibweise verwenden, für Fälle additiv geschriebener siehe auch die Beispiele weiter unten.

Definition 2.31. *Ein* Gruppe *ist eine Menge G mit einer Verknüfung mit folgenden Eigenschaften:*

1. *Die Verknüpfung ist assoziativ:* $(ab)c = a(bc)$ *für alle* $a, b, c \in G$.
2. *Es gibt ein neutrales Element* $e \in G$
3. *Für jedes Element* $g \in G$ *gibt es ein Element* $h \in G$ *mit* $hg = e$.

Das Element h heisst auch das Inverse Element zu g.

Falls nicht jedes Element invertierbar ist, aber wenigstens ein neutrales Element vorhanden ist, spricht man von einem *Monoid*. Hat man nur eine Verknüpfung, spricht man oft von einer *Halbruppe*.

Definition 2.32. *Eine Gruppe G heisst abelsch, wenn* $ab = ba$ *für alle* $a, b \in G$.

Additiv geschrieben Gruppen werden immer als abelsch angenommen, multiplikativ geschrieben Gruppen können abelsch oder nichtabelsch sein.

Beispiele von Gruppen

Beispiel. Die Menge $\mathbb{Z}$ mit der Addition ist eine additive Gruppe mit dem neutralen Element 0. Das additive Inverse eines Elementes a ist $-a$. ○

Beispiel. Die von Null verschiedenen Elemente $\mathbb{k}^*$ eines Zahlekörpers bilden bezüglich der Multiplikation eine Gruppe mit neutralem Element 1. Das multiplikative Inverse eines Elementes $a \in \mathbb{k}$ mit $a \neq 0$ ist $a^{-1} = \frac{1}{a}$. ○

Beispiel. Die Vektoren $\mathbb{k}^n$ bilden bezüglich der Addition eine Gruppe mit dem Nullvektor als neutralem Element. Betrachtet man $\mathbb{k}^n$ als Gruppe, verliert man die Multiplikation mit Skalaren aus den Augen. $\mathbb{k}^n$ als Gruppe zu bezeichnen ist also nicht falsch, man verliert dadurch aber ○

Beispiel. Die Menge aller quadratischen $n \times n$-Matrizen $M_n(\mathbb{k})$ ist eine Gruppe bezüglich der Addition mit der Nullmatrix als neutralem Element. Bezügich der Matrizenmultiplikation ist $M_n(\mathbb{k})$ aber keine Gruppe, da sich die singulären Matrizen nicht inverieren lassen. Die Menge der invertierbaren Matrizen

$$\mathrm{GL}_n(\mathbb{k}) = \{A \in M_n(\mathbb{k}) \mid A \text{ invertierbar}\}$$

ist bezüglich der Multiplikation eine Gruppe. Die Gruppe $\mathrm{GL}_n(\mathbb{k})$ ist eine echte Teilmenge von $M_n(\mathbb{k})$, die Addition und Multiplikation führen im Allgemeinen aus der Gruppe heraus, es gibt also keine Mögichkeit, in der Gruppe $\mathrm{GL}_n(\mathbb{k})$ diese Operationen zu verwenden. ○

Einige einfache Rechenregeln in Gruppen

Die Struktur einer Gruppe hat bereits eine Reihe von Einschränkungen zur Folge. Zum Beispiel sprach die Definition des neutralen Elements e nur von Produkten der Form $ex = x$, nicht von Produkten xe. Und die Definition des inversen Elements h von g hat nur verlangt, dass $gh = e$, es wurde nichts gesagt über das Produkt hg.

Satz 2.33. *Ist G eine Gruppe mit neutralem Element e, dann gilt*

1. *$xe = x$ für alle $x \in G$*
2. *Es gibt nur ein neutrales Element. Wenn also $f \in G$ mit $fx = x$ für alle $x \in G$, ist dann folgt $f = e$.*
3. *Wenn $hg = e$ gilt, dann auch $gh = e$ und h ist durch g eindeutig bestimmt.*

Beweis. Wir beweisen als Erstes den ersten Teil der Eigenschaft 3. Sei h die Inverse von g, also $hg = e$. Sei weiter i die Inverse von h, also $ih = e$. Damit folgt jetzt

$$g = eg = (ih)g = i(hg) = ie.$$

Wende man dies auf das Produkt gh an, folgt

$$gh = (ie)h = i(eh) = ih = e$$

Es ist also nicht nur $hg = e$ sondern immer auch $gh = e$.

Für eine Inverse h von g folgt

$$ge = g(hg) = (gh)g = eg = g,$$

dies ist die Eigenschaft 1.

Sind f und e neutrale Elemente, dann folgt

$$f = fe = e$$

aus der Eigenschaft 1.

Schliesslich sei x ein beliebiges Inverses von g, dann ist $xg = e$, dann folgt $x = xe = x(gh) = (xg)h = eh = h$, es gibt also nur ein Inverses von g. □

Diesem Problem sind wir zum Beispiel auch in Abschnitt 2.1.3 begegnet, wo wir nur gezeigt haben, dass $AA^{-1} = E$ ist. Da aber die invertierbaren Matrizen eine Gruppe bilden, folgt jetzt aus dem Satz automatisch, dass auch $A^{-1}A = E$.

Homomorphismen

Lineare Abbildung zwischen Vektorräumen zeichnen sich dadurch aus, dass sie die algebraische Struktur des Vektorraumes respektieren. Für eine Abbildung zwischen Gruppen heisst dies, dass die Verknüpfung, das neutrale Element und die Inverse respektiert werden müssen.

Definition 2.34. *Ein Abbildung $\varphi\colon G \to H$ zwischen Gruppen heisst ein* Homomorphismus*, wenn $\varphi(g_1g_2) = \varphi(g_1)\varphi(g_2)$ für alle $g_1, g_2 \in G$ gilt.*

Der Begriff des Kerns einer linearen Abbildung lässt sich ebenfalls auf die Gruppensituation erweitern. Auch hier ist der Kern der Teil der Gruppe, er unter dem Homomorphismus "unsichtbar" wird.

Definition 2.35. *Ist $\varphi\colon G \to H$ ein Homomorphisus, dann ist*

$$\ker\varphi = \{g \in G \mid \varphi(g) = e\}$$

eine Untergruppe.

Normalteiler

Der Kern eines Homomorphismus ist nicht nur eine Untergruppe, er erfüllt noch eine zusätzliche Bedingung. Für jedes $g \in G$ und $h \in \ker\varphi$ gilt

$$\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g^{-1}) = \varphi(g)\varphi(g^{-1}) = \varphi(gg^{-1}) = \varphi(e) = e \qquad \Rightarrow \qquad ghg^{-1} \in \ker\varphi.$$

Der Kern wird also von der Abbildung $h \mapsto ghg^{-1}$, der *Konjugation* in sich abgebildet.

Definition 2.36. *Eine Untergruppe $H \subset G$ heisst ein* Normalteiler*, geschrieben $H \triangleleft G$ wenn $gHg^{-1} \subset H$ für jedes $g \in G$.*

Die Konjugation selbst ist ebenfalls keine Unbekannte, sie ist uns bei der Basistransformationsformel schon begegnet. Die Tatsache, dass $\ker\varphi$ unter Konjugation erhalten bleibt, kann man also interpretieren als eine Eigenschaft, die unter Basistransformation erhalten bleibt.

Faktorgruppen

Ein Unterraum $U \subset V$ eines Vektorraumes gibt Anlass zum Quotientenraum, der dadurch entsteht, dass man die Vektoren in U zu 0 kollabieren lässt. Eine ähnliche Konstruktion könnte man für eine Untergruppe $H \subset G$ versuchen. Man bildet also wieder die Mengen von Gruppenelementen, die sich um ein Elemente in H unterscheiden. Man kann diese Mengen in der Form gH mit $g \in G$ schreiben.

Man möchte jetzt aber auch die Verknüpfung für solche Mengen definieren, natürlich so, dass $g_1H \cdot g_2H = (g_1g_2)H$ ist. Da die Verknüpfung nicht abelsch sein muss, entsteht hier ein Problem. Für $g_1 = e$ folgt, dass $Hg_2H = g_2H$ sein muss. Das geht nur, wenn $Hg_2 = g_2H$ oder $g_2Hg_2^{-1} = H$ ist, wenn also H ein Normalteiler ist.

Definition 2.37. *Für eine Gruppe G mit Normalteiler $H \triangleleft G$ ist die Menge*

$$G/H = \{gH \mid g \in G\}$$

eine Gruppe mit der Verknüpfung $g_1H \cdot g_2H = (g_1g_2)H$. G/H heisst Faktorgruppe *oder* Quotientengruppe.

Für abelsche Gruppen ist die Normalteilerbedingung keine zusätzliche Einschränkung, jeder Untergruppe ist auch ein Normalteiler.

Beispiel. Die ganzen Zahlen $\mathbb{Z}$ bilden eine abelsche Gruppe und die Menge der Vielfachen von n $n\mathbb{Z} \subset \mathbb{Z}$ ist eine Untergruppe. Da $\mathbb{Z}$ abelsch ist, ist $n\mathbb{Z}$ ein Normalteiler und die Faktorgruppe $\mathbb{Z}/n\mathbb{Z}$ ist wohldefiniert. Nur die Elemente

$$0 + n\mathbb{Z}, 1 + n\mathbb{Z}, 2 + n\mathbb{Z}, \dots (n-1) + n\mathbb{Z}$$

sind in der Faktorgruppe verschieden. Die Gruppe $\mathbb{Z}/n\mathbb{Z}$ besteht also aus den Resten bei Teilung durch n. Diese Gruppe wird in Kapitel 4 genauer untersucht. ○

Das Beispiel suggeriert, dass man sich die Elemente von G/H als Reste vorstellen kann.

Darstellungen

Abstrakt definierte Gruppen können schwierig zu verstehen sein. Oft hilft es, wenn man eine geometrische Darstellung der Gruppenoperation finden kann. Die Gruppenelemente werden dann zu umkehrbaren linearen Operationen auf einem geeigneten Vektorraum.

Definition 2.38. *Eine Darstellung einer Gruppe G ist ein Homomorphismus $G \to \mathrm{GL}_n(\mathbb{R})$.*

Beispiel. Die Gruppen $\mathrm{GL}_n(\mathbb{Z})$, $\mathrm{SL}_n(\mathbb{Z})$ oder $\mathrm{SO}(n)$ sind alle Teilmengen von $\mathrm{GL}_n(\mathbb{R})$. Die Einbettungsabbildung $G \hookrightarrow \mathrm{GL}_n(\mathbb{R})$ ist damit automatisch eine Darstellung, sie heisst auch die *reguläre Darstellung* der Gruppe G. ○

In Kapitel 6 wird gezeigt, dass Permutationen einer endlichen eine Gruppe bilden und wie sie durch Matrizen dargestellt werden können.

2.3.2 Ringe und Moduln

Die ganzen Zahlen haben ausser der Addition mit neutralem Element 0 auch noch eine Multiplikation mit dem neutralen Element 1. Die Multiplikation ist aber nicht immer invertierbar und zwar nicht nur für 0. Eine ähnliche Situation haben wir bei $M_n(\Bbbk)$ angetroffen. $M_n(\Bbbk)$ ist eine zunächst eine Gruppe bezüglich der Addition, hat aber auch noch eine Multiplikation, die nicht immer umkehrbar ist. Diese Art von Struktur nennt man einen Ring.

Definition eines Rings

Definition 2.39. *Eine Menge R mit einer additiven Operation $+$ mit neutralem Element 0 und einer multiplikativ geschriebenen Operation $\cdot$ heisst ein* Ring*, wenn folgendes gilt.*

1. *R ist eine Gruppe bezüglich der Addition.*
2. *$R \setminus \{0\}$ ist eine Halbgruppe.*
3. *Es gelten die* Distributivgesetze

$$a(b+c) = ab + ac \qquad \text{und} \qquad (a+b)c = ac + bc$$

für beliebige Elemente $a, b, c \in R$.

Die Distributivgesetze stellen sicher, dass man in R beliebig ausmultiplizieren kann. Man kann also so rechnen kann, wie man sich das gewohnt ist. Es stellt auch sicher, dass die Multiplikation mit 0 immer 0 ergibt, denn es ist

$$r0 = r(a-a) = ra - ra = 0.$$

Man beachte, dass weder verlangt wurde, dass die Multiplikation ein neutrales Element hat oder kommutativ ist. Der Ring $\mathbb{Z}$ erfüllt beide Bedingungen. Die Beispiele weiter unten werden zeigen, dass es auch Ringe gibt, in denen die Multiplikation nicht kommutativ ist, die Multiplikation kein neutrales Element hat oder beides.

Definition 2.40. *Ein Ring R heisst ein Ring mit Eins, wenn die Multiplikation ein neutrales Element hat.*

Definition 2.41. *Ein Ring R heisst kommutativ, wenn die Multiplikation kommutativ ist.*

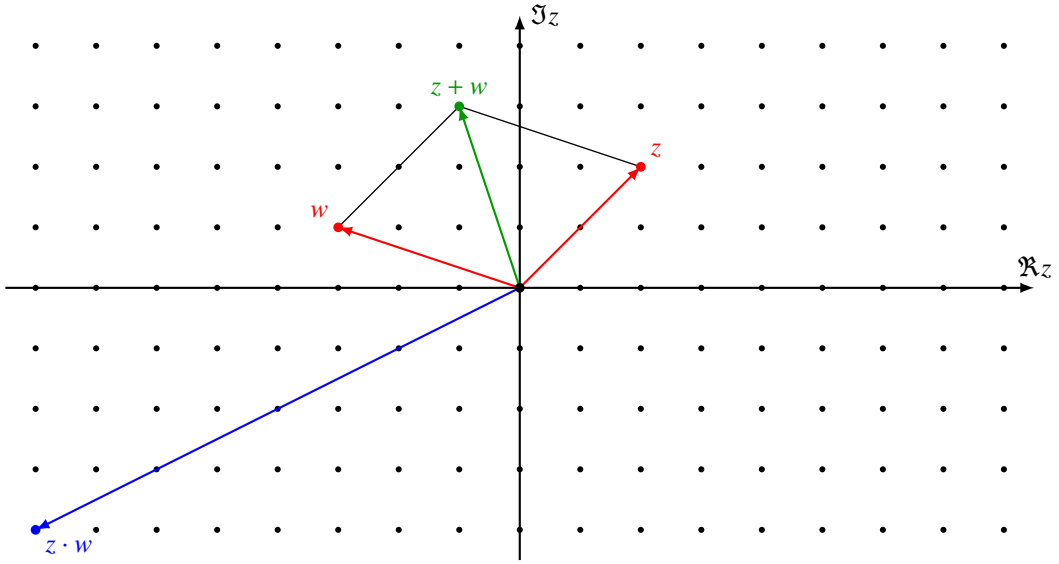


Abbildung 2.3: Der Ring der ganzen Gausschen Zahlen besteht aus den ganzahligen Gitterpunkten in der Gausschen Zahlenebene

Beispiele von Ringen

Beispiel. Alle Zahlenkörper aus Kapitel 1 sind kommutative Ringe mit Eins. ○

Beispiel. Die Menge $c(\mathbb{Z})$ der Folgen $(a_n)_{n\in\mathbb{N}}$ mit Folgengliedern in $\mathbb{Z}$ wird eine Ring, wenn man die Addition und Multiplikation elementweise definiert, also

Addition:	$a+b$ ist die Folge mit Folgengliedern	$(a+b)_n = a_n b_n$	für alle $n \in \mathbb{N}$
Multiplikation:	$a \cdot b$ ist die Folge mit Folgengliedern	$(a \cdot b)_n = a_n b_n$	für alle $n \in \mathbb{N}$

für $a, b \in c(\mathbb{Z})$. Die Algebra ist kommutativ und hat die konstante Folge $u_n = 1\ \forall n$ als Eins.

Wir betrachten jetzt ein Unterring $c_0(\mathbb{Z}) \subset c(\mathbb{Z})$ bestehend aus den Folgen, die nur für endlich viele Folgenglieder von 0 verschieden sind. Für eine Folge $a \in c_0(\mathbb{Z})$ gibt es eine Zahl N derart, dass $a_n = 0$ für $n \geq N$. Die konstante Folge $u_n = 1$, die in $c(\mathbb{Z})$ erfüllt diese Bedingung nicht, die Eins des Ringes $c(\mathbb{Z})$ ist also nicht in $c_0(\mathbb{Z})$. $c_0(\mathbb{Z})$ ist immer noch ein Ring, aber er hat kein Eins. ○

Beispiel. Die Menge

$$\mathbb{Z} + i\mathbb{Z} = \{a + bi \mid a, b \in \mathbb{Z}\} = \mathbb{Z}[i] \subset \mathbb{C}$$

ist eine Teilmenge von $\mathbb{C}$ und erbt natürlich die arithmetischen Operationen. Die Summe zweier solcher Zahlen $a + bi \in \mathbb{Z}[i]$ und $c + di \in \mathbb{Z}[i]$ ist $(a + bi) + (c + di) = (a + c) + (b + d)i \in \mathbb{Z}[i]$, weil $a + c \in \mathbb{Z}$ und $b + d \in \mathbb{Z}$ ganze Zahlen sind. Ebenso ist das Produkt dieser Zahlen $(a + bi)(c + di) = (ac - bd) + (ad + bc)i \in \mathbb{Z}[i]$ weil Realteil $ac - bd \in \mathbb{Z}$ und der Imaginärteil $ad + bc \in \mathbb{Z}$ ganze Zahlen sind. Die Menge $\mathbb{Z}[i]$ ist also ein kommutative Ring mit Eins, er heisst der Ring der ganzen *Gaussschen Zahlen*. ○

Beispiel. Die Menge der Matrizen $M_n(\mathbb{Z})$ ist ein Ring mit Eins. Für $n > 1$ ist er nicht kommutativ. Der Ring $M_2(\mathbb{Z})$ enthält den Teilring

$$G = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \;\middle|\; a, b \in \mathbb{Z} \right\} = \mathbb{Z} + \mathbb{Z}J \subset M_2(\mathbb{Z}).$$

Da die Matrix J die Relation $J^2 = -E$ erfüllt, ist der Ring G nichts anderes als der Ring der ganzen Gaussschen Zahlen. Der Ring $\mathbb{Z}[i]$ ist also ein Unterring des Matrizenrings $M_2(\mathbb{Z})$. ○

Einheiten

In einem Ring mit Eins sind normalerweise nicht alle von 0 verschiedenen Elemente intertierbar. Die Menge der von 0 verschiedenen Elemente in R wir mit R^* bezeichnet. Die Menge der invertierbaren Elemente verdient einen besonderen Namen.

Definition 2.42. *Ist R ein Ring mit Eins, dann heissen die Elemente von*

$$U(R) = \{r \in R \mid r \text{ in } R \text{ invertierbar}\}.$$

die Einheiten *von R.*

Satz 2.43. *$U(R)$ ist eine Gruppe, die sogenannte* Einheitengruppe.

Beispiel. Die Menge $M_2(\mathbb{Z})$ ist ein Ring mit Eins, die Einheitengruppe besteht aus den invertierbaren 2×2-Matrizen. Aus der Formel für

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

zeigt, dass $U(M_2(\mathbb{Z})) = \mathrm{SL}_2(\mathbb{Z})$. ○

Beispiel. Die Einheitengruppe von $M_n(\mathbb{k})$ ist die allgemeine lineare Gruppe $U(M_n(\mathbb{k})) = \mathrm{GL}_n(\mathbb{k})$. ○

Nullteiler

Ein möglicher Grund, warum ein Element $r \in R$ nicht invertierbar ist, kann sein, dass es ein Element $s \in R$ gibt mit $rs = 0$. Wäre nämlich t ein inverses Element, dann wäre $0 = t0 = t(rs) = (tr)s = s$.

Definition 2.44. *Ein Element $r \in R^*$ heisst ein* Nullteiler *in R, wenn es ein $s \in R^*$ gibt mit $rs = 0$ Ein Ring ohne Nullteiler heisst* nullteilerfrei.

In $\mathbb{R}$ ist man sich gewohnt zu argumentieren, dass wenn ein Produkt $ab = 0$ ist, dann muss einer der Faktoren $a = 0$ oder $b = 0$ sein. Dieses Argument funktioniert nur, weil $\mathbb{R}$ ein nullteilerfreier Ring ist. In $M_2(\mathbb{R})$ ist dies nicht mehr möglich. Die beiden Matrizen

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Rightarrow \qquad AB = 0$$

sind Nullteiler in $M_2(\mathbb{Z})$.

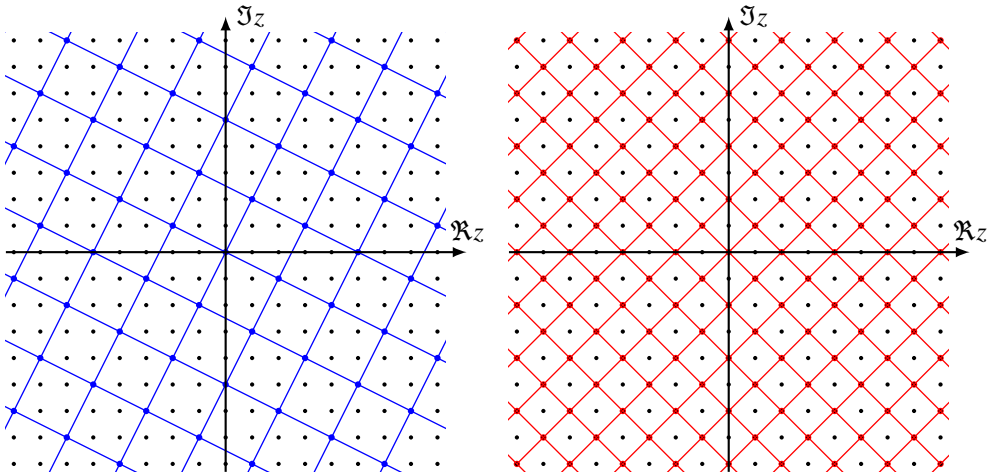


Abbildung 2.4: Ideale im Ring der ganzen Gaussschen Zahlen $\mathbb{Z}[i]$. Für jedes Element $r \in \mathbb{Z}[i]$ ist die Menge $r\mathbb{Z}[i]$ ein ein Ideal in $\mathbb{Z}[i]$. Links das Ideal $(1+2i)\mathbb{Z}[i]$ (blau), rechts das Ideal $(1+i)\mathbb{Z}[i]$ (rot).

Homomorphismus

Eine Abbildung zwischen Ringen muss die algebraische Struktur respektieren, wenn sich damit Eigenschaften vom einen Ring auf den anderen transportieren lassen sollen.

Definition 2.45. *Eine Abbildung $\varphi : R \to S$ zwischen Ringen heisst ein* Homomorphismus *oder* Ringhomomorphismus*, wenn φ ein Gruppenhomomorphismus der additiven Gruppen der Ringe ist und ausserdem gilt*

$$\varphi(r_1 r_2) = \varphi(r_1)\varphi(r_2).$$

Der Kern ist die Menge

$$\ker\varphi = \{r \in R \mid \varphi(r) = 0\}$$

Wieder hat der Kern zusätzliche Eigenschaften. Er ist natürlich bezüglich der additiven Struktur des Ringes ein Normalteiler, aber weil die additive Gruppe ja abelsch ist, ist das keine wirkliche Einschränkung. Für ein beliebiges Element $r \in R$ und $k \in \ker\varphi$ gilt

$$\begin{aligned}\varphi(kr) &= \varphi(k)\varphi(r) = 0 \cdot \varphi(r) = 0\\ \varphi(rk) &= \varphi(r)\varphi(k) = \varphi(r) \cdot 0 = 0.\end{aligned}$$

Für den Kern gilt also, dass $\ker\varphi \cdot R \subset \ker\varphi$ und $R \cdot \ker\varphi \subset \ker\varphi$.

Ideale

Bei der Betrachtung der additiven Gruppe des Ringes $\mathbb{Z}$ der ganzen Zahlen wurde bereits die Untergruppe $n\mathbb{Z}$ diskutiert und die Faktorgruppe $\mathbb{Z}/n\mathbb{Z}$ der Reste konstruiert. Reste können aber auch multipliziert werden, es muss also auch möglich sein, der Faktorgruppe eine multiplikative Struktur zu verpassen.

Sei jetzt also $I \subset R$ ein Unterring. Die Faktorgruppe R/I hat bereits die additive Struktur, es muss aber auch die Multiplikation definiert werden. Die Elemente $r_1 + I$ und $r_2 + I$ der Faktorgruppe R/I haben das Produkt

$$(r_1 + I)(r_2 + I) = r_1 r_2 + r_1 I + I r_2 + II.$$

Dies stimmt nur dann mit $r_1 r_2 + I$ überein, wenn $r_1 I \subset I$ und $r_2 I \subset I$ ist.

Definition 2.46. *Ein Unterring $I \subset R$ heisst ein* Ideal*, wenn für jedes $r \in R$ gilt $rI \subset I$ und $Ir \subset I$ gilt. Die Faktorgruppe R/I erhält eine natürliche Ringstruktur, R/I heisst der* Quotientenring*.*

Beispiel. Die Menge $n\mathbb{Z} \subset \mathbb{Z}$ besteht aus den durch n teilbaren Zahlen. Multipliziert man durch n teilbare Zahlen mit einer ganzen Zahl, bleiben sie durch n teilbar, $n\mathbb{Z}$ ist also ein Ideal in $\mathbb{Z}$. Der Quotientenring ist der Ring der Reste bei Teilung durch n, er wird in Kapitel 4 im Detail untersucht. ○

Ein Ideal $I \subset R$ drückt als die Idee "gemeinsamer Faktoren" auf algebraische Weise aus und der Quotientenring R/I beschreibt das, was übrig bleibt, wenn man diese Faktoren ignoriert.

Beispiel. In Abbildung 2.4 sind zwei Ideale im Ring der ganzen Gaussschen Zahlen dargestellt. Die blauen Punkte sind $I_1 = (1 + 2i)\mathbb{Z}$ und die roten Punkte sind $I_2 = (1 + i)\mathbb{Z}$. Die Faktorgruppen R/I_1 und R/I_2 fassen jeweils Punkte, die sich um ein Element von I_1 bzw. I_2 unterscheiden, zusammen.

Im Falle von I_2 gibt es nur zwei Arten von Punkten, nämlich die roten und die schwarzen, der Quotientenring hat daher nur zwei Elemente, $R/I_2 = \{0 + I_2, 1 + I_2\}$. Wegen $1 + 1 = 0$ in diesem Quotientenring, ist $R/I_2 = \mathbb{Z}/2\mathbb{Z}$.

Im Falle von I_1 gibt es fünf verschiedene Punkte, als Menge ist

$$R/I_1 = \{0 + I_1, 1 + I_1, 2 + I_1, 3 + I_1, 4 + I_1\}.$$

Die Rechenregeln sind also dieselben wie im Ring $\mathbb{Z}/5\mathbb{Z}$. In gewisser Weise verhält sich die Zahl $1 + 2i$ in den ganzen Gaussschen Zahlen bezüglich Teilbarkeit ähnlich wie die Zahl 5 in den ganzen Zahlen. ○

2.3.3 Algebren

Die Skalar-Multiplikation eines Vektorraums ist in einem Ring nicht vorhanden. Die Menge der Matrizen $M_n(\mathbb{k})$ ist sowohl ein Ring als auch ein Vektorraum. Man nennt eine $\mathbb{k}$-*Algebra* oder *Algebra über* $\mathbb{k}$ ein Ring A, der auch eine $\mathbb{k}$-Vektorraum ist. Die Multiplikation des Ringes muss dazu mit der Skalarmultiplikation verträglich sein. Dazu müssen Assoziativgesetze

$$\lambda(\mu a) = (\lambda\mu)a \qquad \text{und} \qquad \lambda(ab) = (\lambda a)b$$

für $a, b \in A$ und $\lambda, \mu \in \mathbb{k}$ und eine Regel der Form

$$a(\lambda b) = \lambda(ab) \tag{2.12}$$

gelten. Die Bedingung (2.12) ist eine Folge der Forderung, dass die Multiplikation eine lineare Abbildung sein soll. Dies bedeutet, dass

$$a(\lambda b + \mu c) = \lambda(ab) + \mu(ac), \tag{2.13}$$

woraus (2.12) für $\mu = 0$ folgt. Die Regel (2.13) beinhaltet aber auch das Distributivgesetz. $M_n(\mathbb{k})$ ist eine Algebra.

Die Algebra der Funktionen $\mathbb{k}^X$

Sie X eine Menge und $\mathbb{k}^X$ die Menge aller Funktionen $X \to \mathbb{k}$. Auf $\mathbb{k}^X$ kann man Addition, Multiplikation mit Skalaren und Multiplikation von Funktionen punktweise definieren. Für zwei Funktion $f, g \in \mathbb{k}^X$ und $\lambda \in \mathbb{k}$ definiert man

$$\begin{aligned} \text{Summe } f+g\text{:} &\quad (f+g)(x) = f(x) + g(x) \\ \text{Skalare } \lambda f\text{:} &\quad (\lambda f)(x) = \lambda f(x) \\ \text{Produkt } f \cdot g\text{:} &\quad (f \cdot g)(x) = f(x)g(x) \end{aligned}$$

Man kann leicht nachprüfen, dass die Menge der Funktionen $\mathbb{k}^X$ mit diesen Verknüfungen die Struktur einer $\mathbb{k}$-Algebra erhält.

Die Algebra der Funktionen $\mathbb{k}^X$ hat auch ein Einselement: die konstante Funktion

$$1 : [a,b] \to \mathbb{k} : x \mapsto 1$$

mit Wert 1 erfüllt

$$(1 \cdot f)(x) = 1(x)f(x) = f(x) \qquad \Rightarrow \qquad 1 \cdot f = f,$$

die Eigenschaft einer Eins in der Algebra.

Die Algebra der stetigen Funktionen $C([a,b])$

Die Menge der stetigen Funktionen $C([a,b])$ ist natürlich eine Teilmenge aller Funktionen: $C([a,b]) \subset \mathbb{R}^{[a,b]}$ und erbt damit auch die Algebraoperationen. Man muss nur noch sicherstellen, dass die Summe von stetigen Funktionen, das Produkt einer stetigen Funktion mit einem Skalar und das Produkt von stetigen Funktionen wieder eine stetige Funktion ist. Eine Funktion ist genau dann stetig, wenn an jeder Stelle der Grenzwert mit dem Funktionswert übereinstimmt. Genau dies garantieren die bekannten Rechenregeln für stetige Funktionen. Für zwei stetige Funktionen $f, g \in C([a,b])$ und einen Skalar $\lambda \in \mathbb{R}$ gilt

$$\begin{aligned} \text{Summe:} &\quad \lim_{x\to x_0}(f+g)(x) = \lim_{x\to x_0}(f(x)+g(x)) = \lim_{x\to x_0} f(x) + \lim_{x\to x_0} g(x) = f(x_0)+g(x_0) = (f+g)(x_0) \\ \text{Skalare:} &\quad \lim_{x\to x_0}(\lambda f)(x) = \lim_{x\to x_0}(\lambda f(x)) = \lambda \lim_{x\to x_0} f(x) = \lambda f(x_0) = (\lambda f)(x_0) \\ \text{Produkt:} &\quad \lim_{x\to x_0}(f\cdot g)(x) = \lim_{x\to x_0} f(x)\cdot g(x) = \lim_{x\to x_0} f(x) \cdot \lim_{x\to x_0} g(x) = f(x_0)g(x_0) = (f\cdot g)(x_0). \end{aligned}$$

für jeden Punkt $x_0 \in [a,b]$. Damit ist $C([a,b])$ eine $\mathbb{R}$-Algebra. Die Algebra hat auch eine Eins, da die konstante Funktion $1(x) = 1$ stetig ist.

2.3.4 Körper

Die Multiplikation ist in einer Algebra nicht immer umkehrbar. Die Zahlenkörper von Kapitel 1 sind also sehr spezielle Algebren, man nennt sie Körper. In diesem Abschnitt sollen die wichtigsten Eigenschaften von Körpern zusammengetragen werden.

XXX TODO

2.4 Hadamard-Algebra

Das Matrizenprodukt ist nicht die einzige Möglichkeit, ein Produkt auf Vektoren oder Matrizen zu definieren. In diesem Abschnitt soll das Hadamard-Produkt beschrieben werden, welches zu einer kommutativen-Algebra-Struktur führt.

2.4.1 Hadamard-Produkt

Im Folgenden werden wir $\mathbb{k}^n = M_{n\times 1}(\mathbb{k})$ setzen und den Fall der Vektoren nicht mehr separat diskutieren. Die Addition und Multiplikation mit Skalaren ist in $M_{m\times n}(\mathbb{k})$ komponentenweise definiert. Wir können natürlich auch ein Produkt komponentenweise definieren, dies ist das Hadamard-Produkt.

Definition 2.47. *Das* Hadamard-Produkt *zweier Matrizen $A, B \in M_{m\times n}(\mathbb{k})$ ist definiert als die Matrix $A \odot B$ mit den Komponenten*

$$(A \odot B)_{ij} = (A)_{ij}(B)_{ij}.$$

Wir nennen $M_{m\times n}(\mathbb{k})$ mit der Multiplikation $\odot$ auch die Hadamard-Algebra $H_{m\times n}(\mathbb{k})$.

Dies ist jedoch nur interessant, wenn $M_{m\times n}(\mathbb{k})$ mit diesem Produkt eine interessante algebraische Struktur erhält. Dazu müssen die üblichen Verträglichkeitsgesetze zwischen den Vektorraumoperationen von $M_{m\times n}(\mathbb{k})$ und dem neuen Produkt gelten, wir erhalten dann eine Algebra. Da alle Operationen elementweise definiert sind, muss man auch alle Rechengesetze nur elementweise prüfen. Es gilt

$$\begin{aligned}
A \odot (B \odot C) = (A \odot B) \odot C &\quad\Leftrightarrow\quad a_{ij}(b_{ij}c_{ij}) = (a_{ij}b_{ij})c_{ij}\\
A \odot (B + C) = A \odot B + A \odot C &\quad\Leftrightarrow\quad a_{ij}(b_{ij} + c_{ij}) = a_{ij}b_{ij} + a_{ij}c_{ij}\\
(A + B) \odot C = A \odot C + B \odot C &\quad\Leftrightarrow\quad (a_{ij} + b_{ij})c_{ij} = a_{ij}c_{ij} + b_{ij}c_{ij}\\
(\lambda A) \odot B = \lambda(A \odot B) &\quad\Leftrightarrow\quad (\lambda a_{ij})b_{ij} = \lambda(a_{ij}b_{ij})\\
A \odot (\lambda B) = \lambda(A \odot B) &\quad\Leftrightarrow\quad a_{ij}(\lambda b_{ij}) = \lambda(a_{ij}b_{ij})
\end{aligned}$$

für alle i, j.

Das Hadamard-Produkt ist kommutativ, da die Multiplikation in $\mathbb{k}$ kommuativ ist. Das Hadamard-Produkt kann auch für Matrizen mit Einträgen in einem Ring definiert werden, in diesem Fall ist es möglich, dass die entsehende Algebra nicht kommutativ ist.

Die Hadamard-Algebra hat auch ein Eins-Elemente, nämlich die Matrix, die aus lauter Einsen besteht.

Definition 2.48. *Die sogenannte* Einsmatrix *U ist die Matrix*

$$U = \begin{pmatrix} 1 & 1 & \dots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \dots & 1 \end{pmatrix} \in M_{m\times n}(\mathbb{k})$$

mit lauter Einträgen $1 \in \mathbb{k}$.

Die Hadamard-Algebra ist ein Spezialfall der Algebra der Funktionen $\mathbb{k}^X$. Ordnet man dem Vektor $v \in \mathbb{k}^n$ mit den Komponenten v_i die Abbildung

$$v\colon [n] \to \mathbb{k} : i \mapsto v_i$$

zu, dann geht die Addition von Vektoren in die Addition von Funktionen über, die Multiplikation von Skalaren mit Vektoren geht in die Multiplikation von Funktionen mit Skalaren über und die Hadamard-Multiplikation geht über in das Produkt von Funktionen.

Auch die Hadamard-Algebra $H_{m\times n}(\mathbb{k})$ kann als Funktionenalgebra betrachtet werden. Einer Matrix $A \in H_{m\times n}(\mathbb{k})$ ordnet man die Funktion

$$a\colon [m]\times[n] : (i,j) \mapsto a_{ij}$$

zu. Dabei gehen die Algebraoperationen von $H_{m\times n}(\mathbb{k})$ über in die Algebraoperationen der Funktionenalgebra $\mathbb{k}^{[m]\times[n]}$. Aus der Einsmatrix der Hadamard-Algebra wird dabei zur konstanten Funktion 1 auf $[m]\times[n]$.

2.4.2 Hadamard-Produkt und Matrizenalgebra

Es ist nur in Ausnahmefällen, Hadamard-Produkt und Matrizen-Produkt gleichzeitig zu verwenden. Das liegt daran, dass die beiden Produkte sich überhaupt nicht vertragen.

Unverträglichkeit von Hadamard- und Matrizen-Produkt

Das Hadamard-Produkt und das gewöhnliche Matrizenprodukt sind in keiner Weise kompatibel. Die beiden Matrizen

$$A = \begin{pmatrix} 3 & 4 \\ 4 & 5 \end{pmatrix} \qquad \text{und} \qquad B = \begin{pmatrix} -5 & 4 \\ 4 & -3 \end{pmatrix}$$

sind inverse Matrizen bezüglich des Matrizenproduktes, also $AB = E$. Für das Hadamard-Produkt gilt dagegen

$$A \odot B = \begin{pmatrix} -15 & 16 \\ 16 & -15 \end{pmatrix}.$$

Die Inverse einer Matrix A Bezüglich des Hadamard-Produktes hat die Einträge a_{ij}^{-1}. Die Matrix E ist bezüglich des gewöhnlichen Matrizenproduktes invertierbar, aber sie ist bezüglich des Hadamard-Produktes nicht invertierbar.

Einbettung der Hadamard-Algebra ein eine Matrizenalgebra

Hadamard-Algebren können als Unteralgebren einer Matrizenalgebra betrachtet werden. Der Operator diag bildet Vektoren ab in Diagonalmatrizen nach der Regel

$$\operatorname{diag}\colon \mathbb{k}^n \to M_n(\mathbb{k}) : \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \mapsto \begin{pmatrix} v_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & v_n \end{pmatrix}$$

Das Produkt von Diagonalmatrizen ist besonders einfach. Für zwei Vektoren $a, b \in \mathbb{k}^n$

$$a \odot b = \begin{pmatrix} a_1b_1 \\ \vdots \\ a_nb_n \end{pmatrix} \mapsto \begin{pmatrix} a_1b_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_nb_n \end{pmatrix} = \begin{pmatrix} a_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_n \end{pmatrix} \begin{pmatrix} b_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & b_n \end{pmatrix}.$$

Das Hadamard-Produkt der Vektoren geht also über in das gewöhnliche Matrizenprodukt der Diagonalmatrizen.

Für die Hadamard-Matrix ist die Einbettung etwas komplizierter. Wir machen aus einer Matrix erst einen Vektor, den wir dann mit dem diag in eine Diagonalmatrix umwandeln:

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & \end{pmatrix} \mapsto \begin{pmatrix} a_{11} \\ \vdots \\ a_{1n} \\ a_{21} \\ \vdots \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix}$$

Bei dieser Abbildung geht die Hadamard-Multiplikation wieder in das gewöhnliche Matrizenprodukt über.

Beispiel: Faltung und Fourier-Theorie

2.4.3 Weitere Verknüpfungen

Transposition

Das Hadamard-Produkt verträgt sich mit der Transposition:

$$(A \odot B)^t = A^t \odot B^t.$$

Insbesondere ist das Hadamard-Produkt zweier symmetrischer Matrizen auch wieder symmetrisch.

Frobeniusnorm

Das Hadamard-Produkt in der Hadamard-Algebra $H_{m\times n}(\mathbb{R})$ nimmt keine Rücksicht auf die Dimensionen einer Matrix und ist nicht unterscheidbar von $\mathbb{R}^{m\times n}$ mit dem Hadamard-Produkt. Daher darf auch der Begriff einer mit den algebraischen Operationen verträglichen Norm nicht von von den Dimensionen abhängen. Dies führt auf die folgende Definition einer Norm.

Definition 2.49. *Die* Frobenius-Norm *einer Matrix $A \in H_{m\times n}\mathbb{R})$ mit den Einträgen $(a_{ij}) = A$ ist*

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Das Frobenius-Skalarprodukt *zweier Matrizen $A, B \in H_{m\times n}(\mathbb{R})$ ist*

$$\langle A, B\rangle_F = \sum_{i,j} a_{ij} b_{ij} = \operatorname{Spur} A^t B$$

und es gilt $\|A\|_F = \sqrt{\langle A, A\rangle}$.

Für komplexe Matrizen muss

Definition 2.50. *Die* komplexe Frobenius-Norm *einer Matrix* $A \in H_{m\times n}(\mathbb{C})$ *ist*

$$\|A\| = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\sum_{i,u} \overline{a}_{ij} a_{ij}}$$

das komplexe Frobenius-Skalarprodukt *zweier Matrizen* $A, B \in H_{m\times n}(\mathbb{C})$ *ist das Produkt*

$$\langle A, B\rangle_F = \sum_{i,j} \overline{a}_{ij} b_{ij} = \operatorname{Spur}(A^* B)$$

und es gilt $\|A\|_F = \sqrt{\langle A, A\rangle}$.

Skalarprodukt

Übungsaufgaben

2.1. Gegeben ist die Matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & a_{1n} \\ 1 & 0 & 0 & \dots & 0 & a_{2n} \\ 0 & 1 & 0 & \dots & 0 & a_{3n} \\ 0 & 0 & 1 & \dots & 0 & a_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & a_{nn} \end{pmatrix}$$

a) Berechnen Sie $\det A$

b) Finden Sie die inverse Matrix A^{-1}

c) Nehmen Sie an, dass $a_{in} \in \mathbb{Z}$. Formulieren Sie eine Bedingung an die Koeffizienten a_{in}, die garantiert, dass A^{-1} eine Matrix mit ganzzahligen Koeffizienten ist.

Lösung. a) Die Determinante ist am einfachsten mit Hilfe des Entwicklungssatzes durch Entwicklung nach der ersten Zeile zu bestimmen:

$$\det A = \begin{vmatrix} 0 & 0 & 0 & \dots & 0 & a_{1n} \\ 1 & 0 & 0 & \dots & 0 & a_{2n} \\ 0 & 1 & 0 & \dots & 0 & a_{3n} \\ 0 & 0 & 1 & \dots & 0 & a_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & a_{nn} \end{vmatrix} = (-1)^{n+1} a_{1n} \det E_n = -1^{n+1} a_{1n}.$$

b) Die inverse Matrix kann am einfachsten mit Hilfe des Gauss-Algorithmus gefunden werden. Dazu schreiben wir die Matrix A in die linke Hälfte eines Tableaus und die Einheitsmatrix in die rechte Hälfte und führen den Gauss-Algorithmus durch.

0	0	0	…	a_{1n}	1	0	0	…	0
1	0	0	…	a_{2n}	0	1	0	…	0
0	1	0	…	a_{3n}	0	0	1	…	0
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋱	⋮
0	0	0	…	a_{nn}	0	0	0	…	1

Die Arbeit wird wesentlich vereinfacht, wenn wir zunächst die erste Zeile ganz nach unten schieben:

$$\rightarrow \left(\begin{array}{ccccc|ccccc} 1 & 0 & \dots & 0 & a_{2n} & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & a_{3n} & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & a_{nn} & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & a_{1n} & 1 & 0 & 0 & \dots & 0 \end{array}\right)$$

Mit einer einzigen Gauss-Operationen kann man jetzt die inverse Matrix finden. Dazu muss man zunächst durch das Pivot-Elemente a_{1n} dividieren, und dann in der Zeile k das $a_{k+1,n}$-fache der letzten Zeile subtrahieren. Dies hat nur eine Auswirkung auf die erste Spalte in der rechten Hälfte:

$$\rightarrow \left(\begin{array}{ccccc|ccccc} 1 & 0 & \dots & 0 & 0 & -\frac{a_{2n}}{a_{1n}} & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & -\frac{a_{3n}}{a_{1n}} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & -\frac{a_{nn}}{a_{1n}} & 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & 1 & \frac{1}{a_{1n}} & 0 & 0 & \dots & 0 \end{array}\right)$$

Die inverse Matrix von A ist also

$$A^{-1} = \begin{pmatrix} -\frac{a_{2n}}{a_{1n}} & 1 & 0 & \dots & 0 \\ -\frac{a_{3n}}{a_{1n}} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{nn}}{a_{1n}} & 0 & 0 & \dots & 1 \\ \frac{1}{a_{1n}} & 0 & 0 & \dots & 0 \end{pmatrix} \tag{2.14}$$

c) Aus der Darstellung (2.14) der Inversen A^{-1} können wir ablesen, dass A^{-1} nur dann eine ganzzahlige Matrix sein kann, wenn a_{1n} invertierbar ist, also $a_{1n} = \pm 1$. ◯

2.2. Nach Aufgabe 2.1 hat die Matrix

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & a \\ 0 & 1 & b \end{pmatrix} \in M_2(\mathbb{Z}) \quad \text{die Inverse} \quad A^{-1} = \begin{pmatrix} -b & 1 & 0 \\ -a & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \in M_2(\mathbb{Z}).$$

Kann man A^{-1} als Linearkombination der Matrizen E, A und A^2 schreiben?

Lösung. Wir berechnen zunächst A^2:

$$A^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & a \\ 0 & 1 & b \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & a \\ 0 & 1 & b \end{pmatrix} = \begin{pmatrix} 0 & 1 & b \\ 0 & a & 1+ab \\ 1 & b & a+b^2 \end{pmatrix}$$

Gesucht sind jetzt die Koeffizienten λ_i einer Linearkombination

$$A^{-1} = \lambda_0 E + \lambda_1 A + \lambda_2 A^2.$$

Die drei Matrizen auf der rechten Seite haben in der ersten Spalte nur Nullen und Einsen, so dass wir an der ersten Spalten von A^+ unmittelbar ablesen können, welche Werte wir für λ_i verwenden müssen. Wir finden $\lambda_0 = -b$, $\lambda_1 = -a$ und $\lambda_2 = 1$. Wir setzen dies ein:

$$\begin{aligned} -bE - aA + A^2 &= -b\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} - a\begin{pmatrix}0&0&1\\1&0&a\\0&1&b\end{pmatrix} + \begin{pmatrix}0&1&b\\0&a&1+ab\\1&b&a+b^2\end{pmatrix} \\ &= \begin{pmatrix}-b&1&b-a\\-a&-b+a&-a^2+1+ab\\1&-a+b&-b-ab+a+b^2\end{pmatrix} = \begin{pmatrix}-b&1&b-a\\-a&a-b&1+a(b-a)\\1&b-a&(1-b)(a-b)\end{pmatrix} \end{aligned}$$

Diese Matrix kann nur dann mit A^{-1} übereinstimmen, wenn $a - b = 0$ ist, als $a = b$. ◯

Kapitel 3

Polynome

Ein *Polynom* ist ein Ausdruck der Form

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \cdots a_2 X^2 + a_1 X + a_0. \tag{3.1}$$

Ursprünglich stand das Symbol X als Platzhalter für eine Zahl. Die Polynomgleichung $Y = p(X)$ drückt dann einen Zusammenhang zwischen den Grössen X und Y aus. Zum Beispiel drückt

$$H = -\frac{1}{2} g T^2 + v_0 T + h_0 = p(T) \tag{3.2}$$

im Schwerefeld der Erde nahe der Oberfläche einen Zusammenhang zwischen der Zeit T und der Höhe H eines frei fallenden Körpers aus. Setzt man einen Wert für T in (3.2) ein, erhält man den zugehörigen Wert für H. Man stellt sich hier also vor, dass T eigentlich eine Zahl ist und dass (3.1) nur ein "unfertiger" Ausdruck oder ein "Programm" für eine Berechnung ist. In dieser eher arithmetischen Sichtweise ist es aber eigentlich egal, dass in (3.1) nur einfache Multiplikationen und Additionen vorkommen. In einem Programm könnten ja auch beliebig komplizierte Operationen verwendet werden, warum also diese Beschränkung.

Für die nachfolgenden Betrachtungen stellen wir uns X daher nicht mehr einfach als einen Platzhalter für eine Zahl vor, sondern als ein neues algebraisches Objekt, für das man die Rechenregeln erst noch definieren muss. In diesem Kapteil sollen die Regeln zum Beispiel sicherstellen, dass man mit Polynomen so rechnen kann, wie wenn X eine Zahl wäre. Es sollen also zum Beispiel die Regeln

$$aX = Xa \qquad (a+b)X = aX + bX \qquad a + X = X + a \tag{3.3}$$

gelten. In dieser algebraischen Sichtweise können je nach den gewählten algebraischen Rechenregeln für X interessante rechnerische Strukturen abgebildet werden. Ziel dieses Kapitels ist zu zeigen, wie man die Rechenregeln für X mit Hilfe von Matrizen allgemein darstellen kann. Diese Betrachtungsweise wird später in Anwendungen ermöglichen, handliche Realisierungen für das Rechnen mit Grössen zu finden, die polynomielle Gleichungen erfüllen. Ebenso sollen in späteren Kapiteln die Regeln (3.3) erweitert oder abgelöst werden um weitere Anwendungen zu erschliessen.

Bei der Auswahl der zusätzlichen algebraischen Regeln muss man sehr vorsichtig vorgehen. Nimmt man zum Beispiel an, dass man durch X teilen kann, dann würde dies in der arithmetischen Sichtweise bereits ausschliessen, dass man für X die Zahl 0 einsetzen kann. Aber auch eine Regel wie $X^2 \geq 0$, die für alle reellen Zahlen gilt, würde die Anwendungsmöglichkeiten zu stark einschränken.

Es gibt zwar keine reelle Zahl, die man in das Polynom $p(X) = X^2 + 1$ einsetzen könnte, so dass es den Wert 0 annimmt. Man könnte X aber als ein neues Objekt ausserhalb von $\mathbb{R}$ betrachten, welches die Gleichung $X^2+1 = 0$ erfüllt. In den komplexen Zahlen $\mathbb{C}$ gibt es mit der imaginären Einheit $i \in \mathbb{C}$ tatsächlich ein Zahl mit der Eigenschaft $i^2 = -1$ und damit eine Objekt, welches die Ungleichung $X^2 \geq 0$ verletzt.

Für das Symbol X sollen also die "üblichen" Rechenregeln gelten. Dies ist natürlich nur sinnvoll, wenn man auch mit den Koeffizienten $a_0, \ldots, a_n$ rechnen kann. Sie müssen also Elemente einer algebraischen Struktur sein, in der mindestens die Addition und die Multiplikation definiert sind. Die ganzen Zahlen $\mathbb{Z}$ kommen dafür in Frage, aber auch die rationalen oder reellen Zahlen $\mathbb{Q}$ und $\mathbb{R}$. Man kann sogar noch weiter gehen: man kann als Koeffizienten auch Vektoren oder sogar Matrizen zulassen. Polynome können addiert werden, indem die Koeffizienten addiert werden. Polynome können aber auch multipliziert werden, was auf die Faltung der Koeffizienten hinausläuft:

$$\begin{aligned} p(X) &= a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 \\ q(X) &= b_m X^m + b_{m-1} X^{m-1} + \cdots + b_1 X + b_0 \\ p(X)q(X) &= a_n b_m X^{n+m} + (a_n b_{m-1} + a_{n-1} b_m) X^{n+m-1} + \cdots + \sum_{i+j=k} a_i b_j X^k + \cdots + (a_1 b_0 + a_0 b_1) X + a_0 b_0 \end{aligned} \tag{3.4}$$

Dies ist aber nur möglich, wenn die Koeffizienten selbst miteinander multipliziert werden können, wenn also die Koeffizienten mindestens Elemente einer Algebra sind.

3.1 Definitionen

In diesem Abschnitt stellen wir einige grundlegende Definitionen für das Rechnen mit Polynomen zusammen.

3.1.1 Skalare

Wie schon in der Einleitung angedeutet sind Polynome nur dann sinnvoll, wenn man mit den Koeffizienten gewisse Rechenoperationen durchführen kann. Wir brauchen mindestens die Möglichkeit, Koeffizienten zu addieren. Wenn wir uns vorstellen, dass wir X durch eine Zahl ersetzen können, dann brauchen wir zusätzlich die Möglichkeit, einen Koeffizienten mit einer Zahl zu multiplizieren.

Die Struktur, die wir hier beschrieben haben, hängt davon ab, was wir uns unter einer "Zahl" vorstellen. Wir bezeichnen die Menge, aus der die "Zahlen" kommen können mit R und nennen sie die Menge der Skalare. Wenn wir uns vorstellen, dass man die Elemente von R an Stelle von X in das Polynom einsetzen kann, dann muss es möglich sein, in R zu Multiplizieren und zu Addieren, und es müssen die üblichen Rechenregeln der Algebra gelten, R muss also ein Ring sein. Wir werden im folgenden meistens voraussetzen, dass R sogar kommutativ ist und eine 1 hat.

Definition 3.1. *Sei R ein Ring. Die Menge*

$$R[X] = \{p(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots a_1 X + a_0 \mid a_k \in R, n \in \mathbb{N}\}$$

heisst die Menge der Polynome *mit Koeffizienten in R oder* Polynome über *R. Polynome können addiert werden, indem Koeffizienten mit gleichem Index addiert werden:*

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0$$

$$q(X) = b_n X^n + b_{n-1} X^{n-1} + \cdots + b_1 X + b_0$$
$$p(X) + q(X) = (a_n + b_n)X^n + (a_{n-1} + b_{n-1})X^{n-1} + \cdots + (a_1 + b_1)X + (a_0 + b_0)$$

Die Multiplikation ist durch die Formel (3.4) *definiert.*

Ein Polynom heisst *normiert* oder auch *monisch*, wenn der höchste Koeffizient oder auch *Leitkoeffizient* des Polynomus 1 ist, also $a_n = 1$. Wenn man in R durch a_n dividieren kann, dann kann man aus dem Polynom $p(X) = a_n X^n + \ldots$ mit Leitkoeffizient a_n das normierte Polynom

$$\frac{1}{a_n} p(X) = \frac{1}{a_n}(a_n X^n + \cdots + a_0) = X^n + \frac{a_{n-1}}{a_n} X^{n-1} + \cdots + \frac{a_0}{a_n}$$

machen. Man sagt auch, das Polynom $p(X)$ wurde normiert.

Die Tatsache, dass zwei Polynome nicht gleich viele von 0 verschiedene Koeffizienten haben müssen, verkompliziert die Beschreibung der Rechenoperationen ein wenig. Wir werden daher im Folgenden oft für ein Polynom

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots a_1 X + a_0$$

annehmen, dass alle Koeffizienten $a_{n+1}, a_{n+2}, \ldots$ implizit mit Wert 0 definiert sind. Wir werden uns also erlauben,

$$p(X) = \sum_k a_k X^k = \sum_{k=0}^{\infty} a_k X^k$$

zu schreiben, wobei in der ersten Form das Summenzeichen bedeuten soll, dass nur über diejenigen Indizes k summiert wird, für die a_k definiert ist.

3.1.2 Der Polynomring

Die Menge $R[X]$ aller Polynome über R wird zu einem Ring, wenn man die Rechenoperationen Addition und Multiplikation so definiert, wie man das in der Schule gelernt hat. Die Summe von zwei Polynomen

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0$$
$$q(X) = b_m X^m + b_{m-1} X^{m-1} + \cdots + b_1 X + b_0$$

ist

$$p(X) + q(X) = \sum_k (a_k + b_k) X^k,$$

wobei die Summe wieder so zu interpretieren ist, über alle Terme summiert wird, für die mindestens einer der Summanden von 0 verschieden ist.

Für das Produkt verwenden wir die Definition

$$p(X)q(X) = \sum_k \sum_l a_k b_l X^{k+l},$$

die natürlich mit Formel (3.4) gleichbedeutend ist. Die Polynom-Multiplikation und Addition sind nur eine natürliche Erweiterung der Rechenregeln, die man schon in der Schule lernt, es ist daher

nicht überraschend, dass die bekannten Rechenregeln auch für Polynome gelten. Das Distributivgesetz

$$p(X)(u(X)+v(X)) = p(X)u(X)+p(X)v(X) \qquad (p(X)+q(X))u(X) = p(X)u(X)+q(X)u(X)$$

zum Beispiel sagt ja nichts anderes, als dass man ausmultiplizieren kann. Oder die Assoziativgesetze

$$p(X)+q(X)+r(X) = (p(X)+q(X))+r(X) = p(X)+(q(X)+r(X))$$
$$p(X)q(X)r(X) = (p(X)q(X))r(X) = p(X)(q(X)r(X))$$

für die Multiplikation besagt, das es keine Rolle spielt, in welcher Reihenfolge man die Additionen oder Multiplikationen ausführt.

3.1.3 Grad

Definition 3.2. *Der* Grad *eines Polynoms $p(X)$ ist die höchste Potenz von X, die im Polynom vorkommt. Das Polynom*

$$p(X) = a_nX^n + a_{n-1}X^{n-1} + \dots a_1X + a_0$$

hat den Grad n, wenn $a_n \neq 0$ ist. Der Grad von p wird mit $\deg p$ *bezeichnet. Konstante Polynome $p(X) = a_0$ mit $a_0 \neq 0$ hat den Grad* 0*. Der Grad des Nullpolynoms $p(X) = 0$ ist definiert als $-\infty$.*

Der Grad eines Polynoms ist sinnvoll in dem Sinn, dass er sich mit den Rechenoperationen gut verträgt. Damit lässt sich weiter unten auch die spezielle Wahl des Grades des Nullpolynoms begründen. Es gelten nämlich die folgenden Rechenregeln.

Lemma 3.3. *Sind p und q Polynome mit Koeffizienten in R und $0 \neq \lambda \in R$, dann gilt*

$$\deg(pq) \leq \deg p + \deg q \tag{3.5}$$
$$\deg(p+q) \leq \max(\deg p, \deg q) \tag{3.6}$$
$$\deg(\lambda p) \leq \deg p \tag{3.7}$$

Die Formel (3.7) ist eigentlich ein Spezialfall von (3.5). Die Zahl $\lambda \in R$ kann man als Polynom vom Grad 0 betrachten, wofür natürlich (3.5) gilt, also $\deg(\lambda p) \leq \deg \lambda + \deg p$.

Beweis. Wir schreiben die Polynome wieder in der Form

$$p(X) = a_nX^n + a_{n-1}X^{n-1} + \dots + a_1X + a_0 \qquad \Rightarrow \qquad \deg p = n$$
$$q(X) = b_mX^m + b_{m-1}X^{m-1} + \dots + b_1X + b_0 \qquad \Rightarrow \qquad \deg q = m.$$

Dann kann der höchste Koeffizient in der Summe $p+q$ nicht weiter oben sein als die grössere von den beiden Zahlen n und m angibt, dies beweist (3.5). Ebenso kann der höchste Koeffizient im Produkt nach der Formel (3.4) nicht weiter oben als bei $n+m$ liegen, dies beweist beweist (3.6). Es könnte aber passieren, dass $a_nb_m = 0$ ist, d. h. es ist durchaus möglich, dass der Grad kleiner ist. Schliesslich kann der höchsten Koeffizient von $\lambda p(X)$ nicht grösser als der höchste Koeffizient von $p(X)$ sein, was (3.7) beweist. □

Etwas enttäuschend an diesen Rechenregeln ist, dass der Grad eines Produktes nicht exakt die Summe der Grade hat. Der Grund ist natürlich, dass es in gewissen Ringen R passieren kann, dass das Produkt $a_n \cdot b_m = 0$ ist. Zum Beispiel ist im Ring der 2×2 Matrizen das Produkt der Elemente

$$a_n = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{und} \quad b_m = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Rightarrow \qquad a_n b_m = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.8}$$

Diese unangehme Situation tritt immer ein, wenn es von Null verschiedene Elemente gibt, deren Produkt 0 ist. In Matrizenringen ist das der Normalfall, man kann diesen Fall also nicht einfach ausschliessen. In den Zahlenmengen wie $\mathbb{Z}$, $\mathbb{Q}$ und $\mathbb{R}$ passiert das natürlich nie.

Definition 3.4. *Ein Ring R heisst* nullteilerfrei*, wenn für zwei Elemente $a, b \in R$ aus $ab = 0$ immer geschlossen werden kann, dass $a = 0$ oder $b = 0$. Ein von 0 verschiedenes Element $a \in R$ heisst Nullteiler, wenn es eine $b \in R$ mit $b \neq 0$ gibt derart dass $ab = 0$.*

Die beiden Matrizen in (3.8) sind Nullteiler im Ring $M_2(\mathbb{Z})$ der 2×2-Matrizen. Der Matrizenring $M_2(\mathbb{Z})$ ist also nicht nullteilerfrei.

In einem nullteilerfreien Ring gelten die Rechenregeln für den Grad jetzt exakt:

Lemma 3.5. *Sei R ein nullteilerfreier Ring und p und q Polynome über R und $0 \neq \lambda \in R$. Dann gilt*

$$\deg(pq) = \deg p + \deg q \tag{3.9}$$

$$\deg(p + q) \leq \max(\deg p, \deg q) \tag{3.10}$$

$$\deg(\lambda p) = \deg p \tag{3.11}$$

Beweis. Der Fall, dass der höchste Koeffizient verschwindet, weil a_n, b_m oder λ Nullteiler sind, kann unter den gegebenen Voraussetzungen nicht eintreten, daher werden die in Lemma 3.3 gefunden Ungleichungen für Produkte exakt. □

Die Gleichung (3.11) kann im Fall $\lambda = 0$ natürlich nicht gelten. Betrachten wir λ wieder als ein Polynom, dann folgt aus (3.9), dass

$$\begin{aligned} \lambda \neq 0 \quad &\Rightarrow \quad \deg(\lambda p) = \deg \lambda + \deg p = 0 + \deg p \\ \lambda = 0 \quad &\Rightarrow \quad \deg(0p) = \deg 0 + \deg p = \deg 0 \end{aligned}$$

Diese Gleichung kann also nur aufrechterhalten werden, wenn die "Zahl" $\deg 0$ die Eigenschaft besitzt, dass man immer noch $\deg 0$ bekommt, wenn man irgend eine Zahl $\deg p$ hinzuaddiert. Wenn also

$$\deg 0 + \deg p = \deg 0 \qquad \forall \deg p \in \mathbb{Z}$$

gilt. So eine Zahl gibt es in den ganzen Zahlen nicht. Wenn man zu einer ganzen Zahl eine andere ganze Zahl hinzuaddiert, ändert sich fast immer etwas. Man muss daher $\deg 0 = -\infty$ setzen und festlegen, dass $-\infty + n = -\infty$ für beliebige ganze Zahlen n gilt.

Definition 3.6. *Die Polynome vom Grad $\leq n$ mit Koeffizienten in R bilden die Teilmenge*

$$R^{(n)}[X] = \{p \in R[X] \mid \deg p \leq n\}.$$

Die Mengen $R^{(n)}[X]$ bilden eine Filtrierung *des Polynomrings $R[X]$, d. h. sie sind ineinander geschachtelt*

$$\begin{array}{ccccccccccccccc} R^{(-\infty)}[X] & \subset & R^{(0)}[X] & \subset & & R^{(1)}[X] & & \subset \ \ldots \ \subset & R^{(k)}[X] & \subset & R^{(k+1)}[X] & \subset \ \ldots \ \subset & R[X] \\ \| & & \| & & & \| & & & & & & & & \\ \{0\} & \subset & R & \subset & \{a_1 X + a_0 \mid a_k \in R\} & & \subset \ \ldots & & & & & & \end{array}$$

und ihre Vereinigung ist $R[X]$.

Die Formeln für den Grad können wir auch mit den Mengen $R^{(k)}[X]$ ausdrücken:

$$\begin{aligned} \deg(p+q) \le \max(\deg p, \deg q) \quad &\Rightarrow \quad R^{(k)} + R^{(l)} \subset R^{(\max(k,l))} = R^{(k)}[X] \cup R^{(l)}[X]. \\ \deg(p \cdot q) = \deg p + \deg q \quad &\Rightarrow \quad R^{(k)}[X] \cdot R^{(l)}[X] = R^{(k+l)}[X]. \end{aligned}$$

3.1.4 Teilbarkeit

Im Ring der ganzen Zahlen sind nicht alle Divisionen ohne Rest ausführbar, so entsteht das Konzept der Teilbarkeit. Der Divisionsalgorithmus, den man in der Schule lernt, liefert zu beliebigen ganzen Zahlen $a, b \in \mathbb{Z}$ den Quotienten q und den Rest r derart, dass $a = qb + r$. Der Algorithmus basiert auf der Zehnersystemdarstellung

$$\begin{aligned} a &= a_n 10^n + a_{n-1} 10^{n-1} + \cdots + a_1 10^1 + a_0 \\ b &= b_m 10^n + b_{m-1} 10^{n-1} + \cdots + b_1 10^1 + b_0 \end{aligned}$$

und ermittelt den Quotienten, indem er mit den einzelnen Stellen a_k und b_k arbeitet. Er ist also eigentlich ein Algorithmus für die Polynome

$$\begin{aligned} a &= a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X^1 + a_0 \\ b &= b_m X^n + b_{m-1} X^{n-1} + \cdots + b_1 X^1 + b_0, \end{aligned}$$

mit dem einzigen Unterschied, dass statt X mit der festen Zahl $X = 10$ gearbeitet wird. Der Teilungsalgorithmus für Polynome lässt sich aber leicht rekonstruieren.

Polynomdivision

Wir zeigen den Polynomdivisionsalgorithmus an einem konkreten Beispiel. Gesucht sind Quotient $q \in \mathbb{Z}[X]$ und Rest $r \in \mathbb{Z}[X]$ der beiden Polynome

$$\begin{aligned} a(X) &= X^4 - X^3 - 7X^2 + X + 6 \\ b(X) &= X^2 + X + 1, \end{aligned} \tag{3.12}$$

für die also gilt $a = bq + r$. Die Division ergibt

$$\begin{array}{rl} X^4 - X^3 - 7X^2 + X + 6 & : X^2 + X + 1 = X^2 - 2X - 6 = q \\ \underline{-(X^4 + X^3 + X^2)} & \\ -2X^3 - 8X^2 + X & \\ \underline{-(-2X^3 - 2X^2 - 2X)} & \\ -6X^2 + 3X + 6 & \\ \underline{-(-6X^2 - 6X - 6)} & \\ 9X + 12 = r & \end{array}$$

Durch nachrechnen kann man überprüfen, dass tatsächlich

$$\begin{aligned} bq &= X^4 - X^3 - 7X^2 - 8X - 6 \\ bq + r &= X^4 - X^3 - 7X^2 + X + 6 = a \end{aligned}$$

gilt.

Das Beispiel (3.13) war besonders einfach, weil der führende Koeffizient des Divisorpolynomes 1 war. Für $b = 2X^2 + X + 1$ funktioniert der Algorithmus dagegen nicht mehr. Jedes für q in Frage kommende Polynom vom Grad 2 muss von der Form $q = q_2X^2 + q_1X + q_0$ sein. Multipliziert man mit b, erhält man $bq = 2q_2X^4 + (2q_1 + q_2)X^3 + \ldots$. Insbesondere ist es nicht möglich mit ganzzahligen Quotienten $q_k \in \mathbb{Z}$ auch nur der ersten Koeffizienten von a zu erhalten. Dazu müsste nämlich $a_n = 1 = 2q_2$ oder $q_2 = \frac{1}{2} \notin \mathbb{Z}$ sein. Der Divisionsalgorithmus funktioniert also nur dann, wenn die Division durch den führenden Koeffizienten des Divisorpolynomes b immer ausführbar ist. Im Beispiel (3.13) war das der Fall, weil der führende Koeffizient 1 war. Für beliebige Polynome $b \in R[X]$ ist das aber nur der Fall, wenn die Koeffizienten in Tat und Wahrheit einem Körper entstammen.

Im Folgenden betrachten wir daher nur noch Polynomringe mit Koeffizienten in einem Körper $\Bbbk$. In $\mathbb{Q}[X]$ ist die Division $a : b$ für die Polynome

$$\begin{aligned} a(X) &= X^4 - X^3 - 7X^2 + X + 6 \\ b(X) &= X^2 + X + 1, \end{aligned} \tag{3.13}$$

problemlos durchführbar:

$$\begin{array}{rl}
X^4 - X^3 - 7X^2 + X + 6 : 2X^2 + X + 1 = \frac{1}{2}X^2 - \frac{3}{4}X - \frac{27}{8} = q \\
\underline{-(X^4 + \frac{1}{2}X^3 + \frac{1}{2}X^2)} & \\
-\frac{3}{2}X^3 - \frac{15}{2}X^2 + X & \\
\underline{-(-\frac{3}{2}X^3 - \frac{3}{4}X^2 - \frac{3}{4}X)} & \\
-\frac{27}{4}X^2 + \frac{7}{4}X + 6 & \\
\underline{-(-\frac{27}{4}X^2 - \frac{27}{8}X - \frac{27}{8})} & \\
\frac{41}{8}X + \frac{75}{8} = r &
\end{array}$$

Der Algorithmus funktioniert selbstverständlich genauso in $\mathbb{R}[X]$ oder $\mathbb{C}[X]$, und ebenso in den in Kapitel 4 studierten endlichen Körpern.

Euklidische Ringe und Faktorzerlegung

Der Polynomring $\Bbbk[X]$ hat noch eine weitere Eigenschaft, die ihn von einem gewöhnlichen Ring unterschiedet. Der Polynomdivisionsalgorithmus findet zu zwei Polynomen $f, g \in \Bbbk[X]$ den Quotienten $q \in \Bbbk[X]$ und den Rest $r \in \Bbbk[X]$ mit $f = qg + r$, wobei ausserdem $\deg r < \deg g$ ist.

Definition 3.7. *Ein* euklidischer Ring *R ist ein nullteilerfreier Ring mit einer Gradfunktion* $\deg\colon R \setminus \{0\} \to \mathbb{N}$ *mit folgenden Eigenschaften*

1. *Für $x, y \in R$ gilt* $\deg(xy) \geq \deg(x)$.
2. *Für alle $x, y \in R$ gibt es $q, r \in R$ mit $x = qy + r$ mit* $\deg(y) > \deg(x)$

Bedingung 2 ist die Division mit Rest.

Die ganzen Zahlen $\mathbb{Z}$ bilden einen euklidischen Ring mit der Gradfunktion $\deg(z) = |z|$ für $z \in \mathbb{Z}$. Aus dem Divisionsalgorithmus für ganze Zahlen leiten sich alle grundlegenden Eigenschaften über Teilbarkeit und Primzahlen ab. Eine Zahl x ist teilbar durch y, wenn $x = qy$ mit $q \in \mathbb{Z}$, es gibt Zahlen $p \in \mathbb{Z}$, die keine Teiler haben und jede Zahl kann auf eindeutige Art und Weise in ein Produkt von Primfaktoren zerlegt werden.

Irreduzible Polynome

Das Konzept der Primzahl lässt sich wie folgt in den Polynomring übertragen.

Definition 3.8. *Ein Polynom $f \in R[X]$ heisst irreduzibel, es keine Faktorisierung $f = gh$ in Faktoren $g, h \in R[X]$ mit* $\deg(g) > 0$ *und* $\deg(h) > 0$.

Beispiel. Polynome ersten Grades $aX + b$ sind immer irreduzibel, da sie bereits minimalen Grad haben.

Sei jetzt $f = X^2 + bX + c$ ein quadratisches Polynom in $\mathbb{Q}[X]$. Wenn es faktorisierbar sein soll, dann müssen die Faktoren Polynome ersten Grades sein, also $f = (X - x_1)(X - x_2)$ mit $x_i \in \mathbb{Q}$. Die Zahlen x_i die einzigen möglichen Lösungen für x_i können mit der Lösungsformel für die quadratische Gleichung

$$x_i = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - c}$$

gefunden werden. Die Faktorisierung ist also genau dann möglich, wenn $b^2/4 - c$ ein Quadrat in $\mathbb{Q}$. In $\mathbb{R}$ ist das Polynom faktorisierbar, wenn $b^2 - 4c \geq 0$ ist. In $\mathbb{C}$ gibt es keine Einschränkung, die Wurzel zu ziehen, in $\mathbb{C}$ gibt es also keine irreduziblen Polynome im Grad 2. ○

Faktorisierung in einem Polynomring

Ein Polynomring ist ganz offensichtlich auch ein euklidischer Ring. Wir erwarten daher die entsprechenden Eigenschaften auch in einem Polynomring. Allerdings ist eine Faktorzerlegung nicht ganz eindeutig. Wenn das Polynom $f \in \mathbb{Z}[X]$ die Faktorisierung $f = g \cdot h$ mit $g, h\mathbb{Z}[X]$ hat, dann ist $rg \cdot r^{-1}h$ ebenfalls eine Faktorisierung für jedes $r = \pm 1$. Dasselbe gilt in $\mathbb{Q}$ für jedes $r \in \mathbb{Q}^*$. Faktorisierung ist also nur eindeutig bis auf Elemente der Einheitengruppe des Koeffizientenringes. Diese Mehrdeutigkeit kann in den Polynomringen $\Bbbk[X]$ überwunden werden, indem die Polynome normiert werden.

Satz 3.9. *Ein normiertes Polynom $f \in \Bbbk[X]$ kann in normierte Faktoren $g_1, \ldots, g_k \in \Bbbk[X]$ zerlegt werden, so dass $f = g_1 \cdot \ldots \cdot g_k$, wobei die Faktoren irreduzibel sind. Zwei solche Faktorisierungen unterscheiden sich nur durch die Reihenfolge der Faktoren. Ein Polynom $f \in \Bbbk[X]$ kann in ein Produkt $a_n g_1 \cdot \ldots \cdot g_k$ zerlegt werden, wobei die normierten Faktoren g_i bis auf die Reihenfolge eindeutig sind.*

3.1.5 Formale Potenzreihen

XXX TODO

3.2 Polynome als Vektoren

Ein Polynom

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots a_1 X + a_0$$

mit Koeffizienten in einem Ring R ist spezifiziert, wenn die Koeffizienten a_k bekannt sind. Die Potenzen von X dienen hier nur dazu, die verschiedenen Koeffizienten zu unterscheiden. Das Polynom

$p(X)$ vom Grad n ist also auch gegeben durch den $n+1$-dimensionalen Vektor

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \\ a_n \end{pmatrix} \in R^n.$$

Diese Darstellung eines Polynoms gibt auch die Addition von Polynomen und die Multiplikation von Polynomen mit Skalaren aus R korrekt wieder. Die Abbildung von Vektoren auf Polynome

$$\varphi\colon R^n \to R[X] : \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} \mapsto a_n X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0$$

erfüllt also

$$\varphi(\lambda a) = \lambda\varphi(a) \qquad \text{und} \qquad \varphi(a+b) = \varphi(a) + \varphi(b)$$

und ist damit eine lineare Abbildung. Umgekehrt kann man auch zu jedem Polynom $p(X)$ vom Grad $\leq n$ einen Vektor finden, der von φ auf das Polynom $p(X)$ abgebildet wird. Die Abbildung φ ist also ein Isomorphismus

$$\varphi\colon \{p \in R[X] \mid \deg(p) \leq n\} \xrightarrow{\equiv} R^{n+1}$$

zwischen der Menge der Polynome vom Grad $\leq n$ auf R^{n+1}. Für alle Rechnungen, bei denen es nur um Addition von Polynomen oder um Multiplikation mit Skalaren geht, ist also diese vektorielle Darstellung mit Hilfe von φ eine zweckmässige Darstellung.

In zwei Bereichen ist die Beschreibung von Polynomen mit Vektoren allerdings ungenügend: einerseits können Polynome können beliebig hohen Grad haben, während Vektoren in R^{n+1} höchstens $n+1$ Komponenten haben können. Andererseits geht bei der vektoriellen Beschreibung die multiplikative Struktur vollständig verloren.

3.2.1 Polynome beliebigen Grades

Ein Polynom

$$q(X) = b_m X^m + b_{m-1}X^{m-1} + \cdots + b_1 X + b_0$$

vom Grad $m < n$ kann dargestellt werden als ein Vektor

$$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{m-1} \\ b_m \\ 0 \\ \vdots \end{pmatrix} \in R^{n+1}$$

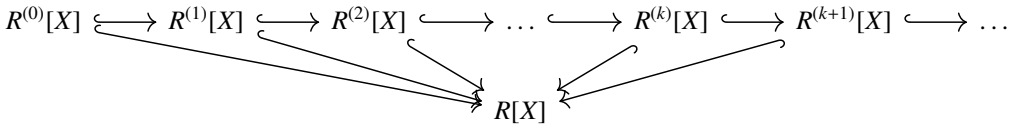
mit der Eigenschaft, dass die Komponenten mit Indizes $m+1, \ldots n$ verschwinden. Polynome vom Grad $m < n$ bilden einen Unterraum der Polynome vom Grad n. Wir können auch die $m+1$-dimensionalen Vektoren in den $n+1$-dimensionalen Vektoren einbetten, indem wir die Vektoren

durch “auffüllen” mit Nullen auf die richtige Länge bringen. Es gibt also eine lineare Abbildung

$$R^{m+1} \to R^{n+1} : \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix} \mapsto \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \\ 0 \\ \vdots \end{pmatrix}.$$

Die Moduln R^k sind also alle ineinandergeschachtelt, können aber alle auf konsistente Weise mit der Abbildung φ in den Polynomring $R[X]$ abgebildet werden.

$$\{0\} \longrightarrow R \longrightarrow R^2 \longrightarrow \ldots \longrightarrow R^k \longrightarrow R^{k+1} \longrightarrow \ldots$$



3.2.2 Multiplikative Struktur

3.3 Polynommultiplikation mit Matrizen

3.4 Minimalpolynom

Kapitel 4

Endliche Körper

Aus den ganzen Zahlen $\mathbb{Z}$ entsteht ein Körper, indem wir Brüche bilden alle von 0 verschiedenen Nenner zulassen. Der Körper der rationalen Zahlen $\mathbb{Q}$ enthält unendliche viele Zahlen und hat zusätzlich die sogenannte archimedische Eigenschaft, nämliche dass es zu zwei positiven rationalen Zahlen a und b immer eine ganze Zahl n gibt derart, dass $na > b$. Dies bedeutet auch, dass es in den rationalen Zahlen beliebig grosse Zahlen gibt. Man kann aus den ganzen Zahlen aber auch eine Reihe von Körpern ableiten, die diese Eigenschaft nicht haben. Nicht überraschend werden die ersten derartigen Körper, die wir in Abschnitt 4.2 konstruieren werden, endlich viele Elemente haben. Als Hilfsmittel für die Definition der Division in diesem Körper wird als Vorbereitung in Abschnitt 4.1 der euklidische Algorithmus vorgestellt, wobei auch eine besonders zum Thema dieses Buches passende Beschreibung in Matrixform angegeben wird. Zu diesen sogenannten Galois-Körpern können wir dann weitere Elemente hinzufügen, wie das in Abschnitt 4.3 gezeigt wird. Diese Technik, die auch für den Körper $\mathbb{Q}$ funktioniert, erlaubt dafür zu sorgen, dass in einem Körper gewisse algebraische Gleichungen lösbar werden.

4.1 Der euklidische Algorithmus

Der euklidische Algorithmus bestimmt zu zwei gegebenen ganzen Zahlen a und b den grössten gemeinsamen Teiler g. Zusätzlich findet er ganze Zahlen s und t derart, dass

$$sa + tb = g.$$

In diesem Abschnitt soll der Algorithmus zunächst für ganze Zahlen vorgestellt werden, bevor er auf Polynome verallgemeinert und dann in Matrixform niedergeschrieben wird.

4.1.1 Ganze Zahlen

Gegeben sind zwei ganze Zahlen a und b und wir dürfen annehmen, dass $a \geq b$. Gesucht ist der grösste gemeinsame Teiler g von a und b. Wir schreiben $g|a$ für "g ist Teiler von a" oder "g teilt a", gesucht ist also die grösste ganze Zahl g derart, dass $g|a$ und $g|b$.

Ist $b|a$, dann ist offenbar b der grösste gemeinsame Teiler von a und b. Im Allgemeinen wird der grösste gemeinsame Teiler aber kleiner sein. Wir teilen daher a durch b, was nur mit Rest möglich ist. Es gibt ganze Zahlen q, der Quotient, und r, der Rest, derart, dass

$$a = qb + r \qquad \Rightarrow \qquad r = a - qb. \tag{4.1}$$

Nach Definition des Restes ist $r < b$. Da der grösste gemeinsame Teiler sowohl a als auch b teilt, muss er wegen (4.1) auch r teilen. Somit haben wir das Problem, den grössten gemeinsamen Teiler von a und b zu finden, auf das "kleinere" Problem zurückgeführt, den grössten gemeinsamen Teiler von b und r zu finden.

Um den eben beschriebenen Schritt zu wiederholen, wählen wir die folgende Notation. Wir schreiben $a_0 = a$ und $b_0 = b$. Im ersten Schritt finden wird q_0 und r_0 derart, dass $a_0 - q_0 b_0 = r_0$. Dann setzen wir $a_1 = b_0$ und $b_1 = r_0$. Mit a_1 und b_1 wiederholen wir den Divisionsschritt, der einen neuen Quotienten q_1 und einen neuen Rest r_1 liefert mit $a_1 - q_1 b_1 = r_1$. So entstehen vier Folgen von Zahlen a_k, b_k, q_k und r_k derart, dass in jedem Schritt gilt

$$a_k - q_k b_k = r_k \qquad g|a_k \qquad g|b_k \qquad a_k = b_{k-1} \qquad b_k = r_{k-1}$$

Der Algorithmus bricht im Schritt n ab, wenn $r_{n+1} = 0$. Der letzte nicht verschwindende Rest r_n muss daher der grösste gemeinsame Teiler sein: $g = r_n$.

Beispiel. Wir bestimmen den grössten gemeinsamen Teiler von 76415 und 23205 mit Hilfe des eben beschriebenen Algorithmus. Wir schreiben die gefundenen Zahlen in eine Tabelle:

k	a_k	b_k	q_k	r_k
0	76415	23205	3	6800
1	23205	6800	3	2805
2	6800	2805	2	1190
3	2805	1190	2	425
4	1190	425	2	340
5	425	340	1	85
6	340	85	4	0

Der Algorithmus bricht also mit dem letzten Rest $r_n = 85$ ab, dies ist der grösste gemeinsame Teiler. ○

Die oben protokollierten Werte von q_k werden für die Bestimmung des grössten gemeinsamen Teilers nicht benötigt. Wir können sie aber verwenden, um die Zahlen s und t zu bestimmen.

Beispiel. Wir drücken die Reste im obigen Beispiel durch die Zahlen a_k, b_k und q_k aus und setzen sie in den Ausdruck $g = a_5 - q_5 b_5$ ein, bis wir einen Ausdruck in a_0 und b_0 für g finden:

$$\begin{aligned}
r_5 &= a_5 - q_5 b_5 = a_5 - 1 \cdot b_5 & g &= a_5 - 1 \cdot b_5 = b_4 - 1 \cdot r_4 \\
r_4 &= a_4 - q_4 b_4 = a_4 - 2 \cdot b_4 & &= b_4 - (a_4 - 2b_4) = -a_4 + 3b_4 = -b_3 + 3r_3 \\
r_3 &= a_3 - q_3 b_3 = a_3 - 2 \cdot b_3 & &= -b_3 + 3(a_3 - 2b_3) = 3a_3 - 7b_3 = 3b_2 - 7r_2 \\
r_2 &= a_2 - q_2 b_2 = a_2 - 2 \cdot b_2 & &= 3b_2 - 7(a_2 - 2b_2) = -7a_2 + 17b_2 = -7b_1 + 17r_1 \\
r_1 &= a_1 - q_1 b_1 = a_1 - 3 \cdot b_1 & &= -7b_1 + 17(a_1 - 3b_1) = 17a_1 - 58b_1 = 17b_0 - 58r_0 \\
r_0 &= a_0 - q_0 b_0 = a_0 - 3 \cdot b_0 & &= 17b_0 - 58(a_0 t - 3b_0) = -58a_0 + 191b_0
\end{aligned}$$

Tatsächlich gilt

$$-58 \cdot 76415 + 191 \cdot 23205 = 85,$$

die Zahlen $t = -58$ und $s = 191$ sind also genau die eingangs versprochenen Faktoren. ○

4.1.2 Matrixschreibweise

Die Durchführung des euklidischen Algorithmus lässt sich besonders elegant in Matrixschreibweise dokumentieren. In jedem Schritt arbeitet man mit zwei ganzen Zahlen a_k und b_k, die wir als zweidimensionalen Spaltenvektor betrachten können. Der Algorithmus macht aus a_k und b_k die neuen Zahlen $a_{k+1} = b_k$ und $b_{k+1} = r_k = a_k - q_k b_k$, dies kann man als

$$\begin{pmatrix} a_{k+1} \\ b_{k+1} \end{pmatrix} = \begin{pmatrix} b_k \\ r_k \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_k \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix}$$

schreiben. Der Algorithmus bricht ab, wenn die zweite Komponente des Vektors = 0 ist, in der ersten steht dann der grösste gemeinsame Teiler. Hier ist die Durchführung des Algorithmus in Matrix-Schreibweise:

$$\begin{aligned}
\begin{pmatrix} 23205 \\ 6800 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} 76415 \\ 23205 \end{pmatrix} \\
\begin{pmatrix} 6800 \\ 2805 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} 23205 \\ 6800 \end{pmatrix} \\
\begin{pmatrix} 2805 \\ 1190 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 6800 \\ 2805 \end{pmatrix} \\
\begin{pmatrix} 1190 \\ 425 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 2805 \\ 1190 \end{pmatrix} \\
\begin{pmatrix} 425 \\ 340 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 1190 \\ 425 \end{pmatrix} \\
\begin{pmatrix} 340 \\ 85 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 425 \\ 340 \end{pmatrix} \\
\begin{pmatrix} 85 \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & -4 \end{pmatrix} \begin{pmatrix} 340 \\ 85 \end{pmatrix} = \begin{pmatrix} g \\ 0 \end{pmatrix}.
\end{aligned}$$

Definition 4.1. *Wir kürzen*

$$Q(q_k) = \begin{pmatrix} 0 & 1 \\ 1 & -q_k \end{pmatrix}$$

ab.

Mit dieser Definition lässt sich der euklidische Algorithmus wie folgt beschreiben.

Algorithmus 4.2 (Euklid)**.** *Der Algorithmus operiert auf zweidimensionalen Zustandsvektoren $x \in \mathbb{Z}^2$ wie folgt:*

1. *Initialisiere den Zustandsvektor mit den ganzen Zahlen a und b: $x = \begin{pmatrix} a \\ b \end{pmatrix}$*
2. *Bestimme den Quotienten q als die grösste ganze Zahl, für die $qx_2 \leq x_1$ gilt.*
3. *Berechne den neuen Zustandsvektor als $Q(q)x$.*
4. *Wiederhole Schritte 2 und 3 bis die zweite Komponente des Zustandsvektors verschwindet. Die erste Komponente ist dann der gesuchte grösste gemeinsame Teiler.*

Auch die Berechnung der Zahlen s und t lässt sich jetzt leichter verstehen. Nach Algorithmus 4.2 ist

$$\begin{pmatrix} g \\ 0 \end{pmatrix} = Q(q_n)Q(q_{n-1})\cdots Q(q_0)\begin{pmatrix} a \\ b \end{pmatrix}.$$

Schreiben wir $Q = Q(q_n)Q(q_{n-1})\cdots Q(q_0)$, dann enthält die Matrix Q in der erste Zeile die ganzen Zahlen s und t, mit denen sich der grösste gemeinsame Teiler aus a und b darstellen lässt:

$$Q = \begin{pmatrix} s & t \\ q_{21} & q_{22} \end{pmatrix} \qquad \Rightarrow \qquad \begin{cases} g = sa + tb \\ 0 = q_{21}a + q_{22}b. \end{cases}$$

Beispiel. Wir verifizieren die Behauptung durch Nachrechnen:

$$\begin{aligned} Q &= \begin{pmatrix} 0 & 1 \\ 1 & -q_n \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -q_{n-1} \end{pmatrix}\cdots\begin{pmatrix} 0 & 1 \\ 1 & -q_0 \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -4 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}}\underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}}\underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix}}\begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 1 & -1 \\ -4 & 5 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}}\underbrace{\begin{pmatrix} 1 & -2 \\ -3 & 7 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix}} \\ &= \begin{pmatrix} 3 & -7 \\ -14 & 33 \end{pmatrix}\begin{pmatrix} -3 & 10 \\ 7 & -23 \end{pmatrix} = \begin{pmatrix} -58 & 191 \\ 273 & -899 \end{pmatrix}. \end{aligned}$$

In der zweiten Zeile findet man Zahlen, die a und b zu 0 kombinieren:

$$273 \cdot 76415 - 899 \cdot 23205 = 0,$$

in der ersten stehen die Zahlen $s = -58$ und $t = 191$ und tatsächlich ergibt

$$ta + sb = -58 \cdot 76415 + 191 \cdot 23205 = 85 = g$$

den grössten gemeinsamen Teiler von 76415 und 23205. ○

Die Wirkung der Matrix

$$Q(q) = \begin{pmatrix} 0 & 1 \\ 1 & -q \end{pmatrix}$$

lässt sich mit genau einer Multiplikation und einer Addition berechnen. Dies ist die Art von Matrix, die wir für die Implementation der Wavelet-Transformation anstreben.

4.1.3 Vereinfachte Durchführung

Die Durchführung des euklidischen Algorithmus mit Hilfe der Matrizen $Q(q_k)$ ist etwas unhandlich. In diesem Abschnitt sollen die Matrizenprodukte daher in einer Form dargestellt werden, die leichter als Programm zu implementieren ist.

In Abschnitt 4.1.3 wurde gezeigt, dass das Produkt der aus den Quotienten q_k gebildeten Matrizen $Q(q_k)$ berechnet werden müssen. Dazu beachten wir zunächst, dass die Multiplikation mit der Matrix $Q(q_k)$ die zweite Zeile in die erste Zeile verschiebt:

$$Q(q_k)\begin{pmatrix} u & v \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_k \end{pmatrix}\begin{pmatrix} u & v \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ u - q_k c & v - q_k d \end{pmatrix}.$$

Die Matrizen

$$Q_k = Q(q_k)Q(q_{k-1})\dots Q(q_0)$$

haben daher jeweils für aufeinanderfolgende Werte vo k eine Zeile gemeinsam. Wir bezeichnen die Einträge der ersten Zeile der Matrix Q_k mit c_k und d_k. Es gilt dann

$$Q_k = \begin{pmatrix} c_k & d_k \\ c_{k+1} & d_{k+1} \end{pmatrix} = Q(q_k)\begin{pmatrix} c_{k-1} & d_{k-1} \\ c_k & d_k \end{pmatrix}$$

Daraus ergeben sich die Rekursionsformeln

$$\begin{aligned} c_{k+1} &= c_{k-1} - q_k c_k \\ d_{k+1} &= d_{k-1} - q_k d_k. \end{aligned} \tag{4.2}$$

Die Auswertung des Matrizenproduktes von links nach rechts beginnt mit der Einheitsmatrix, es ist

$$Q_0 = Q(q_0)I = \begin{pmatrix} 0 & 1 \\ 1 & -q_0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

woraus man ablesen kann, dass

$$Q_{-1} = \begin{pmatrix} c_{-1} & d_{-1} \\ c_0 & d_0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.3}$$

gesetzt werden muss.

Mit diesen Notationen kann man den Algorithmus jetzt in der früher verwendeten Tabelle durchführen, die man um die zwei Spalten c_k und d_k hinzufügt und die Werte in dieser Spalte mit Hilfe der Rekursionsformeln (4.2) aus den initialen Werten (4.3) berechnet.

Beispiel. Wir erweitern das Beispiel von Seite 72 zur Bestimmung des grössten gemeinsamen Teilers von 76415 und 23205 zur Berechnung der Koeffizienten c_k und d_k Wir schreiben die gefundenen Zahlen in eine Tabelle:

k	a_k	b_k	q_k	r_k	c_k	d_k
					1	0
0	76415	23205	3	6800	0	1
1	23205	6800	3	2805	1	−3
2	6800	2805	2	1190	−3	10
3	2805	1190	2	425	7	−23
4	1190	425	2	340	−17	56
5	425	340	1	85	41	−135
6	340	85	4	0	−58	191
7	85	0			273	−899

Aus den letzten zwei Spalten der Tabelle kann man ablesen, dass

$$\begin{aligned} -58 \cdot 76415 + 191 \cdot 23205 &= 85 \\ 273 \cdot 76415 - 899 \cdot 23205 &= 0, \end{aligned}$$

wie erwartet. Die gesuchten Zahlen s und t sind also $s = -58$ und $t = 191$. ○

Die Matrizen Q_k kann man auch as der Tabelle ablesen, sie bestehen aus den vier Elementen in den Zeilen k und $k+1$ in den Spalten c_k und d_k. Auf jeder Zeile gilt $b_k = c_k a_0 + d_k b_0$, für $k > 0$ ist dies $c_k a_0 + d_k b_0 = r_{k-1}$.

Bis jetzt gingen wir immer davon aus, dass $a > b$ ist. Dies ist jedoch nicht nötig, wie die Durchführung des Algorithmus für das obige Beispiel mit vertauschten Werten von a und b zeigt. Wir bezeichnen die Elemente zur Unterscheidung von der ursprünglichen Durchführung mit einem Strich:

k	a'_k	b'_k	q'_k	r'_k	c'_k	d'_k
					1	0
0	23205	76415	0	23205	0	1
1	76415	23205	3	6800	1	0
2	23205	6800	3	2805	−3	1
3	6800	2805	2	1190	10	−3
4	2805	1190	2	425	−23	7
5	1190	425	2	340	56	−17
6	425	340	1	85	−135	41
7	340	85	4	0	191	−58
8	85	0			−899	273

Da für $a < b$ der erste Quotient $q'_0 = 0$ ist, werden die ersten neuen Elemente $c'_1 = 1 = d_0$ und $d'_1 = 0 = c_0$ sein. Die nachfolgenden Quotienten sind genau die gleichen, also $q_k = q'_{k+1}$ und damit werden auch

$$c_k = d'_{k+1} \qquad \text{und} \qquad d_k = c'_{k+1}$$

sein. Man findet also die gleichen Einträge in einer Tabelle, die eine Zeile mehr hat und in der die letzten zwei Spalten gegenüber der ursprünglichen Tabelle vertauscht wurden.

4.1.4 Polynome

Der Ring $\mathbb{Q}[X]$ der Polynome in der Variablen X mit rationalen Koeffizienten[1] verhält sich bezüglich Teilbarkeit ganz genau gleich wie die ganzen Zahlen. Insbesondere ist der euklidische Algorithmus genauso wie die Matrixschreibweise auch für Polynome durchführbar.

Beispiel. Wir berechnen als Beispiel den grössten gemeinsamen Teiler der Polynome

$$a = X^4 - 2X^3 - 7X^2 + 8X + 12, \qquad b = X^4 + X^3 - 7X^2 - X + 6.$$

Wir erstellen wieder die Tabelle der Reste

k	a_k	b_k	q_k	r_k
0	$X^4 - 2X^3 - 7X^2 + 8X + 12$	$X^4 + X^3 - 7X^2 - X + 6$	1	$-3X^3 + 9X + 6$
1	$X^4 + X^3 - 7X^2 - X + 6$	$-3X^3 + 9X + 6$	$-\frac{1}{3}X - \frac{1}{3}$	$-4X^2 + 4X + 8$
2	$-3X^3 + 9X + 6$	$-4X^2 + 4X + 8$	$\frac{3}{4}X + \frac{3}{4}$	0

[1] Es kann auch ein beliebiger anderer Körper für die Koeffizienten verwendet werden. Es gelten sogar ähnlich interessante Gesetzmässigkeiten, wenn man für die Koeffizienten ganze Zahlen zulässt. Dann wird das Problem der Faktorisierung allerdings verkompliziert durch das Problem der Teilbarkeit der Koeffizienten. Dieses Problem entfällt, wenn man die Koeffizienten aus einem Bereich wählt, in dem Teilbarkeit kein Problem ist, also in einem Körper.

Daraus kann man ablesen, dass $-4x^2 + 4x + 8$ grösster gemeinsamer Teiler ist. Normiert auf einen führenden Koeffizienten 1 ist dies das Polynom $x^2 - x + 2 = (x + 2)(x - 1)$.

Wir berechnen auch noch die Polynome s und t. Dazu müssen wir die Matrizen $Q(q_k)$ miteinander multiplizieren:

$$\begin{aligned} Q &= Q(q_2)Q(q_1)Q(q_0) \\ &= \begin{pmatrix} 0 & 1 \\ 1 & -\frac{3}{4}(X+1) \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & \frac{1}{3}(X+1) \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{3}(X+1) & -\frac{1}{3}(X-2) \\ -\frac{1}{4}(X^2+2X-3) & \frac{1}{4}(X^2-X-6) \end{pmatrix}. \end{aligned}$$

In der ersten Zeile finden wir die Polynome $t(X)$ und $s(X)$, mit denen

$$\begin{aligned} ta + sb &= \frac{1}{3}(X+1)(X^4 - 2X^3 - 7X^2 + 8X + 12) - \frac{1}{3}(X-2)(X^4 + X^3 - 7X^2 - X + 6) \\ &= -4X^2 + 4X + 8 \end{aligned}$$

und dies ist tatsächlich der gefundene grösste gemeinsame Teiler. Die zweite Zeile von Q gibt uns die Polynomfaktoren, mit denen a und b gleich werden:

$$\begin{aligned} q_{21}a + q_{22}b &= -\frac{1}{4}(X^2+2X-3)(X^4 - 2X^3 - 7X^2 + 8X + 12) + \frac{1}{4}(X^2-X-6)(X^4 + X^3 - 7X^2 - X + 6) \\ &= 0. \end{aligned}$$

○

Man kann natürlich den grössten gemeinsamen Teiler auch mit Hilfe einer Faktorisierung der Polynome a und b finden:

Faktorisierung von a:	$a =$	$(X-3)(X-2)$	$(X+1)(X+2)$		
Faktorisierung von b:	$b =$	$(X-2)(X-1)(X+1)$		$(X+3)$	
gemeinsame Faktoren:	$g =$	$(X-2)$	$(X+1)$		$= X^2 - X + 2$
	$v = a/g =$	$(X-3)$		$(X+2)$	$= X^2 - X - 6$
	$u = b/g =$		$(X-1)$	$(X+3)$	$= X^2 + 2X - 3$

Aus den letzten zwei Zeilen folgt $ua - vb = ab/g - ab/g = 0$, wie erwartet.

4.1.5 Das kleinste gemeinsame Vielfache

Das kleinste gemeinsame Vielfache zweier Zahlen a und b ist

$$\mathrm{kgV}(a,b) = \frac{ab}{\mathrm{ggT}(a,b)}.$$

Wir suchen nach einen Algorithmus, mit dem man das kleinste gemeinsame Vielfache effizient berechnen kann.

Die Zahlen a und b sind beide Vielfache des grössten gemeinsamen Teilers $g = \mathrm{ggT}(a,b)$, es gibt also Zahlen u und v derart, dass $a = ug$ und $b = vg$. Wenn t ein gemeinsamer Teiler von u und v ist, dann ist tg ein grösserer gemeinsamer Teiler von a und b. Dies kann nicht sein, also müssen u und v teilerfremd sein. Das kleinste gemeinsame Vielfache von a und b ist dann $ugv = av = ub$. Die

Bestimmung des kleinsten gemeinsamen Vielfachen ist also gleichbedeutend mit der Bestimmung der Zahlen u und v.

Die definierende Eigenschaften von u und v kann man in Matrixform als

$$\begin{pmatrix} a \\ b \end{pmatrix} = \underbrace{\begin{pmatrix} u & ? \\ v & ? \end{pmatrix}}_{=K} \begin{pmatrix} \mathrm{ggT}(a,b) \\ 0 \end{pmatrix} \tag{4.4}$$

geschrieben werden, wobei wir die Matrixelemente ? nicht kennen. Diese Elemente müssen wir auch nicht kennen, um u und v zu bestimmen.

Bei der Bestimmung des grössten gemeinsamen Teilers wurde der Vektor auf der rechten Seite von (4.4) bereits gefunden. Die Matrizen $Q(q_i)$, die die einzelne Schritte des euklidischen Algorithmus beschreiben, ergeben ihn als

$$\begin{pmatrix} \mathrm{ggT}(a,b) \\ 0 \end{pmatrix} = Q(q_n)Q(q_{n-1})\dots Q(q_1)Q(q_0)\begin{pmatrix} a \\ b \end{pmatrix}.$$

Indem wir die Matrizen $Q(q_n)$ bis $Q(q_0)$ auf die linke Seite der Gleichung schaffen, erhalten wir

$$\begin{pmatrix} a \\ b \end{pmatrix} = Q(q_0)^{-1}Q(q_1)^{-1}\dots Q(q_{n-1})^{-1}Q(q_n)^{-1}\begin{pmatrix} \mathrm{ggT}(a,b) \\ 0 \end{pmatrix}.$$

Eine mögliche Lösung für die Matrix K in (4.4) ist der die Matrix

$$K = Q(q_0)^{-1}Q(q_1)^{-1}\dots Q(q_{n-1})^{-1}Q(q_n).$$

Insbesondere ist die Matrix K die Inverse der früher gefundenen Matrix Q.

Die Berechnung der Matrix K als Inverse von Q ist nicht sehr effizient. Genauso wie es möglich war, das Produkt Q der Matrizen $Q(q_k)$ iterativ zu bestimmen, muss es auch eine Rekursionsformel für das Produkt der inversen Matrizen $Q(q_k)^{-1}$ geben.

Schreiben wir die gesuchte Matrix

$$K_k = Q(q_0)^{-1}\dots Q(q_{k-1})^{-1} = \begin{pmatrix} e_k & e_{k-1} \\ f_k & f_{k-1} \end{pmatrix},$$

dann kann man K_k durch die Rekursion

$$K_{k+1} = K_kQ(q_k)^{-1} = K_kK(q_k) \qquad \text{mit} \qquad K_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I \tag{4.5}$$

berechnen. Die Inverse von $Q(q)$ ist

$$K(q) = Q(q)^{-1} = \frac{1}{\det Q(q)}\begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} \quad \text{denn} \quad K(q)Q(q) = \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -q \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Da die zweite Spalte von $K(q)$ die erste Spalte einer Einheitsmatrix ist, wird die zweite Spalte des Produktes $AK(q)$ immer die erste Spalte von A sein. In K_{k+1} ist daher nur die erste Spalte neu, die zweite Spalte ist die erste Spalte von K_k.

Aus der Rekursionsformel (4.5) für die Matrizen K_k kann man jetzt eine Rekursionsbeziehung für die Folgen e_k und f_k ablesen, es gilt

$$e_{k+1} = q_ke_k + e_{k-1}$$

$$f_{k+1} = q_k f_k + f_{k-1}$$

für $k = 0, 1, \ldots, n$. Damit können e_k und f_k gleichzeitig mit den Zahlen c_k und d_k in einer Tabelle berechnen.

Beispiel. Wir erweitern das Beispiel von Seite 75 um die beiden Spalten zur Berechnung von e_k und f_k:

k	a_k	b_k	q_k	r_k	c_k	d_k	e_k	f_k
					1	0	0	1
0	76415	23205	3	6800	0	1	1	0
1	23205	6800	3	2805	1	−3	3	1
2	6800	2805	2	1190	−3	10	10	3
3	2805	1190	2	425	7	−23	23	7
4	1190	425	2	340	−17	56	56	17
5	425	340	1	85	41	−135	135	41
6	340	85	4	0	−58	191	191	58
7	85	0			273	−899	899	273

Der grösste gemeinsame Teiler ist $\mathrm{ggT}(a, b) = 85$. Aus der letzten Zeile der Tabelle kann man jetzt die Zahlen $u = e_7 = 899$ und $v = f_7 = 273$ ablesen, und tatsächlich ist

$$a = 76415 = 899 \cdot 85 \qquad \text{und} \qquad b = 23205 = 273 \cdot 85.$$

Daraus kann man dann auch das kleinste gemeinsame Vielfache ablesen, es ist

$$\mathrm{kgV}(a, b) = \mathrm{kgV}(76415, 23205) = \left\{\begin{matrix} ub = 899 \cdot 23205 \\ va = 273 \cdot 76415 \end{matrix}\right\} = 20861295. \qquad \bigcirc$$

Der erweiterte Algorithmus kann auch dazu verwendet werden, das kleinste gemeinsame Vielfache zweier Polynome zu berechnen. Dies wird zum Beispiel bei der Decodierung des Reed-Solomon-Codes in Kapitel 15 verwendet.

Polynome

Im Beispiel auf Seite 76 wird der grösste gemeinsame Teiler der Polynome

$$a = X^4 - 2X^3 - 7X^2 + 8X + 12, \qquad b = X^4 + X^3 - 7X^2 - X + 6$$

berechnet. Dies kann jetzt erweitert werden für die Berechnung des kleinsten gemeinsamen Vielfachen.

Beispiel. Die Berechnungstabelle nur für die Spalten e_k und f_k ergibt

k	q_k	e_k	f_k
		0	1
0	1	1	0
1	$-\frac{1}{3}X - \frac{1}{3}$	1	1
2	$\frac{3}{4}X + \frac{3}{4}$	$-\frac{1}{3}X + \frac{2}{3}$	$-\frac{1}{3}X - \frac{1}{3}$
		$-\frac{1}{4}X^2 + \frac{1}{4}X + \frac{3}{2}$	$-\frac{1}{4}X^2 - \frac{1}{2}X + \frac{3}{4}$

Daraus kann man ablesen, dass

$$u = -\frac{1}{4}X^2 + \frac{1}{4}X + \frac{3}{2} \qquad \text{und} \qquad v = -\frac{1}{4}X^2 - \frac{1}{2}X + \frac{3}{4}.$$

Daraus ergibt sich das kleinste gemeinsame Vielfache auf zwei verschiedene Weisen:

$$\operatorname{ggT}(a,b) = \left\{ \begin{matrix} (-\frac{1}{4}X^2 + \frac{1}{4}X + \frac{3}{2}) \cdot (X^4 - 2X^3 - 7X^2 + 8X + 12) \\ (-\frac{1}{4}X^2 - \frac{1}{2}X + \frac{3}{4}) \cdot (X^4 + X^3 - 7X^2 - X + 6) \end{matrix} \right\} = -\frac{1}{4}X^6 + \frac{7}{2}X^4 - \frac{49}{4}X^2 + 9.$$

Die beiden Berechnungsmöglichkeiten stimmen wie erwartet überein. ○

Anwendung: Decodierung des Reed-Solomon-Codes

Der Reed-Solomon-Code verwendet Polynome zur Codierung der Daten, dies wird in Kapitel 15 im Detail beschrieben. Bei der Decodierung muss der Faktor u für zwei gegebene Polynome $n(X)$ und $r(X)$ bestimmt werden. Allerdings ist das Polynom $r(X)$ nicht vollständig bekannt, nur die ersten paar Koeffizienten sind gegeben. Dafür weiss man zusätzlich, wieviele Schritte genau der Euklidische Algorithmus braucht. Daraus lässt sich genügend Information gewinnen, um die Faktoren u und v zu bestimmen. Das Video `https://youtu.be/uOLW43OIZJ0` von Edmund Weitz erklärt die Theorie hinter dieser Teilaufgabe anhand von Beispielen.

Beispiel. Wir berechnen also die Faktoren u und v für die beiden Polynome

$$\begin{aligned} n(X) &= X^{12} + 12 \\ r(X) &= 7X^{11} + 4X^{10} + X^9 + 12X^8 + 2X^7 + 12X^6 + w(X) \end{aligned}$$

in $\mathbb{F}_{13}[X]$, wobei $w(X)$ ein unbekanntes Polynom vom Grad 5 ist. Man weiss zusätzlich noch, dass der euklidische Algorithmus genau drei Schritte braucht, es gibt also genau drei Quotienten, die in die Berechnung der Zahlen e_k und f_k einfliessen.

Im ersten Schritt des euklidischen Algorithmus ist der Quotient $n(X)/r(X)$ zu bestimmen, der Grad 1 haben muss.

$$\begin{aligned} a_0 &= n(X) = X^{12} + 12 \\ b_0 &= r(X) = 7X^{11} + 4X^{10} + X^9 + 12X^8 + 2X^7 + 12X^6 + \ldots \\ q_0 &= 2X + 10 \\ r_0 = a_0 - b_0 \cdot q_0 &= 10X^{10} + 5X^9 + 6X^8 + 8X^7 + \ldots \\ a_1 &= 7X^{11} + 4X^{10} + X^9 + 12X^8 + 2X^7 + 12X^6 + \ldots \\ b_1 &= 10X^{10} + 5X^9 + 6X^8 + 8X^7 + \ldots \\ q_1 &= 2X + 2 \\ r_1 = a_1 - b_1 q_1 &= 5X^9 + 10X^8 + \ldots \\ a_2 &= 10X^{10} + 5X^9 + 6X^8 + 8X^7 + \ldots \\ b_2 &= 5X^9 + 10X^8 + \ldots \\ q_2 &= 2X + 10 \end{aligned}$$

Aus den Polynomen q_k können jetzt die Faktoren u und v bestimmt werden:

k	q_k	e_k	f_k
		0	1
0	$2X + 10$	1	0
1	$2X + 2$	$2X + 10$	1
2	$2X + 10$	$4X^2 + 11X + 8$	$2X + 2$
		$8X^3 + 10X^2 + 11X + 12$	$4X^2 + 11X + 8$

Die Faktorisierung des Polynoms

$$u = 8X^3 + 10X^2 + 11X + 12$$

kann bestimmt werden, indem man alle Zahlen $1, 2, \ldots, 12 \in \mathbb{F}_{13}$ einsetzt. Man findet so die Nullstellen 3, 4 und 8, also muss das Polynom u faktorisiert werden können als

$$u = 8(X - 3)(X - 4)(X - 8) = 8X^3 - 120X^2 + 544X - 768 = 8X^3 + 10X^2 + 11X + 12. \qquad \bigcirc$$

4.2 Galois-Körper

Ein Körper $\Bbbk$ enthält mindestens die Zahlen 0 und 1. Die Null ist nötig, damit $\Bbbk$ eine Gruppe bezüglich der Addition ist, die immer ein neutrales Element, geschrieben 0 enthält. Die Eins ist nötig, damit $\Bbbk^* = \Bbbk \setminus \{0\}$ eine Gruppe bezüglich der Multiplikation ist, die immer eine neutrales Element, geschrieben 1 enthält. Durch wiederholte Addition entstehen auch die Zahlen $2 = 1 + 1$, $3 = 2 + 1$ und so weiter. Es sieht also so aus, als ob ein Körper immer unendliche viele Elemente enthalten müsste. Wie können also endliche Körper entstehen?

In diesem Abschnitt sollen die sogenannten Galois-Körper $\mathbb{F}_p$ mit genau p Elementen konstruiert werden, die es für jede Primzahl p gibt. Sie sind die Basis für weitere endliche Körper, die eine beliebige Primzahlpotenz p^n von Elementen haben und die die Basis wichtiger kryptographischer Algorithmen sind.

4.2.1 Arithmetik modulo p

Damit aus den Zahlen $0, 1, 2, \ldots$ ein endlicher Körper werden kann, muss die Folge sich wiederholen. Schreiben wir $a_0 = 0, a_1 = 1, \ldots$ für die Folge, dann muss es also ein Folgenelement a_k geben und ein n derart, dass $a_{k+n} = a_k$. Dies bedeutet, dass $k + n = k$ sein muss. Subtrahiert man k auf beiden Seiten, dann folgt, dass $n = 0$ sein muss. Damit ein endlicher Körper entsteht, muss also die Menge

$$\{0, 1, 2, \ldots, n - 1\}$$

eine Gruppe bezüglich der Addition sein, und

$$\{1, 2, \ldots, n - 1\}$$

eine Gruppe bezüglich der Multiplikation.

Restklassenring

Wir definieren die Grundoperationen in einer Menge, die mit den Zahlen $\{0, 1, 2, \ldots, n - 1\}$ identifiziert werden kann.

Definition 4.3. *Die Zahlen $a, b \in \mathbb{Z}$ heissen* kongruent modulo *n, geschrieben*

$$a \equiv b \mod n,$$

wenn $a - b$ durch n teilbar ist, also $n|(a - b)$.

Die Zahlen mit gleichem Rest sind Äquivalenzklassen der Kongruenz modulo n. Die Zahlen mit Rest k modulo n bilden die *Restklasse*

$$[\![k]\!] = \{\dots, k - 2n, k - n, k, k + n, k + 2n, \dots\} \subset \mathbb{Z}.$$

Sie bilden eine endliche Menge, die man mit den Resten $0, 1, \dots, n - 1$ identifizieren kann.

Definition 4.4. *Die Menge $\mathbb{Z}/n\mathbb{Z}$ besteht aus den Restklassen $[\![0]\!], [\![1]\!], \dots, [\![n-1]\!]$, die auch einfach $0, 1, \dots, n - 1$ geschrieben werden.*

Beim Rechnen mit Resten modulo n können Vielfache von n ignoriert werden. Zum Beispiel gilt

$$\begin{aligned} 48 \equiv -1 \mod 7 \qquad & 48 = -1 \quad \text{in } \mathbb{Z}/7\mathbb{Z} \\ 3 \cdot 5 = 15 \equiv 1 \mod 7 \qquad & 3 \cdot 5 = 1 \quad \text{in } \mathbb{Z}/7\mathbb{Z}. \end{aligned}$$

Das Beispiel zeigt, dass man mindestens in $\mathbb{Z}/7\mathbb{Z}$ mit Resten ganz ähnlich rechnen kann wie in $\mathbb{Q}$. In $\mathbb{Z}/7\mathbb{Z}$ scheinen 3 und 5 multiplikative inverse zu sein.

Tatsächlich kann man auf den Restklassen eine Ringstruktur definieren. Dazu muss man sicherstellen, dass die Auswahl eines Repräsentanten keinen Einfluss auf den Rest hat. Der Rest a kann jede Zahl der Form $a + kn$ darstellen. Ebenso kann der Rest b jede zahl der Form $b + ln$ darstellen. Deren Summe ist $a + b + (k + l)n \equiv a + b \mod n$. Der Repräsentant des Restes hat also keinen Einfluss auf die Summe.

Ebenso ist das Produkt der beiden Repräsentaten $(a + kn) \cdot (b + ln) = ab + (al + bk)n + kln^2 = ab + (al + bk + kln)n \equiv ab \mod n$ für jede Wahl von k und l. Auch die Multiplikation ist also unabhängig vom gewählten Repräsentanten.

Definition 4.5. *Die Menge $\mathbb{Z}/n\mathbb{Z}$ ist ein Ring, heisst der* Restklassenring modulo *n.*

Division in $\mathbb{Z}/n\mathbb{Z}$

Um einen endlichen Körper zu erhalten, muss die Menge

$$\mathbb{Z}/n\mathbb{Z} \setminus \{[\![0]\!]\} = \{[\![1]\!], [\![2]\!], \dots [\![n - q]\!]\}$$

eine Gruppe bezüglich der Multiplikation sein. Insbesondere darf kein Produkt $a \cdot b$ mit Faktoren in $\mathbb{Z}/n\mathbb{Z} \setminus \{[\![0]\!]\}$ zu Null werden. Für $n = 15$ funktioniert dies nicht, das Produkt $3 \cdot 5 \equiv 0 \mod 15$. Man nennt von Null verschiedene Faktoren, deren Produkt Null ist, einen *Nullteiler*. Falls sich $n = p_1 \cdot p_2$ in zwei Faktoren zerlegen lässt, dann sind p_1 und p_2 Nullteiler in $\mathbb{Z}/n\mathbb{Z}$. Ein Körper kann also nur entstehen, wenn n eine Primzahl ist.

Definition 4.6. *Ist p eine Primzahl, dann heisst $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ der Galois-Körper der Ordnung p.*

Diese Definition ist nur gerechtfertigt, wenn $\mathbb{F}_p^*$ tatsächlich eine Gruppe ist, wenn also jede Zahl zwischen 1 und $p - 1$ ein Inverses bezüglich der Multiplikation hat. Zu einem Rest $a \in \mathbb{F}_p^*$ muss also

ein Rest b gefunden werden, so dass $ab \equiv 1 \mod p$. Dies ist gleichbedeutend mit Zahlen b und n derart, dass

$$ab + np = 1. \tag{4.6}$$

In (4.6) sind a und p gegeben, gesucht sind b und n.

In Abschnitt 4.1 wurde gezeigt, wie der euklidische Algorithmus eine Gleichung der Form (4.6) lösen kann, wenn die beiden gegebenen Zahlen a und p teilerfremd sind. Dies ist aber dadurch garantiert, dass p eine Primzahl ist und $1 \leq a < p$. Die multiplikative Inverse von a in $\mathbb{F}_p^*$ kann also mit Hilfe des euklidischen Algorithmus effizient gefunden werden.

Beispiel. Die kleinste Primzahl grösser als 2021 ist $p = 2063$. Was ist die Inverse von 2021 in $\mathbb{F}_{2063}$?

Wir führen den euklidischen Algorithmus für das Paar $(2063, 2021)$ durch und erhalten

k	a_k	b_k	q_k	r_k
0	2063	2021	1	42
1	2021	42	48	5
2	42	5	8	2
3	5	2	2	1
4	2	1	2	0

Die gesuchten Faktoren b und n können aus dem Matrizenprodukt $Q(q_n) \dots Q(q_0)$ gefunden werden:

$$\begin{aligned}
Q &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -8 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -48 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -8 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -48 & 49 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} -48 & 49 \\ 385 & -393 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 385 & -393 \\ -818 & 835 \end{pmatrix} \\
&= \begin{pmatrix} -818 & 835 \\ 2021 & -2063 \end{pmatrix}
\end{aligned}$$

Daraus können wir ablesen, dass

$$-818 \cdot 2021 + 835 \cdot 2063 = 1.$$

Der Rest $-818 \equiv 1245 \mod 2063$ ist also die multiplikative Inverse von 2021 in $\mathbb{F}_{2063}$. ○

Der kleine Satz von Fermat

In $\mathbb{Z}$ wachsen die Potenzen einer Zahl immer weiter an. In einem endlichen Körper kann dies nicht gelten, da nur endlich viele Werte zur Verfügung stehen. Tatsächlich müssen die Potenzen einer von 0 verschiedenen Zahl $a \in \mathbb{F}_p^*$ alle in $\mathbb{F}_p^*$ liegen. Es gibt aber nur $p - 1$ Zahlen in $\mathbb{F}_p^*$, spätestens die Potenz mit Exponent p muss also mit einer früheren Potenz übereinstimmen. Der kleine Satz von Fermat sagt etwas genauer: die p-te Potenz von a ist genau die Zahl a:

Satz 4.7 (Kleiner Satz von Fermat)**.** *In $\mathbb{F}_p$ gilt $a^p = a$ für alle $a \in \mathbb{F}_p^*$.*

Wir beweisen diesen Satz in der folgenden, traditionelleren Formulierung.

Satz 4.8. *Für jede ganze Zahl $a > 0$ gilt $p|(a^p - a)$ genau dann, wenn p eine Primzahl ist.*

Beweis. Wir müssen zeigen, dass p ein Teiler ist von $a^p - a$. Das nachfolgende kombinatorische Argument wird zum Beispiel von Mathologor auf seinem Youtube-Kanal im Video `https://youtu.be/_9fbBSxhkuA` illustriert.

Zum Beiweis interpretieren wir die vorkommenden Zahlen kombinatorisch. Die Zahl a^p ist die Anzahl der verschiedenen Perlenketten der Länge p, die sich aus Glasperlen mit a verschiedenen Farben herstellen lassen. Davon bestehen a Perlenketten aus nur einer einzigen Farbe. Die Zahl $a^p - a$ ist also die Anzahl der Perlenketten der Länge p aus Glasperlen mit a verschiedenen Farben, die mindestens zwei verschiedene Farben verwenden.

Wir stellen jetzt die Frage nach der Anzahl der geschlossenen Perlenketten der Länge p als Glasperlen in a verschiedenen Farben. Aus jeder geschlossenen Perlenkette lassen sich p Perlenketten machen, indem man sie an einer der p Trennstellen zwischen Perlen aufteilt.

Wir müssen uns noch überlegen, unter welchen Voraussetzungen alle diese möglichen Auftrennungen zu verschiedenen Perlenketten führen. Zwei Trennstellen, die k-Perlen auseinander liegen, führen nur dann zur gleichen Perlenkette, wenn die geschlossenen Ketten durch Drehung um k Perlen ineinander übergehen. Dies bedeutet aber auch, dass sich das Farbmuster alle k-Perlen wiederholen muss. Folglich ist k ein Teiler von p. p verschiedene Perlenketten entstehen also immer genau dann, wenn p eine Primzahl ist.

Wir schliessen daraus, dass $a^p - a$ durch p teilbar ist, genau dann, wenn p eine Primzahl ist. □

Der kleine Satz von Fermat kann auch dazu verwendet werden, Potenzen in $\mathbb{F}_p$ zu vereinfachen, wie das folgende Beispiel[2] zeigt.

Beispiel. Man berechnet in $\mathbb{F}_{13}$ die Potenz 11^{666}. Nach dem kleinen Satz von Fermat ist $11^{13} = 11$ oder $11^{12} = 1$, man kann also den Exponenten modulo 12 reduzieren. Weil $666 = 55 \cdot 12 + 6$ erhält man $11^{666} = 11^6$. Da die Potenzen von 11 etwas mühsam zu berechnen sind, kann man sie wegen $11 = -2$ in $\mathbb{F}_{13}$ auch als Potenzen von -2 bekommen. Aber $(-2)^6 = 64 = -1 \in \mathbb{F}_{13}$. ○

In der Form $a^{p-1} = 1$ in $\mathbb{F}_p$ liefert der kleine Satz von Fermat die Inverse von a als a^{p-2}. Dies bedeutet zum Beispiel, dass in $\mathbb{F}_3$ jede von 0 verschiedene Zahl zu sich selbst invers ist: $1 \cdot 1 = 1$ und $2 \cdot 2 = 1$. Diese Art, die Inverse zu bestimmen, ist allerdings nicht effizienter als der euklidische Algorithmus, aber sie ist manchmal für theoretische Überlegungen nützlich.

Der Satz von Wilson

Der Satz von Wilson ermöglicht, die multiplikative Inverse auf eine andere Art zu berechnen. Sie ist zwar nicht unbedingt einfacher, aber manchmal nützlich für theoretische Überlegungen.

Satz 4.9 (Wilson)**.** *Die ganze Zahl $p \geq 2$ ist genau dann eine Primzahl, wenn $(p-1)! \equiv -1 \mod p$.*

Beweis. Wenn p keine Primzahl ist, dann lässt sich p in Faktoren $p = n_1 \cdot n_2 = p$ zerlegen. Beide Faktoren kommen in der Liste $1, 2, \ldots, p-1$ vor. Insbesondere haben $p = n_1 n_2$ und $(p-1)!$ mindestens einen der Faktoren n_1 oder n_2 gemeinsam, wir können annehmen, dass n_1 dieser Faktor ist. Es folgt, dass der grösste gemeinsame Teiler von p und $(p-1)!$ grösser als n_1 ist, auch $(p-1)!$ ein Vielfaches von n_1 in $\mathbb{F}_p$. Insbesondere kann $(p-1)!$ nicht $-1 \in \mathbb{F}_p$ sein.

[2] Das Beispiel stammt aus dem Video `https://youtu.be/_9fbBSxhkuA`, welches Mathologer zu Halloween 2018 veröffentlich hat

Ist andererseits p eine Primzahl, dann sind die Zahlen $1, 2, \ldots, p-1$ alle invertierbar in $\mathbb{F}_p$. Die Zahlen 1 und $-1 \equiv p-1 \mod p$ sind zu sich selbst invers, da $1 \cdot 1 = 1$ und $(-1) \cdot (-1) = 1$. Wenn eine Zahl a zu sich selbst invers ist in $\mathbb{F}_p$, dann ist $a^2 - 1 = 0$ in $\mathbb{F}_p$. Daher ist auch $(a+1)(a-1) = 0$, in $\mathbb{F}_p$ muss daher einer der Faktoren 0 sein, also $a = -1$ oder $a = 1$ in $\mathbb{F}_p$.

Zu jeder Zahl $a \in \{2, \ldots, p-2\}$ liegt die Inverse a^{-1} ebenfalls in diesen Bereich und ist verschieden von a: $a^{-1} \neq a$. Das Produkt der Zahlen $2 \cdot 3 \cdot \ldots \cdot (p-2)$ besteht also aus zueinander inversen Paaren. Es folgt

$$2 \cdot 3 \cdot \ldots \cdot (p-2) = 1.$$

Multipliziert man dies mit $p - 1 = -1 \in \mathbb{F}_p$, folgt die Behauptung des Satzes. □

Mit dem Satz von Wilson kann man die Inverse einer beliebigen Zahl $a \in \mathbb{F}_p$ finden. Dazu verwendet man, dass a einer der Faktoren in $(p-1)!$ ist. Lässt man diesen Faktor weg, erhält man eine Zahl

$$b = 1 \cdot 2 \cdot \ldots \cdot \hat{a} \cdot \ldots \cdot (p-1),$$

wobei der Hut bedeutet, dass der Faktor a weggelassen werden soll. Nach dem Satz von Wilson ist $ab = -1$ in $\mathbb{F}_p$, also ist $-b$ die multiplikative Inverse von a.

Beispiel. Die Inverse von $2 \in \mathbb{F}_7$ ist

$$\begin{aligned} a^{-1} &= -\underbrace{1 \cdot 3 \cdot 4}_{} \cdot \underbrace{5 \cdot 6}_{} \\ &= -5 \cdot 2 = -3 = 4 \end{aligned}$$

Tatsächlich ist $2 \cdot 4 = 8 \equiv 1 \mod 7$. ○

4.2.2 Charakteristik

In diesem Abschnitt zeigen wir, dass jeder Körper $\mathbb{k}$ eine Erweiterung entweder von $\mathbb{Q}$ oder eines endlichen Körpers $\mathbb{F}_p$ ist.

Primkörper

Sei $\mathbb{k}$ ein Körper. Er enthält mindestens die Zahlen 0 und 1 und alle Vielfachen davon. Wenn alle Vielfachen in $\mathbb{k}$ von 0 verschieden sind, dann bilden Sie ein Bild der ganzen Zahlen $\mathbb{Z} \subset \mathbb{k}$. Damit müssen dann aber auch alle Brüche in $\mathbb{k}$ enhalten sein, es folgt also, dass $\mathbb{Q} \subset \mathbb{k}$ sein muss.

Wenn andererseits eines der Vielfachen von 1 in $\mathbb{k}$ verschwindet, dann wissen wir aus Abschnitt 4.2.1, dass der Körper $\mathbb{F}_p$ in $\mathbb{k}$ enthalten sein muss. Dies ist der kleinste Teilkörper, der in $\mathbb{k}$ enthalten ist.

Definition 4.10. *Der kleinste Teilkörper eines Körpers $\mathbb{k}$ heisst der* Primkörper *von $\mathbb{k}$.*

Der Primkörper erlaubt jetzt, die Charakteristik eines Körpers $\mathbb{k}$ zu definieren.

Definition 4.11. *Die Charakteristik eines Körpers $\mathbb{k}$ ist p, wenn der Primkörper $\mathbb{F}_p$ ist. Falls der Primkörper $\mathbb{Q}$ ist, ist die Charakteristik* 0*.*

Die Charakteristik hat wichtige Auswirkungen darauf, wie in einem Körper gerechnet wird. Endliche Körper enthalten immer einen Körper von Primzahl-Ordnung und haben damit immer Primcharakteristik. Ein Körper mit Charakteristik 0 enthält immer unendliche viele Elemente.

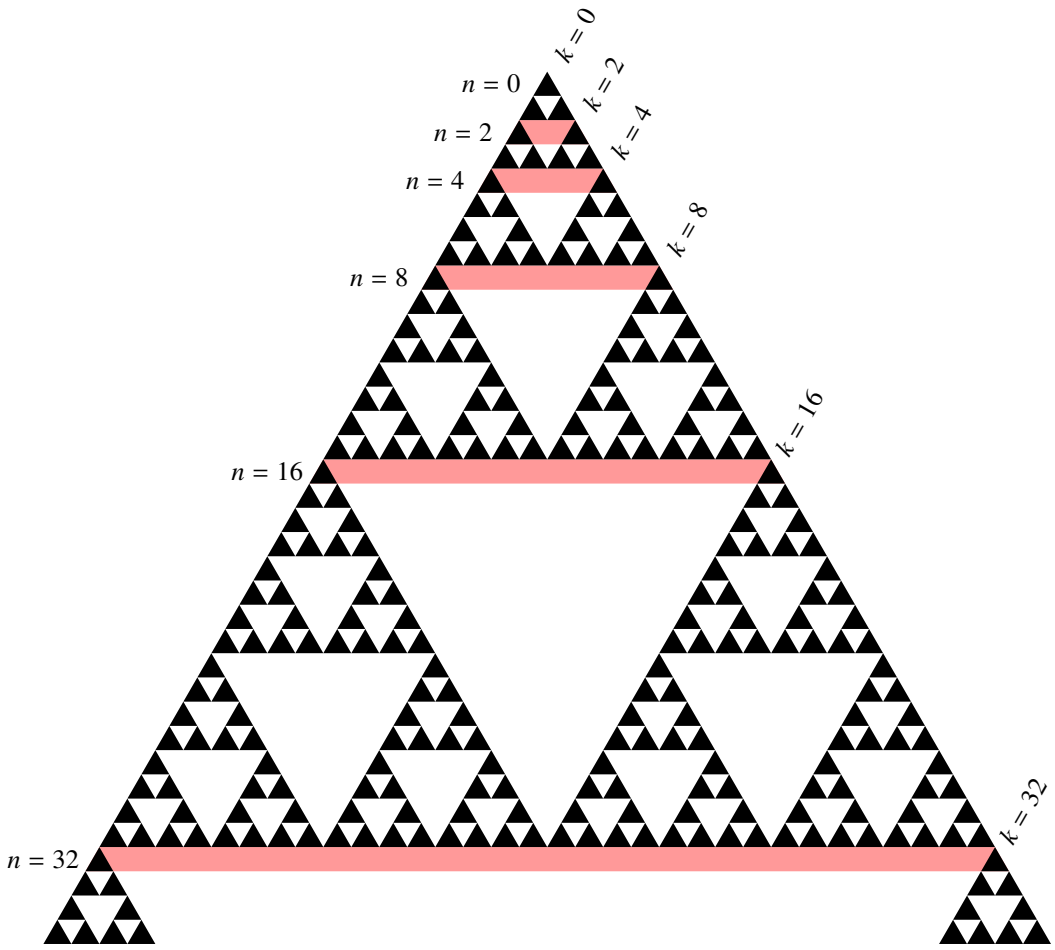


Abbildung 4.1: Binomialkoeffizienten modulo 2 im Pascal-Dreieck. Auf den rot hinterlegten Zeilen, die zu Exponenten der Form 2^k gehören, sind alle Koeffizienten ausser dem ersten und letzten durch 2 teilbar.

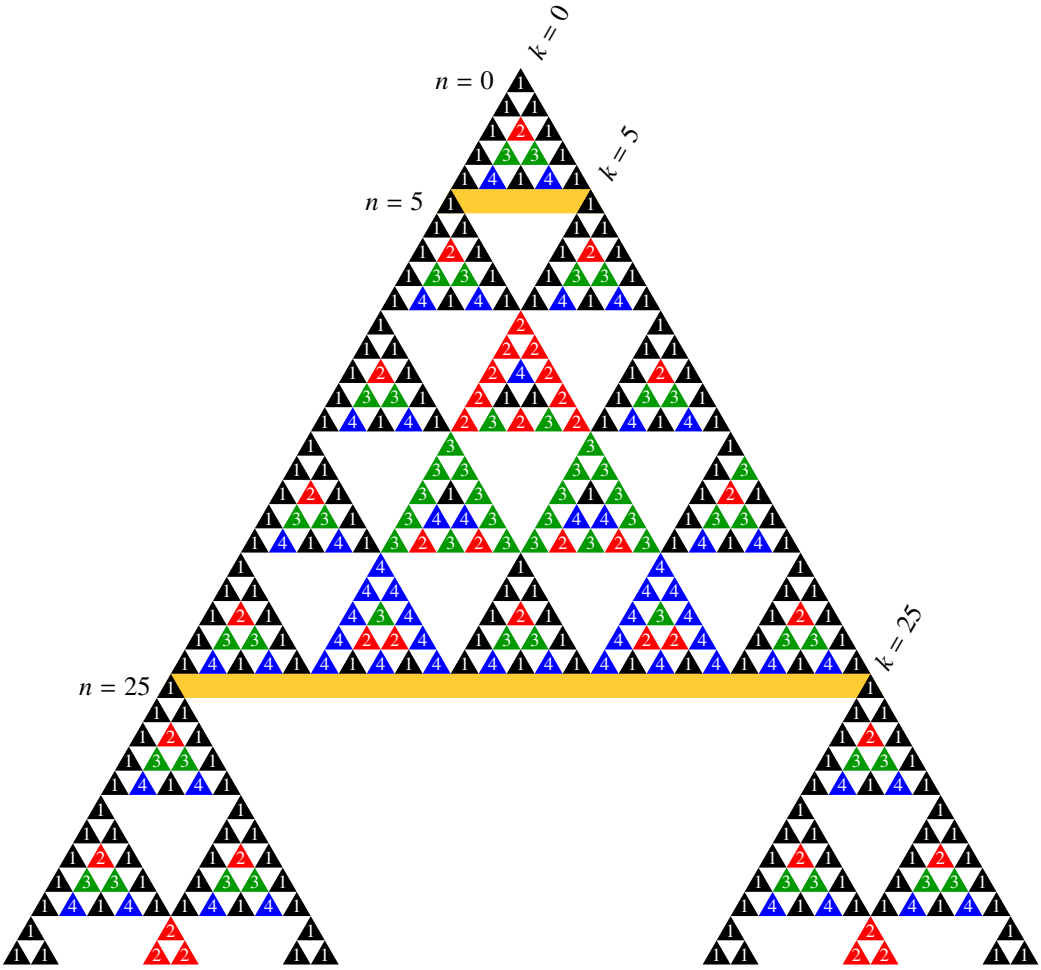


Abbildung 4.2: Binomialkoeffizienten modulo 5 im Pascal-Dreieck. Die von 0 verschiedenen Reste werden durch Farben dargestellt: 1 = schwarz, 2 = rot, 3 = grün, 4 = blau. Auf den gelb hinterlegten Zeilen, die zu Exponenten der Form 5^k gehören, sind alle Koeffizienten ausser dem ersten und letzten durch 5 teilbar.

Teilbarkeit von Binomialkoeffizienten

Die Abbildung 4.1 zeigt den Rest bei Teilung durch 2 der Binomialkoeffizienten. Man kann daraus ablesen, dass $\binom{n}{m} \equiv 0 \mod 2$ für $n = 2^k$ und $0 < m < n$. Abbildung 4.2 zeigt das Pascal-Dreieck auch noch für $p = 5$. Hier ist auch schön die Selbstähnlichkeit des Pascal-Dreiecks erkennbar. Ersetzt man die "5er-Dreiecke" durch ein volles Dreieck mit der Farbe des kleinen Dreiecks an seiner Spitze, entsteht wieder das ursprüngliche Pascal-Dreieck. Dabei gehen die Zeilen aus lauter Nullen ausser an den Enden ineinander über.

Satz 4.12. *Sei p eine Primzahl, dann ist*

$$\binom{p}{m} \equiv 0 \mod p$$

für $0 < m < n$.

Beweis. Für den Binomialkoeffizienten gilt

$$\binom{p}{m} = \frac{p \cdot (p-1) \cdot (p-2) \cdot \ldots \cdot (p-m+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot m}.$$

Für $m < p$ kann keiner der Faktoren im Nenner p sein, der Faktor p im Zähler kann also nicht weggekürzt werden, so dass der Binomialkoeffizient durch p teilbar sein muss. □

Satz 4.13. *Sei p eine Primzahl, dann ist*

$$\binom{p^k}{m} \equiv 0 \mod p \tag{4.7}$$

für $0 < m < p^k$

Beweis. Wir wissen aus Satz 4.12, dass

$$(a+b)^p = a^p + b^p. \tag{4.8}$$

Wir müssen zeigen, dass $(a+b)^{p^k} = a^{p^k} + b^{p^k}$ gilt. Wir verwenden vollständige Induktion, (4.8) ist die Induktionsverankerung. Wir nehmen jetzt im Sinne der Induktionsannahme an, dass (4.7) für ein bestimmtes k gilt. Dann ist

$$(a+b)^{p^{k+1}} = (a+b)^{p^k \cdot p} = ((a+b)^{p^k})^p = (a^{p^k} + b^{p^k})^p = a^{p^k \cdot p} + b^{p^k \cdot p} = a^{p^{k+1}} + b^{p^{k+1}},$$

also die Behauptung für $k+1$. Damit ist (4.7) für alle k bewiesen. □

Die Aussage von Satz 4.13 kann man auch im Körper $\mathbb{F}_p$ formulieren:

Satz 4.14. *In $\mathbb{F}_p$ gilt*

$$\binom{p^k}{m} = 0$$

für beliebige $k > 0$ und $0 < m < p^k$.

Frobenius-Automorphismus

Die Abbildung $x \mapsto x^n$ ist weit davon entfernt, sich mit den algebraischen Strukturen zu vertragen. Zum Beispiel kann man nicht erwarten, dass $(a+b)^n = a^n + b^n$, denn nach der binomischen Formel

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = a^n + \binom{n}{1} a^{n-1} b + \cdots + \binom{n}{n-1} a b^{n-1} + b^n \tag{4.9}$$

gibt es zwischen den Termen an den Enden des Ausdrucks noch viele Zwischenterme, die normalerweise nicht verschwinden.

Ganz anders sieht die Situation aus, wenn $n = p$ ist. Nach Satz 4.14 verschwinden die Binomialkoeffizienten der Zwischenterme der Summe (4.9) als Elemente von $\mathbb{F}_p$. Daher gilt

Satz 4.15 (Frobenius-Automorphismus)**.** *In einem Körper $\Bbbk$ der Charakteristik p ist die Abbildung $x \mapsto x^p$ ist ein Automorphismus, der den Primkörper $\mathbb{F}_p \subset \Bbbk$ fest lässt.*

Beweis. Wir müssen uns nur noch davon überzeugen, dass $\mathbb{F}_p \subset \Bbbk$ fest bleibt. Nach dem kleine Satz von Fermat 4.7 ist $a^p = a$ für alle $a \in \mathbb{F}_p$, der Frobenius-Automorphismus lässt also alle Elemente von $\mathbb{F}_p$ fest. □

Definition 4.16. *Der Automorphismus $x \mapsto x^p$ heisst* Frobenius-Automorphismus.

4.3 Wurzeln

Im Körper $\mathbb{Q}$ kann man zum Beispiel die Wurzel aus 2 nicht ziehen. Das Problem haben wir in Abschnitt 1.4 dadurch gelöst, dass wir $\mathbb{Q}$ zu den reellen Zahlen $\mathbb{R}$ erweitert haben. Es ist aber auch möglich, nur die Zahl $\sqrt{2}$ hinzuzufügen, so entsteht der Körper $\mathbb{Q}(\sqrt{2})$. Das Problem dabei ist, was denn eigentlich $\sqrt{2}$ überhaupt ist. Solange man die reellen Zahlen nicht hat, hat man auch $\sqrt{2}$ nicht. Das Problem wird akut bei den endlichen Körpern wie zum Beispiel $\mathbb{F}_3$, da man diese nicht in $\mathbb{R}$ einbetten kann, also keine bekannte Menge von Zahlen existiert, in der wir die Wurzel $\sqrt{2}$ finden könnte.

Im Altertum fiel dieses Problem zunächst den Pythagoreern auf. Wenn $\sqrt{2}$ kein Bruch ist, was ist es dann? Im 15. Jahrhundert stellte sich dieses Problem bei den Versuchen, die kubische Gleichung allgemein zu lösen, erneut. Hier war es die Wurzel $\sqrt{-1}$, die den reellen Zahlen hinzuzufügen war. In $\mathbb{R}$ hat $\sqrt{-1}$ sicher keinen Platz, also wo existert es denn überhaupt? Auch der von Descartes eingeführte, eher unglückliche Begriff "imaginäre Zahl" illustriert dieses Dilemma.

Inzwischen hat man sich daran gewöhnt, dass man einfach ein neues Symbol wählt, die algebraischen Regeln postuliert, nach denen damit zu rechnen ist, und dann hofft oder besser beweist, dass keine Widersprüche auftreten. Auf diese Weise kann man einem Körper $\Bbbk$ eine beliebige Nullstelle α eines Polynoms $f \in \Bbbk[X]$ mit Koeffizienten in $\Bbbk$ hinzufügen und so den Körper $\Bbbk(\alpha)$ konstruieren. Trotzdem bleibt die Frage offen: was *ist* denn eigentlich α?

In diesem Abschnitt werden Wurzeln wie folgt konstruiert. Zunächst wird in Abschnitt 4.3.2 gezeigt, dass man immer eine Matrix M_α finden kann, welche genau die algebraischen Eigenschaften einer Nullstelle α eines Polynoms hat. Die Frage "Was ist α?" erhält also die Antwort "Eine Matrix". Mit diesem Bild lassen sich alle Körperoperationen realisieren, die Inverse kann zum Beispiel als die inverse Matrix mit dem Gauss-Algorithmus berechnet werden. In einem zweiten Schritt zeigen wir dann, dass man die Rechnung noch etwas vereinfachen kann, wenn man in Polynomringen arbeitet. Schliesslich zeigen wir dann im Abschnitt 4.3.3, wie man den Prozess iterieren kann und so für

beliebige Polynome immer einen Körper finden kann, der alle Nullstellen enthält. Wir beginnen in Abschnitt 4.3.1 damit, die Polynome, die für die Konstruktion in Frage kommen, etwas genauer zu charakterisieren.

4.3.1 Irreduzible Polynome

Die Zahlen, die man dem Körper hinzufügen möchte, müssen Nullstellen eines Polynoms sein. Wir gehen daher davon aus, dass $f \in \Bbbk[X]$ ein Polynom mit Koeffizienten in $\Bbbk$ ist, dessen Nullstelle α hinzugefügt werden sollen. Das Ziel ist natürlich, dass diese Erweiterung vollständig beschrieben werden kann durch das Polynom, ganz ohne Bezug zum Beispiel auf einen numerischen Wert der Nullstelle, der ohnehin nur in $\mathbb{C}$ sinnvoll wäre.

Nehmen wir jetzt an, dass sich das Polynom f faktorisieren lässt. Dann gibt es Polynome $g, h \in \Bbbk[X]$ derart, dass $f = g \cdot h$. Die Polynome g und h haben geringeren Grad als f. Setzt man die Nullstelle α ein, erhält man $0 = f(\alpha) = g(\alpha)h(\alpha)$, daher muss einer der Faktoren verschwinden, also $g(\alpha) = 0$ oder $h(\alpha) = 0$. Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass $g(\alpha) = 0$. Die Operation des Hinzufügens der Nullstelle α von f muss also genauso gut mit g ausgeführt werden können. Indem wir diese Überlegung auf g anwenden können wir schliessen, dass es ein Polynom $m \in \Bbbk[X]$ kleinstmöglichen Grades geben muss, welches α als Nullstelle hat. Zusätzlich kann verlangt werden, dass das Polynom normiert ist.

Definition 4.17. *Ein Polynom $f \in \Bbbk[X]$ heisst* irreduzibel*, wenn es sich nicht in zwei Faktoren $g, h \in \Bbbk[X]$ mit $f = gh$ zerlegen lässt.*

Für die Konstruktion des Körpers $\Bbbk(\alpha)$ muss daher ein irreduzibles Polynom verwendet werden.

Beispiel. Das Polynom $f(X) = X^2 - 2$ ist in $\mathbb{Q}[X]$, es hat die beiden Nullstellen $\sqrt{2}$ und $-\sqrt{2}$. Beide Nullstellen haben die exakt gleichen algebraischen Eigenschaften, sie sind mit algebraischen Mitteln nicht zu unterscheiden. Nur die Vergleichsrelation ermöglicht, die negative Wurzel von der positiven zu unterscheiden. Das Polynom kann in $\mathbb{Q}$ nicht faktorisiert werden, denn die einzig denkbare Faktorisierung ist $(X - \sqrt{2})(X + \sqrt{2})$, die Faktoren sind aber keine Polynome in $\mathbb{Q}[X]$. Also ist $f(X) = X^2 - 2$ ein irreduzibles Polynom über $\mathbb{Q}$.

Man kann das Polynom aber auch als Polynom in $\mathbb{F}_{23}[X]$ betrachten. Im Körper $\mathbb{F}_{23}$ kann man durch probieren zwei Nullstellen finden:

$$5^2 = 25 \equiv 2 \mod 23$$
$$\text{und} \quad 18^2 = 324 \equiv 2 \mod 23.$$

Und tatsächlich ist in $\mathbb{F}_{23}[X]$

$$(X - 5)(X - 18) = X^2 - 23X + 90 \equiv X^2 - 2 \mod 23,$$

über $\mathbb{F}_{23}$ ist das Polynom $X^2 - 2$ also reduzibel. ○

Beispiel. Die Zahl

$$\alpha = \frac{1 + i\sqrt{3}}{2}$$

ist eine Nullstelle des Polynoms $f(X) = X^3 - 1 \in \mathbb{Z}[X]$. α enthält aber nur Quadratwurzeln, man würde also eigentlich erwarten, dass α Nullstelle eines quadratischen Polynoms ist. Tatsächlich ist $f(X)$ nicht irreduzibel, es ist nämlich

$$X^3 - 1 = (X - 1)(X^2 + X + 1).$$

Da α nicht Nullstelle des ersten Faktors ist, muss es Nullstelle des Polynoms $m(X) = X^2 + X + 1$ sein. Der zweite Faktor ist irreduzibel.

Das Polynom $m(X)$ kann man aber auch als Polynom in $\mathbb{F}_7$ ansehen. Dann kann man aber zwei Nullstellen finden,

$$\begin{aligned} X = 2 \quad &\Rightarrow \qquad 2^2 + 2 + 1 = 4 + 2 + 1 \equiv 0 \mod 7 \\ X = 4 \quad &\Rightarrow \quad 4^2 + 4 + 1 = 16 + 4 + 1 = 21 \equiv 0 \mod 7. \end{aligned}$$

Dies führt auf die Faktorisierung

$$(X-2)(X-4) \equiv (X+5)(X+3) = X^2 + 8X + 15 \equiv X^2 + X + 1 \mod 7.$$

Das Polynom $X^2 + X + 1$ ist daher über $\mathbb{F}_7$ reduzibel und das Polynom $X^3 - 1 \in \mathbb{F}_7$ zerfällt daher in Linearfaktoren $X^3 - 1 = (X + 6)(X + 3)(X + 5)$. ○

4.3.2 Körpererweiterungen

Nach den Vorbereitungen von Abschnitt 4.3.1 können wir jetzt definieren, wie die Körpererweiterung konstruiert werden soll.

Erweiterung mit einem irreduziblen Polynom

Sei $m \in \Bbbk[X]$ ein irreduzibles Polynome über $\Bbbk$ mit dem Grad $\deg m = n$, wir dürfen es als normiert annehmen und schreiben es in der Form

$$m(X) = m_0 + m_1 X + m_2 X^2 + \dots m_{n-1} X^{n-1} + X^n.$$

Wir möchten den Körper $\Bbbk$ um eine Nullstelle α von m erweitern. Da es in $\Bbbk$ keine Nullstelle von m gibt, konstruieren wir $\Bbbk(\alpha)$ auf abstrakte Weise, ganz so wie das mit der imaginären Einheit i gemacht wurde. Die Zahl α ist damit einfach ein neues Symbol, mit dem man wie in der Algebra üblich rechnen kann. Die einzige zusätzliche Eigenschaft, die von α verlangt wird, ist dass $m(\alpha) = 0$. Unter diesen Bedingungen können beliebige Ausdrücke der Form

$$a_0 + a_1\alpha + a_2\alpha^2 + \dots a_k\alpha^k \tag{4.10}$$

gebildet werden. Aus der Bedingung $m(\alpha) = 0$ folgt aber, dass

$$\alpha^n = -m_{n-1}\alpha^{n-1} - \dots - m_2\alpha^2 - m_1\alpha - m_0. \tag{4.11}$$

Alle Potenzen mit Exponenten $\geq n$ in (4.10) können daher durch die rechte Seite von (4.11) ersetzt werden. Als Menge ist daher

$$\Bbbk(\alpha) = \{a_0 + a_1\alpha + a_2\alpha^2 + \dots + a_{n-1}\alpha^{n-1} \mid a_i \in \Bbbk\}$$

ausreichend. Die Addition von solchen Ausdrücken und die Multiplikation mit Skalaren aus $\Bbbk$ machen $\Bbbk(\alpha) \cong \Bbbk^n$ zu einem Vektorraum, die Operationen können auf den Koeffizienten komponentenweise ausgeführt werden.

Matrixrealisierung der Multiplikation mit α

Die schwierige Operation ist die Multiplikation mit α. Dazu stellen wir zusammen, wie die Multiplikation mit α auf den Basisvektoren von $\Bbbk(\alpha)$ wirkt:

$$\alpha\colon \Bbbk^n \to \Bbbk^n : \begin{cases} 1 \mapsto \alpha \\ \alpha \mapsto \alpha^2 \\ \alpha^2 \mapsto \alpha^3 \\ \quad\vdots \\ \alpha^{n-2} \mapsto \alpha^{n-1} \\ \alpha^{n-1} \mapsto \alpha^n = -m_0 - m_1\alpha - m_2\alpha^2 - \cdots - m_{n-1}\alpha^{n-1} \end{cases}$$

Diese lineare Abbildung hat die Matrix

$$M_\alpha = \begin{pmatrix} 0 & & & & & -m_0 \\ 1 & 0 & & & & -m_1 \\ & 1 & 0 & & & -m_2 \\ & & 1 & \ddots & & \vdots \\ & & & \ddots & 0 & -m_{n-2} \\ & & & & 1 & -m_{n-1} \end{pmatrix}.$$

Aufgrund der Konstruktion die Lineare Abbildung $m(M_\alpha)$, die man erhält, wenn man die Matrix M_α in das Polynom m einsetzt, jeden Vektor in $\Bbbk(\alpha)$ zu Null machen. Als Matrix muss daher $m(M_\alpha) = 0$ sein. Dies kann man auch mit einem Computeralgebra-System nachprüfen.

Beispiel. In einem früheren Beispiel haben wir gesehen, dass $\alpha = \frac{1}{2}(-1 + \sqrt{3})$ eine Nullstelle des irreduziblen Polynomes $m(X) = X^2 + X + 1$ ist. Die zugehörige Matrix M_α ist

$$M_\alpha = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \qquad \Rightarrow \qquad M_\alpha^2 = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}, \quad M_\alpha^3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Wir können auch verifizieren, dass

$$m(M_\alpha) = M_\alpha^2 + M_\alpha + I = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Die Matrix ist also eine mögliche Realisierung für das "mysteriöse" Element α. Es hat alle algebraischen Eigenschaften von α. ○

Die Menge $\Bbbk(\alpha)$ kann durch die Abbildung $\alpha \mapsto M_\alpha$ mit der Menge aller Matrizen

$$\Bbbk(M_\alpha) = \left\{ a_0 I + a_1 M_\alpha + a_2 M_\alpha^2 + \cdots + a_{n-1} M_\alpha^{n-1} \;\middle|\; a_i \in \Bbbk \right\}$$

in eine Eins-zu-eins-Beziehung gebracht werden. Diese Abbildung ist ein Algebrahomomorphismus. Die Menge $\Bbbk(M_\alpha)$ ist also das Bild des Körpers $\Bbbk(\alpha)$ in der Matrizenalgebra $M_n(\Bbbk)$.

Inverse

Im Moment wissen wir noch nicht, wie wir α^{-1} berechnen sollten. Wir können aber auch die Matrizendarstellung verwenden. Für Matrizen wissen wir selbstverständlich, wie Matrizen invertiert werden können. Tatsächlich kann man die Matrix M_α direkt invertieren:

$$M_\alpha^{-1} = \frac{1}{m_0}\begin{pmatrix} -m_1 & m_0 & & & & \\ -m_2 & 0 & m_0 & & & \\ -m_3 & & 0 & m_0 & & \\ \vdots & & & \ddots & \ddots & \\ -m_{n-1} & 0 & 0 & & 0 & m_0 \\ -1 & 0 & 0 & & 0 & 0 \end{pmatrix},$$

wie man durch Ausmultiplizieren überprüfen kann:

$$\frac{1}{m_0}\begin{pmatrix} -m_1 & m_0 & & & & \\ -m_2 & 0 & m_0 & & & \\ -m_3 & & 0 & m_0 & & \\ \vdots & & & \ddots & \ddots & \\ -m_{n-1} & 0 & 0 & & 0 & m_0 \\ -1 & 0 & 0 & & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & & & & & -m_0 \\ 1 & 0 & & & & -m_1 \\ & 1 & 0 & & & -m_2 \\ & & 1 & \ddots & & \vdots \\ & & & \ddots & 0 & -m_{n-2} \\ & & & & 1 & -m_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Die Invertierung in $\mathbb{k}(M_\alpha)$ ist damit zwar geklärt, aber es wäre viel einfacher, wenn man die Inverse auch in $\mathbb{k}(\alpha)$ bestimmen könnte.

Die Potenzen von M_α^k haben in der ersten Spalte genau in Zeile $k+1$ eine 1, alle anderen Einträge in der ersten Spalte sind 0. Die erste Spalte eines Elementes $a(\alpha) = a_0 + a_1\alpha + a_2\alpha^2 + a_{n-1}\alpha^{n-1}$ besteht daher genau aus den Elementen a_i. Die Inverse des Elements a kann daher wie folgt gefunden werden. Zunächst wird die Matrix $a(M_\alpha)$ gebildet und invertiert. Wir schreiben $B = a(M_\alpha)^{-1}$. Aus den Einträgen der ersten Spalte kann man jetzt die Koeffizienten

$$b_0 = (B)_{11}, b_1 = (B)_{21}, b_2 = (B)_{31}, \dots, b_{n-1} = (B)_{n,1}$$

ablesen und daraus das Element

$$b(\alpha) = b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1}$$

bilden. Da $b(M_\alpha) = B$ die inverse Matrix von $a(M_\alpha)$ ist, muss $b(\alpha)$ das Inverse von $a(\alpha)$ sein.

Beispiel. Wir betrachten das Polynom

$$m(X) = X^3 + 2X^2 + 2X + 3 \in \mathbb{F}_7[X],$$

es ist irreduzibel. Sei α eine Nullstelle von m, wir suchen das inverse Element zu

$$a(\alpha) = 1 + 2\alpha + 2\alpha^2 \in \mathbb{F}_7(\alpha).$$

Die Matrix $a(M_\alpha)$ bekommt die Form

$$A = \begin{pmatrix} 1 & 1 & 6 \\ 2 & 4 & 5 \\ 2 & 5 & 1 \end{pmatrix}.$$

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

·	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Abbildung 4.3: Additions- und Multiplikationstabelle für das Rechnen im Galois-Körper $\mathbb{F}_7$. Die multiplikative Inverse eines Elements in $a \in \mathbb{F}_7^*$ findet man, indem man in der Multiplikationstabelle in der Zeile a die Spalte mit der 1 sucht, diese Spalte ist mit der multiplikativen Inversen von a angeschrieben.

Die Inverse kann man bestimmen, indem man den Gauss-Algorithmus in $\mathbb{F}_7$ durchführt. Die Arithmetik in $\mathbb{F}_7$ ist etwas ungewohnt, insbesondere die Pivot-Division ist etwas mühsam, daher sind in Abbildung 4.3 die Additions- und Multiplikationstabellen zusammengestellt. Mit dieser Rechenhilfe kann jetzt der Gaussalgorithmus leicht durchgeführt werden:

$$\left(\begin{array}{ccc|ccc} 1&1&6&1&0&0\\ 2&4&5&0&1&0\\ 2&5&1&0&0&1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1&1&6&1&0&0\\ 0&2&0&5&1&0\\ 0&3&3&5&0&1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1&1&6&1&0&0\\ 0&1&0&6&4&0\\ 0&0&3&1&2&1 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1&1&6&1&0&0\\ 0&1&0&6&4&0\\ 0&0&1&5&3&5 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1&1&0&6&3&5\\ 0&1&0&6&4&0\\ 0&0&1&5&3&5 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1&0&0&0&6&5\\ 0&1&0&6&4&0\\ 0&0&1&5&3&5 \end{array}\right)$$

Für die Durchführung braucht man die Inversen in $\mathbb{F}_7$ der Pivot-Elemente, sie sind $2^{-1} = 4$ und $3^{-1} = 5$. Im rechten Teil des Tableau steht jetzt die inverse Matrix

$$A^{-1} = B = \begin{pmatrix} 0 & 6 & 5 \\ 6 & 4 & 0 \\ 5 & 3 & 5 \end{pmatrix}.$$

Daraus können wir jetzt das inverse Element

$$b(\alpha) = 6\alpha + 5\alpha^2$$

ablesen. Das Produkt $b(X) \cdot a(X)$ ist

$$\begin{aligned}(1 + 2X + 2X^2)(6X + 5X^2) &= 10X^4 + 22X^3 + 17X^2 + 6X \\ &= 3X^4 + X^3 + 3X^2 + 6X\end{aligned}$$

Diese Polynom muss jetzt mit dem Minimalpolynom $m(X)$ reduziert werden, wir subtrahieren dazu $3Xm(X)$ und erhalten

$$= -5X^3 - 3X^2 - 3X$$

$$= 2X^3 + 4X^2 + 4X$$

Die vollständige Reduktion wird erreicht, indem wir nochmals $2m(X)$ subtrahieren:

$$= -6 \equiv 1 \mod 7,$$

das Element $b(\alpha) = 6\alpha + 5\alpha^2$ ist also das Inverse Element von $a(\alpha) = 1 + 2\alpha + 2\alpha^2$ in $\mathbb{F}_7(\alpha)$. ○

Die Matrixrealisation von $\Bbbk(\alpha)$ führt also auf eine effiziente Berechnungsmöglichkeit für das Inverse eines Elements von $\Bbbk(\alpha)$.

Algebraische Konstruktion

Die Matrixdarstellung von α ermöglicht eine rein algebraische und für die Rechnung besser geeignete Konstruktion. Für jedes Polynom $f \in \Bbbk[X]$ ist $f(M_\alpha) \in M_n(\Bbbk)$. Dies definiert einen Homomorphismus

$$\varphi\colon \Bbbk[X] \to M_n(\Bbbk) : f \mapsto f(M_\alpha).$$

Wir haben früher schon gesehen, dass das Bild dieses Homomorphismus genau die Menge $\Bbbk(M_\alpha)$ ist. Allerdings ist φ nicht injektiv, das Polynom m wird zum Beispiel auf $\varphi(m) = m(M_\alpha) = 0$ abgebildet.

Der Kern von φ besteht aus allen Polynomen $p \in \Bbbk[X]$, für die $p(M_\alpha) = 0$ gilt. Da aber alle Matrizen $E, M_\alpha, \dots, M_\alpha^{n-1}$ linear unabhängig sind, muss ein solches Polynom den gleichen Grad haben we m, und damit ein Vielfaches von m sein. Der Kern besteht daher genau aus den Vielfachen von $m(X)$, $\ker\varphi = m(X)\Bbbk[X]$.

Es ist nicht a priori klar, dass der Quotient R/I für ein Ideal $I \subset R$ ein Körper ist. Hier spielt es eine Rolle, dass das von m erzeugte Ideal maximal ist im folgenden Sinne.

Definition 4.18. *Ein Ideal $I \subset R$ heisst* maximal*, wenn für jedes andere Ideal J mit $I \subset J \subset R$ entweder $I = J$ oder $J = R$ gilt.*

Beispiel. Die Ideale $p\mathbb{Z} \subset \mathbb{Z}$ sind maximal genau dann, wenn p eine Primzahl ist.

TODO: XXX Begründung ○

Satz 4.19. *Der Ring R/I ist genau dann ein Körper, wenn I ein maximales Ideal ist.*

Beweis. □

Ein irreduzibles Polynom $m \in \Bbbk[X]$ erzeugt ein maximales Ideal, somit ist $\Bbbk[X]/m\Bbbk[X] \cong \Bbbk(M_\alpha) \cong \Bbbk(\alpha)$.

Reduktion modulo m

Die algebraische Konstruktion hat gezeigt, dass die arithmetischen Operationen im Körper $\Bbbk(\alpha)$ genau die Operationen in $\Bbbk[X]/m\Bbbk[X]$ sind. Eine Zahl in $\Bbbk(\alpha)$ wird also durch ein Polynom vom $n-1$ dargestellt. Addieren und Subtrahieren erfolgen Koeffizientenweise in $\Bbbk$. Bei der Multiplikation entsteht möglicherwise ein Polynom grösseren Grades, mit dem Polynomdivisionsalgorithmus kann der Rest bei Division durch m ermittelt werden.

Beispiel. Das Polyonom $f = X^5+X^4+X^3+X^2+X^1+1 \in \mathbb{F}_7[X]$ soll modulo $m(X) = X^3+2X^2+2X^2+3$ reduziert werden. Wir führen die Polynomdivision in $\mathbb{F}_7[X]$ durch, die Multiplikationstabelle von $\mathbb{F}_7$ in Abbildung 4.3 ist dabei wieder hilfreich.

$$\begin{aligned}
X^5 + X^4 + X^3 + X^2 + X + 1 &: X^3 + 2X^2 + 2X + 3 = X^2 + 6X + 1 = q \\
-(X^5 + 2X^4 + 2X^3 + 3X^2) & \\
\hline
6X^4 + 6X^3 + 5X^2 + X & \\
-(6X^4 + 5X^3 + 5X^2 + 4X) & \\
\hline
X^3 + 4X + 1 & \\
-(X^3 + 2X^2 + 2X + 3) & \\
\hline
5X^2 + 2X + 5 = r &
\end{aligned}$$

Die Kontrolle

$$\begin{aligned}
(X^2 + 6X + 1) \cdot (X^3 + 2X^2 + 2X + 3) & \\
\hline
X^3 + 2X^2 + 2X + 3 & \\
6X^4 + 5X^3 + 5X^2 + 4X & \\
X^5 + 2X^4 + 2X^3 + 3X^2 & \\
\hline
X^5 + X^4 + X^3 + 3X^2 + 6X + 3 & = q \cdot m \\
+(5X^2 + 2X + 5) & = r \\
\hline
X^5 + X^4 + X^3 + X^2 + X + 1 &
\end{aligned}$$

zeigt $f = qm + r$ und damit die Korrektheit der Rechnung. ○

Die Identität $m(\alpha) = 0$ kann aber auch wie folgt interpretiert werden. Sei der Grad von f mindestens so gross wie der von m, also $l = \deg f \geq \deg m = n$. Indem man mit α^{l-n} multipliziert, erhält man die Relation

$$\alpha^l + m_{n-1}\alpha^{l-1} + m_{n-2}\alpha^{l-2} + \cdots + a_1\alpha^{l-n+1} + a_0\alpha^{l-n} = 0.$$

Ist f_l der führende Koeffizient des Polynoms f, dann ist $f - f_0 m X^{n-l}$ ein Polynom vom Grad $l-1$, welches modulo m mit f übereinstimmt. Indem man dies wiederholt, kann man also die Reduktion finden, ohne den Polynomdivisionsalgorithmus durchzuführen. Man erhält auf diese Weise zwar den Quotienten q nicht, aber den Rest r kann man trotzdem bekommen.

Beispiel. Wir wenden den eben beschriebenen Algorithmus wieder auf das Polynom $f = X^5 + X^4 + X^3 + X^2 + X + 1$ an und erhalten:

$$\begin{aligned}
X^5 + X^4 + X^3 + X^2 + X + 1 & \\
-(X^5 + 2X^4 + 2X^3 + 3X^2 = X^2 m) & \\
\hline
6X^4 + 6X^3 + 5X^2 + X + 1 & \\
-(6X^4 + 5X^3 + 5X^2 + 4X = 6Xm) & \\
\hline
X^3 + 4X + 1 & \\
-(X^3 + 2X^2 + 2X + 3 = m) & \\
\hline
5X^2 + 2X + 5 = r &
\end{aligned}$$

Dies ist derselbe Rest wie wir mit dem Divisionsalgorithmus gefunden haben. ○

Diese Form des Reduktionsalgorithmus ist besonders leicht durchzuführen in einem Körper $\mathbb{F}_2$, da dort die Addition und die Subtraktion der Koeffizienten übereinstimmen. Die Multiplikation mit X ist nichts anders als ein Shift der Koeffizienten.

Multiplikative Inverse

Die schwierigste Operation in $\Bbbk(\alpha)$ ist die Division. Wie bei der Berechnung der Inversion in einem Galois-Körper $\mathbb{F}_p$ kann dafür der euklidische Algorithmus verwendet werden. Sei also $f \in \Bbbk[X]$ ein Polynom vom Grad $\deg f < \deg m$, es soll das multiplikative Inverse gefunden werden. Da m ein irreduzibles Polynom ist, müssen f und m teilerfremd sein. Der euklidische Algorithmus liefert zwei Polynome $s, t \in \Bbbk[X]$ derart, dass

$$sf + tm = 1.$$

Reduzieren wir modulo m, wird daraus $af = 1$ in $\Bbbk[X]/m\Bbbk[X]$. Das Polynom a, reduziert modulo m, ist also die multiplikative Inverse von f.

Bei der praktischen Durchführung des euklidischen Algorithmus ist der letzte Rest r_{n-1} oft nicht 1 sondern ein anderes Element von $\mathbb{F}_p^*$. Die Linearkombination von f und m mit den berechneten Faktoren s und t ist daher auch nicht 1, sondern

$$sf + tm = r_{n-1}.$$

Da aber alle Elemente in $\mathbb{F}_p^*$ invertierbar sind, kann man durch r_{n-1} dividieren, was

$$r_{n-1}^{-1} sf + r_{n-1}^{-1} tm = 1$$

ergibt. Also ist $r_{n-1}^{-1} s$ die gesuchte Inverse in $\mathbb{F}_p(\alpha)$, dies passiert auch im folgenden Beispiel.

Beispiel. Auf Seite 95 haben wir die multiplikative Inverse von $f = 2X^2 + 2X + 1 \in \mathbb{F}_7[X]/m\mathbb{F}_7[X]$ mit $m = X^3 + 2X^2 + 2X + 3$ mit Hilfe von Matrizen berechnet, hier soll sie jetzt nochmals mit dem euklidischen Algorithmus berechnet werden.

Zunächst müssen wir den euklidischen Algorithmus für die beiden Polynome f und m durchführen. Der Quotient $m : f$ ist:

$$\begin{aligned}
&X^3 + 2X^2 + 2X + 3 : 2X^2 + 2X + 1 = 4X + 4 = q_0\\
-&\underline{(X^3 + X^2 + 4X)}\\
&X^2 + 5X + 3\\
-&\underline{(\ X^2 + X + 4)}\\
&4X + 6 = r_0
\end{aligned}$$

Jetzt muss der Quotient $f : r_0$ berechnet werden:

$$\begin{aligned}
&2X^2 + 2X + 1 : 4X + 6 = 4X + 5 = q_1\\
-&\underline{(2X^2 + 3X)}\\
&6X + 1\\
-&\underline{(6X + 2)}\\
&6 = r_1
\end{aligned}$$

Da der Rest $r_1 \in \mathbb{F}_7^*$ liegt, gibt die nächste Division natürlich den Rest 0 und der letzte nicht verschwindende Rest ist $r_1 = 6$:

$$\begin{aligned}
&4X + 6 : 6 = 3X + 1 = q_2\\
-&\underline{(4X)}\\
&0 + 6\\
-&\underline{(6)}\\
&0 = r_2
\end{aligned}$$

Damit ist der euklidische Algorithmus abgeschlossen.

Durch Ausmultiplizieren der Matrizen $Q(-q_i)$ können wir jetzt auch die Faktoren s und t finden.

$$\begin{aligned}
Q = \begin{pmatrix} s & t \\ * & * \end{pmatrix} = Q(q_2)Q(q_1)Q(q_0) &= \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -q_0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \\ 1 & 4X+6 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 3X+2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 3X+3 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \\ 1 & 4X+6 \end{pmatrix}\begin{pmatrix} 1 & 3X+3 \\ 3X+2 & 2X^2+X \end{pmatrix} \\
&= \begin{pmatrix} 3X+2 & 2X^2+X \\ 1+(4X+6)(3X+2) & 3X+3+(4X+6)(2X^2+X) \end{pmatrix} \\
&= \begin{pmatrix} 3X+2 & 2X^2+X \\ 5X^2+5X+6 & X^3+2X^2+2X+6 \end{pmatrix}
\end{aligned}$$

Daraus liest man

$$s = 2X^2 + X \qquad \text{und} \qquad t = 3X + 2$$

ab. Wir überprüfen, ob die Koeffizienten der ersten Zeile tatsächlich m und f zu $r_1 = 6$ kombinieren. Es ist

$$(3X+2)\cdot m + (2X^2+X)\cdot f = (3X+2)(X^3+3X^2+X+2) + (2X^2+X)(2X^2+2X+1) = 6 = r_1$$

Die multiplikative Inverse ist daher $r_1^{-1}(2X^2+X) = 6^{-1}(2X^2+X) = 6(2X^2+X) = 5X^2+6X$, was mit dem Beispiel von Seite 95 übereinstimmt. ○

Besonders einfach ist die Rechung für $\mathbb{k} = \mathbb{F}_2$. Dieser Spezialfall ist für die praktische Anwendung in der Kryptographie von besonderer Bedeutung, daher wird er im In Kapitel 10 genauer untersucht.

4.3.3 Zerfällungskörper

XXX TODO

Übungsaufgaben

4.1. Der Körper $\mathbb{F}_2$ ist besonders einfach, da er nur zwei Elemente 0 und 1 enthält.

a) Bestimmen Sie die Additions- und Multiplikationstabelle für $\mathbb{F}_2$.

b) Lösen Sie das lineare Gleichungssystem

$$\begin{array}{rcrcrcl}
x_1 & + & x_2 & & & = & 0 \\
 & & x_2 & + & x_3 & + & x_4 = 1 \\
x_1 & + & x_2 & + & x_3 & + & x_4 = 1 \\
 & & x_2 & + & x_3 & & = 0
\end{array}$$

über dem Körper $\mathbb{F}_2$ mit dem Gauss-Algorithmus.

c) Bestimmen Sie die Inverse $A^{-1} \in \mathrm{GL}_2(\mathbb{F}_2)$ der Koeffizientenmatrix A des Gleichungssystems.

d) Kontrollieren Sie das Resultat durch Ausmultiplizieren des Produktes AA^{-1}.

Lösung. a) Die Additions- und Multiplikationstabellen sind

+	0	1
0	0	1
0	1	0

·	0	1
0	0	0
0	0	1

Betrachtet als Bitoperationen entspricht die Addition dem XOR, die Multiplikation dem AND.

b) Die Gauss-Tableaux sind

$$\left(\begin{array}{cccc|c} 1&1&0&0&0\\ 0&1&1&1&1\\ 1&1&1&1&1\\ 0&1&1&0&0 \end{array}\right) \to \left(\begin{array}{cccc|c} 1&1&0&0&0\\ 0&1&1&1&1\\ 0&0&1&1&1\\ 0&1&1&0&0 \end{array}\right) \to \left(\begin{array}{cccc|c} 1&1&0&0&0\\ 0&1&1&1&1\\ 0&0&1&1&1\\ 0&0&0&1&1 \end{array}\right)$$

$$\to \left(\begin{array}{cccc|c} 1&1&0&0&0\\ 0&1&1&0&0\\ 0&0&1&0&0\\ 0&0&0&1&1 \end{array}\right) \to \left(\begin{array}{cccc|c} 1&1&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&1 \end{array}\right) \to \left(\begin{array}{cccc|c} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&1 \end{array}\right)$$

In der ersten Zeile stehen die Schritt der Vorwärtsreduktion, in der zweiten die Schritte des Rückwärtseinsetzens. Als Lösung liest man ab

$$x = \begin{pmatrix} 0\\0\\0\\1 \end{pmatrix},$$

die Korrektheit kann man leicht durch Einsetzen überprüfen.

c) Wir wenden erneut den Gauss-Algorithmus an:

$$\left(\begin{array}{cccc|cccc} 1&1&0&0&1&0&0&0\\ 0&1&1&1&0&1&0&0\\ 1&1&1&1&0&0&1&0\\ 0&1&1&0&0&0&0&1 \end{array}\right) \to \left(\begin{array}{cccc|cccc} 1&1&0&0&1&0&0&0\\ 0&1&1&1&0&1&0&0\\ 0&0&1&1&1&0&1&0\\ 0&1&1&0&0&0&0&1 \end{array}\right)$$

$$\to \left(\begin{array}{cccc|cccc} 1&1&0&0&1&0&0&0\\ 0&1&1&1&0&1&0&0\\ 0&0&1&1&1&0&1&0\\ 0&0&0&1&0&1&0&1 \end{array}\right)$$

$$\to \left(\begin{array}{cccc|cccc} 1&1&0&0&1&0&0&0\\ 0&1&1&0&0&0&0&1\\ 0&0&1&0&1&1&1&1\\ 0&0&0&1&0&1&0&1 \end{array}\right)$$

$$\rightarrow \left[\begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right]$$

Daraus liest man die Inverse A^{-1} der Koeffizientenmatrix A ab als

$$A^{-1} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

d) Wir prüfen das Resultat durch Ausmultiplizieren:

$$AA^{-1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Dabei kann man verwenden, dass der Eintrag in Zeile i und Spalte k des Produktes die Anzahl der Positionen ist, wo in der Zeile i von A und in der Spalte j von A^{-1} eine 1 steht.

○

4.2. Die Zahl $p = 47$ ist eine Primzahl, der Ring $\mathbb{Z}/p\mathbb{Z} = \mathbb{F}_{47}$ ist daher ein Körper. Jeder von Null verschiedene Rest $b \in \mathbb{F}_p^*$ hat daher eine multiplikative Inverse. Berechnen Sie die multiplikative Inverse von $b = 11 \in \mathbb{F}_{47}$.

Lösung. Der euklidische Algorithmus muss auf die Zahlen $p = 47$ und $b = 11$ angewendet werden, es ergeben sich die Quotienten und Reste der folgenden Tabelle:

k	a_k	b_k	q_k	r_k
0	47	11	4	3
1	11	3	3	2
2	3	2	1	1
3	2	1	2	0

Wie erwartet ist der grösste gemeinsame Teiler $\mathrm{ggT}(47, 11) = r_2 = 1$. Um die Zahlen s, t zu finden, für die $sp + tb = 1$ gilt, können wir die Matrixform verwenden, wir berechnen dazu

$$\begin{aligned} Q = Q(2)Q(1)Q(3)Q(4) &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -4 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -4 \\ -3 & 13 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} -3 & 13 \\ 4 & -17 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} 4 & -17 \\ -11 & 47 \end{pmatrix}.$$

Daraus kann man ablesen, dass $s = 4$ und $t = -17$, tatsächlich ist $4 \cdot 47 - 17 \cdot 11 = 188 - 187 = 1$. Wir schliessen daraus, dass $-17 = 30 \in \mathbb{F}_{47}$ die multiplikative Inverse von $b = 11$ ist. Die Rechnung $11 \cdot 30 = 330 = 7 \cdot 47 + 1$ zeigt, dass dies der Fall ist.

Alternativ zur Matrixdarstellung kann man die Koeffizienten s und t auch mit Hilfe der erweiterten Tabelle finden:

k	a_k	b_k	q_k	r_k	c_k	d_k
					1	0
0	47	11	4	3	0	1
1	11	3	3	2	1	−4
2	3	2	1	1	−3	13
3	2	1	2	0	4	−17
4	1	0			−11	47

Die gesuchten Zahlen s und t sind rot hervorgehoben. ○

4.3. Berechnen Sie 666^{666} in $\mathbb{F}_{13}$.

Lösung. Zunächst ist die Basis der Potenz $666 = 3$ in $\mathbb{F}_{13}$, es muss also nur 3^{666} berechnet werden. Nach dem kleinen Satz von Fermat ist $3^{12} = 1$ in $\mathbb{F}_{13}$. Wegen $666 = 12 \cdot 50 + 6$ folgt $3^{666} = 3^6 = 729 = 1$ in $\mathbb{F}_{13}$. ○

4.4. Im Rahmen der Aufgabe, die Zehntausenderstelle der Zahl $5^{5^{5^5}}$ zu berechnen muss Michael Penn im Video `https://youtu.be/Xg24FinMiws` bei 12:52 zwei Zahlen x und y finden, so dass,

$$5^5 x + 2^5 y = 1$$

ist. Verwenden Sie die Matrixform des euklidischen Algorithmus.

Lösung. Zunächst berechnen wir die beiden Potenzen

$$5^5 = 3125 \qquad \text{und} \qquad 2^5 = 32.$$

Damit können wir jetzt den Algorithmus durchführen. Die Quotienten und Reste sind

$a_0 = q_0 \cdot b_0 + r_0$	$3125 = 97 \cdot 32 + 21$	$q_0 = 97$	$r_0 = 21$
$a_1 = q_1 \cdot b_1 + r_1$	$32 = 1 \cdot 21 + 10$	$q_1 = 1$	$r_1 = 11$
$a_2 = q_2 \cdot b_2 + r_2$	$21 = 1 \cdot 11 + 10$	$q_2 = 1$	$r_2 = 10$
$a_3 = q_3 \cdot b_3 + r_3$	$11 = 1 \cdot 10 + 1$	$q_3 = 1$	$r_3 = 1$
$a_4 = q_4 \cdot b_4 + r_4$	$10 = 10 \cdot 1 + 0$	$q_4 = 10$	$r_4 = 0$

Daraus kann man jetzt auch die Matrizen $Q(q_k)$ bestimmen und ausmultiplizieren:

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & -10 \end{pmatrix} \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}}_{} \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -97 \end{pmatrix}}_{}$$

$$= \begin{pmatrix} 0 & 1 \\ 1 & -10 \end{pmatrix} \underbrace{\begin{pmatrix} 0 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -97 \\ -1 & 98 \end{pmatrix}}$$

$$= \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -10 \end{pmatrix} \begin{pmatrix} 2 & -195 \\ -3 & 293 \end{pmatrix}}$$

$$= \begin{pmatrix} -3 & 293 \\ 32 & -3125 \end{pmatrix}.$$

Daras kann man jetzt ablesen, dass

$$-3 \cdot 3125 + 293 \cdot 32 = -9375 + 9376 = 1.$$

Die gesuchten Zahlen sind also $x = -3$ und $y = 293$. ○

4.5. Das Polynom $m(X) = X^2 + 2X + 2$ ist als Polynom in $\mathbb{F}_3[X]$ irreduzibel. Dies bedeutet, dass der Ring der Polynome $\mathbb{F}_3[X]/(m(X))$ ein Körper ist. Man bezeichnet ihn auch mit $\mathbb{F}_3(\alpha)$, wobei man sich α als Nullstelle von $m(X)$ oder als die Matrix

$$\alpha = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

vorstellen kann.

a) Stellen Sie die Additions- und Multiplikationstabellen für das Rechnen in $\mathbb{F}_3$ auf.

b) Berechnen Sie α^{-1} in $\mathbb{F}_3(\alpha)$ aus der Bedingung $m(\alpha) = 0$.

c) Verwenden Sie den euklidischen Algorithmus, um $(1 + \alpha)^{-1}$ in $\mathbb{F}_3(\alpha)$ zu bestimmen.

d) Berechnen Sie α^3.

Lösung. a) Die Additions- und Multiplikationstabelle von $\mathbb{F}_3$ ist

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

·	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

b) Aus $m(\alpha) = \alpha^2 + 2\alpha + 2 = 0$ folgt

$$\begin{aligned} \alpha + 2 + 2\alpha^{-1} &= 0 \\ 2\alpha^{-1} &= 2\alpha + 1 \\ \alpha^{-1} &= \alpha + 2. \end{aligned}$$

Als Matrix kann man

$$\alpha^{-1} = \alpha + 2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \equiv \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \mod 3$$

schreiben und durch Nachrechnen verifizieren dass, tatsächlich gilt

$$\alpha\alpha^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mod 3.$$

c) Für den euklidischen Algorithmus müssen wir wiederholt eine Polynomdivision in $\mathbb{F}_3[X]$ durchführen. Im ersten Schritt ist es

$$\begin{aligned} (X^2 + 2X + 2) : (X + 1) &= X + 1 = q_1 \\ -\underline{(X^2 + \ X)} & \\ X + 2 & \\ -\underline{(X + 1)} & \\ 1 \ &= r_1 \end{aligned}$$

Die nächste Division ist $(X + 1) : 1$, die als Quotient $q_2 = X + 1$ und den Rest $r_2 = 0$ hat. Mit der Matrixform des euklidischen Algorithmus kann man jetzt auch die Koeffizienten s und t bestimmen, die beide Polynome in $\mathbb{F}_3[X]$ sind:

$$Q = Q(q_2)Q(q_1) = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 2X + 2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 2X + 2 \end{pmatrix} = \begin{pmatrix} 1 & 2X + 2 \\ 2X + 2 & X^2 + X + 2 \end{pmatrix}.$$

Die gesuchten Polynome sind $s = 1$ und $t = 2X + 2$ und man kann nachrechnen, dass

$$\begin{aligned} s \cdot m(X) + t \cdot (X + 1) &= X^2 + 2X + 2 + (2X + 2) \cdot (X + 1) \\ &= X^2 + 2X + 2 + 2X^2 + 4X + 2 \\ &= 3X^2 + 6X + 1 \\ &\equiv 1 \mod 3. \end{aligned}$$

Natürlich kann man s und t auch mit der erweiterten Tabelle finden:

k	a_k	b_k	q_k	r_k	c_k	d_k
					1	0
0	$X^2 + 2X + 2$	$X + 1$	$X + 1$	1	0	1
1	$X + 1$	1	$X + 1$	0	1	$2X + 2$
2	1	0		0	$2X + 2$	$X^2 + 2X + 2$

In allen Fällen ist also $(X + 1)^{-1} = 2X + 2$.

d) Wegen $m(\alpha) = \alpha^2 + 2\alpha + 2 = 0$ ist $\alpha^2 = -2\alpha - 2 = \alpha + 1$ und damit

$$\alpha^3 = \alpha \cdot \alpha^2 = \alpha(\alpha + 1) = \alpha^2 + \alpha = 2\alpha + 1.$$

Die restlichen Potenzen von α sind

$$\begin{aligned} \alpha^4 &= \alpha(2\alpha + 1) = 2\alpha^2 + \alpha = 2\alpha + 2 + \alpha = 2 \\ \alpha^5 &= 2\alpha \\ \alpha^6 &= 2\alpha^2 = 2\alpha + 2 \\ \alpha^7 &= 2\alpha^2 + 2\alpha = \alpha + 2 \end{aligned}$$

○

Kapitel 5

Eigenwerte und Eigenvektoren

Die algebraischen Eigenschaften einer Matrix A sind eng mit der Frage nach linearen Beziehungen unter den Potenzen von A^k verbunden. Im Allgemeinen ist die Berechnung dieser Potenzen eher unübersichtlich, es sei denn, die Matrix hat eine spezielle Form. Die Potenzen einer Diagonalmatrix erhält man, indem man die Diagonalelemente potenziert. Auch für Dreiecksmatrizen ist mindestens die Berechnung der Diagonalelemente von A^k einfach. Die Theorie der Eigenwerte und Eigenvektoren ermöglicht, Matrizen in eine solche besonders einfache Form zu bringen.

In Abschnitt 5.1 werden die grundlegenden Definitionen der Eigenwerttheorie in Erinnerung gerufen. Damit kann dann in Abschnitt 5.2 gezeigt werden, wie Matrizen in besonders einfache Form gebracht werden können. Die Eigenwerte bestimmen auch die Eigenschaften von numerischen Algorithmen, wie in den Abschnitten **??** und **??** dargestellt wird. Für viele Funktionen kann man auch den Wert $f(A)$ berechnen, unter geeigneten Voraussetzungen an den Spektralradius. Dies wird in Abschnitt 5.4 beschrieben.

5.1 Grundlagen

Die Potenzen A^k sind besonders einfach zu berechnen, wenn die Matrix Diagonalform hat, wenn also $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ ist. In diesem Fall ist $Ae_k = \lambda_k e_k$ für jeden Standardbasisvektor e_k. Statt sich auf Diagonalmatrizen zu beschränken könnten man also auch Vektoren v suchen, für die gilt $Av = \lambda v$, die also von A nur gestreckt werden. Gelingt es, eine Basis aus solchen sogenanten *Eigenvektoren* zu finden, dann kann man die Matrix A durch Basiswechsel in diese Form bringen.

5.1.1 Kern und Bild von Matrixpotenzen

In diesem Abschnitt ist $A \in M_n(\mathbb{k})$, A beschreibt eine lineare Abbildung $f\colon \mathbb{k}^n \to \mathbb{k}^n$. In diesem Abschnitt sollen Kern und Bild der Potenzen A^k untersucht werden.

Definition 5.1. *Wir bezeichnen Kern und Bild der iterierten Abbildung A^k mit*

$$\mathcal{K}^k(A) = \ker A^k \qquad \textit{und} \qquad \mathcal{J}^k(A) = \operatorname{im} A^k.$$

Durch Iteration wird das Bild immer kleiner. Wegen

$$\mathcal{J}^k(A) = \operatorname{im} A^k = \operatorname{im} A^{k-1}A = \{A^{k-1}Av \mid v \in \mathbb{k}^n\} \subset \{A^{k-1}v \mid v \in \mathbb{k}^n\} = \mathcal{J}^{k-1}(A)$$

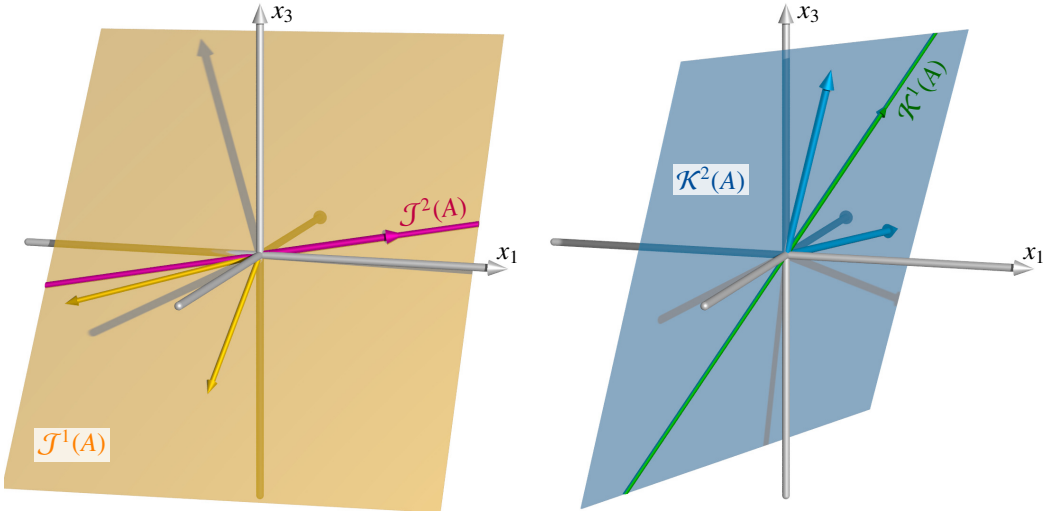


Abbildung 5.1: Iterierte Kerne und Bilder einer 3×3-Matrix mit Rang 2. Die abnehmend geschachtelten iterierten Bilder $\mathcal{J}^1(A) \subset \mathcal{J}^2(A)$ sind links dargestellt, die zunehmen geschachtelten iterierten Kerne $\mathcal{K}^1(A) \subset \mathcal{K}^2(A)$ rechts.

folgt

$$\mathbb{k}^n = \operatorname{im} E = \operatorname{im} A^0 = \mathcal{J}^0(A) \supset \mathcal{J}^1(A) = \operatorname{im} A \supset \mathcal{J}^2(A) \supset \cdots \supset \mathcal{J}^k(A) \supset \mathcal{J}^{k+1}(A) \supset \cdots \supset \{0\}. \tag{5.1}$$

Für die Kerne gilt etwas Ähnliches. Ein Vektor $x \in \mathcal{K}^k(A)$ erfüllt $A^k x = 0$. Dann erfüllt er aber erst recht auch

$$A^{k+1}x = A \underbrace{A^k x}_{=0} = 0,$$

also ist $x \in \mathcal{K}^k(A)$. Es folgt

$$\{0\} = \mathcal{K}^0(A) = \ker A^0 = \ker E \subset \mathcal{K}^1(A) = \ker A \subset \cdots \subset \mathcal{K}^k(A) \subset \mathcal{K}^{k+1}(A) \subset \cdots \subset \mathbb{k}^n. \tag{5.2}$$

Neben diesen offensichtlichen Resultaten kann man aber noch mehr sagen. Es ist klar, dass in beiden Ketten und nur in höchstens n Schritten eine wirkliche Änderung stattfinden kann. Man kann aber sogar genau sagen, wo Änderungen stattfinden:

Satz 5.2. *Ist $A \in M_n(\mathbb{k})$ eine $n \times n$-Matrix, dann gibt es eine Zahl k so, dass*

$$\begin{array}{ccccccccccccc} 0 = \mathcal{K}^0(A) & \subsetneq & \mathcal{K}^1(A) & \subsetneq & \mathcal{K}^2(A) & \subsetneq & \dots & \subsetneq & \mathcal{K}^k(A) & = & \mathcal{K}^{k+1}(A) & = & \dots \\ \mathbb{k}^n = \mathcal{J}^0(A) & \supsetneq & \mathcal{J}^1(A) & \supsetneq & \mathcal{J}^2(A) & \supsetneq & \dots & \supsetneq & \mathcal{J}^k(A) & = & \mathcal{J}^{k+1}(A) & = & \dots \end{array}$$

ist.

Beweis. Es sind zwei Aussagen zu beweisen. Erstens müssen wir zeigen, dass die Dimension von $\mathcal{K}^i(A)$ nicht mehr grösser werden kann, wenn sie zweimal hintereinander gleich war. Nehmen wir daher an, dass $\mathcal{K}^i(A) = \mathcal{K}^{i+1}(A)$. Wir müssen $\mathcal{K}^{i+2}(A)$ bestimmen. $\mathcal{K}^{i+2}(A)$ besteht aus allen Vektoren $x \in \mathbb{k}^n$ derart, dass $Ax \in \mathcal{K}^{i+1}(A) = \mathcal{K}^i(A)$ ist. Daraus ergibt sich, dass $AA^i x = 0$, also ist

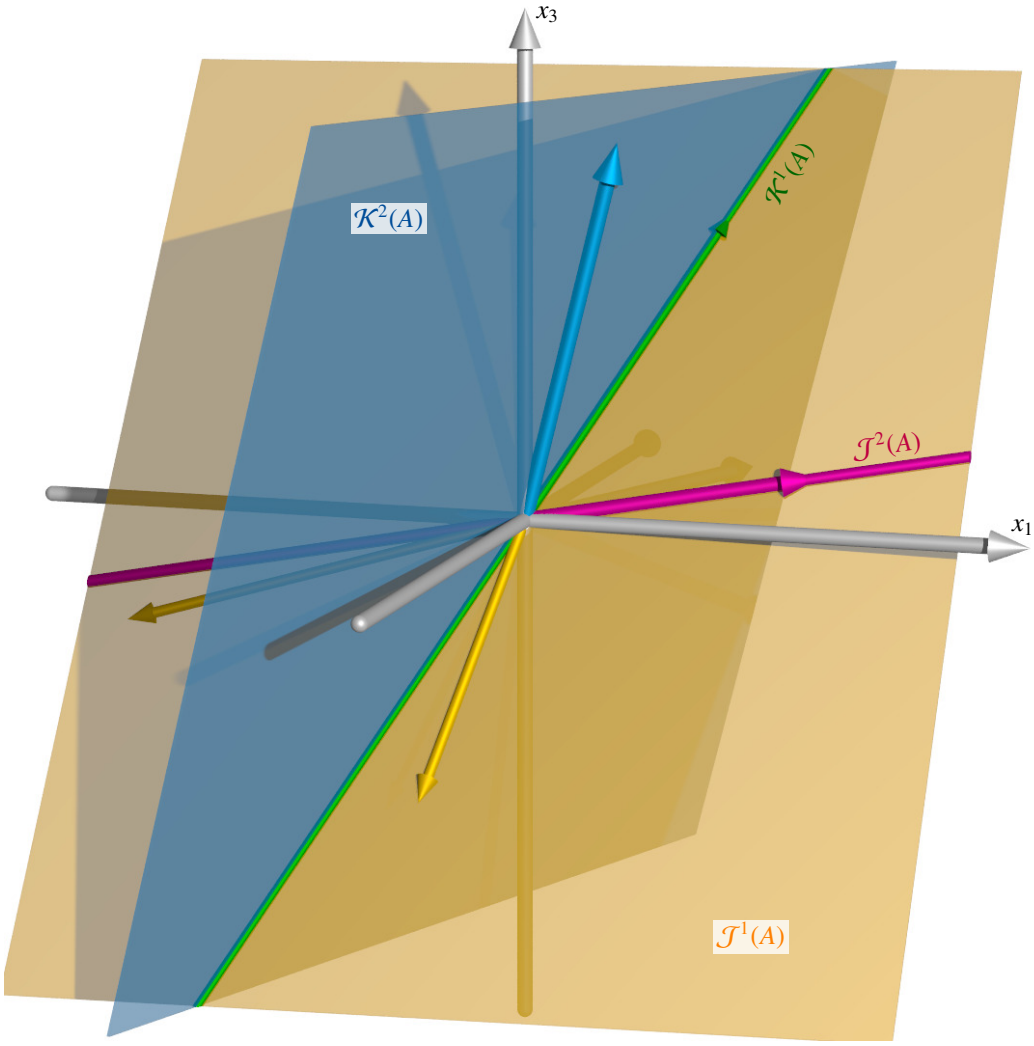


Abbildung 5.2: Iterierte Kerne und Bilder einer 3×3-Matrix mit Rang 2. Da $\dim \mathcal{J}^2(A) = 1$ und $\dim \mathcal{J}^1(A) = 2$ ist, muss es einen Vektor in $\mathcal{J}^1(A)$ geben, der von A auf 0 abgebildet wird, der also auch im Kern $\mathcal{K}^1(A)$ liegt. Daher ist $\mathcal{K}^1(A)$ die Schnittgerade von $\mathcal{J}^1(A)$ und $\mathcal{K}^2(A)$. Man kann auch gut erkennen, dass $\mathbb{R}^3 = \mathcal{K}^1(A) \oplus \mathcal{J}^1(A) = \mathcal{K}^2(A) \oplus \mathcal{J}^2(A)$ ist.

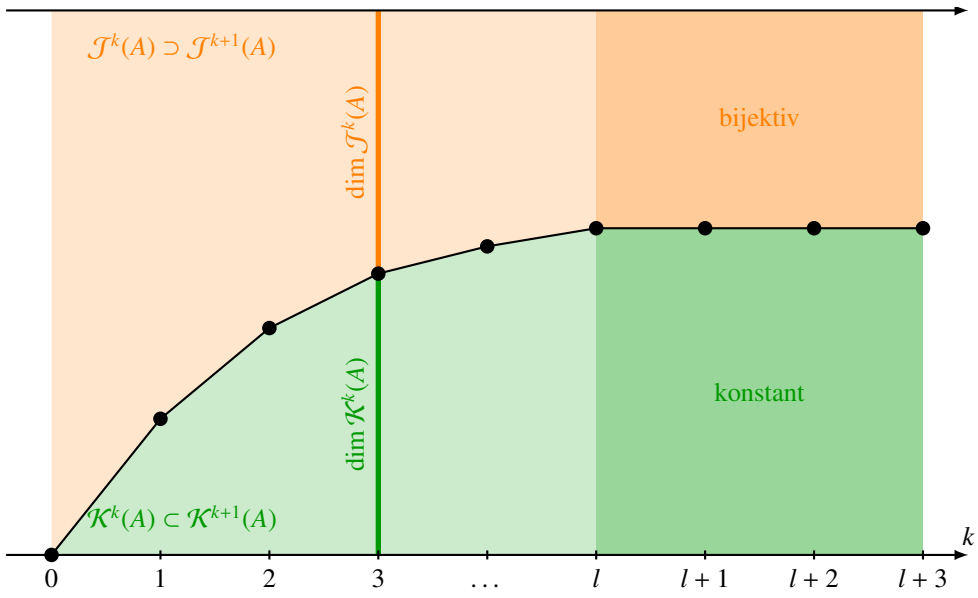


Abbildung 5.3: Entwicklung der Dimension von $\dim \mathcal{K}^k(A)$ (grün) und $\dim \mathcal{J}^k(A)$ (orange) in Abhängigkeit vom Exponenten k. Für $k \geq l$ ändern sich die Dimensionen nicht mehr, A eingeschränkt auf $\mathcal{J}^l(A) = \mathcal{J}(A)$ ist injektiv.

$x \in \mathcal{K}^{i+1}(A)$. Wir erhalten also $\mathcal{K}^{i+2}(A) \subset \mathcal{K}^{i+1} \subset \mathcal{K}^{i+2}(A)$, dies ist nur möglich, wenn beide gleich sind.

Analog kann man für die Bilder vorgehen. Wir nehmen an, dass $\mathcal{J}^i(A) = \mathcal{J}^{i+1}(A)$ und bestimmten $\mathcal{J}^{i+2}(A)$. $\mathcal{J}^{i+2}(A)$ besteht aus all jenen Vektoren, die als Ax mit $x \in \mathcal{J}^{i+1}(A) = \mathcal{J}^i(A)$ erhalten werden können. Es gibt also insbesondere ein $y \in \mathbb{k}^i$ mit $x = A^i y$. Dann ist $Ax = A^{i+1}y \in \mathcal{J}^{i+1}(A)$. Insbesondere besteht $\mathcal{J}^{i+2}(A)$ genau aus den Vektoren von $\mathcal{J}^{i+1}(A)$.

Zweitens müssen wir zeigen, dass die beiden Ketten bei der gleichen Potenz von A konstant werden. Dies folgt jedoch daraus, dass $\dim \mathcal{J}^i(A) = \operatorname{Rang} A^i = n - \dim \ker A^i = n - \dim \mathcal{K}^i(A)$. Der Raum $\mathcal{J}^k(A)$ hört also beim gleichen i auf, kleiner zu werden, bei dem auch $\mathcal{K}^i(A)$ aufhört, grösser zu werden. □

Satz 5.3. *Die Zahl k in Satz 5.2 ist nicht grösser als n, also*

$$\mathcal{K}^n(A) = \mathcal{K}^l(A) \qquad \textit{und} \qquad \mathcal{J}^n(A) = \mathcal{J}^l(A)$$

für $l \geq n$.

Beweis. Nach Satz 5.2 muss die Dimension von $\mathcal{K}^i(A)$ in jedem Schritt um mindestens 1 zunehmen, das ist nur möglich, bis zur Dimension n. Somit können sich $\mathcal{K}^i(A)$ und $\mathcal{J}^i(A)$ für $i > n$ nicht mehr ändern. □

Abbildung 5.3 zeigt die Abhängigkeit der Dimensionen $\dim \mathcal{K}^k(A)$ und $\dim \mathcal{J}^k(A)$ von k. Die Dimension $\dim \mathcal{J}^k(A)$ nimmt ab bis zu $k = l$, danach ändert sie sich nicht mehr und die Einschränkung von A auf $\mathcal{J}^l(A)$ ist injektiv. Die Dimension $\dim \mathcal{K}^k(A)$ nimmt zu bis zu $k = l$, danach ändert sie sich nicht mehr.

Definition 5.4. *Die gemäss Satz 5.2 identischen Unterräume $\mathcal{K}^i(A)$ für $i \geq k$ und die identischen Unterräume $\mathcal{J}^i(A)$ für $i \geq k$ werden mit*

$$\begin{aligned} \mathcal{K} &= \mathcal{K}^i(A) \quad \forall i \geq k \qquad \textit{und} \\ \mathcal{J} &= \mathcal{J}^i(A) \quad \forall i \geq k \end{aligned}$$

bezeichnet.

5.1.2 Invariante Unterräume

Kern und Bild sind der erste Schritt zu einem besseren Verständnis einer linearen Abbildung oder ihrer Matrix. Invariante Räume dienen dazu, eine lineare Abbildung in einfachere Abbildungen zwischen "kleineren" Räumen zu zerlegen, wo sie leichter analysiert werden können.

Definition 5.5. *Sei $f\colon V \to V$ eine lineare Abbildung eines Vektorraums in sich selbst. Ein Unterraum $U \subset V$ heisst* invarianter Unterraum*, wenn*

$$f(U) = \{f(x) \mid x \in U\} \subset U$$

gilt.

Der Kern $\ker A$ einer linearen Abbildung ist trivialerweise ein invarianter Unterraum, da alle Vektoren in $\ker A$ auf $0 \in \ker A$ abgebildet werden. Ebenso ist natürlich $\operatorname{im} A$ ein invarianter Unterraum, denn jeder Vektor wird in $\operatorname{im} A$ abgebildet, insbesondere auch jeder Vektor in $\operatorname{im} A$.

Satz 5.6. *Sei $f\colon V \to V$ eine lineare Abbildung mit Matrix A. Jeder der Unterräume $\mathcal{J}^i(A)$ und $\mathcal{K}^i(A)$ ist ein invarianter Unterraum.*

Beweis. Sei $x \in \mathcal{K}^i(A)$, es gilt also $A^i x = 0$. Wir müssen überprüfen, dass $Ax \in \mathcal{K}^i(A)$. Wir berechnen daher $A^i \cdot Ax = A^{i+1}x = A \cdot A^i x = A \cdot 0 = 0$, was zeigt, dass $Ax \in \mathcal{K}^i(A)$.

Sei jetzt $x \in \mathcal{J}^i(A)$, es gibt also ein $y \in V$ derart, dass $A^i y = x$. Wir müssen überprüfen, dass $Ax \in \mathcal{J}^i(A)$. Dazu berechnen wir $Ax = AA^i y = A^i Ay \in \mathcal{J}^i(A)$, Ax ist also das Bild von Ay unter A^i. □

Korollar 5.7. *Die Unterräume $\mathcal{K}(A) \subset V$ und $\mathcal{J}(A) \subset V$ sind invariante Unterräume.*

Die beiden Unterräume $\mathcal{K}(A)$ und $\mathcal{J}(A)$ sind besonders interessant, da wir aus der Einschränkung der Abbildung f auf diese Unterräume mehr über f lernen können.

Satz 5.8. *Die Einschränkung von f auf $\mathcal{J}(A)$ ist injektiv.*

Beweis. Die Einschränkung von f auf $\mathcal{J}^k(A)$ ist $\mathcal{J}^k(A) \to \mathcal{J}^{k+1}(A)$, nach Definition von $\mathcal{J}^{k+1}(A)$ ist diese Abbildung surjektiv. Da aber $\mathcal{J}^k(A) = \mathcal{J}^{k+1}(A)$ ist, ist $f\colon \mathcal{J}^k(A) \to \mathcal{J}^k(A)$ surjektiv, also ist f auf $\mathcal{J}^k(A)$ auch injektiv. □

Die beiden Unterräume $\mathcal{J}(A)$ und $\mathcal{K}(A)$ sind Bild und Kern der iterierten Abbildung mit Matrix A^k. Das bedeutet, dass $\dim \mathcal{J}(A) + \mathcal{K}(A) = n$. Da $\mathcal{K}(A) = \ker A^k$ und andererseits A injektiv ist auf $\mathcal{J}(A)$, muss $\mathcal{J}(A) \cap \mathcal{K}(A) = 0$. Es folgt, dass $V = \mathcal{J}(A) + \mathcal{K}(A)$.

In $\mathcal{K}(A)$ und $\mathcal{J}(A)$ kann man unabhängig voneinander jeweils eine Basis wählen. Die Basen von $\mathcal{K}(A)$ und $\mathcal{J}(A)$ zusammen ergeben eine Basis von V. Die Matrix A' in dieser Basis wird die Blockform

$$A' = \left(\begin{array}{c|c} A_{\mathcal{K}'} & \\ \hline & A_{\mathcal{J}'} \end{array}\right)$$

haben, wobei die Matrix $A'_{\mathcal{J}}$ invertierbar ist. Die Zerlegung in invariante Unterräume ergibt also eine natürlich Aufteilung der Matrix A in kleiner Matrizen mit zum Teil bekannten Eigenschaften.

5.1.3 Nilpotente Matrizen

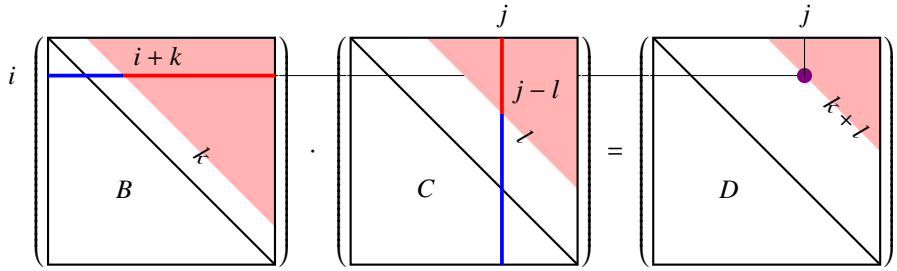
Die Zerlegung von V in die beiden invarianten Unterräume $\mathcal{J}(A)$ und $\mathcal{K}(A)$ reduziert die lineare Abbildung auf zwei Abbildungen mit speziellen Eigenschaften. Es wurde bereits in Satz gezeigt, dass die Einschränkung auf $\mathcal{J}(A)$ injektiv ist. Die Einschränkung auf $\mathcal{K}(A)$ bildet nach Definition alle Vektoren nach k-facher Iteration auf 0 ab, $A^k\mathcal{K}(A) = 0$. Solche Abbildungen haben eine speziellen Namen.

Definition 5.9. *Eine Matrix A heisst nilpotent, wenn es eine Zahl k gibt, so dass $A^k = 0$.*

Beispiel. Obere (oder untere) Dreiecksmatrizen mit Nullen auf der Diagonalen sind nilpotent. Wir rechnen dies wie folgt nach. Die Matrix A mit Einträgen a_{ij}

$$A = \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1n} \\ 0 & 0 & a_{23} & \dots & a_{1,n-1} & a_{2n} \\ 0 & 0 & 0 & \dots & a_{1,n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & a_{n-1,n} \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

erfüllt $a_{ij} = 0$ für $i \geq j$. Wir zeigen jetzt, dass sich bei der Multiplikation die nicht verschwinden Elemente bei der Multiplikation noch rechts oben verschieben. Dazu multiplizieren wir zwei Matrizen B und C mit $b_{ij} = 0$ für $i + k > j$ und $c_{ij} = 0$ für $i + l > j$. In der folgenden graphischen Darstellung der Matrizen sind die Bereiche, wo die Matrixelemente verschwinden, weiss.



Bei der Berechnung des Elementes d_{ij} wird die Zeile i von B mit der Spalte j von C multipliziert. Die blau eingefärbten Elemente in dieser Zeile und Spalte sind 0. Aus der Darstellung ist abzulesen,

dass das Produkt verschwindet, die roten, von 0 verschiedenen Elemente von den blauen Elementen annihiliert werden. Dies passiert immer, wenn $i + k > j - l$ ist, oder $i + (k + l) > j$.

Wir wenden diese Beobachtung jetzt auf die Potenzen A^s an. Für die Matrixelemente von A^s schreiben wir a^s_{ij}. Wir behaupten, dass die Matrixelemente A^s die Bedingung $a^s_{ij} = 0$ für $i + s > j$ erfüllen. Dies ist für $s = 1$ nach Voraussetzung richtig, dies ist die Induktionsvoraussetzung. Nehmen wir jetzt an, dass $a^s_{ij} = 0$ für $i + s > j$, dann folgt aus obiger Rechnung, dass $a^{s+1}_{ij} = 0$ für $i + s + 1 > j$, so dass die Bedingung auch für A^s gilt (Induktionsschritt). Mit vollständiger Induktion folgt, dass $a^s_{ij} = 0$ für $i + s > j$. Insbesondere ist $A^n = 0$, die Matrix A ist nilpotent. ◯

Man kann die Konstruktion der Unterräume $\mathcal{K}^i(A)$ weiter dazu verwenden, eine Basis zu finden, in der eine nilpotente Matrix eine besonders einfach Form erhält.

Satz 5.10. *Sei A eine nilpotente $n \times n$-Matrix mit der Eigenschaft, dass $A^{n-1} \neq 0$. Dann gibt es eine Basis so, dass A die Form*

$$A' = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & & & \\ & & & \ddots & 1 & \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix} \tag{5.3}$$

bekommt.

Beweis. Da $A^{n-1} \neq 0$ ist, gibt es einen Vektor b_n derart, dass $A^{n-1}b_n \neq 0$. Wir konstruieren die Vektoren

$$b_n,\ b_{n-1} = Ab_n,\ b_{n-2} = Ab_{n-1},\ \ldots,\ b_2 = Ab_3,\ b_1 = Ab_2.$$

Aus der Konstruktion folgt $b_1 = A^{n-1}b_n \neq 0$, aber $Ab_1 = A^n b_n = 0$. Aus der Konstruktion der iterierten Kerne $\mathcal{K}^i(A)$ folgt jetzt, dass die Vektoren $b_1, \ldots, b_n$ eine Basis bilden. In dieser Basis hat die Matrix die Form 5.3. □

Definition 5.11. *Wir bezeichnen mit N_n eine Matrix der Form* (5.3)*.*

Mit etwas mehr Sorgfalt kann man auch die Bedingung, dass $A^{n-1} \neq 0$ sein muss, im Satz 5.10 loswerden.

Satz 5.12. *Sei A ein nilpotente Matrix, dann gibt es eine Basis, in der die Matrix aus lauter Nullen besteht ausser in den Einträgen unmittelbar oberhalb der Hauptdiagonalen, wo die Einträge* 0 *oder* 1 *sind. Insbesondere zerfällt eine solche Matrix in Blöcke der Form N_{k_i}, $i = 1, \ldots, l$, wobei $k_1 + \cdots + k_l = n$ sein muss:*

$$A' = \begin{pmatrix} N_{k_1} & & & \\ & N_{k_2} & & \\ & & \ddots & \\ & & & N_{k_l} \end{pmatrix} \tag{5.4}$$

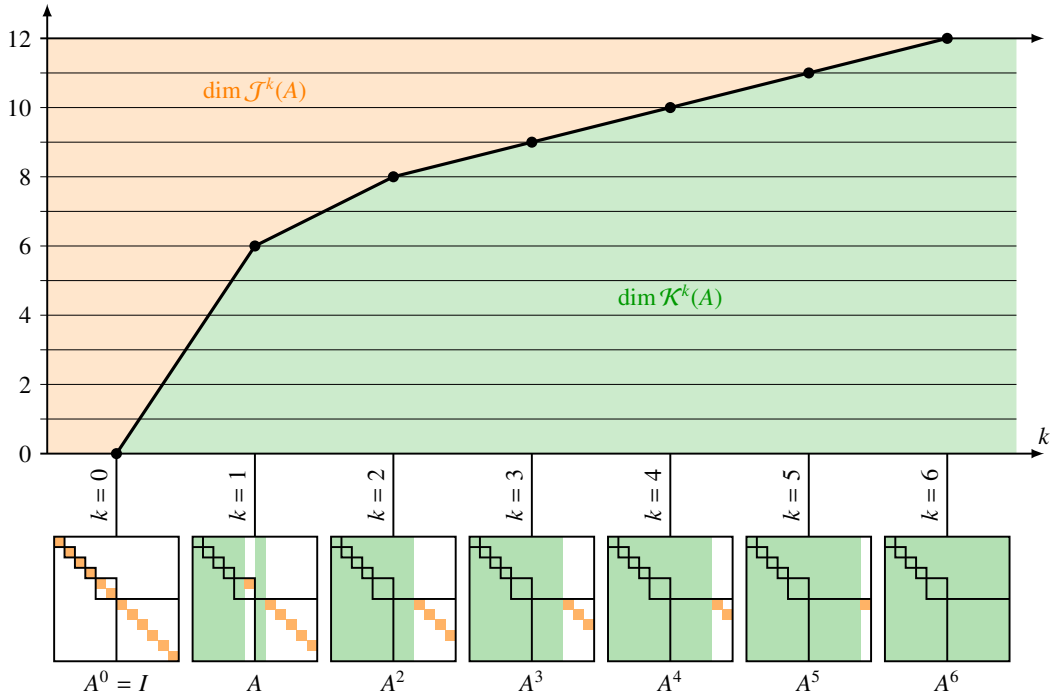


Abbildung 5.4: Entwicklung der Dimensionen von Kern und Bild von A^k in Abhängigkeit von k

Die Einschränkung von f auf den invarianten Unterraum $\mathcal{K}(A)$ ist nilpotent. Die Zerlegung $V = \mathcal{J}(A) \oplus \mathcal{K}(A)$ führt also zu einer Zerlegung der Abbildung f in eine invertierbare Abbildung $\mathcal{J}(A) \to \mathcal{J}(A)$ und eine nilpotente Abbildung $\mathcal{K}(A) \to \mathcal{K}(A)$. Nach Satz 5.12 kann man in $\mathcal{K}(A)$ eine Basis so wählen, dass die Matrix die Blockform (5.6) erhält.

Beispiel. In der Abbildung 5.4 sind die Dimensionen von Kern und Bild der Matrix

$$A = \begin{pmatrix} 0 & & & & & & & & & & & \\ & 0 & & & & & & & & & & \\ & & 0 & & & & & & & & & \\ & & & 0 & & & & & & & & \\ & & & & 0 & 1 & & & & & & \\ & & & & & 0 & & & & & & \\ & & & & & & 0 & 1 & & & & \\ & & & & & & & 0 & 1 & & & \\ & & & & & & & & 0 & 1 & & \\ & & & & & & & & & 0 & 1 & \\ & & & & & & & & & & 0 & \end{pmatrix}$$

dargestellt. Die Matrix A^k ist in den kleinen Quadraten am unteren Rand der Matrix symbolisch dargestellt. Grüne Spalten bestehen aus lauter Nullen, die zugehörigen Standardbasisvektoren werden von diesem A^k auf 0 abgebildet. Die orangen Felder enthalten Einsen, die entsprechenden Standardbasisvektoren bilden daher eine Basis des Bildes von A^k. ○

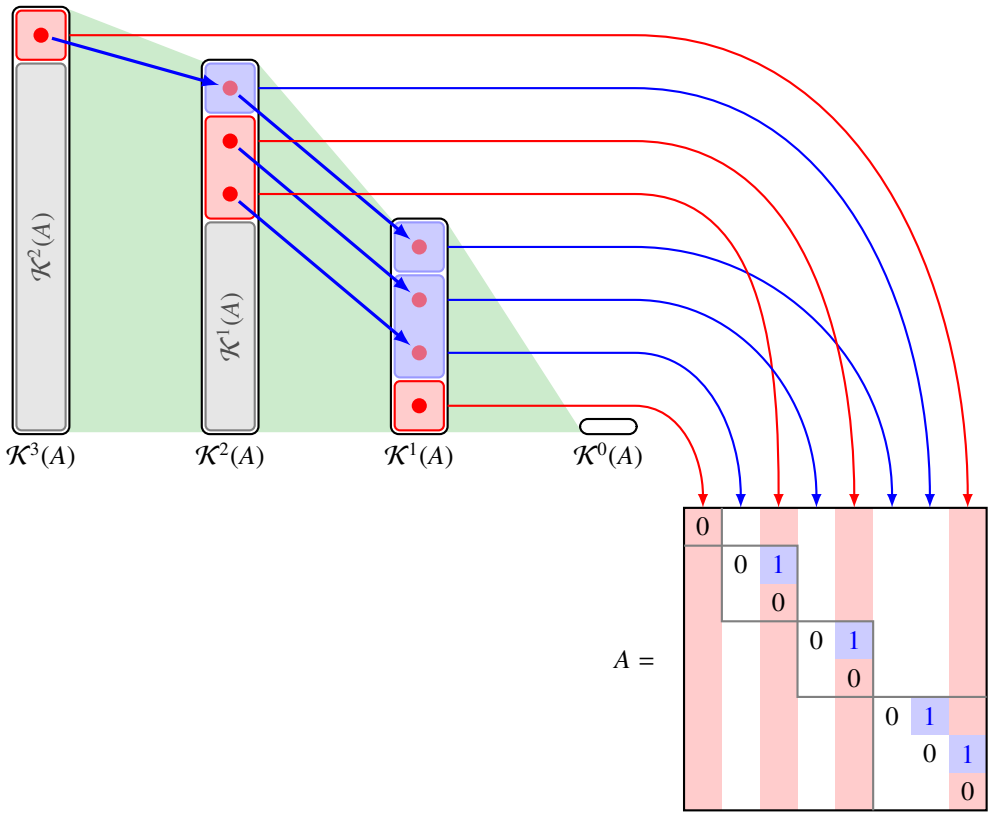


Abbildung 5.5: Konstruktion der Basis für die Jordansche Normalform einer nilpotenten Matrix. Die Vektoren werden in der Reihenfolge von rechts nach links in die Matrix gefüllt.

5.1.4 Basis für die Normalform einer nilpotenten Matrix bestimmen

Die Zerlegung in die invarianten Unterräume $\mathcal{J}^k(f)$ und $\mathcal{K}^k(f)$ ermöglichen, eine Basis zu finden, in der die Matrix von f die Blockform (5.6) hat. In diesem Abschnitt soll die Konstruktion einer solchen Basis etwas ausführlicher beschrieben werden.

Abbildung 5.5 illustriert den Prozess an einer nilpotenten Matrix A mit $A^3 = 0$ Die vertikalen Rechtecke im linken Teil der Abbildung symbolisieren die Unterräume $\mathcal{K}^k(A)$. Es ist bekannt, dass $\mathcal{K}^k(A) \subset \mathcal{K}^{k+1}(A)$ ist, die Einbettung wird in der Abbildung durch graue Rechtecke dargestellt. Es sei wieder l der Exponent, für den $\mathcal{K}^l(A) = \mathbb{k}^n$ wird. Da $\mathcal{K}^{l-1}(A) \neq \mathcal{K}^l(A)$ ist, muss es einen komplementären Unterraum geben, in dem eine Basis gewählt wird. Jeder der Vektoren $b_1, \ldots, b_s$ dieser Basis gibt Anlass zu einem Block der Form N_l, der auf dem Unterraum $\langle b_i, Ab_i, \ldots, A^{l-1}b_i\rangle$ operiert. In der Abbildung ist b_i durch einen roten Punkt symbolisiert und die Bilder $Ab_i, \ldots, A^{l-1}b_i$ werden durch blaue Pfeile untereinander verbunden.

Der Raum $\mathcal{K}^{l-1}(A)$ enthält dann $\mathcal{K}^{l-2}(A)$ und die Vektoren $Ab_1, \ldots, Ab_s$. Es ist aber möglich, dass diese Vektoren nicht den ganzen Raum $\mathcal{K}^{l-1}(A)$ erzeugen. In diesem Fall lassen sich die Vektoren mit Hilfe weiterer Vektoren $b_{s+1}, \ldots, b_{s+r}$ zu einer Basisi von $\mathcal{K}^{l-1}(A)$ ergänzen. Wie vorhin gibt jeder der Vektoren b_{s+i} Anlass zu einem Block der Form N_{l-1}, der auf dem Unterraum $\langle b_{s+i}, Ab_{s+i} \ldots, A^{l-2}b_{s+i}\rangle$ operiert.

Durch Wiederholung dieses Prozesses können schrittweise Basisvektoren b_i erzeugt werden. Die Matrix der Abbildung f in der Basis $\{b_i, Ab_i, \dots, A^k b_i\}$ ist ein Block der Form N_k. Für $0 \le k \le l-1$ sind die Vektoren $A^k b_i$, solange sie von 0 verschieden sind, alle nach Konstruktion linear unabhängig, sie bilden eine Basis von $\mathcal{K}^l(A) = \mathbb{R}^n$.

Beispiel. Die Basis für die Zerlegung der Matrix

$$A = \begin{pmatrix} 3 & 1 & -2 \\ -21 & -7 & 14 \\ -6 & -2 & 4 \end{pmatrix}$$

in Blockform soll nach der oben beschriebenen Methode ermittelt werden. Zunächst kann man nachrechnen, dass $A^2 = 0$ ist. Der Kern von A ist der Lösungsraum der Gleichung $Ax = 0$, da alle Zeilen Vielfache der ersten Zeile sind, recht es zu verlangen, dass die Komponenten x_i der Lösung die Gleichung

$$3x_1 + x_2 - 2x_3 = 0$$

erfüllen. Jetzt muss ein Vektor b_1 ausserhalb von $\mathbb{L}$ gefunden werden, der erste Standardbasisvektor e_1 kann dazu verwendet werden. Es ist auch klar, dass $Ae_1 \neq 0$ ist. Wir verwenden daher die beiden Vektoren

$$b_3 = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad b_2 = Ab_3 = \begin{pmatrix} 3 \\ -21 \\ -6 \end{pmatrix},$$

in dieser Basis hat A die Matrix N_2. Jetzt muss noch ein Basisvektor b_1 gefunden werden, der in $\ker A = \mathbb{L}$ liegt und so, dass b_1 und b_2 linear unabhängig sind. Die zweite Bedingung kann leicht dadurch sichergestellt werden, dass man die erste Komponente von b_1 als 0 wählt. Eine mögliche Lösung ist dann

$$b_1 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$$

Die Matrix

$$B = \begin{pmatrix} 0 & 1 & 3 \\ 2 & 0 & -21 \\ 1 & 0 & -6 \end{pmatrix} \qquad \text{mit Inverser} \qquad B^{-1} = \begin{pmatrix} 0 & -\frac{2}{3} & \frac{7}{3} \\ 0 & -\frac{1}{9} & \frac{2}{9} \\ 1 & \frac{1}{3} & -\frac{2}{3} \end{pmatrix}$$

transformiert die Matrix A auf den Block N_3:

$$B^{-1}AB = B^{-1} \begin{pmatrix} 0 & 0 & 3 \\ 0 & 0 & -21 \\ 0 & 0 & -6 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = N_3. \qquad \bigcirc$$

5.1.5 Eigenwerte und Eigenvektoren

In diesem Abschnitt betrachten wir Vektorräume $V = \mathbb{k}^n$ über einem beliebigen Körper $\mathbb{k}$ und quadratische Matrizen $A \in M_n(\mathbb{k})$. In den meisten Anwendungen wird $\mathbb{k} = \mathbb{R}$ sein. Da aber in $\mathbb{R}$ nicht alle algebraischen Gleichungen lösbar sind, ist es manchmal notwendig, den Vektorraum zu erweitern um zum Beispiel Eigenschaften der Matrix A abzuleiten.

Definition 5.13. *Ein Vektor $v \in V$ heisst* Eigenvektor *von A zum Eigenwert $\lambda \in \mathbb{k}$, wenn $v \neq 0$ und $Av = \lambda v$ gilt.*

Die Bedingung $v \neq 0$ dient dazu, pathologische Situationen auszuschliessen. Für den Nullvektor gilt $A0 = \lambda 0$ für jeden beliebigen Wert von $\lambda \in \mathbb{k}$. Würde man $v = 0$ zulassen, wäre jede Zahl in $\mathbb{k}$ ein Eigenwert, ein Eigenwert von A wäre nichts besonderes. Ausserdem wäre 0 ein Eigenvektor zu jedem beliebigen Eigenwert.

Eigenvektoren sind nicht eindeutig bestimmt, jedes von 0 verschiedene Vielfache von v ist ebenfalls ein Eigenvektor. Zu einem Eigenwert kann man also einen Eigenvektor jeweils mit geeigneten Eigenschaften finden, zum Beispiel kann man für $\mathbb{k} = \mathbb{R}$ Eigenvektoren auf Länge 1 normieren. Im Folgenden werden wir oft die abkürzend linear unabhängige Eigenvektoren einfach als "verschiedene" Eigenvektoren bezeichnen.

Wenn v ein Eigenvektor von A zum Eigenwert λ ist, dann kann man ihn mit zusätzlichen Vektoren $v_2, \dots, v_n$ zu einer Basis $\mathcal{B} = \{v, v_2, \dots, v_n\}$ von V ergänzen. Die Vektoren v_k mit $k = 2, \dots, n$ werden von A natürlich auch in den Vektorraum V abgebildet, können also als Linearkombinationen

$$Av = a_{1k}v + a_{2k}v_2 + a_{3k}v_3 + \dots a_{nk}v_n$$

dargestellt werden. In der Basis $\mathcal{B}$ bekommt die Matrix A daher die Form

$$A' = \begin{pmatrix} \lambda & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}.$$

Bereits ein einzelner Eigenwert und ein zugehöriger Eigenvektor ermöglichen uns also, die Matrix in eine etwas einfachere Form zu bringen.

Definition 5.14. *Für $\lambda \in \mathbb{k}$ heisst*

$$E_\lambda = \{v \mid Av = \lambda v\}$$

der Eigenraum *zum Eigenwert λ.*

Der Eigenraum E_λ ist ein Unterraum von V, denn wenn $u, v \in E_\lambda$, dann ist

$$A(su + tv) = sAu + tAv = s\lambda u + t\lambda v = \lambda(su + tv),$$

also ist auch $su + tv \in E_\lambda$. Der Fall $E_\lambda = \{0\} = 0$ bedeutet natürlich, dass λ gar kein Eigenwert ist.

Satz 5.15. *Wenn* $\dim E_\lambda = n$*, dann ist* $A = \lambda E$.

Beweis. Da V ein n-dimensionaler Vektoraum ist, ist $E_\lambda = V$. Jeder Vektor $v \in V$ erfüllt also die Bedingung $Av = \lambda v$, oder $A = \lambda E$. □

Wenn man die Eigenräume von A kennt, dann kann man auch die Eigenräume von $A + \mu E$ berechnen. Ein Vektor $v \in E_\lambda$ erfüllt

$$Av = \lambda v \qquad \Rightarrow \qquad (A + \mu)v = \lambda v + \mu v = (\lambda + \mu)v,$$

somit ist v ein Eigenvektor von $A + \mu E$ zum Eigenwert $\lambda + \mu$. Insbesondere können wir statt die Eigenvektoren von A zum Eigenwert λ zu studieren, auch die Eigenvektoren zum Eigenwert 0 von $A - \lambda E$ untersuchen.

5.1.6 Verallgemeinerte Eigenräume

Wenn λ ein Eigenwert der Matrix A ist, dann ist ist $A - \lambda E$ injektiv und $\ker(A - \lambda E) \neq 0$. Man kann daher die invarianten Unterräume $\mathcal{K}(A - \lambda E)$ und $\mathcal{J}(A - \lambda E)$.

Beispiel. Wir untersuchen die Matrix

$$A = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 0 & 3 & -1 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Man kann zeigen, dass $\lambda = 1$ ein Eigenwert ist. Wir suchen die Zerlegung des Vektorraums $\mathbb{R}^4$ in invariante Unterräume $\mathcal{K}(A - E)$ und $\mathcal{J}(A - E)$. Die Matrix $B = A - E$ ist

$$B = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 2 & -1 & 1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

und wir berechnen davon die Potenz

$$D = B^4 = (A - E)^4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 4 \\ 0 & 2 & -1 & 4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Daraus kann man ablesen, dass das Bild im D von D die Basis

$$b_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \qquad b_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

hat. Für den Kern von D können wir zum Beispiel die Basisvektoren

$$b_3 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \qquad b_4 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

verwenden.

Als erstes überprüfen wir, ob diese Basisvektoren tatsächlich invariante Unterräume sind. Für $\mathcal{J}(A - E) = \langle b_1, b_2 \rangle$ berechnen wir

$$(A - E)b_1 = \begin{pmatrix} 0 \\ 4 \\ 4 \\ 1 \end{pmatrix} = 4b_2 + b_1,$$

$$(A - E)b_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} = b_2.$$

Dies beweist, dass $\mathcal{J}(A-E)$ invariant ist. In dieser Basis hat die von $A-E$ beschriebene lineare Abbildung auf $\mathcal{J}(A-E)$ die Matrix

$$A_{\mathcal{J}(A-E)} = \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}.$$

Für den Kern $\mathcal{K}(A-E)$ findet man analog

$$\left.\begin{aligned} Ab_3 &= -b_4 \\ Ab_4 &= 0 \end{aligned}\right\} \quad \Rightarrow \quad A_{\mathcal{K}(A-E)} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

In der Basis $\mathcal{B} = \{b_1, b_2, b_3, b_4\}$ hat A die Matrix in Blockform

$$A' = \left(\begin{array}{cc|cc} 2 & 4 & & \\ 0 & 2 & & \\ \hline & & 1 & -1 \\ & & 0 & 1 \end{array}\right),$$

die Blöcke gehören zu den invarianten Unterräumen $\mathcal{K}(A-E)$ und $\mathcal{K}(A-E)$. Die aus $A-E$ gewonnen invarianten Unterräume sind offenbar auch invariante Unterräume für A. ○

Definition 5.16. *Ist A eine Matrix mit Eigenwert λ, dann heisst der invariante Unterraum*

$$\mathcal{E}_\lambda(A) = \mathcal{K}(A-\lambda E)$$

der verallgemeinerte Eigenraum von A.

Es ist klar, dass $E_\lambda(A) = \ker(A-\lambda E) \subset \mathcal{E}_\lambda(A)$.

5.1.7 Zerlegung in invariante Unterräume

Wenn λ kein Eigenwert von A ist, dann ist $A-\lambda E$ injektiv und damit $\ker(A-\lambda E) = 0$. Es folgt, dass $\mathcal{K}^i(A-\lambda E) = 0$ und daher auch $\mathcal{J}^i(A-\lambda E) = V$. Die Zerlegung in invariante Unterräume $\mathcal{J}(A-\lambda E)$ und $\mathcal{K}(A-\lambda E)$ liefert in diesem Falle also nichts Neues.

Für einen Eigenwert λ_1 von A dagegen, erhalten wir die Zerlegung

$$V = \mathcal{E}_{\lambda_1}(A) \oplus \underbrace{\mathcal{J}(A-\lambda_1 E)}_{=V_2},$$

wobei $\mathcal{E}_{\lambda_1}(A) \neq 0$ ist. Die Matrix $A-\lambda_1 E$ ist eingeschränkt auf $\mathcal{E}_{\lambda_1}(A)$ nilpotent. Die Zerlegung in invariante Unterräume ist zwar mit Hilfe von $A-\lambda_1 E$ gewonnen worden, ist aber natürlich auch eine Zerlegung in invariante Unterräume für A. Wir können daher das Problem auf V_2 einschränken und nach einem weiteren Eigenwert λ_2 von A in V_2 suchen, was wieder eine Zerlegung in invariante Unterräume liefert. Indem wir so weiterarbeiten, bis wir den ganzen Raum ausgeschöpft haben, können wir eine Zerlegung des ganzen Raumes V finden, so dass A auf jedem einzelnen Summanden eine sehr einfach Form hat:

Satz 5.17. *Sei V ein $\mathbb{k}$-Vektorraum und f eine lineare Abbildung mit Matrix A derart, dass alle Eigenwerte $\lambda_1, \ldots, \lambda_l$ von A in $\mathbb{k}$ sind. Dann gibt es eine Zerlegung von V in verallgemeinerte Eigenräume*

$$V = \mathcal{E}_{\lambda_1}(A) \oplus \mathcal{E}_{\lambda_2}(A) \oplus \cdots \oplus \mathcal{E}_{\lambda_l}(A).$$

Die Einschränkung von $A-\lambda_i E$ auf den Eigenraum $\mathcal{E}_{\lambda_i}(A)$ ist nilpotent.

5.1.8 Das charakteristische Polynom

Ein Eigenvektor von A erfüllt $Av = \lambda v$ oder gleichbedeutend $(A - \lambda E)v = 0$, er ist also eine nichttriviale Lösung des homogenen Gleichungssystems mit Koeffizientenmatrix $A - \lambda E$. Ein Eigenwert ist also ein Skalar derart, dass $A - \lambda E$ singulär ist. Ob eine Matrix singulär ist, kann mit der Determinante festgestellt werden. Die Eigenwerte einer Matrix A sind daher die Nullstellen von $\det(A - \lambda E)$.

Definition 5.18. *Das* charakteristische Polynom

$$\chi_A(x) = \det(A - xE) = \begin{vmatrix} a_{11} - x & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - x & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - x \end{vmatrix}.$$

der Matrix A ist ein Polynom vom Grad n mit Koeffizienten in $\mathbb{k}$.

Findet man eine Nullstelle $\lambda \in \mathbb{k}$ von $\chi_A(x)$, dann ist die Matrix $A - \lambda E \in M_n(\mathbb{k})$ und mit dem Gauss-Algorithmus kann man auch mindestens einen Vektor $v \in \mathbb{k}^n$ finden, der $Av = \lambda v$ erfüllt. Eine Matrix der Form wie in Satz **??** hat

$$\chi_A(x) = \begin{vmatrix} \lambda - x & 1 & & & & \\ & \lambda - x & 1 & & & \\ & & \lambda - x & & & \\ & & & \ddots & & \\ & & & & \lambda - x & 1 \\ & & & & & \lambda - x \end{vmatrix} = (\lambda - x)^n = (-1)^n (x - \lambda)^n$$

als charakteristisches Polynom, welches λ als einzige Nullstelle hat. Der Eigenraum der Matrix ist aber nur eindimensional, man kann also im Allgemeinen für jede Nullstelle des charakteristischen Polynoms nicht mehr als einen Eigenvektor (d. h. einen eindimensionalen Eigenraum) erwarten.

Wenn das charakteristische Polynom von A keine Nullstellen in $\mathbb{k}$ hat, dann kann es auch keine Eigenvektoren in $\mathbb{k}^n$ geben. Gäbe es nämlich einen solchen Vektor, dann müsste eine der Komponenten des Vektors von 0 verschieden sein, wir nehmen an, dass es die Komponente in Zeile k ist. Die Komponente v_k kann man auf zwei Arten berechnen, einmal als die k-Komponenten von Av und einmal als k-Komponente von λv:

$$a_{k1}v_1 + \dots + a_{kn}v_n = \lambda v_k.$$

Da $v_k \neq 0$ kann man nach λ auflösen und erhält

$$\lambda = \frac{a_{k1}v_1 + \dots + a_{kn}v_n}{v_k}.$$

Alle Terme auf der rechten Seite sind in $\mathbb{k}$ und werden nur mit Körperoperationen in $\mathbb{k}$ verknüpft, also muss auch $\lambda \in \mathbb{k}$ sein, im Widerspruch zur Annahme.

Durch Hinzufügen von geeigneten Elementen können wir immer zu einem Körper $\mathbb{k}'$ übergehen, in dem das charakteristische Polynom in Linearfaktoren zerfällt. In diesem Körper kann man jetzt das homogene lineare Gleichungssystem mit Koeffizientenmatrix $A - \lambda E$ lösen und damit mindestens einen Eigenvektor v für jeden Eigenwert finden. Die Komponenten von v liegen in $\mathbb{k}'$, und mindestens eine davon kann nicht in $\mathbb{k}$ liegen. Das bedeutet aber nicht, dass man diese Vektoren nicht für theoretische Überlegungen über von $\mathbb{k}'$ unabhängige Eigenschaften der Matrix A machen. Das folgende Beispiel soll diese Idee illustrieren.

Beispiel. Wir arbeiten in diesem Beispiel über dem Körper $\mathbb{k} = \mathbb{Q}$. Die Matrix

$$A = \begin{pmatrix} -4 & 7 \\ -2 & 4 \end{pmatrix} \in M_2(\mathbb{Q})$$

hat das charakteristische Polynom

$$\chi_A(x) = \begin{vmatrix} -4-x & 7 \\ -2 & 4-x \end{vmatrix} = (-4-x)(4-x) - 7 \cdot (-2) = -16 + x^2 + 14 = x^2 - 2.$$

Die Nullstellen sind $\pm\sqrt{2}$ und damit nicht in $\mathbb{Q}$. Wir gehen daher über zum Körper $\mathbb{Q}(\sqrt{2})$, in dem sich zwei Nullstellen $\lambda = \pm\sqrt{2}$ finden lassen. Zu jedem Eigenwert lässt sich auch ein Eigenvektor $v_{\pm\sqrt{2}} \in \mathbb{Q}(\sqrt{2})^2$, und unter Verwendung dieser Basis bekommt die Matrix $A' = TAT^{-1}$ Diagonalform. Die Transformationsmatrix T enthält Matrixelemente aus $\mathbb{Q}(\sqrt{2})$, die nicht in $\mathbb{Q}$ liegen. Die Matrix A lässt sich also über dem Körper $\mathbb{Q}(\sqrt{2})$ diagonalisieren, nicht aber über dem Körper $\mathbb{Q}$.

Da A' Diagonalform hat mit $\pm\sqrt{2}$ auf der Diagonalen, folgt $A'^2 = 2E$, die Matrix A' erfüllt also die Gleichung

$$A'^2 - E = \chi_A(A) = 0. \tag{5.5}$$

Dies is ein Spezialfall des Satzes von Cayley-Hamilton **??** welcher besagt, dass jede Matrix A eine Nullstelle ihres charakteristischen Polynoms ist: $\chi_A(A) = 0$. Die Gleichung 5.5 wurde zwar in $\mathbb{Q}(\sqrt{2})$ hergeleitet, aber in ihr kommen keine Koeffizienten aus $\mathbb{Q}(\sqrt{2})$ vor, die man nicht auch in $\mathbb{Q}$ berechnen könnte. Sie gilt daher ganz allgemein. ○

Beispiel. Die Matrix

$$A = \begin{pmatrix} 32 & -41 \\ 24 & -32 \end{pmatrix} \in M_2(\mathbb{R})$$

über dem Körper $\mathbb{k} = \mathbb{R}$ hat das charakteristische Polynom

$$\det(A - xE) = \begin{vmatrix} 32-x & -41 \\ 25 & -32-x \end{vmatrix} = (32-x)(-32-x) - 25 \cdot (-41) = x^2 - 32^2 + 1025 = x^2 + 1.$$

Die charakteristische Gleichung $\chi_A(x) = 0$ hat in $\mathbb{R}$ keine Lösungen, daher gehen wir zum Körper $\mathbb{k}' = \mathbb{C}$ über, in dem dank dem Fundamentalsatz der Algebra alle Nullstellen zu finden sind, sie sind $\pm i$. In $\mathbb{C}$ lassen sich dann auch Eigenvektoren finden, man muss dazu die folgenden linearen Gleichungssyteme lösen:

$$\left|\begin{array}{cc} 32-i & -41 \\ 25 & -32-i \end{array}\right| \to \left|\begin{array}{cc} 1 & t \\ 0 & 0 \end{array}\right| \qquad \left|\begin{array}{cc} 32+i & -41 \\ 25 & -32+i \end{array}\right| \to \left|\begin{array}{cc} 1 & \bar{t} \\ 0 & 0 \end{array}\right|,$$

wobei wir $t = -41/(32-i) = -41(32+i)/1025 = -1.28 - 0.04i = (64-1)/50$ abgekürzt haben. Die zugehörigen Eigenvektoren sind

$$v_i = \begin{pmatrix} t \\ i \end{pmatrix} \qquad v_{-i} = \begin{pmatrix} \bar{t} \\ i \end{pmatrix}$$

Mit den Vektoren v_i und v_{-i} als Basis kann die Matrix A als komplexe Matrix, also mit komplexem T in die komplexe Diagonalmatrix $A' = \mathrm{diag}(i, -i)$ transformiert werden. Wieder kann man sofort ablesen, dass $A'^2 + E = 0$, und wieder kann man schliessen, dass für die relle Matrix A ebenfalls $\chi_A(A) = 0$ gelten muss. ○

5.2 Normalformen

In den Beispielen im vorangegangenen wurde wiederholt der Trick verwendet, den Koeffizientenkörper so zu erweitern, dass das charakteristische Polynom in Linearfaktoren zerfällt und für jeden Eigenwert Eigenvektoren gefunden werden können. Diese Idee ermöglicht, eine Matrix in einer geeigneten Körpererweiterung in eine besonders einfache Form zu bringen, das Problem dort zu lösen. Anschliessend kann man sich darum kümmern in welchem Mass die gewonnenen Resultate wieder in den ursprünglichen Körper transportiert werden können.

5.2.1 Diagonalform

Sei A eine beliebige Matrix mit Koeffizienten in $\mathbb{k}$ und sei $\mathbb{k}'$ eine Körpererweiterung von $\mathbb{k}$ derart, dass das charakteristische Polynom in Linearfaktoren

$$\chi_A(x) = (x-\lambda_1)^{k_1} \cdot (x-\lambda_2)^{k_2} \cdot \dots \cdot (x-\lambda_m)^{k_m}$$

mit Vielfachheiten k_1 bis k_m zerfällt, $\lambda_i \in \mathbb{k}'$. Zu jedem Eigenwert λ_i gibt es sicher einen Eigenvektor, wir wollen aber in diesem Abschnitt zusätzlich annehmen, dass es eine Basis aus Eigenvektoren gibt. In dieser Basis bekommt die Matrix Diagonalform, wobei auf der Diagonalen nur Eigenwerte vorkommen können. Man kann die Vektoren so anordnen, dass die Diagonalmatrix in Blöcke der Form $\lambda_i E$ zerfällt

$$A' = \begin{pmatrix} \boxed{\lambda_1 E} & & & \\ & \boxed{\lambda_2 E} & & \\ & & \ddots & \\ & & & \boxed{\lambda_m E} \end{pmatrix}$$

Über die Grösse eines solchen $\lambda_i E$-Blockes können wir zum jetzigen Zeitpunkt noch keine Aussagen machen.

Die Matrizen $A - \lambda_k E$ enthalten jeweils einen Block aus lauter Nullen. Das Produkt all dieser Matrizen ist daher

$$(A-\lambda_1 E)(A-\lambda_2 E)\cdots(A-\lambda_m E) = 0.$$

Über dem Körper $\mathbb{k}'$ gibt es also das Polynom $m(x) = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_m)$ mit der Eigenschaft $m(A) = 0$. Dies ist auch das Polynom von kleinstmöglichem Grad, denn für jeden Eigenwert muss ein entsprechender Linearfaktor in so einem Polynom vorkommen. Das Polynom $m(x)$ ist daher das Minimalpolynom der Matrix A. Da jeder Faktor in $m(x)$ auch ein Faktor von $\chi_A(x)$ ist, folgt wieder $\chi_A(A) = 0$. Ausserdem ist über dem Körper $\mathbb{k}'$ das Polynom $m(x)$ ein Teiler des charakteristischen Polynoms $\chi_A(x)$.

5.2.2 Jordan-Normalform

Die Eigenwerte einer Matrix A können als Nullstellen des charakteristischen Polynoms gefunden werden. Da der Körper $\mathbb{k}$ nicht unbedingt algebraische abgeschlossen ist, zerfällt das charakteristische Polynom nicht unbedingt in Linearfaktoren, die Nullstellen sind nicht unbedingt in $\mathbb{k}$. Wir können aber immer zu einem grösseren Körper $\mathbb{k}'$ übergehen, in dem das charakteristische Polynom in Linearfaktoren zerfällt. Wir nehmen im Folgenden an, dass

$$\chi_A(x) = (x-\lambda_1)^{k_1} \cdot (x-\lambda_2)^{k_2} \cdot \dots \cdot (x-\lambda_l)^{k_l}$$

ist mit $\lambda_i \in \mathbb{k}'$.

Nach Satz 5.17 liefern die verallgemeinerten Eigenräume $V_i = \mathcal{E}_{\lambda_i}(A)$ eine Zerlegung von V in invariante Eigenräume

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_l,$$

derart, dass $A - \lambda_i E$ auf V_i nilpotent ist. Wählt man in jedem der Unterräume V_i eine Basis, dann zerfällt die Matrix A in Blockmatrizen

$$A' = \begin{pmatrix} \boxed{A_1} & & & \\ & \boxed{A_2} & & \\ & & \ddots & \\ & & & \boxed{A_l} \end{pmatrix} \tag{5.6}$$

wobei, A_i Matrizen mit dem einzigen Eigenwert λ_i sind.

Nach Satz 5.12 kann man in den Unterräume die Basis zusätzlich so wählen, dass die entstehenden Blöcke $A_i - \lambda_i E$ spezielle nilpotente Matrizen aus lauter Null sind, die höchstens unmittelbar über der Diagonalen Einträge 1 haben kann. Dies bedeutet, dass sich immer eine Basis so wählen lässt, dass die Matrix A_i zerfällt in sogenannte Jordan-Blöcke.

Definition 5.19. *Ein m-dimensionaler* Jordan-Block *ist eine $m \times m$-Matrix der Form*

$$J_m(\lambda) = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & & & \\ & & & \ddots & & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}.$$

Eine Jordan-Matrix *ist eine Blockmatrix Matrix*

$$J = \begin{pmatrix} \boxed{J_{m_1}(\lambda)} & & & \\ & \boxed{J_{m_2}(\lambda)} & & \\ & & \ddots & \\ & & & \boxed{J_{m_p}(\lambda)} \end{pmatrix}$$

mit $m_1 + m_2 + \cdots + m_p = m$.

Da Jordan-Blöcke obere Dreiecksmatrizen sind, ist das charakteristische Polynom eines Jordan-Blocks oder einer Jordan-Matrix besonders einfach zu berechnen. Es gilt

$$\chi_{J_m(\lambda)}(x) = \det(J_m(\lambda) - xE) = (\lambda - x)^m$$

für einen Jordan-Block $J_m(\lambda)$. Für eine $m \times m$-Jordan-Matrix J mit Blöcken $J_{m_1}(\lambda)$ bis $J_{m_p}(\lambda)$ ist

$$\chi_{J(\lambda)}(x) = \chi_{J_{m_1}(\lambda)}(x)\chi_{J_{m_2}(\lambda)}(x) \cdot \cdots \cdot \chi_{J_{m_p}(\lambda)}(x) = (\lambda - x)^{m_1}(\lambda - x)^{m_2} \cdot \cdots \cdot (\lambda - x)^{m_p} = (\lambda - x)^m.$$

Satz 5.20. *Über einem Körper $\mathbb{k}' \supset \mathbb{k}$, über dem das charakteristische Polynom $\chi_A(x)$ in Linearfaktoren zerfällt, lässt sich immer eine Basis finden derart, dass die Matrix A zu einer Blockmatrix wird, die aus lauter Jordan-Matrizen besteht. Die Dimension der Jordan-Matrix zum Eigenwert λ_i ist die Vielfachheit des Eigenwerts im charakteristischen Polynom.*

Beweis. Es ist nur noch die Aussage über die Dimension der Jordan-Blöcke zu beweisen. Die Jordan-Matrizen zum Eigenwert λ_i werden mit J_i bezeichnet und sollen $m_i \times m_i$-Matrizen sein. Das charakteristische Polynom jedes Jordan-Blocks ist dann $\chi_{J_i}(x) = (\lambda_i - x)^{m_i}$. Das charakteristische Polynom der Blockmatrix mit diesen Jordan-Matrizen als Blöcken ist das Produkt

$$\chi_A(x) = (\lambda_1 - x)^{m_1}(\lambda_2 - x)^{m_2} \cdots (\lambda_p - x)^{m_p}$$

mit $m_1 + m_2 + \cdots + m_p$. Die Blockgrösse m_i ist also auch die Vielfachheit von λ_i im charakteristischen Polynom $\chi_A(x)$. □

Satz 5.21 (Cayley-Hamilton)**.** *Ist A eine $n \times n$-Matrix über dem Körper $\mathbb{k}$, dann gilt $\chi_A(A) = 0$.*

Beweis. Zunächst gehen wir über zu einem Körper $\mathbb{k}' \supset \mathbb{k}$, indem das charakteristische Polynom $\chi_A(x)$ in Linearfaktoren $\chi_A(x) = (\lambda_1 - x)^{m_1}(\lambda_2 - x)^{m_2} \ldots (\lambda_p - x)^{m_p}$ zerfällt. Im Vektorraum $\mathbb{k}'$ kann man eine Basis finden, in der die Matrix A in Jordan-Matrizen $J_1, \ldots, J_p$ zerfällt, wobei J_i eine $m_i \times m_i$-Matrix ist. Für den Block mit der Nummer i erhalten wir $(J_i - \lambda_i E)^{m_i} = 0$. Setzt man also den Block J_i in das charakteristische Polynom $\chi_A(x)$ ein, erhält man

$$\chi_A(J_i) = (\lambda_1 E - J_1)^{m_1} \cdot \ldots \cdot \underbrace{(\lambda_i E - J_i)^{m_i}}_{=0} \cdot \ldots \cdot (\lambda_i E - J_p)^{m_p} = 0.$$

Jeder einzelne Block J_i wird also zu 0, wenn man ihn in das charakteristische Polynome $\chi_A(x)$ einsetzt. Folglich gilt auch $\chi_A(A) = 0$.

Die Rechnung hat zwar im Körper $\mathbb{k}'$ stattgefunden, aber die Berechnung $\chi_A(A)$ kann in $\mathbb{k}$ ausgeführt werden, also ist $\chi_A(A) = 0$. □

Aus dem Beweis kann man auch noch eine strengere Bedingung ableiten. Auf jedem verallgemeinerten Eigenraum $\mathcal{E}_{\lambda_i}(A)$ ist $A_i - \lambda_i$ nilpotent, es gibt also einen minimalen Exponenten q_i derart, dass $(A_i - \lambda_i E)^{q_i} = 0$ ist. Wählt man eine Basis in jedem verallgemeinerten Eigenraum derart, dass A_i eine Jordan-Matrix ist, kann man wieder zeigen, dass für das Polynom

$$m_A(x) = (x - \lambda_1 x)^{q_1}(x - \lambda_2 x)^{q_2} \cdot \ldots \cdot (x - \lambda_p x)^{q_p}$$

gilt $m_A(A) = 0$. $m_A(x)$ ist das *Minimalpolynom* der Matrix A.

Satz 5.22 (Minimalpolynom)**.** *Über dem Körper $\mathbb{k}' \subset \mathbb{k}$, über dem das charakteristische Polynom $\chi_A(x)$ in Linearfaktoren zerfällt, ist das Minimalpolynom von A das Polynom*

$$m(x) = (x - \lambda_1)^{q_1}(x - \lambda_2)^{q_2} \cdots \ldots \cdots (x - \lambda_p)^{q_p}$$

wobei q_i der kleinste Index ist, für den die q_i-te Potenz derEinschränkung von $A - \lambda_i E$ auf den verallgemeinerten Eigenraum $\mathcal{E}_{\lambda_i}(A)$ verschwindet. Es ist das Polynom geringsten Grades über $\mathbb{k}'$, welches $m(A) = 0$ erfüllt.

5.2.3 Reelle Normalform

Wenn eine reelle Matrix A komplexe Eigenwerte hat, ist die Jordansche Normalform zwar möglich, aber die zugehörigen Basisvektoren werden ebenfalls komplexe Komponenten haben. Für eine rein reelle Rechnung ist dies nachteilig, da der Speicheraufwand dadurch verdoppelt und der Rechenaufwand für Multiplikationen vervierfacht wird.

Die nicht reellen Eigenwerte von A treten in konjugiert komplexen Paaren λ_i und $\overline{\lambda}_i$ auf. Wir betrachten im Folgenden nur ein einziges Paar $\lambda = a + ib$ und $\overline{\lambda} = a - ib$ von konjugiert komplexen Eigenwerten mit nur je einem einzigen $n \times n$-Jordan-Block J und $\overline{J}$. Ist $\mathcal{B} = \{b_1, \ldots, b_n\}$ die Basis für den Jordan-Block J, dann kann man die Vektoren $\overline{\mathcal{B}} = \{\overline{b}_1, \ldots, \overline{b}_n\}$ als Basis für $\overline{J}$ verwenden. Die vereinigte Basis $C = \mathcal{B} \cup \overline{\mathcal{B}} = \{b_1, \ldots, b_n, \overline{b}_1, \ldots, \overline{b}_n\}$ erzeugen einen $2n$-dimensionalen Vektorraum, der direkte Summe der beiden von $\mathcal{B}$ und $\overline{\mathcal{B}}$ erzeugen Vektorräume $V = \langle \mathcal{B} \rangle$ und $\overline{V} = \langle \overline{\mathcal{B}} \rangle$ ist. Es ist also

$$U = \langle C \rangle = V \oplus \overline{V}.$$

Wir bezeichnen die lineare Abbildung mit den Jordan-Blöcken J und $\overline{J}$ wieder mit A.

Auf dem Vektorraum U hat die lineare Abbildung in der Basis C die Matrix

$$A = \begin{pmatrix} J & 0 \\ 0 & \overline{J} \end{pmatrix} = \begin{pmatrix} \lambda & 1 & & & & & & & & \\ & \lambda & 1 & & & & & & & \\ & & \lambda & \ddots & & & & & & \\ & & & \ddots & 1 & & & & & \\ & & & & \lambda & & & & & \\ & & & & & \overline{\lambda} & 1 & & & \\ & & & & & & \overline{\lambda} & 1 & & \\ & & & & & & & \overline{\lambda} & \ldots & \\ & & & & & & & & \ldots & 1 \\ & & & & & & & & & \overline{\lambda} \end{pmatrix}.$$

Die Jordan-Normalform bedeutet, dass

$$\begin{aligned} Ab_1 &= \lambda b_1 & A\overline{b}_1 &= \overline{\lambda b_1} \\ Ab_2 &= \lambda b_2 + b_1 & A\overline{b}_2 &= \overline{\lambda b_2} + \overline{b_1} \\ Ab_3 &= \lambda b_3 + b_2 & A\overline{b}_3 &= \overline{\lambda b_3} + \overline{b_2} \\ &\vdots & &\vdots \\ Ab_n &= \lambda b_n + b_{n-1} & A\overline{b}_n &= \overline{\lambda b_n} + \overline{b_{n-1}} \end{aligned}$$

Für die Linearkombinationen

$$c_i = \frac{b_i + \overline{b}_i}{\sqrt{2}}, \quad d_i = \frac{b_i - \overline{b}_i}{i\sqrt{2}} \tag{5.7}$$

folgt dann für $k > 1$

$$\begin{aligned} Ac_k &= \frac{Ab_k + A\overline{b}_k}{2} & Ad_k &= \frac{Ab_k - A\overline{b}_k}{2i} \\ &= \frac{1}{\sqrt{2}}(\lambda b_k + b_{k-1} + \overline{\lambda b}_k + \overline{b}_{k-1}) & &= \frac{1}{i\sqrt{2}}(\lambda b_k + b_{k-1} - \overline{\lambda b}_k - \overline{b}_{k-1}) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2}}(\alpha b_k + i\beta b_k + \alpha\overline{b}_k - i\beta\overline{b}_k) + c_{k-1} && = \frac{1}{i\sqrt{2}}(\alpha b_k + i\beta b_k - \alpha\overline{b}_k + i\beta\overline{b}_k) + d_{k-1}\\
&= \alpha\frac{b_k + \overline{b}_k}{\sqrt{2}} + i\beta\frac{b_k - \overline{b}_k}{\sqrt{2}} + c_{k-1} && = \alpha\frac{b_k - \overline{b}_k}{i\sqrt{2}} + i\beta\frac{b_k + \overline{b}_k}{i\sqrt{2}} + d_{k-1}\\
&= \alpha c_k - \beta d_k + c_{k-1} && = \alpha d_k + \beta c_k + d_{k-1}.
\end{aligned}
$$

Für $k = 1$ fallen die Terme c_{k-1} und d_{k-1} weg. In der Basis $\mathcal{D} = \{c_1, d_1, \ldots, c_n, d_n\}$ hat die Matrix also die *reelle Normalform*

$$
A_{\text{reell}} = \begin{pmatrix}
\alpha & \beta & 1 & 0 & & & & & & \\
-\beta & \alpha & 0 & 1 & & & & & & \\
 & & \alpha & \beta & 1 & 0 & & & & \\
 & & -\beta & \alpha & 0 & 1 & & & & \\
 & & & & \alpha & \beta & & & & \\
 & & & & -\beta & \alpha & & & & \\
 & & & & & & \alpha & \beta & 1 & 0 \\
 & & & & & & -\beta & \alpha & 0 & 1 \\
 & & & & & & & & \alpha & \beta \\
 & & & & & & & & -\beta & \alpha
\end{pmatrix}. \tag{5.8}
$$

Wir bestimmen noch die Transformationsmatrix, die A in die reelle Normalform bringt. Dazu beachten wir, dass die Vektoren c_k und d_k in der Basis $\mathcal{B}$ nur in den Komponenten k und $n + k$ von 0 verschiedene Koordinaten haben, nämlich

$$
c_k = \frac{1}{\sqrt{2}}\begin{pmatrix} \vdots \\ 1 \\ \vdots \\ \hline \vdots \\ 1 \\ \vdots \end{pmatrix} \qquad \text{und} \qquad d_k = \frac{1}{i\sqrt{2}}\begin{pmatrix} \vdots \\ 1 \\ \vdots \\ \hline \vdots \\ -1 \\ \vdots \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} \vdots \\ -i \\ \vdots \\ \hline \vdots \\ i \\ \vdots \end{pmatrix}
$$

gemäss (5.7). Die Umrechnung der Koordinaten von der Basis $\mathcal{B}$ in die Basis $\mathcal{D}$ wird daher durch die Matrix

$$
S = \frac{1}{\sqrt{2}}\left(\begin{array}{cccccccccc}
1 & -i & & & & & & & & \\
 & & 1 & -i & & & & & & \\
 & & & & 1 & -i & & & & \\
 & & & & & & \cdots & \cdots & & \\
 & & & & & & & & 1 & -i \\
\hline
1 & i & & & & & & & & \\
 & & 1 & i & & & & & & \\
 & & & & 1 & i & & & & \\
 & & & & & & \cdots & \cdots & & \\
 & & & & & & & & 1 & i
\end{array}\right)
$$

vermittelt. Der Nenner $\sqrt{2}$ wurde so gewählt, dass die Zeilenvektoren der Matrix S als komplexe Vektoren orthonormiert sind, die Matrix S ist daher unitär und hat die Inverse

$$S^{-1} = S^* = \frac{1}{\sqrt{2}} \left(\begin{array}{ccccc|ccccc} 1 & & & & & 1 & & & & \\ i & & & & & -i & & & & \\ & 1 & & & & & 1 & & & \\ & i & & & & & -i & & & \\ & & 1 & & & & & 1 & & \\ & & i & & & & & -i & & \\ & & & \cdots & & & & & \cdots & \\ & & & \cdots & & & & & \cdots & \\ & & & & 1 & & & & & 1 \\ & & & & i & & & & & -i \end{array}\right).$$

Insbesondere folgt jetzt

$$A = S^{-1} A_{\text{reell}} S = S^* A_{\text{reell}} S \qquad \text{und} \qquad A_{\text{reell}} = S A S^{-1} = S A S^*.$$

5.3 Analytische Funktionen einer Matrix

Eine zentrale Motivation in der Entwicklung der Eigenwerttheorie war das Bestreben, Potenzen A^k auch für grosse k effizient zu berechnen. Mit der Jordan-Normalform ist dies auch gelungen, wenigstens über einem Körper, in dem das charakteristische Polynom in Linearfaktoren zerfällt. Die Berechnung von Potenzen war aber nur der erste Schritt, das Ziel in diesem Abschnitt ist, $f(A)$ für eine genügend grosse Klasse von Funktionen f berechnen zu können.

5.3.1 Polynom-Funktionen

In diesem Abschnitt ist $B \in M_n(\mathbb{k})$ und $\mathbb{k}' \supset \mathbb{k}$ ein Körper, über dem das charakteristische Polynome $\chi_A(x)$ in Linearfaktoren

$$\chi_A(x) = (\lambda_1 - x)^{m_1} (\lambda_2 - x)^{m_2} \cdot \ldots \cdot (\lambda_p - x)^{m_p}$$

zerfällt.

Für jedes beliebige Polynome $p(X) \in \mathbb{k}[X]$ der Form

$$p(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots a_1 x + a_0$$

kann man auch

$$p(A) = a_n A^n + a_{n-1} A^{n-1} + \ldots a_1 A + a_0 E$$

berechnen. In der Jordan-Normalform können die Potenzen A^k leicht zusammengestellt werden, sobald man die Potenzen von Jordan-Blöcken berechnet hat.

Satz 5.23. *Die k-te Potenz von $J_n(\lambda)$ ist die Matrix mit*

$$J_n(\lambda)^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \binom{k}{3}\lambda^{k-3} & \ldots & \binom{k}{n-1}\lambda^{k-n+1} \\ 0 & \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \ldots & \binom{k}{n-2}\lambda^{k-n+2} \\ 0 & 0 & \lambda^k & \binom{k}{1}\lambda^{k-1} & \ldots & \binom{k}{n-3}\lambda^{k-n+3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \lambda^k \end{pmatrix} \tag{5.9}$$

mit den Matrixelementen

$$(J_n(\lambda)^k)_{ij} = \binom{k}{j-i}\lambda^{k-j+i}.$$

Die Binomialkoeffizienten verschwinden für $j < i$ und $j > i + k$.

Beweis. Die Herkunft der Binomialkoeffizienten wird klar, wenn man

$$J_n(\lambda) = \lambda E + N_n$$

schreibt, wobei N_n die Matrix (5.3) ist. Die Potenzen von N_n haben die Matrix-Elemente

$$(N_n^k)_{ij} = \delta_{i,j-k} = \begin{cases} 1 & j - i = k \\ 0 & \text{sonst,} \end{cases}$$

sie haben also Einsen genau dort, wo in der die Potenz λ^k steht. Die kt-te Potenz von $J_n(\lambda)$ kann dann mit dem binomischen Satz berechnet werden:

$$J_n(\lambda)^k = \sum_{l=0}^{k} \binom{k}{l} \lambda^l N_n^{k-l},$$

dies ist genau die Form (5.3.1). □

Wir haben bereits gesehen, dass $\chi_A(A) = 0$, ersetzt man also das Polynom $p(X)$ durch $p(X) + \chi_A(X)$, dann ändert sich am Wert

$$(p + \chi_A)(A) = p(A) + \chi_A(A) = p(A)$$

nichts. Man kann also nicht erwarten, dass verschiedene Polynome $p(X)$ zu verschiedenen Matrizen $p(A)$ führen. Doch welche Unterschiede zwischen Polynomen wirken sich genau aus?

Satz 5.24. *Für zwei Polynome $p(X)$ und $q(X)$ ist genau dann $p(A) = q(A)$, wenn das Minimalpolynom von A die Differenz $p - q$ teilt.*

Beweis. Wenn $p(A) = q(A)$, dann ist $h(X) = p(X) - q(X)$ ein Polynom mit $h(A) = 0$, daher muss $h(X)$ vom Minimalpolynom geteilt werden. Ist andererseits $p(X) - q(X) = m(X)t(X)$, dann ist $p(A) - q(A) = m(A)t(A) = 0 \cdot t(A) = 0$, also $p(A) = q(A)$. □

Über einem Körper $\mathbb{k}' \supset \mathbb{k}$, über dem das charakteristische Polynom in Linearfaktoren zerfällt, kann man das Minimalpolynom aus der Jordanschen Normalform ableiten. Es ist

$$m(X) = (\lambda_1 - X)^{q_1}(\lambda_2 - X)^{q_2} \cdot \ldots \cdot (\lambda_p - X)^{q_p},$$

wobei q_i die Dimension des grössten Jordan-Blocks ist, der in der Jordan-Normalform vorkommt. Zwei Polynome $p_1(X)$ und $p_2(X)$ haben genau dann den gleichen Wert, wenn die Differenz $p_1(X) - p_2(X)$ genau die Nullstellen $\lambda_1, \ldots, \lambda_p$ mit Vielfachheiten $q_1, \ldots, q_p$ hat.

Beispiel. Wir betrachten die Matrix

$$A = \begin{pmatrix} 1 & 9 & -4 \\ -1 & 3 & 0 \\ -2 & 0 & 3 \end{pmatrix}$$

mit dem charakteristischen Polynom

$$\chi_A(x) = -x^3 + 7x^2 - 16x + 12 = -(x-3)(x-2)^2.$$

Daraus kann man bereits ablesen, dass das Minimalpolynom $m(X)$ von A entweder $(X-2)(X-3)$ oder $(X-2)^2(X-3)$ ist. Es genügt also nachzuprüfen, ob $p(A) = 0$ für das Polynom $p(X) = (X-2)(X-3) = X^2 - 5X + 6$ ist. Tatsächlich sind die Potenzen von A:

$$A^2 = \begin{pmatrix} 0 & 36 & -16 \\ -4 & 0 & 4 \\ -8 & -18 & 17 \end{pmatrix}, \qquad A^3 = \begin{pmatrix} -4 & 108 & -48 \\ -12 & -36 & 28 \\ -24 & -126 & 83 \end{pmatrix}$$

und daraus kann man jetzt $P(A)$ berechnen:

$$p(A) = \begin{pmatrix} 0 & 36 & -16 \\ -4 & 0 & 4 \\ -8 & -18 & 17 \end{pmatrix} - 5\begin{pmatrix} 1 & 9 & -4 \\ -1 & 3 & 0 \\ -2 & 0 & 3 \end{pmatrix} + 6\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -9 & 4 \\ 1 & -9 & 4 \\ 2 & -18 & 8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}\begin{pmatrix} 1 & -9 & 4 \end{pmatrix} \tag{5.10}$$

Also ist tatsächlich $(X-2)^2(X-3)$ das Minimalpolynom.

Das Quadrat des Polynoms $p(X)$ ist $p(X)^2 = (X-2)^2(X-3)^2$, es hat das Minimalpolynom als Teiler, also muss $p(A)^2 = 0$ sein. Die Gleichung (5.10) ermöglicht, das Quaddrat $p(A)^2$ leichter zu berechnen:

$$p(A)^2 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & -9 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}}_{=0} \begin{pmatrix} 1 & -9 & 4 \end{pmatrix} = 0,$$

wie zu erwarten war.

Wenn sich zwei Polynome nur um das charakteristische Polynom unterscheiden, dann haben sie den gleichen Wert auf A. Das Polynom $p_1(X) = X^3$ unterschiedet sich vom Polynom $p_2(X) = 7X^2 - 16X + 12$ um das charakteristische Polynom, welches wir bereits als das Minimalpolynom von A erkannt haben. Die dritte Potenz A^3 von A muss sich daher auch mit $p_2(X)$ berechnen lassen:

$$7\begin{pmatrix} 0 & 36 & -16 \\ -4 & 0 & 4 \\ -8 & -18 & 17 \end{pmatrix} - 16\begin{pmatrix} 1 & 9 & -4 \\ -1 & 3 & 0 \\ -2 & 0 & 3 \end{pmatrix} + 12\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -4 & 108 & -48 \\ -12 & -36 & 28 \\ -24 & -126 & 83 \end{pmatrix} = A^3. \qquad \bigcirc$$

Satz 5.25. *Wenn A diagonalisierbar ist über einem geeignet erweiterten Körper $\mathbb{k}'$, dann haben zwei Polynome $p(X)$ und $q(X)$ in $\mathbb{k}[X]$ genau dann den gleichen Wert auf A, also $p(A) = q(A)$, wenn $p(\lambda) = q(\lambda)$ für alle Eigenwerte λ von A.*

Über dem Körper der komplexen Zahlen ist die Bedingung, dass die Differenz $d(X) = p_1(X) - p_2(X)$ vom Minimalpolynom geteilt werden muss, gleichbedeutend damit, dass $p_1(X)$ und $p_2(X)$ den gleichen Wert und gleiche Ableitungen bis zur Ordnung $q_i - 1$ haben in allen Eigenwerten λ_i, wobei q_i der Exponent von $\lambda_i - X$ im Minimalpolynom von A ist.

Das Beispiel illustriert auch noch ein weiteres wichtiges Prinzip. Schreiben wir das Minimalpolynom von A in der Form

$$m(X) = X^k + a_{k-1}X^{k-1} + \cdots + a_1X + a_0,$$

dann kann man wegen $m(A) = 0$ die Potenzen A^i mit $i \geq k$ mit der Rekursionsformel

$$A^i = A^{i-k}A^k = A^{i-k}(-a_{k-1}A^{k-1} + \cdots + a_1A + a_0E)$$

in einer Linearkombination kleinerer Potenzen reduzieren. Jedes Polynom vom Grad $\geq k$ kann also reduizert werden in ein Polynom vom Grad $< k$ mit dem gleichen Wert auf A.

Satz 5.26. *Sei A eine Matrix über $\Bbbk$ mit Minimalpolynom $m(X)$. Zu jedem $p(X) \in \Bbbk[X]$ gibt es ein Polynom $q(X) \in \Bbbk[X]$ vom Grad* $\deg q < \deg m$ *mit $p(A) = q(A)$.*

5.3.2 Approximation von $f(A)$

Die Quadratwurzelfunktion $x \mapsto \sqrt{x}$ lässt sich nicht durch ein Polynom darstellen, es gibt also keine direkte Möglichkeit, $\sqrt{A}$ für eine beliebige Matrix zu definieren. Wir können versuchen, die Funktion durch ein Polynom zu approximieren. Damit dies geht, müssen wir folgende zwei Fragen klären:

1. Wie misst man, ob ein Polynom eine Funktion gut approximiert?
2. Was bedeutet es genau, dass zwei Matrizen "nahe beeinander" sind?
3. In welchem Sinne müssen Polynome "nahe" beeinander sein, damit auch die Werte auf A nahe beeinander sind.

Wir wissen bereits, dass nur die Werte und gewisse Ableitungen des Polynoms $p(X)$ in den Eigenwerten einen Einfluss auf $p(A)$ haben. Es genügt also, Approximationspolynome zu verwenden, welche in der Nähe der Eigenwerte "gut genug" approximieren. Solche Polynome gibt es dank dem Satz von Stone-Weierstrass immer:

Satz 5.27 (Stone-Weierstrass)**.** *Ist $I \subset \mathbb{R}$ kompakt, dann lässt sich jede stetige Funktion durch eine Folge $p_n(x)$ beliebig genau approximieren.*

Wir haben schon gezeigt, dass es dabei auf die höheren Potenzen gar nicht ankommt, nach Satz 5.26 kann man ein approximierendes Polynom immer durch ein Polynom von kleinerem Grad als das Minimalpolynom ersetzen.

Definition 5.28. *Die* Norm *einer Matrix M ist*

$$\|M\| = \max\{|Mx| \mid x \in \mathbb{R}^n \wedge |x| = 1\}.$$

Für einen Vektor $x \in \mathbb{R}^n$ gilt $|Mx| \leq \|M\| \cdot |x|$.

Beispiel. Die Matrix

$$M = \begin{pmatrix} 0 & 2 \\ \frac{1}{3} & 0 \end{pmatrix}$$

hat Norm

$$\|M\| = \max_{|x|=1} |Mx| = \max_{t\in\mathbb{R}} \sqrt{2^2 \cos^2 t + \frac{1}{3^2}\sin^2 t} = 2.$$

Da aber

$$M^2 = \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \qquad \Rightarrow \qquad \|M^2\| = \frac{2}{3}$$

ist, wird eine Iteration mit Ableitungsmatrix M trotzdem konvergieren, weil der Fehler nach jedem zweiten Schritt um den Faktor $\frac{2}{3}$ kleiner geworden ist. ○

Beispiel. Wir berechnen die Norm eines Jordan-Blocks.

○

5.3.3 Potenzreihen

Dies führt uns auf die Grösse

$$\pi(M) = \limsup_{n\to\infty} \|M^n\|^{\frac{1}{n}}. \tag{5.11}$$

Ist $\pi(M) > 1$, dann gibt es Anfangsvektoren v für die Iteration, für die $M^k v$ über alle Grenzen wächst. Ist $\pi(M) < 1$, dann wird jeder Anfangsvektor v zu einer Iterationsfolge $M^k v$ führen, die gegen 0 konvergiert. Die Kennzahl $\pi(M)$ erlaubt also zu entscheiden, ob ein Iterationsverfahren konvergent ist.

Die Berechnung von $\pi(M)$ als Grenzwert ist sehr unhandlich. Viel einfacher ist der Begriff des Spektralradius.

Definition 5.29. *Der* Spektralradius *der Matrix M ist der Betrag des betragsgrössten Eigenwertes.*

5.3.4 Gelfand-Radius und Eigenwerte

In Abschnitt **??** ist der Gelfand-Radius mit Hilfe eines Grenzwertes definiert worden. Nur dieser Grenzwert ist in der Lage, über die Konvergenz eines Iterationsverfahrens Auskunft zu geben. Der Grenzwert ist aber sehr mühsam zu berechnen. Es wurde angedeutet, dass der Gelfand-Radius mit dem Spektralradius übereinstimmt, dem Betrag des betragsgrössten Eigenwertes. Dies hat uns ein vergleichsweise einfach auszuwertendes Konvergenzkriterium geliefert. In diesem Abschnitt soll diese Identität zunächst an Spezialfällen und später ganz allgemein gezeigt werden.

Spezialfall: Diagonalisierbare Matrizen

Ist eine Matrix A diagonalisierbar, dann kann Sie durch eine Wahl einer geeigneten Basis in Diagonalform

$$A' = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

gebracht werden, wobei die Eigenwerte λ_i möglicherweise auch komplex sein können. Die Bezeichnungen sollen so gewählt sein, dass λ_1 der betragsgrösste Eigenwert ist, dass also

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Wir nehmen für die folgende, einführende Diskussion ausserdem an, dass sogar $|\lambda_1| > |\lambda_2|$ gilt.

Unter den genannten Voraussetzungen kann man jetzt den Gelfand-Radius von A berechnen. Dazu muss man $|A^n v|$ für einen beliebigen Vektor v und für beliebiges n berechnen. Der Vektor v lässt sich in der Eigenbasis von A zerlegen, also als Summe

$$v = v_1 + v_2 + \cdots + v_n$$

schreiben, wobei v_i Eigenvektoren zum Eigenwert λ_i sind oder Nullvektoren. Die Anwendung von A^k ergibt dann

$$A^k v = A^k v_1 + A^k v_2 + \cdots + A^k v_n = \lambda_1^k v_1 + \lambda_2^k v_2 + \cdots + \lambda_n^k v_n.$$

Für den Grenzwert braucht man die Norm von $A^k v$, also

$$\begin{aligned} |A^k v| &= |\lambda_1^k v_1 + \lambda_2^k v_2 + \cdots + \lambda_3 v_3| \\ \Rightarrow \qquad \frac{|A^k v|}{\lambda_1^k} &= \left| v_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n \right|. \end{aligned} \tag{5.12}$$

Da alle Quotienten $|\lambda_i/\lambda_1| < 1$ sind für $i \geq 2$, konvergieren alle Terme auf der rechten Seite von (5.12) ausser dem ersten gegen 0. Folglich ist

$$\lim_{k\to\infty} \frac{|A^k v|}{|\lambda_1|^k} = |v_1| \qquad \Rightarrow \qquad \lim_{k\to\infty} \frac{|A^k v|^{\frac{1}{k}}}{|\lambda_1|} = \lim_{k\to\infty} |v_1|^{\frac{1}{k}} = 1.$$

Dies gilt für alle Vektoren v, für die $v_1 \neq 0$ ist. Der maximale Wert dafür wird erreicht, wenn man für v einen Eigenvektor der Länge 1 zum Eigenwert λ_1 einsetzt, dann ist $v = v_1$. Es folgt dann

$$\pi(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to\infty} |A^k v|^{\frac{1}{k}} = |\lambda_1| = \varrho(A).$$

Damit ist gezeigt, dass im Spezialfall einer diagonalisierbaren Matrix der Gelfand-Radius tatsächlich der Betrag des betragsgrössten Eigenwertes ist.

Blockmatrizen

Wir betrachten jetzt eine $(n+m)\times(n+m)$-Blockmatrix der Form

$$A = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix} \tag{5.13}$$

mit einer $n \times n$-Matrix B und einer $m \times m$-Matrix C. Ihre Potenzen haben ebenfalls Blockform:

$$A^k = \begin{pmatrix} B^k & 0 \\ 0 & C^k \end{pmatrix}.$$

Ein Vektor v kann in die zwei Summanden v_1 bestehen aus den ersten n Komponenten und v_2 bestehen aus den letzten m Komponenten zerlegen. Dann ist

$$A^k v = B^k v_1 + C^k v_2. \qquad \Rightarrow \qquad |A^k v| \leq |B^k v_1| + |C^k v_2| \leq \pi(B)^k |v_1| + \pi(C)^k |v_2|.$$

Insbesondere haben wir das folgende Lemma gezeigt:

Lemma 5.30. *Eine diagonale Blockmatrix A* (5.13) *Blöcken B und C hat Gelfand-Radius*

$$\pi(A) = \max(\pi(B), \pi(C))$$

Selbstverständlich lässt sich das Lemma auf Blockmatrizen mit beliebig vielen diagonalen Blöcken verallgemeinern.

Für Diagonalmatrizen der genannten Art sind aber auch die Eigenwerte leicht zu bestimmen. Hat B die Eigenwerte $\lambda_i^{(B)}$ mit $1 \leq i \leq n$ und C die Eigenwerte $\lambda_j^{(C)}$ mit $1 \leq j \leq m$, dann ist das charakteristische Polynom der Blockmatrix A natürlich

$$\chi_A(\lambda) = \chi_B(\lambda)\chi_C(\lambda),$$

woraus folgt, dass die Eigenwerte von A die Vereinigung der Eigenwerte von B und C sind. Daher gilt auch für die Spektralradius die Formel

$$\varrho(A) = \max(\varrho(B), \varrho(C)).$$

Jordan-Blöcke

Nicht jede Matrix ist diagonalisierbar, die bekanntesten Beispiele sind die Matrizen

$$J_n(\lambda) = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & \ddots & & \\ & & & \ddots & 1 & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{5.14}$$

wobei $\lambda \in \mathbb{C}$ eine beliebige komplexe Zahl ist. Wir nennen diese Matrizen *Jordan-Matrizen*. Es ist klar, dass $J_n(\lambda)$ nur den n-fachen Eigenwert λ hat und dass der erste Standardbasisvektor ein Eigenvektor zu diesem Eigenwert ist.

In der linearen Algebra lernt man, dass jede Matrix durch Wahl einer geeigneten Basis als Blockmatrix der Form

$$A = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{n_2}(\lambda_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & J_{n_l}(\lambda_l) \end{pmatrix}$$

geschrieben werden kann[1]. Die früheren Beobachtungen über den Spektralradius und den Gelfand-Radius von Blockmatrizen zeigen uns daher, dass nur gezeigt werden muss, dass nur die Gleichheit des Gelfand-Radius und des Spektral-Radius von Jordan-Blöcken gezeigt werden muss.

Iterationsfolgen

Satz 5.31. *Sei A eine $n \times n$-Matrix mit Spektralradius $\varrho(A)$. Dann ist $\varrho(A) < 1$ genau dann, wenn*

$$\lim_{k\to\infty} A^k = 0.$$

Ist andererseits $\varrho(A) > 1$, dann ist

$$\lim_{k\to\infty} \|A^k\| = \infty.$$

Beweis. Wie bereits angedeutet reicht es, diese Aussagen für einen einzelnen Jordan-Block mit Eigenwert λ zu beweisen. Die k-te Potenz von $J_n(\lambda)$ ist

$$J_n(\lambda)^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \dots & \binom{k}{n-1}\lambda^{k-n+1} \\ 0 & \lambda^k & \binom{k}{1}\lambda^{k-1} & \dots & \binom{k}{n-2}\lambda^{k-n+2} \\ 0 & 0 & \lambda^k & \dots & \binom{k}{n-k+3}\lambda^{k-n+3} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda^k \end{pmatrix}.$$

Falls $|\lambda| < 1$ ist, gehen alle Potenzen von λ exponentiell schnell gegen 0, während die Binomialkoeffizienten nur polynomiell schnell anwachsen. In diesem Fall folgt also $J_n(\lambda) \to 0$.

Falls $|\lambda| > 1$ divergieren bereits die Elemente auf der Diagonalen, also ist $\|J_n(\lambda)^k\| \to \infty$ mit welcher Norm auch immer man man die Matrix misst. □

[1] Sofern die Matrix komplexe Eigenwerte hat muss man auch komplexe Basisvektoren zulassen.

Aus dem Beweis kann man noch mehr ablesen. Für $\varrho(A) < 1$ ist die Norm $\|A^k\| \leq M\varrho(A)^k$ für eine geeignete Konstante M, für $\varrho(A) > 1$ gibt es eine Konstante m mit $\|A^k\| \geq m\varrho(A)^k$.

Der Satz von Gelfand

Der Satz von Gelfand ergibt sich jetzt als direkte Folge aus dem Satz 5.31.

Satz 5.32 (Gelfand)**.** *Für jede komplexe $n \times n$-Matrix A gilt*

$$\pi(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \varrho(A).$$

Beweis. Der Satz 5.31 zeigt, dass der Spektralradius ein scharfes Kriterium dafür ist, ob $\|A^k\|$ gegen 0 oder ∞ konvergiert. Andererseits ändert ein Faktor t in der Matrix A den Spektralradius ebenfalls um den gleichen Faktor, also $\varrho(tA) = t\varrho(A)$. Natürlich gilt auch

$$\pi(tA) = \lim_{k\to\infty} \|t^k A^k\|^{\frac{1}{k}} = \lim_{k\to\infty} t\|A^k\|^{\frac{1}{k}} = t \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = t\pi(A).$$

Wir betrachten jetzt die Matrix

$$A(\varepsilon) = \frac{A}{\varrho(A) + \varepsilon}.$$

Der Spektralradius von $A(\varepsilon)$ ist

$$\varrho(A(\varepsilon)) = \frac{\varrho(A)}{\varrho(A) + \varepsilon},$$

er ist also > 1 für negatives ε und < 1 für positives ε. Aus dem Satz 5.31 liest man daher ab, dass $\|A(\varepsilon)^k\|$ genau dann gegen 0 konvergiert, wenn $\varepsilon > 0$ ist und divergiert genau dann, wenn $\varepsilon < 0$ ist.

Aus der Bemerkung nach dem Beweis von Satz 5.31 schliesst man daher, dass es im Fall $\varepsilon > 0$ eine Konstante M gibt mit

$$\begin{aligned} \|A(\varepsilon)^k\| \leq M\varrho(A(\varepsilon))^k \quad &\Rightarrow \quad \|A(\varepsilon)^k\|^{\frac{1}{k}} \leq M^{\frac{1}{k}}\varrho(A(\varepsilon)) \\ &\Rightarrow \quad \pi(A) \leq \varrho(A(\varepsilon)) \underbrace{\lim_{k\to\infty} M^{\frac{1}{k}}}_{=1} = \varrho(A(\varepsilon)) = \varrho(A) + \varepsilon. \end{aligned}$$

Dies gilt für beliebige $\varepsilon > 0$, es folgt daher $\pi(A) \leq \varrho(A)$.

Andererseits gibt es für $\varepsilon < 0$ eine Konstante m mit

$$\begin{aligned} \|A(\varepsilon)^k\| \geq m\varrho(A(\varepsilon))^k \quad &\Rightarrow \quad \|A(\varepsilon)^k\|^{\frac{1}{k}} \geq m^{\frac{1}{k}}\varrho(A(\varepsilon)) \\ &\Rightarrow \quad \pi(A) \geq \varrho(A(\varepsilon)) \underbrace{\lim_{k\to\infty} m^{\frac{1}{k}}}_{=1} = \varrho(A(\varepsilon)) = \varrho(A) + \varepsilon. \end{aligned}$$

Dies gilt für beliebige $\varepsilon > 0$, es folgt daher $\pi(A) \geq \varrho(A)$. Zusammen mit $\pi(A) \leq \varrho(A)$ folgt $\pi(A) = \varrho(A)$. □

5.4 Spektraltheorie

Aufgabe der Spektraltheorie ist, Bedingungen an eine Matrix A und eine Funktion $f(z)$ zu finden, unter denen es möglich ist, $f(A)$ auf konsistente Art und Weise zu definieren. Weiter müssen Methoden entwickelt werden, mit denen $f(A)$ berechnet werden kann. Für ein Polynom $p(z)$ ist $p(A)$ durch einsetzen definiert. Für Funktionen, die sich nicht durch ein Polynom darstellen lassen, muss eine Approximation der Funktion durch Polynome verwendet werden. Sei also $p_n(z)$ eine Folge von Polynomen, die als Approximation der Funktion $f(z)$ verwendet werden soll. Das Ziel ist, $f(A)$ als den Grenzwert der Matrixfolge $p_n(A)$ zu definieren.

Zunächst ist nicht klar, wie eine solche Folge gewählt werden muss. Es muss eine Teilmenge von $K \subset \mathbb{C}$ spezifiziert werden, auf der die Funktionenfolge $p_n(z)$ konvergieren muss, damit auch die Konvergenz der Matrizenfolge $p_n(A)$ garantiert ist. Auch die Art der Konvergenz von $p_n(z)$ auf der Menge K ist noch unklar. Da der Abstand zweier Matrizen A und B in der Operatornorm mit der grössten Abweichung $\|(A - B)v\|$ für Einheitsvektoren v gemessen wird, ist es einigermassen plausibel, dass die grösse Abweichung zwischen zwei Polynomen $|p(z) - q(z)|$ auf der Menge K kleine sein muss, wenn $\|p(A) - q(A)\|$ klein sein soll. Da die Differenz $p(z) - q(z)$ für beliebige Polynome, die sich nicht nur um eine Konstante unterscheiden, mit z über alle Grenzen wächst, muss K beschränkt sein. Gesucht ist also eine kompakte Menge $K \subset \mathbb{C}$ und eine Folge $p_n(z)$ von Polynomen, die auf K gleichmässig gegen $f(z)$ konvergieren. Die Wahl von K muss sicherstellen, dass für jede gleichmässig konvergente Folge von Polynomen $p_n(z)$ auch die Matrizenfolge $p_n(A)$ konvergiert.

Es wird sich zeigen, dass die Menge K das Spektrum von A ist, also eine endliche Teilmenge von $\mathbb{C}$. Jede Funktion kann auf so einer Menge durch Polynome exakt wiedergegeben werden. Es gibt insbesondere Folgen von Polynomen, die eingeschränkt auf das Spektrum gleich sind, also $p_n(z) = p_m(z)$ für alle $z \in K$, die aber ausserhalb des Spektrums alle verschieden sind. Als Beispiel kann die Matrix

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

herangezogen werden. Ihr Spektrum ist $\mathrm{Sp}(N) = \{0\} \subset \mathbb{C}$. Zwei Polynome stimmen genau dann auf $\mathrm{Sp}(N)$ überein, wenn der konstante Koeffizient gleich ist. Die Polynome $p(z) = z$ und $q(z) = z^2$ stimmen daher auf dem Spektrum überein. Für die Matrizen gilt aber $p(N) = N$ und $q(N) = N^2 = 0$, die Matrizen stimmen also nicht überein. Es braucht also zusätzliche Bedingungen an die Matrix A, die sicherstellen, dass $p(A) = 0$ ist, wann immer $p(z) = 0$ für $z \in \mathrm{Sp}(A)$ gilt.

In diesem Abschnitt sollen diese Fragen untersucht werden. In Abschnitt 5.4.1 wird gezeigt, wie sich Funktionen durch Polynome approximieren lassen, woraus sich dann Approximationen von $f(A)$ für diagonalisierbare Matrizen mit reellen Eigenwerten ergeben.

Der Satz von Stone-Weierstrass, der in Abschnitt 5.4.2 dargestellt wird, ist ein sehr allgemeines Approximationsresultat, welches nicht nur zeigt, dass die Approximation unter sehr natürlichen Voraussetzungen beliebig genau möglich ist, sondern uns im komplexen Fall auch weitere Einsicht dafür geben kann, welche Voraussetzungen an eine komplexe Matrix gestellt werden müssen, damit man damit rechnen kann, dass die Approximation zu einer konsistenten Definition von $f(A)$ führt.

5.4.1 Approximation durch Polynome

Die der Berechnung von $f(A)$ für eine beleibige stetige Funktion, die sich nicht als Potenzreihe schreiben lässt, verwendet Approximationen von Polynomen. Die numerische Mathematik hat eine

grosse Menge von solchen Approximationsverfahren entwickelt, wovon zwei kurz (ohne Beweise) vorgestellt werden sollen.

Das Legendre-Interpolationspolynom

Zu vorgegebenen, verschiedenen Zahlen $z_i \in \mathbb{C}$, $0 \le i \le n$, die auch die *Stützstellen* genannt werden, gibt es immer ein Polynom vom Grade n, welches in den z_i vorgegebene Werte $f(z_i)$ annimmt. Ein solches Polynom lässt sich im Prinzip mit Hilfe eines linearen Gleichungssystems finden, man kann aber auch direkt eine Lösung konstruieren. Dazu bildet man erst die Polynome

$$\begin{aligned} l(z) &= (z-z_0)(z-z_1)\dots(z-z_n) \qquad \text{und} \\ l_i(z) &= (z-z_0)\dots\widehat{(z-z_i)}\dots(z-z_n). \end{aligned}$$

Darin bedeutet der Hut, dass dieser Term weggelassen werden soll. Für $z \neq z_i$ ist $l_i(z) = l(z)/(z-z_i)$. Die Polynome

$$k_i(z) = \frac{l_i(z)}{l_i(z_i)} = \frac{(z-z_0)\dots\widehat{(z-z_i)}\dots(z-z_n)}{(z_i-z_0)\dots\widehat{(z_i-z_i)}\dots(z_i-z_n)}$$

haben die Eigenschaft $k_i(z_j) = \delta_{ij}$. Damit lässt sich jetzt ein Polynom

$$p(z) = \sum_{j=0}^{n} f(z_j)\frac{l_j(z)}{l_j(z_j)}$$

vom Grad n konstruieren, welches die Werte

$$p(z_i) = \sum_{j=0}^{n} f(z_j)\frac{l_j(z_i)}{l_j(z_j)} = \sum_{j=0}^{n} f(z_j)\delta_{ij} = f_(z_i)$$

annimmt. Das Polynom $p(z)$ heisst das *Legendre-Interpolationspolynom*.

Zwar lässt sich also für eine endliche Menge von komplexen Zahlen immer ein Polynom finden, welches vorgeschriebene Wert in allen diesen Zahlen annimmt, doch ist die Stabilität für grosse n eher beschränkt.

Gleichmassige Approximation mit Bernstein-Polynomen

Das Legendre-Interpolationspolynom nimmt in den Stützstellen die verlangten Werte an, aber ausserhalb der Stützstellen ist nicht garantiert, dass man eine gute Approximation einer Funktion $f(z)$ erhält.

Für die Approximation auf einem reellen Interval $[a, b]$ hat Sergei Natanowitsch Bernstein ein Dazu werden zuerst die reellen Bernsteinpolynome vom Grad n durch

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}.$$

definiert. Als Approximationspolynom für die auf dem Interval $[0, 1]$ definierte, stetige Funktion $f(t)$ kann man dann

$$B_n(f)(t) = \sum_{i=0}^{n} B_{i,n}(t) f\left(\frac{i}{n}\right)$$

verwenden. Die Polynome $B_n(f)(t)$ konvergieren gleichmässig auf $[0,1]$ gegen die Funktion $f(t)$. Über die Konvergenz ausserhalb des reellen Intervalls wird nichts ausgesagt. Die Approximation mit Bernstein-Polynomen ist daher nur sinnvoll, wenn man weiss, dass die Eigenwerte der Matrix reell sind, was im wesentlichen auf diagonalisierbare Matrizen führt.

Für ein anderes Interval $[a,b]$ kann man ein Approximationspolynom erhalten, indem man die affine Transformation $s \mapsto (s-a)/(b-a)$ von $[a,b]$ auf $[0,1]$ verwendet.

5.4.2 Der Satz von Stone-Weierstrasss

Der Satz von Stone-Weierstrass behandelt im Gegensatz zu den in Abschnitt 5.4.1 besprochenen Approximationsmethoden nicht nur Funktionen von reellen Variablen durch Polynome. Vielmehr kann das Definitionsgebiet irgend eine abgeschlossene und beschränkte Teilmenge eines reellen oder komplexen Vektorraumes sein und die Funktionen können Polynome aber auch viel allgemeinere Funktionen verwendet werden, wie zum Beispiel die Funktionen $x \mapsto \cos nx$ und $x \mapsto \sin nx$ definiert auf dem Intervall $[0, 2\pi]$. In diesem Fall liefert der Satz von Stone-Weierstrass die Aussage, dass sich jede stetige periodische Funktion gleichmässig durch trigonometrische Polynome approximieren lässt.

Die Aussage des Satz von Stone-Weierstrass über reelle Funktionen lässt sich nicht auf komplexe Funktionen erweitern. Von besonderem Interesse ist jedoch, dass der Beweis des Satz zeigt, warum solche Aussagen für komplexe Funktionen nicht mehr zutreffen. Im Falle der Approximation von komplexen Funktionen $f(z)$ durch Polynome zwecks Definition von $f(A)$ werden sich daraus Bedingungen an die Matrix ableiten lassen, die eine konsistente Definition überhaupt erst ermöglichen werden.

Punkte trennen

Aus den konstanten Funktionen lassen sich durch algebraische Operationen nur weitere konstante Funktionen erzeugen. Die konstanten Funktionen sind also nur dann eine genügend reichhaltige Menge, wenn die Menge K nur einen einzigen Punkt enthält. Damit sich Funktionen approximieren lassen, die in zwei Punkten verschiedene Werte haben, muss es auch unter den zur Approximation zur Verfügung stehenden Funktionen solche haben, deren Werte sich in diesen Punkten unterscheiden. Diese Bedingung wird in der folgenden Definition formalisiert.

Definition 5.33. *Sei K eine beliebige Menge und A eine Menge von Funktionen $K \to \mathbb{C}$. Man sagt, A* trennt die Punkte von *K, wenn es für jedes Paar von Punkten $x, y \in K$ eine Funktion $f \in A$ gibt derart, dass $f(x) \neq f(y)$.*

Man kann sich die Funktionen f, die gemäss dieser Definition die Punkte von K trennen, als eine Art Koordinaten der Punkte in K vorstellen. Die Punkte der Teilmenge $K \subset \mathbb{R}^n$ werden zum Beispiel von den Koordinatenfunktionen $x \mapsto x_i$ getrennt. Wir schreiben für die i-Koordinate daher auch als Funktion $x_i(x) = x_i$. Zwei verschiedene Punkte $x, y \in K$ unterscheiden sich in mindestens einer Koordinate. Für diese Koordinate sind dann die Werte der zugehörigen Koordinatenfunktion $x_i = x_i(x) \neq x_i(y) = y_i$ verschieden, die Funktionen $x_1(x)$ bis $x_n(x)$ trennen also die Punkte.

Beispiel. Wir betrachten einen Kreis in der Ebene, also die Menge

$$S^1 = \{(x_1, x_2) \mid x_1^2 + x_2^2 = 1\}$$

S^1 ist eine abgeschlossene und beschränkte Menge in $\mathbb{R}^2$. Die Funktion $x \mapsto x_1$ trennt die Punkte nicht, denn zu jedem Punkt $(x_1, x_2) \in S^2$ gibt es den an der ersten Achse gespiegelten Punkt

$\sigma(x) = (x_1, -x_2)$, dessen erste Koordinate den gleichen Wert hat. Ebenso trennt die Koordinatenfunktion $x \mapsto x_2$ die Punkte nicht. Die Menge $A = \{x_1(x), x_2(x)\}$ bestehend aus den beiden Koordinatenfunktionen trennt dagegen die Punkte von S^1, da die Punkte sich immer in mindestens einem Punkt unterscheiden.

Man könnte auch versuchen, den Kreis in Polarkoordinaten zu beschreiben. Die Funktion $\varphi(x)$, die jedem Punkt $x \in S^1$ den Polarwinkel zuordnet, trennt sicher die Punkte des Kreises. Zwei verschiedene Punkte auf dem Kreis haben verschieden Polarwinkel. Die Menge $\{\varphi\}$ trennt also die Punkte von S^1. Allerdings ist die Funktion nicht stetig, was zwar der Definition nicht widerspricht aber ein Hindernis für spätere Anwendungen ist. ○

Der Satz von Stone-Weierstrass für reelle Funktionen

Die Beispiele von Abschnitt 5.4.1 haben bezeigt, dass sich reellwertige Funktionen einer reellen Variable durch Polynome beliebig genau approximieren lassen. Es wurde sogar eine Methode vorgestellt, die eine auf einem Intervall gleichmässig konvergente Polynomefolge produziert. Die Variable $x \in [a, b]$ trennt natürlich die Punkte, die Algebra der Polynome in der Variablen x enthält also sicher Funktionen, die in verschiedenen Punkten des Intervalls auch verschiedene Werte annehmen. Nicht ganz so selbstverständlich ist aber, dass sich daraus bereits ergibt, dass jede beliebige Funktion sich als Polynome in x approximieren lässt. Dies ist der Inhalt des folgenden Satzes von Stone-Weierstrass.

Satz 5.34 (Stone-Weierstrass)**.** *Enthält eine $\mathbb{R}$-Algebra A von stetigen, rellen Funktionen auf einer kompakten Menge K die konstanten Funktionen und trennt sie Punkte, d. h. für zwei verschiedene Punkte $x, y \in K$ gibt es immer eine Funktion $f \in A$ mit $f(x) \neq f(y)$, dann ist jede stetige, reelle Funktion auf K gleichmässig approximierbar durch Funktionen in A.*

Für den Beweis des Satzes wird ein Hilfsresultat benötigt, welches wir zunächst ableiten. Es besagt, dass sich die Wurzelfunktion $t \mapsto \sqrt{t}$ auf dem Interval $[0, 1]$ gleichmässig von unten durch Polynome approximieren lässt, die in Abbildung 5.7 dargestellt sind.

Satz 5.35. *Die rekursiv definierte Folge von Polynomen*

$$u_{n+1}(t) = u_n(t) + \frac{1}{2}(t - u_n(t)^2), \qquad u_0(t) = 0 \tag{5.15}$$

ist monoton wachsend und approximiert die Wurzelfunktion $t \mapsto \sqrt{t}$ gleichmässig auf dem Intervall $[0, 1]$.

Beweis. Wer konstruieren zunächst das in Abbildung 5.6 visualierte Verfahren, mit dem für jede Zahl $a \in [0, 1]$ die Wurzel $\sqrt{a}$ berechnet werden kann. Sei $u < \sqrt{a}$ eine Approximation der Wurzel. Die Approximation ist der exakte Wert der Lösung, wenn $a - u^2 = 0$. In jedem anderen Fall muss u um einen Betrag d vergrössert werden. Natürlich muss immer noch $u + d < \sqrt{a}$ sein. Man kann die maximal zulässige Korrektur d geometrisch abschätzen, wie dies in Abbildung 5.6 skizziert ist. Die maximale Steigung des Graphen der Funktion $u \mapsto u^2$ ist 2, daher darf man u maximal um die Hälfte der Differenz $a - u^2$ (grün) vergrössern, also $d = \frac{1}{2}(a - u^2)$. Die Rekursionsformel

$$u_{n+1} = u_n + d = u_n + \frac{1}{2}(a - u_n^2)$$

mit dem Startwert $u_0 = 0$ liefert daher eine Folge, die gegen $\sqrt{a}$ konvergiert. □

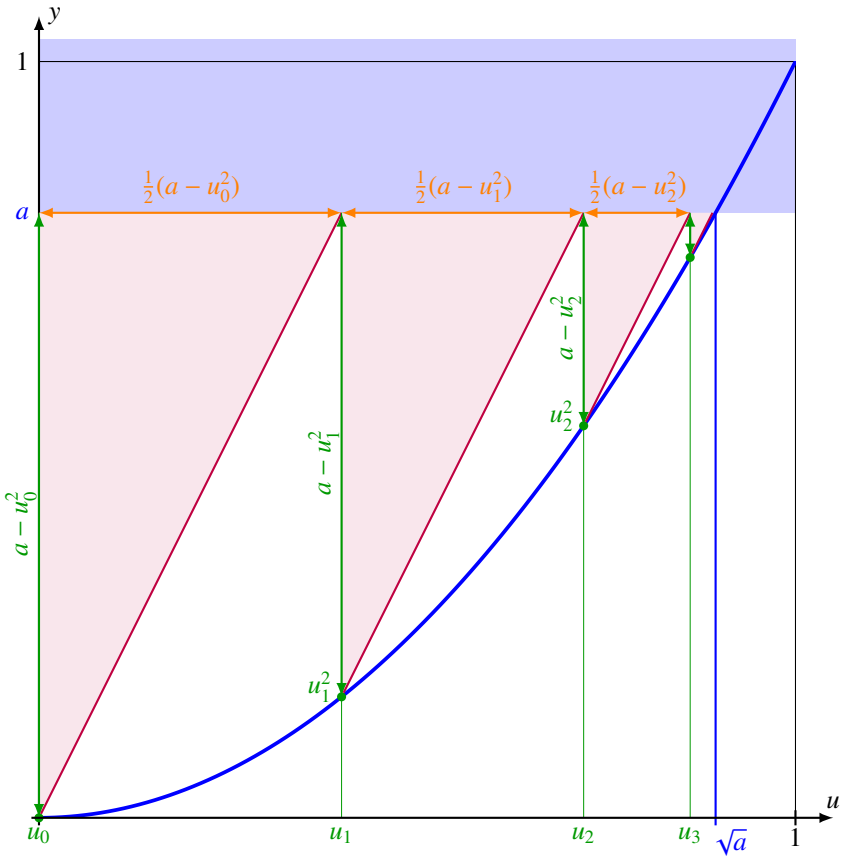


Abbildung 5.6: Konstruktion einer monoton wachsenden Approximationsfolge für $\sqrt{a}$

Beweis des Satzes von Stone-Weierstrass. Da A eine Algebra ist, ist mit jeder Funktion $f \in A$ für jedes Polynome $p \in \mathbb{R}[X]$ auch $p(f)$ eine Funktion in A.

1. Schritt: Für jede Funktion $f \in A$ lässt sich auch $|f|$ durch Funktionen in A beliebig genau durch eine monoton wachsende Folge von Funktionen approximieren.

   Da A eine Algebra ist, ist $f^2 \in A$. Sei ausserdem $m^2 = \sup\{f(x)^2 \mid x \in K\}$, so dass f^2/m^2 eine Funktion mit Werten im Intervall $[0,1]$ ist. Die Funktionen $f_n(x) = mu_n(f(x)^2/m^2)$ sind ebenfalls in A und approximieren gleichmässig $\sqrt{f(x)^2} = |f(x)|$.

2. Schritt: Für zwei Funktionen $f, g \in A$ gibt es eine monoton wachsende Folge, die $\max(f,g)$ gleichmässig beliebig genau approximiert und eine monoton fallende Folge, die $\min(f,g)$ gleichmässig beliebig genau approximiert.

   Diese Folgen können aus der Approximationsfolge für den Betrag einer Funktion und den Identitäten

$$\begin{aligned} \max(f,g) &= \frac{1}{2}(f+g+|f-g|) \\ \min(f,g) &= \frac{1}{2}(f+g-|f-g|) \end{aligned} \tag{5.16}$$

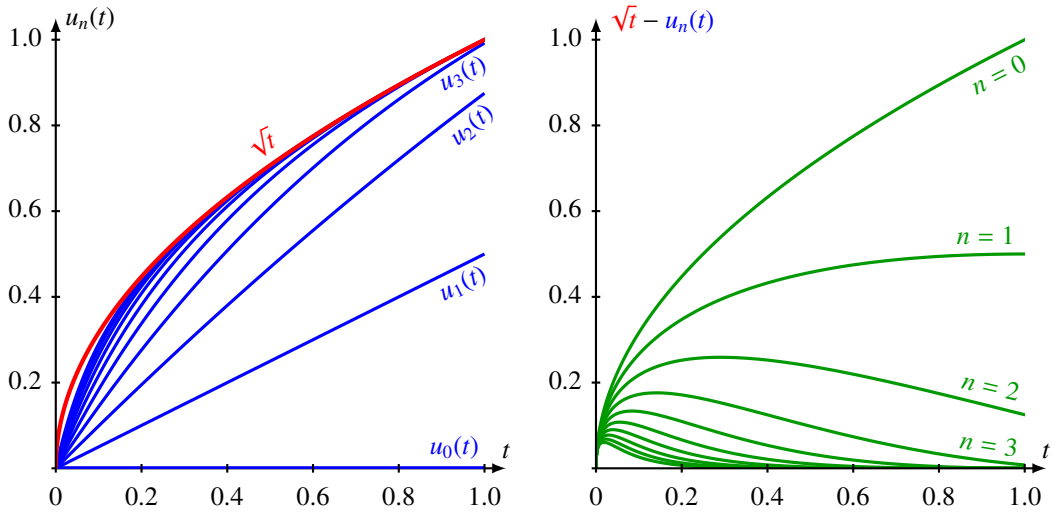


Abbildung 5.7: Monoton wachsende Approximation der Funktion $t \mapsto \sqrt{t}$ mit Polynomen $u_n(t)$ nach (5.15) (links) und der Fehler der Approximation (rechts).

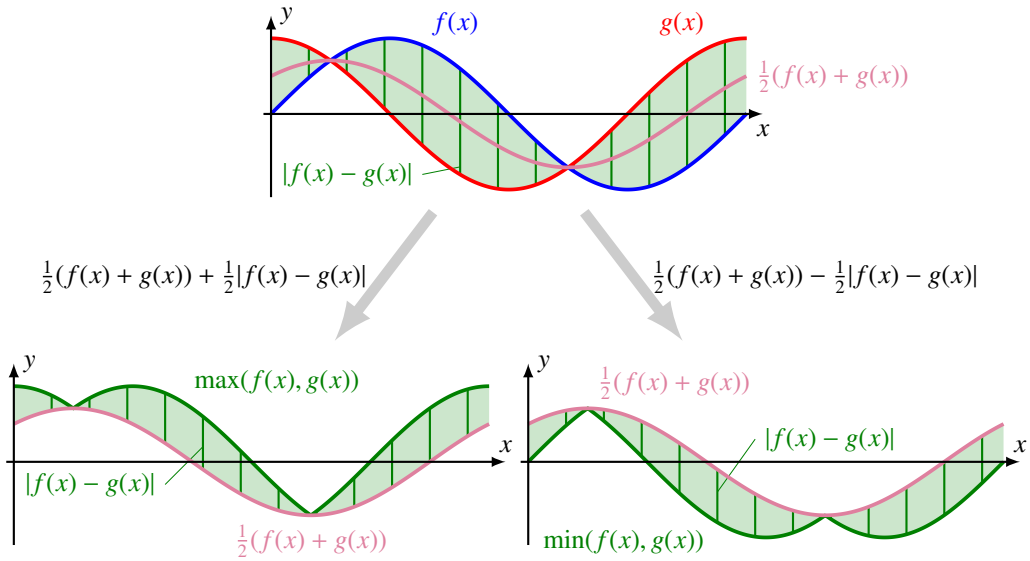


Abbildung 5.8: Graphische Erklärung der Identitäten (5.16) für $\max(f(x), g(x))$ und $\min(f(x), g(x))$. Die purpurrote Kurve stellt den Mittelwert von $f(x)$ und $g(x)$ dar, die vertikalen grünen Linien haben die Länge der Differenz $|f(x) - g(x)|$. Das Maximum erhält man, indem man den halben Betrag der Differenz zum Mittelwert hinzuaddiert, das Minimum erhält man durch Subtraktion der selben Grösse.

gefunden werden, die in Abbildung 5.8 graphisch erklärt werden.

3. Schritt: Zu zwei beliebigen Punkten $x, y \in K$ und Werten $\alpha, \beta \in \mathbb{R}$ gibt es immer eine Funktion in A, die in den Punkten x, y die vorgegebenen Werte α bzw. β annimmt. Da A die Punkte trennt, gibt es eine Funktion f_0 mit $f_0(x) \neq f_0(y)$. Dann ist die Funktion

$$f(t) = \beta + \frac{f_0(t) - f_0(y)}{f_0(x) - f_0(y)}(\alpha - \beta)$$

wohldefiniert und nimmt die verlangten Werte an.

4. Schritt: Zu jeder stetigen Funktion $f\colon K \to \mathbb{R}$, jedem Punkt $x \in K$ und jedem $\varepsilon > 0$ gibt es eine Funktion $g \in A$ derart, dass $g(x) = f(x)$ und $g(y) \leq f(y) + \varepsilon$ für alle $y \in K$.

   Zu jedem $z \in K$ gibt es eine Funktion in A mit $h_z(x) = f(x)$ und $h_z(z) \leq f(z) + \frac{1}{2}\varepsilon$. Wegen der Stetigkeit von h_z gibt es eine Umgebung V_z von z, in der immer noch gilt $h_z(y) \leq f(y) + \varepsilon$ für $y \in V_z$. Wegen der Kompaktheit von K kann man endlich viele Punkte z_i wählen derart, dass die V_{z_i} immer noch K überdecken. Dann erfüllt die Funktion $g(z) = \inf h_{z_i}$ die Bedingungen $g(x) = f(x)$ und für $z \in V_{z_i}$

$$g(z) = \inf_j h_{z_j}(z) \leq h_{z_i}(z) \leq f(z) + \varepsilon.$$

   Ausserdem ist $g(z)$ nach dem zweiten Schritt beliebig genau durch Funktionen in A approximierbar.

5. Schritt: Jede stetige Funktion $f\colon K \to \mathbb{R}$ kann beliebig genau durch Funktionen in A approximiert werden. Sei $\varepsilon > 0$.

   Nach dem vierten Schritt gibt es für jedes $y \in K$ eine Funktion g_y derart, dass $g_y(y) = f(y)$ und $g_y(x) \leq f(x) + \varepsilon$ für $x \in K$. Da g_y stetig ist, gilt ausserdem $g_y(x) \geq f(x) - \varepsilon$ in einer Umgebung U_y von y. Da K kompakt ist, kann man endlich viele y_i derart, dass die U_{y_i} immer noch ganz K überdecken. Die Funktion $g = \sup g_{y_i}$ erfüllt dann überall $g(x) \leq f(x) + \varepsilon$, weil jede der Funktionen g_y diese Ungleichung erfüllt. Ausserdem gilt für $x \in V_{x_j}$

$$g(x) = \sup_i g_{x_i}(x) \geq g_{x_j}(x) \geq f(x) - \varepsilon.$$

   Somit ist

$$|f(x) - g(x)| \leq \varepsilon.$$

   Damit ist $f(x)$ beliebig nahe an der Funktion $g(x)$, die sich beliebig genau durch Funktionen aus A approximieren lässt. □

Im ersten Schritt des Beweises ist ganz entscheidend, dass man die Betragsfunktion konstruieren kann. Daraus leiten sich dann alle folgenden Konstruktionen ab.

Anwendung auf symmetrische und hermitesche Matrizen

Für symmetrische und hermitesche Matrizen A ist bekannt, dass die Eigenwerte reell sind, also das Spektrum $\mathrm{A} \subset \mathbb{R}$ ist. Für eine Funktion $\mathbb{R} \to \mathbb{R}$ lässt sich nach dem Satz 5.34 immer eine Folge p_n von approximierenden Polynomen in x finden, die auf $\mathrm{Sp}(A)$ gleichmässig konvergiert. Die Matrix $f(A)$ kann dann definiert werden also der Grenzwert

$$f(A) = \lim_{n\to\infty} p_n(A).$$

Da diese Matrizen auch diagonalisierbar sind, kann man eine Basis aus Eigenvektoren verwenden. Die Wirkung von $p_n(A)$ auf einem Eigenvektor v zum Eigenwert λ ist

$$p_n(A)v = (a_kA^k+a_{k-1}A^{k-1}+\cdots+a_2A^2+a_1A+a_0I)v = (a_k\lambda^k+a_{k-1}\lambda^{k-1}+\cdots+a_2\lambda^2+a_1\lambda+a_0)v = p_n(\lambda)v.$$

Im Grenzwert wirkt $f(A)$ daher durch Multiplikation eines Eigenvektors mit $f(\lambda)$, die Matrix $f(A)$ hat in der genannten Basis die Diagonalform

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \qquad \Rightarrow \qquad f(A) = \begin{pmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_n) \end{pmatrix}.$$

Satz 5.36. *Ist A symmetrische oder selbstadjungiert Matrix und f eine Funktion auf dem Spektrum* $\mathrm{Sp}(A)$ *von A. Dann gibt es genau eine Matrix $f(A)$, die Grenzwert jeder beliebigen Folge $p_n(A)$ für Polynomfolgen, die* $\mathrm{Sp}(A)$ *gleichmässig gegen f konvergieren.*

Unmöglichkeit der Approximation von $z \mapsto \overline{z}$ in $\mathbb{C}[z]$

Der Satz 5.34 von Stone-Weierstrass für reelle Funktionen gilt nicht für komplexe Funktionen. In diesem Abschnitt zeigen wir, dass sich die Funktion $z \mapsto \overline{z}$ auf der Einheitskreisscheibe $K = \{z \in \mathbb{C} \mid |z| \leq 1\}$ nicht gleichmässig durch Polynome $p(z)$ mit komplexen Koeffizienten approximieren lässt.

Wäre eine solche Approximation möglich, dann könnte man $\overline{z}$ auch durch eine Potenzreihe

$$\overline{z} = \sum_{k=0}^{\infty} a_k z^k$$

darstellen. Das Wegintegral beider Seiten über den Pfad $\gamma(t) = e^{it}$ in der komplexen Ebene ist

$$\begin{aligned} \oint_\gamma z^k\,dz &= \int_0^{2\pi} e^{ikt} ie^{it}\,dt = i\int_0^{2\pi} e^{it(k+1)}\,dt = i\Big[\frac{1}{i(k+1)}e^{it(k+1)}\Big]_0^{2\pi} = 0 \\ \oint_\gamma \sum_{k=0}^{\infty} a_k z^k\,dz &= \sum_{k=0}^{\infty} a_k \oint_\gamma z^k\,dz = \sum_{k=0}^{\infty} a_k \cdot 0 = 0 \\ \oint_\gamma \overline{z}\,dz &= \int_0^{2\pi} e^{it} ie^{it}\,dt = i\int_0^{2\pi} dt = 2\pi i, \end{aligned}$$

dabei wurde $\overline{\gamma}(t) = e^{-it}$ verwendet. Insbesondere widersprechen sich die beiden Integrale. Die ursprüngliche Annahmen, $\overline{z}$ lasse sich durch Polynome gleichmässig approximieren, muss daher verworfen werden.

Der Satz von Stone-Weierstrass für komplexe Funktionen

Der Satz von Stone-Weierstrass kann nach dem vorangegangene Abschnitt also nicht gelten. Um den Beweis des Satzes 5.34 auf komplexe Zahlen zu übertragen, muss im ersten Schritt ein Weg gefunden werden, den Betrag einer Funktion zu approximieren.

Im reellen Fall geschah dies, indem zunächst eine Polynom-Approximation für die Quadratwurzel konstruiert wurde, die dann auf das Quadrat einer Funktion angewendet wurde. Der Betrag einer

komplexen Zahl z ist aber nicht allein aus z berechenbar, man braucht in irgend einer Form Zugang zu Real- und Imaginärteil. Zum Beispiel kann man Real- und Imaginärteil als $\Re z = \frac{1}{2}(z + \overline{z})$ und $\Im z = \frac{1}{2}(z - \overline{z})$ bestimmen. Kenntnis von Real- und Imaginärteil ist als gleichbedeutend mit der Kenntnis der komplex Konjugierten $\overline{z}$. Der Betrag lässt sich daraus als $|z|^2 = z\overline{z}$ finden. Beide Beispiele zeigen, dass man den im Beweis benötigten Betrag nur dann bestimmen kann, wenn mit jeder Funktion aus A auch die komplex konjugierte Funktion zur Verfügung steht.

Satz 5.37 (Stone-Weierstrass)**.** *Enthält eine $\mathbb{C}$-Algebra A von stetigen, komplexwertigen Funktionen auf einer kompakten Menge K die konstanten Funktionen, trennt sie Punkte und ist ausserdem mit jeder Funktion $f \in A$ auch die komplex konjugiert Funktion $\overline{f} \in A$, dann lässt sich jede stetige, komplexwertige Funktion auf K gleichmässig durch Funktionen aus A approximieren.*

Mit Hilfe der konjugiert komplexen Funktion lässt sich immer eine Approximation für die Betragsfunktion finden, so dass sich der Beweis des reellen Satzes von Stone-Weierstrass übertragen lässt.

5.4.3 Normale Matrizen

Aus dem Satz von Stone-Weierstrass für komplexe Matrizen kann man jetzt einen Spektralsätze für eine etwas grössere Klasse von Matrizen ableiten, als im Satz 5.36 möglich war. Der Satz besagt, dass für eine beliebige Funktion f auf dem Spektrum $\mathrm{Sp}(A)$ eine Folge von auf $\mathrm{Sp}(A)$ gleichmässig konvergenten, approximierenden Polynomen $p_n(z, \overline{z})$ gefunden werden kann. Doch wie soll jetzt aus dieser Polynomfolge ein Kandidat von $f(A)$ gefunden werden?

Zunächst stellt sich die Frage, was für die Variable $\overline{z}$ eingesetzt werden soll. 1×1-Matrizen sind notwendigerweise diagonal, also muss man in diesem Fall die Matrix $\overline{A}$ für die Variable $\overline{z}$ eingesetzt werden. Dies erklärt aber noch nicht, wie für $n \times n$-Matrizen vorzugehen ist, wenn $n > 1$ ist.

Die Notwendigkeit, die Variable $\overline{z}$ hinzuzunehmen ergab sich aus der Anforderung, dass der Betrag aus $|z|^2 = z\overline{z}$ konstruiert werden können muss. Insbesondere muss beim Einsetzen eine Matrix entstehen, die nur positive Eigenwerte hat. Für eine beliebige komplexe $n \times n$-Matrix A ist aber $A\overline{A}$ nicht notwendigerweise positiv, wie das Beispiel

$$A = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \qquad \Rightarrow \qquad A\overline{A} = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I$$

zeigt. Eine positive Matrix entsteht dagegen immer, wenn man statt A die Adjungierte $A^* = \overline{A}^t$ verwendet.

Die Substitution von A für z und A^* für $\overline{z}$ in einem Polynom $p(z, \overline{z})$ ist nicht unbedingt eindeutig. Schon das Polynom $p(z, \overline{z}) = z\overline{z}$ kann man auch als $\overline{z}z$ schreiben. Damit die Substition eindeutig wird, muss man also fordern, dass $AA^* = A^*A$ ist.

Definition 5.38. *Eine Matrix $A \in M_n(\mathbb{C})$ heisst* normal*, wenn $AA^* = A^*A$ gilt.*

Beispiele normaler Matrizen

1. Hermitesche und Antihermitesche Matrizen sind normal, denn solche Matrizen erfüllen $A^* = \pm A$ und damit $AA^* = \pm A^2 = A^*A$.

2. Symmetrische und antisymmetrische Matrizen sind normal, denn aus $A = A^t$ folgt $A^* = \overline{A}^t$ und damit

$$AA^* = A\overline{A}^t =$$

$$A^*A =$$

3. Unitäre Matrizen U sind normal, das $UU^* = I = U^*U$ gilt.

4. Orthogonale Matrizen sind normal wegen $O(n) = U(n) \cap M_n(\mathbb{R})$.

Jede Matrix lässt sich durch Wahl einer geeigneten Basis in Jordansche Normalform bringen. Allerdings sind Jordan-Blöcke keine normalen Matrizen, wie der folgende Satz zeigt.

Satz 5.39. *Eine Dreiecksmatrix ist genau dann normal, wenn sie diagonal ist.*

Beweis. Sei A eine obere Dreiecksmatrix, das Argument für eine untere Dreiecksmatrix funktioniert gleich. Wir berechnen ein Diagonalelement für beide Produkte AA^* und A^*A. Dazu brauchen wir die Matrixelemente von A und A^*. Bezeichnen wir die Matrixelemente von A mit a_{ij}, dann hat A^* die Matrixelemente $(A^*)_{ij} = \overline{a}_{ji}$. Damit kann man die Diagonalelemente der Produkte als

$$(AA^*)_{ii} = \sum_{j=1}^{n} a_{ij}\overline{a}_{ij} = \sum_{j=i}^{n} |a_{ij}|^2$$
$$(A^*A)_{ii} = \sum_{j=1}^{n} \overline{a}_{ji}a_{ji} = \sum_{j=1}^{i} |a_{ji}|^2$$

ausrechnen. Der obere Ausdruck ist die quadrierte Länge der Zeile i der Matrix A, der untere ist die quadrierte Länge der Spalte i. Da die Matrix eine obere Dreiecksmatrix ist, hat die erste Spalte höchstens ein einziges von 0 verschiedenes Element. Daher kann auch die erste Zeile höchstens dieses eine Elemente haben. Die Matrix hat daher Blockstruktur mit einem 1×1-Block in der linken obere Ecke und einem $n-1$-dimensionalen Block für den Rest. Durch Wiederholen des Arguments für den $(n-1) \times (n-1)$-Block kann man so schrittweise schliessen, dass die Matrix A diagonal sein muss. □

Satz 5.40. *Sind A und B normale Matrizen und $AB^* = B^*A$, dann sind auch $A + B$ und AB normal.*

Beweis. Zunächst folgt aus $AB^* = B^*A$ auch $A^*B = (B^*A)^* = (AB^*)^* = BA^*$. Der Beweis erfolgt durch Nachrechnen:

$$(A + B)(A + B)^* = AA^* + AB^* + BA^* + BB^*$$
$$(A + B)^*(A + B) = A^*A + A^*B + B^*A + B^*B$$

Die ersten und letzten Terme auf der rechten Seite stimmen überein, weil A und B normal sind. Die gemischten Terme stimmen überein wegen der Vertauschbarkeit von A und B^*.

Für das Produkt rechnet man

$$(AB)(AB)^* = ABB^*A^* = AB^*BA^* = B^*AA^*B = B^*A^*AB = (AB)^*(AB),$$

was zeigt, dass auch AB normal ist. □

Äquivalente Bedingungen

Es gibt eine grosse Zahl äquivalenter Eigenschaften für normale Matrizen. Die folgenden Eigenschaften sind äquivalent:

1. Die Matrix A ist mit einer unitären Matrix diagonalisierbar
2. Es gibt eine orthonormale Basis von Eigenvektoren von A für $\mathbb{C}^n$
3. Für jeden Vektor $x \in \mathbb{C}^n$ gilt $\|Ax\| = \|A^*x\|$
4. Die Forbenius-Norm der Matrix A kann mit den Eigenwerten λ_i von A berechnet werden: $\operatorname{Spur}(A^*A) = \sum_{i=1}^n |\lambda_i|^2$
5. Der hermitesche Teil $\frac{1}{2}(A + A^*)$ und der antihermitesche Teil $\frac{1}{2}(A - A^*)$ von A vertauschen.
6. A^* ist ein Polynom vom Grad $n - 1$ in A.
7. Es gibt eine unitäre Matrix U derart, dass $A^* = AU$
8. Es gibt eine Polarzerlegugn $A = UP$ mit einer unitären Matrix U und einer postiv semidefiniten Matrix P, die untereinander vertauschen.
9. Es gibt eine Matrix N mit verschiedenen Eigenwerten, mit denen A vertauscht.
10. Wenn A die (absteigend geordneten) singulärwerte σ_i und die absteigend geordneten Eigenwerte λ_i hat, dann it $\sigma_i = |\lambda_i|$.

Übungsaufgaben

5.1. Verwenden Sie die Matrixdarstellung komplexer Zahlen, um i^i zu berechnen.

Hinweis. Verwenden Sie die Eulersche Formel um $\log J$ zu bestimmen.

Lösung. Wir berechnen J^J mit Hilfe des Logarithmus als $J^J = \exp(J \log J)$. Zunächst erinnern wir an die Eulersche Formel

$$\exp tJ = \sum_{k=0}^{\infty} \frac{t^k J^k}{k!} = \sum_{i=0}^{\infty} \frac{t^{2i}(-1)^i}{(2i)!} \cdot I + \sum_{i=0}^{\infty} \frac{t^{2i+1}(-1)^i}{(2i+1)!} \cdot J = \cos t \cdot I + \sin t \cdot J.$$

Daraus liest man ab, dass

$$\log \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = tJ$$

gilt. Für die Matrix J heisst das

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \cos \frac{\pi}{2} & -\sin \frac{\pi}{2} \\ \sin \frac{\pi}{2} & \cos \frac{\pi}{2} \end{pmatrix} \qquad \Rightarrow \qquad \log J = \frac{\pi}{2} J. \tag{5.17}$$

Als nächstes müssen wir $J \log J$ berechnen. Aus (5.17) folgt

$$J \log J = J \cdot \frac{\pi}{2} J = -\frac{\pi}{2} \cdot I.$$

Darauf ist die Exponentialreihe auszuwerten, also

$$J^J = \exp(J \log J) = \exp(-\frac{\pi}{2} I) = \exp\begin{pmatrix} -\frac{\pi}{2} & 0 \\ 0 & -\frac{\pi}{2} \end{pmatrix} = \begin{pmatrix} e^{-\frac{\pi}{2}} & 0 \\ 0 & e^{-\frac{\pi}{2}} \end{pmatrix} = e^{-\frac{\pi}{2}} I.$$

Als komplexe Zahlen ausgedrückt folgt also $i^i = e^{-\frac{\pi}{2}}$. ○

5.2. Seien z und w komplexe Zahlen derart, dass $z = e^w$, d. h. w ist ein Wert des Logarithmus von z. Zeigen Sie, dass die Zahlen $w + 2\pi ik$ für $k \in \mathbb{Z}$ ebenfalls Logarithmen von z sind. Dies zeigt, dass eine komlexe Zahl unendlich viele verschiedene Logarithmen haben kann, die Logarithmusfunktion ist im Komplexen nicht eindeutig.

Lösung. Aus der Eulerschen Formel folgt

$$e^{w+2\pi ik} = e^w \cdot e^{2\pi ik} = e^w(\underbrace{\cos 2\pi k}_{=1} + i \underbrace{\sin 2\pi k}_{=0}) = e^w = z.$$ ○

5.3. Finden Sie eine Basis von $\mathbb{Q}^4$ derart, dass die Matrix A

$$A = \begin{pmatrix} -13 & 5 & -29 & 29 \\ -27 & 11 & -51 & 51 \\ -3 & 1 & -2 & 5 \\ -6 & 2 & -10 & 13 \end{pmatrix}$$

Jordansche Normalform hat.

Lösung. Zunächst muss man die Eigenwerte finden. Dazu kann man das charakteristische Polynom berechnen, man findet nach einiger Rechnung oder mit Hilfe einer Software für symbolische Rechnung:

$$\chi_A(\lambda) = x^4 - 9x^3 + 30x^2 - 44x + 24 = (x-3)^3(x-2),$$

Eigenwerte sind also $\lambda = 3$ und $\lambda = 2$.

Der Eigenwert $\lambda = 2$ ist ein einfacher Eigenwert, der zugehörige Eigenraum ist daher eindimensional. Ein Eigenvektor kann mit Hilfe des linearen Gleichungssystems

$$\begin{array}{|cccc|} \hline -13-\lambda & 5 & -29 & 29 \\ -27 & 11-\lambda & -51 & 51 \\ -3 & 1 & -2-\lambda & 5 \\ -6 & 2 & -10 & 13-\lambda \\ \hline \end{array} \rightarrow \begin{array}{|cccc|} \hline -16 & 5 & -29 & 29 \\ -27 & 8 & -51 & 51 \\ -3 & 1 & -5 & 5 \\ -6 & 2 & -10 & 10 \\ \hline \end{array} \rightarrow \begin{array}{|cccc|} \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \\ \hline \end{array}$$

gefunden werden. Daraus liest man den Eigenvektor

$$b_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \qquad Ab_1 = \begin{pmatrix} -13 & 5 & -29 & 29 \\ -27 & 11 & -51 & 51 \\ -3 & 1 & -2 & 5 \\ -6 & 2 & -10 & 13 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 3 \\ 3 \end{pmatrix} = 3b_1$$

ab. Diesen Vektor können wir auch finden, indem wir $\mathcal{J}(A-2I)$ bestimmen. Die vierte Potenz von $A-2I$ ist

$$(A-2I)^4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 2 & -1 \end{pmatrix}, \tag{5.18}$$

der zugehörige Bildraum ist wieder aufgespannt von b_1.

Aus (5.18) kann man aber auch eine Basis

$$b_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad b_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad b_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

für den Kern $\mathcal{K}(A-2I)$ ablesen. Da $\lambda = 2$ der einzige andere Eigenwert ist, muss $\mathcal{K}(A-2I) = \mathcal{J}(A-3I)$ sein. Dies lässt sich überprüfen, indem wir die vierte Potenz von $A-2I$ berechnen, sie ist

$$(A-2I)^4 = \begin{pmatrix} 79 & -26 & 152 & -152 \\ 162 & -53 & 312 & -312 \\ 12 & -4 & 23 & -23 \\ 24 & -8 & 46 & -46 \end{pmatrix}.$$

Die Spaltenvektoren lassen sich alle durch die Vektoren b_2, b_3 und b_4 ausdrücken, also ist $\mathcal{J}(A-2I) = \langle b_2, b_3, b_4 \rangle$.

Indem die Vektoren b_i als Spalten in eine Matrix T schreibt, kann man jetzt berechnen, wie die Matrix der linearen Abbildung in dieser neuen Basis aussieht, es ist

$$A' = T^{-1}AT \left(\begin{array}{c|ccc} 3 & 0 & 0 & 0 \\ \hline 0 & -13 & 5 & 29 \\ 0 & -27 & 11 & 51 \\ 0 & -3 & 1 & 8 \end{array}\right),$$

wir haben also tatsächlich die versprochene Blockstruktur.

Der 3×3-Block

$$A_1 = \begin{pmatrix} -13 & 5 & 29 \\ -27 & 11 & 51 \\ -3 & 1 & 8 \end{pmatrix}$$

in der rechten unteren Ecke hat den dreifachen Eigenwert 2, und die Potenzen von $A_1 - 2I$ sind

$$A_1 - 2I \begin{pmatrix} -15 & 5 & 29 \\ -27 & 9 & 51 \\ -3 & 1 & 6 \end{pmatrix}, \qquad (A_1-2I)^2 = \begin{pmatrix} 3 & -1 & -6 \\ 9 & -3 & -18 \\ 0 & 0 & 0 \end{pmatrix}, \qquad (A_1-2I)^3 = 0.$$

Für die Jordan-Normalform brauchen wir einen von 0 verschiedenen Vektor im Kern von $(A_1-2I)^2$, zum Beispiel den Vektor mit den Komponenten $1, 3, 1$. Man beachte aber, dass diese Komponenten jetzt in der neuen Basis $b_2, \ldots, b_4$ zu verstehen sind, d. h. der Vektor, den wir suchen, ist

$$c_3 = b_1 + 3b_2 + b_3 = \begin{pmatrix} 1 \\ 3 \\ 1 \\ 2 \end{pmatrix}.$$

Jetzt berechnen wir die Bilder von c_3 unter $A - 2I$:

$$c_2 = \begin{pmatrix} 29 \\ 51 \\ 6 \\ 12 \end{pmatrix}, \qquad c_1 = \begin{pmatrix} -6 \\ -18 \\ 0 \\ 0 \end{pmatrix}.$$

Die Basis b_1, c_1, c_2, c_3 ist also eine Basis, in der die Matrix A Jordansche Normalform annimmt.

Die Umrechnung der Matrix A in die Basis $\{b_1, c_1, c_2, c_3\}$ kann mit der Matrix

$$T_1 = \begin{pmatrix} 0 & -6 & 29 & 1 \\ 0 & -18 & 51 & 3 \\ 1 & 0 & 6 & 1 \\ 1 & 0 & 12 & 2 \end{pmatrix}, \qquad T_1^{-1} = \frac{1}{216} \begin{pmatrix} 0 & 0 & 432 & -216 \\ 33 & -23 & -36 & 36 \\ 18 & -6 & 0 & 0 \\ -108 & 36 & -216 & 216 \end{pmatrix}$$

erfolgen und ergibt die Jordansche Normalform

$$A' = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

wie erwartet. ○

5.4. Berechnen Sie $\sin At$ für die Matrix

$$A = \begin{pmatrix} \omega & 1 \\ 0 & \omega \end{pmatrix}.$$

Kontrollieren Sie Ihr Resultat, indem Sie den Fall $\omega = 0$ gesondert ausrechnen.

Hinweis. Schreiben Sie $A = \omega I + N$ mit einer nilpotenten Matrix.

Lösung. Man muss At in die Potenzreihe

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \dots$$

für die Sinus-Funktion einsetzen. Mit der Schreibweise $A = \omega I + N$, wobei $N^2 = 0$ können die Potenzen etwas leichter berechnet werden:

$$\begin{aligned} A^0 &= I \\ A^1 &= \omega I + N \\ A^2 &= \omega^2 I + 2\omega N \\ A^3 &= \omega^3 I + 3\omega^2 N \\ A^4 &= \omega^4 I + 4\omega^3 N \\ &\vdots \end{aligned}$$

$$A^k = \omega^k I + k\omega^{k-1} N$$

Damit kann man jetzt $\sin At$ berechnen:

$$\begin{aligned}
\sin At &= At - \frac{A^3t^3}{3!} + \frac{A^5t^5}{5!} - \frac{A^7t^7}{7!} \dots \\
&= \Big(\omega t - \frac{\omega^3t^3}{3!} + \frac{\omega^5t^5}{5!} - \frac{\omega^7t^7}{7!} + \dots\Big) I + \Big(t - \frac{3\omega^2t^3}{3!} + \frac{5\omega^4t^5}{5!} - \frac{7\omega^6t^7}{7!} + \dots\Big) N \\
&= I \sin \omega t + tN\Big(1 - \frac{\omega^2t^2}{2!} + \frac{\omega^4t^4}{4!} - \frac{\omega^6t^6}{6!} + \dots\Big) \\
&= I \sin \omega t + tN \cos \omega t. \qquad (5.19)
\end{aligned}$$

Im Fall $\omega = 0$ ist $A = N$ und $A^2 = 0$, so dass

$$\sin At = tN,$$

dies stimmt mit (5.19) für $\omega = 0$ überein, da $\cos \omega t = \cos 0 = 1$ in diesem Fall. ○

5.5. Rechnen Sie nach, dass die Matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

normal ist.

a) Berechnen Sie die Eigenwerte, indem Sie das charakteristische Polynom von A und seine Nullstellen bestimmen.

b) Das Polynom

$$p(z, \overline{z}) = \frac{(3 - \sqrt{3}) z\overline{z} - 9(1 - \sqrt{3})}{6}$$

hat die Eigenschaft, dass

$$p(\lambda, \lambda) = |\lambda|$$

für alle drei Eigenwerte von A. Verwenden Sie dieses Polynom, um $B = |A|$ zu berechen.

c) Überprüfen Sie Ihr Resultat, indem Sie mit einem Computeralgebra-Programm die Eigenwerte von B bestimmen.

Lösung. Die Matrix A ist von der Form $2I + O$ mit $O \in \mathrm{SO}(3)$, für solche Matrizen wurde gezeigt, dass sie normal sind. Man kann aber auch direkt nachrechnen:

$$AA^t = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 5 \end{pmatrix}$$

$$A^tA = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 5 \end{pmatrix}$$

Es gilt also $AA^t = A^tA$, die Matrix ist also normal.

a) Das charakteristische Polynom ist

$$\chi_A(\lambda) = \begin{vmatrix} 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 1 \\ 1 & 0 & 2-\lambda \end{vmatrix} = (2-\lambda)^3 + 1 \tag{5.20}$$
$$= -\lambda^3 - 6\lambda^2 + 12\lambda + 9.$$

Mit einem Taschenrechner kann man die Nullstellen finden, aber man kann das auch die Form (5.20) des charakteristischen Polynoms direkt faktorisieren:

$$\begin{aligned}\chi_A(\lambda) &= (2-\lambda)^3 + 1 \\ &= ((2-\lambda)+1)((2-\lambda)^2 - (2-\lambda) + 1) \\ &= (3-\lambda)(\lambda^2 - 3\lambda + 4 - 2 + \lambda + 1) \\ &= (3-\lambda)(\lambda^2 - 2\lambda + 3)\end{aligned}$$

Daraus kann man bereits einen Eigenwert $\lambda = 3$ ablesen, die weiteren Eigenwerte sind die Nullstellen des zweiten Faktors, die man mit der Lösungsformel für quadratische Gleichungen finden kann:

$$\lambda_\pm = \frac{3 \pm \sqrt{9-12}}{2} = \frac{3}{2} \pm \frac{\sqrt{-3}}{2} = \frac{3}{2} \pm i\frac{\sqrt{3}}{2}$$

b) Wir müssen $z = A$ und $\overline{z} = A^t$ im Polynom $p(z, \overline{z})$ substituieren und erhalten

$$B = \frac{3-\sqrt{3}}{6}\begin{pmatrix} 5 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 5 \end{pmatrix} + \frac{\sqrt{3}-1}{2} I$$
$$= \begin{pmatrix} 2.1547005 & 0.42264973 & 0.42264973 \\ 0.4226497 & 2.15470053 & 0.42264973 \\ 0.4226497 & 0.42264973 & 2.15470053 \end{pmatrix}$$

c) Tatsächlich gibt die Berechnung der Eigenwerte den einfachen Eigenwert $\mu_0 = 3 = |\lambda_0|$ und den doppelten Eigenwert $\mu_\pm = \sqrt{3} = 1.7320508 = |\lambda_\pm|$. ○

5.6. Man findet eine Basis, in der die Matrix

$$A = \begin{pmatrix} -5 & 2 & 6 & 0 \\ -11 & 12 & -3 & -15 \\ -7 & 0 & 9 & 4 \\ 0 & 5 & -7 & -8 \end{pmatrix}$$

die relle Normalform bekommt.

Lösung. Das charakteristische Polynom der Matrix ist

$$\chi_A(\lambda) = \lambda^4 - 8\lambda^3 + 42\lambda^2 - 104\lambda + 169 = (\lambda^2 - 4\lambda + 13)^2.$$

Es hat die doppelten Nullstellen

$$\lambda_\pm = 2 \pm \sqrt{4-13} = 2 \pm \sqrt{-9} = 2 \pm 3i.$$

Zur Bestimmung der Basis muss man jetzt zunächst den Kern von $A_+ = A - \lambda_+ I$ bestimmen, zum Beispiel mit Hilfe des Gauss-Algorithmus, man findet

$$b_1 = \begin{pmatrix} 1+i \\ 2+2i \\ i \\ 1 \end{pmatrix}.$$

Als nächstes braucht man einen Vektor $b_1 \in \ker A_+^2$, der b_1 auf $b_1 + \lambda_+ b_2$ abbildet. Durch Lösen des Gleichungssystems $Ab_2 - \lambda_+ b_2 = b_1$ findet man

$$b_2 = \begin{pmatrix} 2-i \\ 3 \\ 2 \\ 0 \end{pmatrix} \qquad \text{und damit weiter} \qquad \overline{b}_1 = \begin{pmatrix} 1-i \\ 2-2i \\ -i \\ 1 \end{pmatrix}, \quad \overline{b}_2 = \begin{pmatrix} 2+i \\ 3 \\ 2 \\ 0 \end{pmatrix}.$$

Als Basis für die reelle Normalform von A kann man jetzt die Vektoren

$$c_1 = b_1 + \overline{b}_1 = \begin{pmatrix} 2 \\ 4 \\ 0 \\ 2 \end{pmatrix}, \quad d_1 = \frac{1}{i}(b_1 - \overline{b}_1) = \begin{pmatrix} 2 \\ 4 \\ 2 \\ 0 \end{pmatrix}, \quad c_2 = b_2 + \overline{b}_2 = \begin{pmatrix} 4 \\ 6 \\ 4 \\ 0 \end{pmatrix}, \quad d_2 = \frac{1}{i}(b_2 - \overline{b}_2) = \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

verwenden. In dieser Basis hat A die Matrix

$$A' = \begin{pmatrix} 2 & 3 & 1 & 0 \\ -3 & 2 & 0 & 1 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & -3 & 2 \end{pmatrix},$$

wie man einfach nachrechnen kann. ○

Kapitel 6

Permutationen

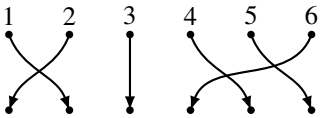
Die Berechnung der Determinante einer Matrix macht ausgedehnten Gebrauch von der Tatsache, dass die Vertauschung von zwei Zeilen oder Spalten das Vorzeichen des Wertes der Determinanten dreht. In diesem Kapitel sollen die Permutationen der Zeilen abstrakt untersucht werden. Wir erhalten so eine abstrakte Permutationsgruppe. Ihre Elemente lassen sich auch durch spezielle Matrizen beschreiben, eine Darstellung dieser Gruppe, die auch unmittelbar zu einer Formel für die Determinante einer Matrix führt.

6.1 Permutationen einer endlichen Menge

Eine endliche Anzahl n von Objekten können auf $n!$ Arten angeordnet werden. Als Objektmenge nehmen wir $[n] = \{1, \ldots, n\}$. Die Operation, die die Objekte in eine bestimmte Reihenfolge bringt, ist eine Abbildung $\sigma\colon [n] \to [n]$. Eine Permutation ist eine umkehrbare Abbildung $[n] \to [n]$. Die Menge S_n aller umkehrbaren Abbildungen $[n] \to [n]$ mit der Verknüpfung von Abbildungen als Operation heisst die die *symmetrische Gruppe*. Die identische Abbildung $\sigma(x) = x$ ist das *neutrale Element* der Gruppe S_n und wir auch mit e bezeichnet.

6.1.1 Permutationen als $2 \times n$-Matrizen

Eine Permutation kann als $2 \times n$-Matrix geschrieben werden:



$$= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix}$$

Das neutrale Element hat die Matrix

$$e = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$$

aus zwei identischen Zeilen.

Die Verknüpfung zweier solcher Permutationen kann leicht graphisch dargestellt werden: dazu werden die beiden Permutationen untereinander geschrieben und Spalten der zweiten Permutation in

der Reihen folge der Zahlen in der zweiten Zeile der ersten Permutation angeordnet. Die zusammengesetzte Permutation kann dann in der zweiten Zeile der zweiten Permutation abgelesen werden:

$$\sigma_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix}$$

$$\sigma_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 1 & 2 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 & 3 & 5 & 6 & 4 \\ 4 & 3 & 5 & 1 & 2 & 6 \end{pmatrix} \qquad \sigma_2\sigma_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 5 & 1 & 2 & 6 \end{pmatrix}$$

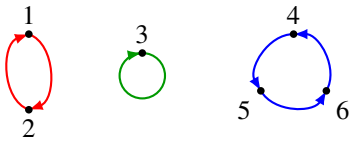
Die Inverse einer Permutation kann erhalten werden, indem die beiden Zeilen vertauscht werden und dann die Spalten wieder so angeordnet werden, dass die Zahlen in der ersten Zeile ansteigend sind:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix} \quad \Rightarrow \quad \sigma^{-1} = \begin{pmatrix} 2 & 1 & 3 & 5 & 6 & 4 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 6 & 4 & 5 \end{pmatrix}.$$

6.1.2 Zyklenzerlegung

Eine Permutation $\sigma \in S_n$ kann auch mit sogenanten Zyklenzerlegung analysiert werden. Zum Beispiel:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix} =$$



Definition 6.1. *Ein Zyklus Z ist eine unter σ invariante Teilmenge von $[n]$ minimaler Grösse. Die Zyklenzerlegung ist eine Zerlegung von $[n]$ in Zyklen*

$$[n] = \cup_{i=1}^{k} Z_i,$$

wobei jede Menge Z_i ein Zyklus ist.

Der folgende Algorithmus findet die Zyklenzerlegung einer Permutation.

Satz 6.2. *Sei $\sigma \in S_n$ eine Permutation. Der folgende Algorithmus findet die Zyklenzerlegung von σ:*

1. $i = 1$
2. *Wähle das erste noch nicht verwendete Element*
$$s_i = \min\Big([n] \setminus \bigcup_{j<i} Z_j\Big)$$
3. *Bestimme alle Elemente, die aus s_i durch Anwendung von σ entstehen:*
$$Z_i = \{s_i, \sigma(s_i), \sigma(\sigma(s_i)), \dots\} = \{\sigma^k(s_i) \mid k \geq 0\}.$$
4. *Falls $\bigcup_{j\leq i} Z_j \neq [n]$, erhöhe i um 1 und fahre weiter bei 2.*

Mit Hilfe der Zyklenzerlegung von σ lassen sich auch gewisse Eigenschaften von σ ableiten. Sei also $[n] = Z_1 \cup \cdots \cup Z_k$ die Zyklenzerlegung. Für jedes Element $x \in S_i$ gilt $\sigma^{|S_i|}(x) = x$. Die kleinste Zahl m, für die $\sigma^m = e$ ist, das kleinste gemeinsame Vielfache der Zyklenlängen:

$$m = \mathrm{kgV}(|Z_1|, |Z_2|, \dots, |Z_k|).$$

6.1.3 Konjugierte Elemente in S_n

Zwei Elemente $g_1, g_2 \in G$ einer Gruppe heissen konjugiert, wenn es ein Element $c \in G$ gibt derart, dass $cg_1c^{-1} = g_2$. Bei Matrizen hat dies bedeutet, dass die beiden Matrizen durch Basiswechsel auseinander hervorgehen. Dasselbe lässt sich auch im Kontext der symmetrischen Gruppe sagen.

Seien σ_1 und σ_2 zwei konjugierte Permutationen in S_n. Es gibt also eine Permutation $\gamma \in S_n$ derart, dass $\sigma_1 = \gamma\sigma_2\gamma^{-1}$ oder $\gamma^{-1}\sigma_1\gamma = \sigma_2$. Dann gilt auch für die Potenzen

$$\sigma_1^k = \gamma\sigma_2^k\gamma^{-1}. \tag{6.1}$$

Ist Z_i ein Zyklus von σ_2 und $x \in Z_i$, dann ist $Z_i = \{x, \sigma_2(x), \sigma_2^2(x), \dots\}$. Die Menge $\gamma(Z_i)$ besteht dann aus dem Elementen $\gamma(Z_i) = \{\gamma(x), \gamma(\sigma_2(x)), \gamma(\sigma_2^2(x)), \dots\}$. Aus der Formel (6.1) folgt $\sigma_1^k\gamma = \gamma\sigma_2^k$, also

$$\gamma(Z_i) = \{\gamma(x), \sigma_1(\gamma(x)), \sigma_1^2(\gamma(x)), \dots\},$$

Also ist $\gamma(Z_i)$ ein Zyklus von σ_1. Die Permutation γ bildet also Zyklen von σ_2 auf Zyklen von σ_1 ab. Es folgt daher der folgende Satz:

Satz 6.3. *Sind $\sigma_1, \sigma_2 \in S_n$ konjugiert $\sigma_1 = \gamma\sigma_2\gamma^{-1}$ mit dem $\gamma \in S_n$. Wenn $Z_1, \dots, Z_k$ die Zyklen von σ_2 sind, dann sind $\gamma(Z_1), \dots, \gamma(Z_k)$ die Zyklen von σ_1.*

Die Zyklenzerlegung kann mit der Jordan-Normalform **??** einer Matrix verglichen werden. Durch einen Basiswechsel, welcher durch eine "Konjugation" von Matrizen ausgedrückt wir, kann die Matrix in eine besonders übersichtliche Form gebracht werden. Wenn σ die Zyklenzerlegung $Z_1, \dots, Z_k$ mit Zyklenlängen $l_i = |Z_i|$, dann kann man die Menge $[n]$ wie folgt in Teilmengen

$$\begin{aligned}
X_1 &= \{1, \dots, l_1\},\\
X_2 &= \{l_1 + 1, \dots, l_1 + l_2\},\\
X_i &= \{l_1 + \cdots + l_{i-1} + 1, \dots, l_1 + \cdots + l_i\}\\
X_k &= \{l_1 + \cdots + l_{k-1} + 1, \dots n\}
\end{aligned}$$

zerlegen. Sei σ_2 die Permutation, die in jeder der Mengen X_i durch zyklische Vertauschung der Elemente wirkt. Indem man die Elemente von Z_i in der Reihenfolge, in der sie durch σ_1 erreicht werden, auf die Elemente X_i abbildet, findet man eine Permutation, die Zyklen von σ_1 in Zyklen von σ_2 überführt.

Satz 6.4. *Wenn zwei Elemente $\sigma_1, \sigma_2 \in S_n$ Zyklenzerlegungen mit den gleichen Zyklenlängen haben, dann sind sie konjugiert.*

Ein Element $\sigma \in S_n$ ist also bis auf eine Permutation vollständig durch die Länge der Zyklen von σ charakterisiert.

6.2 Permutationen und Transpositionen

Im vorangegangenen Abschnitt haben wir Permutationen durch die Zyklenzerlegung charakterisiert. Es zeigt sich aber, dass sich eine Permutation in noch elementarere Bausteine zerlegen lässt, die Transpositionen.

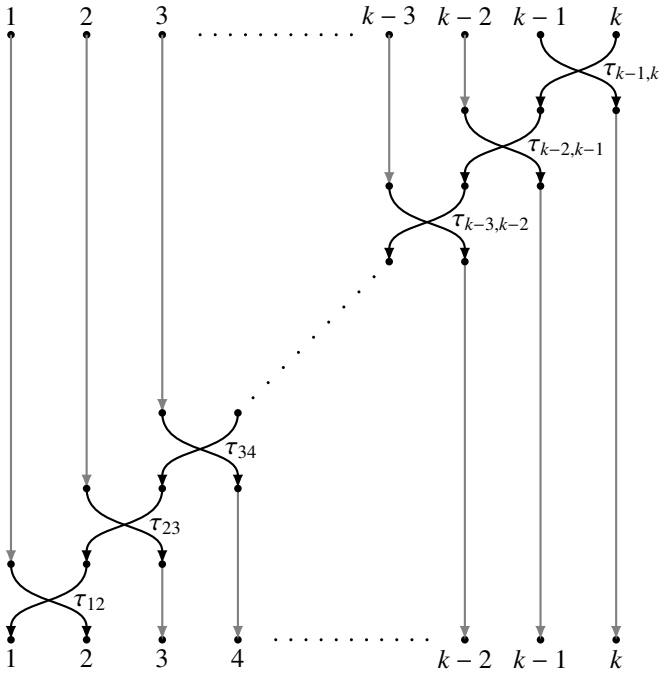
Definition 6.5. *Einen Transposition $\tau \in S_n$ ist ein Permutation, die genau zwei Elemente vertauscht. Die Transposition τ_{ij} ist definiert durch*

$$\tau_{ij}(x) = \begin{cases} i & x = j \\ j & x = i \\ x & \text{sonst.} \end{cases}$$

Eine Transposition hat genau einen Zyklus der Länge 2, alle anderen Zyklen haben die Länge 1.

6.2.1 Zyklus und Permutationen aus Transpositionen

Sei σ die zyklische Vertauschung der Elemente $1, \ldots, k \in [n]$, also die Permutation, die $1 \to 2 \to 3 \to \cdots \to k-2 \to k-1 \to k \to 1$ abbildet. Dieser Zyklus lässt sich wie folgt aus Transpositionen zusammensetzen:



Es ist also

$$\sigma = \tau_{12}\tau_{23}\tau_{34}\ldots\tau_{k-3,k-2}\tau_{k-2,k-1}\tau_{k-1,k}.$$

Satz 6.6. *Jede Permutation $\sigma \in S_n$ lässt sich als ein Produkt von Transpositionen schreiben. Jeder Zyklus der Länge k lässt sich aus $k-1$ Transpositionen zusammensetzen. Eine Permutation mit einer Zerlegung in Zyklen der Längen $l_1, \ldots, l_p$ kann als Produkt von $l_1 + \cdots + l_p - p$ Transpositionen geschrieben werden.*

6.2.2 Signum einer Permutation

Die Anzahl Transpositionen, die benötigt werden, um eine Permutation zu beschreiben, ist nicht fest. Wenn σ mit k Transpositionen geschrieben werden kann und γ mit l, dann hat $\gamma\sigma\gamma^{-1}$ die gleiche

Zyklenzerlegung, kann aber mit $k + 2l$ Transpositionen geschrieben werden. Die Anzahl Transpositionen, die zur Darstellung einer Permutation nötig ist, ändert sich aber immer nur um eine gerade Zahl. Die Anzahl ist also keine Invariante einer Permutation, aber ob die Anzahl gerade ist oder nicht, ist sehr wohl eine charkterisierende Eigenschaft einer Permutation.

Definition 6.7. *Das* Vorzeichen *oder* Signum *einer Permutation σ ist die Zahl* $\operatorname{sgn}(\sigma) = (-1)^k$*, wenn σ als Produkt von k Transpositionen geschrieben werden kann.*

Die inverse Permutation σ^{-1} hat das gleiche Signum wie σ. Wenn nämlich $\sigma = \tau_1\tau_2\dots\tau_k$ geschrieben werden kann, dann ist $\sigma^{-1} = \tau_k\dots\tau_2\tau_1$, sowohl σ wie σ^{-1} können also mit der gleichen Zahl von Transpositionen geschrieben werden, sie haben also auch das gleiche Vorzeichen.

Die Abbildung $S_n \to \{\pm\}$, die einer Permutation das Signum zuordnet, ist ein Homomorphismus von Gruppen, d. h.

$$\operatorname{sgn}(\sigma_1\sigma_2) = \operatorname{sgn}(\sigma_1)\operatorname{sgn}(\sigma_2)$$

da ganz offensichtlich $\sigma_1\sigma_2$ mit $k_1 + k_2$ Transpositionen geschrieben kann, wenn σ_i mit k_i Transpositionen geschrieben werden kann.

Das Signum definiert in der symmetrischen Gruppe eine Teilmenge bestehnd aus den Permutationen mit Signum +1.

Definition 6.8. *Die Teilmenge*

$$A_n = \{\sigma \in S_n \mid \operatorname{sgn}(\sigma) = 1\} \subset S_n.$$

heisst die alternierende Gruppe *der Ordnung n Die Elemente von A_n heissen auch die* geraden *Permutationen, die Elemente von $S_n \setminus A_n$ heissen auch die* ungeraden *Permutationen.*

Die alternierende Gruppe A_n ist tatsächlich eine Untergruppe. Zunächst ist $\operatorname{sgn}(e) = (-1)^0 = 1$, also ist $e \in A_n$. Es wurde schon gezeigt, dass mit jedem Element $\sigma \in A_n$ auch das inverse Element $\sigma^{-1} \in A_n$ ist. Es muss aber noch sichergestellt werden, dass das Produkt von zwei geraden Transpositionen wieder gerade ist:

$$\begin{aligned}\sigma_1, \sigma_2 \in A_n \quad &\Rightarrow \quad \operatorname{sgn}(\sigma_1) = \operatorname{sgn}(\sigma_2) = 1 \\ &\Rightarrow \quad \operatorname{sgn}(\sigma_1\sigma_2) = \operatorname{sgn}(\sigma_1)\operatorname{sgn}(\sigma_2) = 1 \cdot 1 = 1 \quad \Rightarrow \quad \sigma_1\sigma_2 \in A_n.\end{aligned}$$

Damit ist gezeigt, dass die alternierende Gruppe A_n ein Untergruppe von S_n ist.

6.3 Permutationsmatrizen

Die Eigenschaft, dass eine Vertauschung das Vorzeichen kehrt, ist eine wohlebekannte Eigenschaft der Determinanten. In diesem Abschnitt soll daher eine Darstellung von Permutationen als Matrizen gezeigt werden und die Verbindung zwischen dem Vorzeichen einer Permutation und der Determinanten hergestellt werden.

6.3.1 Matrizen

Gegeben sei jetzt eine Permutation $\sigma \in S_n$. Aus σ lässt sich eine lineare Abbildung $\mathbb{k}^n \to \mathbb{k}^n$ konstruieren, die die Standardbasisvektoren permutiert, also

$$f_\sigma : \mathbb{k}^n \to \mathbb{k}^n : \begin{cases} e_1 \mapsto e_{\sigma(1)} \\ e_2 \mapsto e_{\sigma(2)} \\ \vdots \\ e_n \mapsto e_{\sigma(n)} \end{cases}$$

Die Matrix P_σ der linearen Abbildung f_σ hat in Spalte i genau eine 1 in der Zeile $\sigma(i)$, also

$$(P_\sigma)_{ij} = \delta_{j\sigma(i)}.$$

Beispiel. Die zur Permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 6 & 4 \end{pmatrix}$$

gehörige lineare Abbildung f_σ hat die Matrix

$$A_\sigma = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$ ○

Definition 6.9. *Eine Permutationsmatrix ist eine Matrix $P \in M_n(\mathbb{k})$ derart, die in jeder Zeile und Spalte genau eine* 1 *enthalten ist, während alle anderen Matrixelemente* 0 *sind.*

Es ist klar, dass aus einer Permutationsmatrix auch die Permutation der Standardbasisvektoren abgelesen werden kann. Die Verknüpfung von Permutationen wird zur Matrixmultiplikation von Permutationsmatrizen, die Zuordnung $\sigma \mapsto P_\sigma$ ist also ein Homomorphismus $S_n \to M_n(\mathbb{k}^n)$, es ist $P_{\sigma_1\sigma_2} = P_{\sigma_1} P_{\sigma_2}$.

6.3.2 Transpositionen

Transpositionen sind Permutationen, die genau zwei Elemente von $[n]$ vertauschen. Wir ermitteln jetzt die Permutationsmatrix der Transposition $\tau = \tau_{ij}$

$$P_{\tau_{ij}} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & \dots & 1 & & \\ & & & \vdots & & \vdots & & \\ & & & 1 & \dots & 0 & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}$$

Die Permutation σ mit dem Zyklus $1 \to 2 \to \cdots \to l-1 \to l \to 1$ der Länge l kann aus aufeinanderfolgenden Transpositionen zusammengesetzt werden, die zugehörigen Permutationsmatrizen sind

$$
\begin{aligned}
P_\sigma &= P_{\tau_{12}} P_{\tau_{23}} P_{\tau_{34}} \dots P_{\tau_{l-2,l-1}} P_{\tau_{l-1,l}} \\
&= \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \dots \\
&= \begin{pmatrix} 0 & 0 & 1 & 0 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \dots \\
&= \begin{pmatrix} 0 & 0 & 0 & 1 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \\
&\vdots \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}
\end{aligned}
$$

6.3.3 Determinante und Vorzeichen

Die Transpositionen haben Permutationsmatrizen, die aus der Einheitsmatrix entstehen, indem genau zwei Zeilen vertauscht werden. Die Determinante einer solchen Permutationsmatrix ist

$$\det P_\tau = -\det E = -1 = \operatorname{sgn}(\tau).$$

Nach der Produktregel für die Determinante folgt für eine Darstellung der Permutation $\sigma = \tau_1 \dots \tau_l$ als Produkt von Transpositionen, dass

$$\det P_\sigma = \det P_{\tau_1} \dots \det P_{\tau_l} = (-1)^l = \operatorname{sgn}(\sigma).$$

Das Vorzeichen einer Permutation ist also identisch mit der Determinante der zugehörigen Permutationsmatrix.

6.4 Determinante

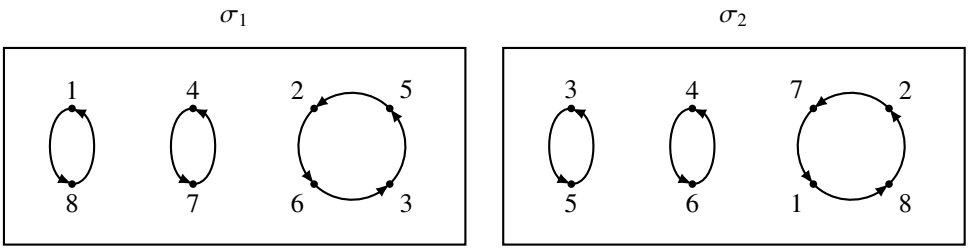
Übungsaufgaben

6.1. Sind die beiden Permutationen

$$\sigma_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 6 & 5 & 7 & 2 & 3 & 4 & 1 \end{pmatrix} \qquad \text{und} \qquad \sigma_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 7 & 5 & 6 & 3 & 4 & 1 & 2 \end{pmatrix}$$

konjugiert in S_8? Wenn ja, finden Sie eine Permutation γ derart, dass $\gamma\sigma_1\gamma^{-1} = \sigma_2$

Lösung. Die Zyklenzerlegungen von σ_1 und σ_2 sind



Da die beiden Permutationen die gleiche Zyklenzerlegung haben, müssen sie konjugiert sein. Die Permutation

$$\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 5 & 1 & 4 & 8 & 7 & 2 & 3 \end{pmatrix}$$

bildet die Zyklenzerlegung ab, also ist $\gamma\sigma_1\gamma^{-1} = \sigma_2$. ○

Kapitel 7

Matrizengruppen

Matrizen können dazu verwendet werden, Symmetrien von geometrischen oder physikalischen Systemen zu beschreiben. Neben diskreten Symmetrien wie zum Beispiel Spiegelungen gehören dazu auch kontinuierliche Symmetrien wie Translationen oder Invarianz einer phyisikalischen Grösse über die Zeit. Solche Symmetrien müssen durch Matrizen beschrieben werden können, die auf stetige oder sogar differenzierbare Art von der Zeit abhängen. Die Menge der Matrizen, die zur Beschreibung solcher Symmetrien benutzt werden, muss also eine zusätzliche Struktur haben, die ermöglicht, sinnvoll über Stetigkeit und Differenzierbarkeit bei Matrizen zu sprechen.

Die Menge der Matrizen bilden zunächst eine Gruppe, die zusätzliche differenziarbare Struktur macht daraus eine sogenannte Lie-Gruppe. Die Ableitungen nach einem Parameter liegen in der sogenannten Lie-Algebra, einer Matrizen-Algebra mit dem antisymmetrischen Lie-Klammer-Produkt $[A, B] = AB - BA$, auch Kommutator genannt. Lie-Gruppe und Lie-Algebra sind eng miteinander verknüpft, so eng, dass sich die meisten Eigenschaften der Gruppe aus den Eigenschaften der Lie-Gruppe aus der Lie-Algebra ableiten lassen. Die Verbindung wird hergestellt durch die Exponentialabbildung. Ziel dieses Kapitels ist, die Grundzüge dieses interessanten Zusammenhangs darzustellen.

7.1 Symmetrien

Der geometrische Begriff der Symmetrie meint die Eigenschaft eines geometrischen Objektes, dass es bei einer Bewegung auf sich selbst abgebildet wird. Das Wort stammt aus dem altgriechischen, wo es *Gleichmass* bedeutet. Spiegelsymmetrische Objekte zeichnen sich zum Beispiel dadurch aus, dass Messungen von Strecken die gleichen Werte ergeben wie die Messungen der entsprechenden gespiegelten Strecken (siehe auch Abbildung 7.1, was die Herkunft des Begriffs verständlich macht. In der Physik wird dem Begriff der Symmetrie daher auch eine erweiterte Bedeutung gegeben. Jede Transformation eines Systems, welche bestimmte Grössen nicht verändert, wird als Symmetrie bezeichnet. Die Gesetze der Physik sind typischerweise unabhängig davon, wo man den den Nullpunkt der Zeit oder das räumlichen Koordinatensystems ansetzt, eine Transformation des Zeitnullpunktes oder des Ursprungs des Koordinatensystems ändert daher die Bewegungsgleichungen nicht, sie ist eine Symmetrie des Systems.

Umgekehrt kann man fragen, welche Symmetrien ein System hat. Da sich Symmetrien zusammensetzen und umkehren lassen, kann man in davon ausgehen, dass die Symmetrietransformationen eine Gruppe bilden. Besonders interessant ist dies im Falle von Transformationen, die durch Matri-

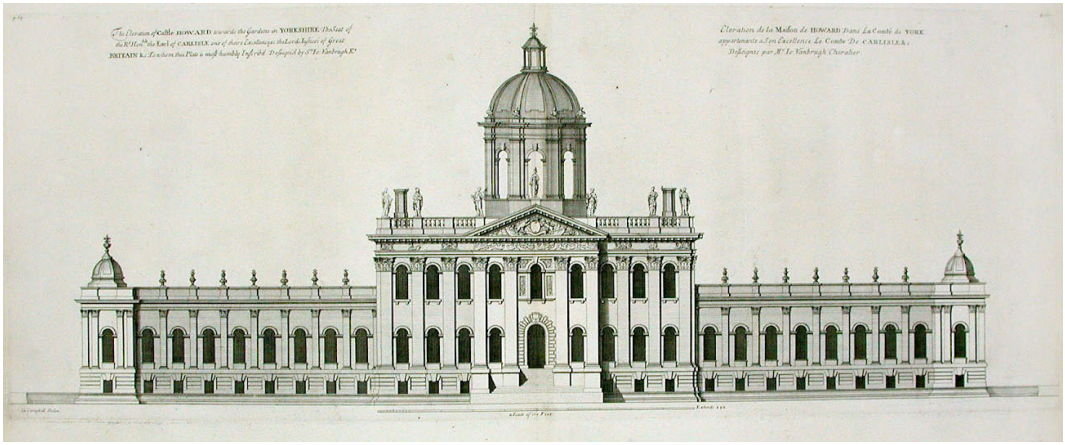


Abbildung 7.1: Das Castle Howard in Yorkshire war in dieser ausgeprägt symmetrischen Form geplant, wurde dann aber in modifizeirter Form gebaut. Messungen zwischen Punkten in der rechten Hälfte des Bildes ergeben die gleichen Werte wie Messungen entsprechenden Strecken in der linken Hälfte, was den Begriff Symmetrie rechtfertigt.

zen beschrieben weren. Eine unter der Symmetrie erhaltene Eigenschaft definiert so eine Untergruppe der Gruppe $\mathrm{GL}_n(\mathbb{R})$ der invertierbaren Matrizen. Die erhaltenen Eigenschaften definieren eine Menge von Gleichungen, denen die Elemente der Untergruppe genügen müssen. Als Lösungsmenge einer Gleichung erhält die Untergruppe damit eine zusätzliche geometrische Struktur, man nennt sie eine differenzierbare Mannigfaltigkeit. Dieser Begriff wird im Abschnitt 7.1.3 eingeführt. Es wird sich zum Beispiel zeigen, dass die Menge der Drehungen der Ebene mit den Punkten eines Kreises parametrisieren lassen, die Lösungen der Gleichung $x^2 + y^2 = 1$ sind.

Eine Lie-Gruppe ist eine Gruppe, die gleichzeitig eine differenzierbare Mannigfaltigkeit ist. Die Existenz von geometrischen Konzepten wie Tangentialvektoren ermöglicht zusätzliche Werkzeuge, mit denen diese Gruppe untersucht und verstanden werden können. Ziel dieses Abschnitts ist, die Grundlagen für diese Untersuchung zu schaffen, die dann im Abschnitt 7.3 durchgeführt werden soll.

7.1.1 Algebraische Symmetrien

Mit Matrizen lassen sich Symmetrien in einem geometrischen Problem oder in einem physikalischen System beschreiben. Man denkt dabei gerne zuerst an geometrische Symmetrien wie die Symmetrie unter Punktspiegelung oder die Spiegelung an der x_1-x_2-Ebene, wie sie zum Beispiel durch die Abbildungen

$$\mathbb{R}^3 \to \mathbb{R}^3 : x \mapsto -x \qquad \text{oder} \qquad \mathbb{R}^3 \to \mathbb{R}^3 : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

dargestellt werden. Beide haben zunächst die Eigenschaft, dass Längen und Winkel und damit das Skalarprodukt erhalten sind. Diese Eigenschaft allein erlaubt aber noch nicht, die beiden Transformationen zu unterscheiden. Die Punktspiegelung zeichnet sich dadurch aus, das alle Geraden und alle Ebenen durch den Ursprung auf sich selbst abgebildet werden. Dies funktioniert für die Ebenenspiegelung nicht, dort bleibt nur die Spiegelungsebene (die x_1-x_2-Ebene im vorliegenden Fall) und

ihre Normale erhalten. Die folgenden Beispiele sollen zeigen, wie solche Symmetriedefinitionen auf algebraische Bedingungen an die Matrixelemente führen.

Zu jeder Abbildung $f\colon \mathbb{R}^n \to \mathbb{R}^n$, unter der ein geometrisches Objekt in $\mathbb{R}^n$ symmetrisch ist, können wir sofort weitere Abbildungen angeben, die ebenfalls Symmetrien sind. Zum Beispiel sind die iterierten Abbildungen $f \circ f$, $f \circ f \circ f$ u. s. w., die wir auch f^n mit $n \in \mathbb{N}$ schreiben werden, ebenfalls Symmetrien. Wenn die Symmetrie auch umkehrbar ist, dann gilt dies sogar für alle $n \in \mathbb{Z}$. Wir erhalten so eine Abbildung $\varphi\colon \mathbb{Z} \to \mathrm{GL}_n(\mathbb{R}) : n \mapsto f^n$ mit den Eigenschaften $\varphi(0) = f^0 = I$ und $\varphi(n+m) = f^{n+m} = f^n \circ f^m = \varphi(n) \circ \varphi(m)$. φ ist ein Homomorphismus der Gruppe $\mathbb{Z}$ in die Gruppe $\mathrm{GL}_n(\mathbb{R})$. Wir nennen dies eine *diskrete Symmetrie*.

7.1.2 Kontinuierliche Symmetrien

Von besonderem Interesse sind kontinuierliche Symmetrien. Dies sind Abbildungen eines Systems, die von einem Parameter abhängen. Zum Beispiel können wir Drehungen der Ebene $\mathbb{R}^2$ um den Winkel α durch Matrizen

$$D_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

beschrieben werden. Ein Kreis um den Nullpunkt bleibt unter jeder dieser Drehungen invariant. Im Gegensatz dazu sind alle $3n$-Ecke mit Schwerpunkt 0 nur invariant unter der einen Drehung $D_{\frac{2\pi}{3}}$ invariant. Die kleinste Menge, die einen vorgegebenen Punkt enthält und unter allen Drehungen D_α invariant ist, ist immer ein Kreis um den Nullpunkt.

Definition 7.1. *Ein Homomorphismus $\varphi\colon \mathbb{R} \to \mathrm{GL}_n(\mathbb{R})$ von der additiven Gruppe $\mathbb{R}$ in die allgemeine lineare Gruppe heisst eine* Einparameter-Untergruppe *von* $\mathrm{GL}_n(\mathbb{R})$.

Die Abbildung

$$\varphi\colon \mathbb{R} \to \mathrm{GL}_n(\mathbb{R}) : \alpha \mapsto D_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

ist also eine Einparameter-Untergruppe von $\mathrm{GL}_2(\mathbb{R})$.

Der harmonische Oszillator

Eine Masse m verbunden mit einer Feder mit der Federkonstanten K schwingt um die Ruhelage $x = 0$ entsprechend der Differentialgleichung

$$m\frac{d^2}{dt^2}x(t) = -Kx(t).$$

Die Kreisfrequenz der Schwingung ist

$$\omega = \sqrt{\frac{K}{m}}.$$

Das System kann als zweidimensionales System im Phasenraum mit den Koordinaten $x_1 = x$ und $x_2 = p = m\dot{x}$ beschrieben werden. Die zweidimensionale Differentialgleichung ist

$$\left.\begin{aligned} \dot{x}(t) &= \frac{1}{m}p(t) \\ \dot{p}(t) &= -Kx(t) \end{aligned}\right\} \quad \Rightarrow \quad \frac{d}{dt}\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ -K & 0 \end{pmatrix}\begin{pmatrix} x(t) \\ p(t) \end{pmatrix}. \tag{7.1}$$

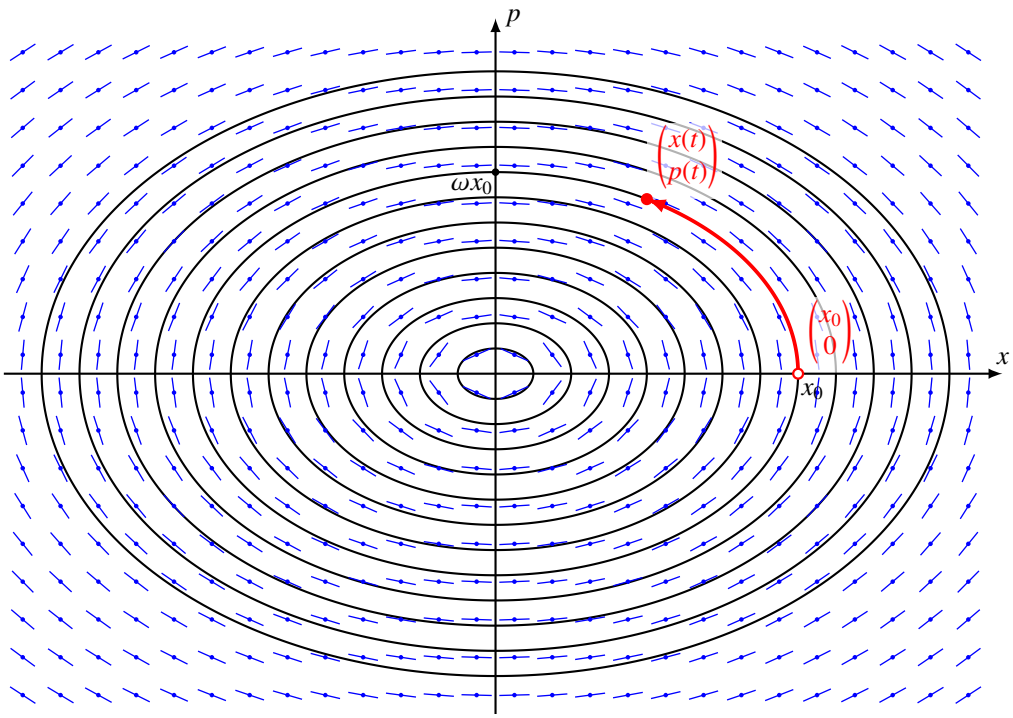


Abbildung 7.2: Die Lösungen der Differentialgleichung (7.1) im Phasenraum sind Ellipsen mit Halbachsenverhältnis ω^{-1}.

Die Lösung der Differentialgleichung für die Anfangsbedingung $x(0) = 1$ und $p(0) = 0$ ist

$$x(t) = \cos\ \omega t \qquad \Rightarrow \qquad p(t) = -\omega \sin \omega t,$$

die Lösung zur Anfangsbedingung $x(0) = 0$ und $p(0) = 1$ ist

$$x(t) = \frac{1}{\omega} \sin \omega t, \qquad p(t) = \cos \omega t.$$

In Matrixform kann man die allgemeine Lösung zur Anfangsbedingun $x(0) = x_0$ und $p(0) = p_0$

$$\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = \underbrace{\begin{pmatrix} \cos \omega t & \frac{1}{\omega} \sin \omega t \\ -\omega \sin \omega t & \cos \omega t \end{pmatrix}}_{= \Phi_t} \begin{pmatrix} x_0 \\ p_0 \end{pmatrix} \tag{7.2}$$

schreiben. Die Matrizen Φ_t bilden eine Einparameter-Untergruppe von $\mathrm{GL}_n(\mathbb{R})$, da

$$\begin{aligned} \Phi_s \Phi_t &= \begin{pmatrix} \cos \omega s & \frac{1}{\omega} \sin \omega s \\ -\omega \sin \omega s & \cos \omega s \end{pmatrix} \begin{pmatrix} \cos \omega t & \frac{1}{\omega} \sin \omega t \\ -\omega \sin \omega t & \cos \omega t \end{pmatrix} \\ &= \begin{pmatrix} \cos \omega s \cos \omega t - \sin \omega s \sin \omega t & \frac{1}{\omega}(\cos \omega s \sin \omega t + \sin \omega s \cos \omega t) \\ -\omega(\sin \omega s \cos \omega t + \cos \omega s \sin \omega t) & \cos \omega s \cos \omega t - \sin \omega s \sin \omega t \end{pmatrix} \\ &= \begin{pmatrix} \cos \omega(s+t) & \frac{1}{\omega} \sin \omega(s+t) \\ -\omega \sin \omega(s+t) & \cos \omega(s+t) \end{pmatrix} = \Phi_{s+t} \end{aligned}$$

gilt. Die Lösungen der Differentialgleichung (7.1) sind in Abbildung 7.2 Die Matrizen Φ_t beschreiben eine kontinuierliche Symmetrie des Differentialgleichungssystems, welches den harmonischen Oszillator beschreibt.

Fluss einer Differentialgleichung

Die Abbildungen Φ_t von (7.2) sind jeweils Matrizen in $\mathrm{GL}_n(\mathbb{R})$. Der Grund dafür ist, dass die Differentialgleichung (7.1) linear ist. Dies hat zur Folge, dass für zwei Anfangsbedingungen $x_1, x_2 \in \mathbb{R}^2$ die Lösung für Linearkombinationen $\lambda x_1 + \mu x_2$ durch Linearkombination der Lösungen erhalten werden kann, also aus der Formel

$$\Phi_t(\lambda x_1 + \mu x_2) = \lambda \Phi_t x_1 + \mu \Phi_t x_2.$$

Dies zeigt, dass Φ_t für jedes t eine lineare Abbildung sein muss.

Für eine beliebige Differentialgleichung kann man immer noch eine Abbildung Φ konstruieren, die aber nicht mehr linear ist. Sei dazu die Differentialgleichung erster Ordnung

$$\frac{dx}{dt} = f(t, x) \qquad \text{mit} \qquad f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n \tag{7.3}$$

gegeben. Für jeden Anfangswert $x_0 \in \mathbb{R}^n$ kann man mindestens für eine gewisse Zeit $t < \varepsilon$ eine Lösung $x(t, x_0)$ finden mit $x(t, x_0) = x_0$. Aus der Theorie der gewöhnlichen Differentialgleichungen ist auch bekannt, dass $x(t, x_0)$ mindestens in der Nähe von x_0 differenzierbar von x_0 abhängt. Dies erlaubt eine Abbildung

$$\Phi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n : (t, x_0) \mapsto \Phi_t(x_0) = x(t, x_0)$$

zu definieren, die sowohl von t als auch von x_0 differenzierbar abhängt. Aus der Definition folgt unmittelbar, dass $\Phi_0(x_0) = x_0$ ist, dass also Φ_0 die identische Abbildung von $\mathbb{R}^n$ ist.

Aus der Definition lässt sich auch ableiten, dass $\Phi_{s+t} = \Phi_s \circ \Phi_t$ gilt. $\Phi_t(x_0) = x(t, x_0)$ ist der Endpunkt der Bahn, die bei x_0 beginnt und sich während der Zeit t entwickelt. $\Phi_s(x(t, x_0))$ ist dann der Endpunkt der Bahn, die bei $x(t, x_0)$ beginnt und sich während der Zeit s entwickelt. Somit ist $\Phi_s \circ \Phi_t(x_0)$ der Endpunkt der Bahn, die bei x_0 beginnt und sich über die Zeit $s + t$ entwickelt. In Formeln bedeutet dies

$$\Phi_{s+t} = \Phi_s \circ \Phi_t.$$

Die Abbildung $t \mapsto \Phi_t$ ist also wieder ein Homomorphismus von der additiven Gruppe $\mathbb{R}$ in eine Gruppe von differenzierbaren Abbildungen $\mathbb{R}^n \to \mathbb{R}^n$.

Definition 7.2. *Die Abbildung*

$$\Phi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n : (t, x_0) \mapsto \Phi_t(x_0) = x(t, x_0)$$

heisst der Fluss *der Differentialgleichung* (7.3)*, wenn für jedes $x_0 \in \mathbb{R}^n$ die Kurve $t \mapsto \Phi_t(x_0)$ eine Lösung der Differentialgleichung ist mit Anfangsbedingung x_0.*

Die Abbildung Φ_t von (7.2) ist also der Fluss der Differentialgleichung des harmonischen Oszillators.

7.1.3 Mannigfaltigkeiten

Eine Differentialgleichung der Form (7.3) stellt einen Zusammenhang her zwischen einem Punkt x und der Tangentialrichtung einer Bahnkurve $f(t, x)$. Die Ableitung liefert die lineare Näherung der Bahkurve

$$x(t_0 + h) = x(t_0) + hf(t_0, x_0) + o(h)$$

für h in einer kleinen Umgebung von 0. Das funktioniert auch, weil $f(t_0, x_0)$ selbst ein Vektor von $\mathbb{R}^n$ ist, in dem die Bahnkurve verläuft.

Diese Idee funktioniert nicht mehr zum Beispiel für eine Differentialgleichung auf einer Kugeloberfläche, weil alle Punkte $x(t_0) + hf(t_0, x_0)$ für alle $h \neq 0$ nicht mehr auf der Kugeloberfläche liegen. Physikalisch äussert sich das ein einer zusätzlichen Kraft, die nötig ist, die Bahn auf der Kugeloberfläche zu halten. Diese Kraft stellt zum Beispiel sicher, dass die Vektoren $f(t, x)$ für Punkte x auf der Kugeloberfläche immer tangential an die Kugel sind. Trotzdem ist der Tangentialvektor oder der Geschwindigkeitsvektor nicht mehr ein Objekt, welches als Teil der Kugeloberfläche definiert werden kann, er kann nur definiert werden, wenn man sich die Kugel als in einen höherdimensionalen Raum eingebettet vorstellen kann.

Um die Idee der Differentialgleichung auf einer beliebigen Fläche konsistent zu machen ist daher notwendig, die Idee einer Tagentialrichtung auf eine Art zu definieren, die nicht von der Einbettung der Fläche in den n-dimensionalen Raum abhängig ist. Das in diesem Abschnitt entwickelte Konzept der *Mannigfaltigkeit* löst dieses Problem.

Karten

Die Navigation auf der Erdoberfläche verwendet das Koordinatensystem der geographischen Länge und Breite. Dieses Koordinatensystem funktioniert gut, solange man sich nicht an den geographischen Polen befindet, denn deren Koordinaten sind nicht mehr eindeutig. Alle Punkte mit geographischer Breite 90° und beliebiger geographischer Länge beschreiben den Nordpol. Auch die Ableitung funktioniert dort nicht mehr. Bewegt man sich mit konstanter Geschwindigkeit über den Nordpol, springt die Ableitung der geographischen Breite von einem positiven Wert auf einen negativen Wert, sie kann also nicht differenzierbar sein. Diese Einschränkungen sind in der Praxis nur ein geringes Problem dar, da die meisten Reisen nicht über die Pole erfolgen.

Der Polarforscher, der in unmittelbarer Umgebung des Poles arbeitet, kann das Problem lösen, indem er eine lokale Karte für das Gebiet um den Pol erstellt. Dafür kann er beliebige Koordinaten verwenden, zum Beispiel auch ein kartesisches Koordinatensystem, er muss nur eine Methode haben, wie er seine Koordinaten wieder auf geographische Länge und Breite umrechnen will. Und wenn er über Geschwindigkeiten kommunizieren will, dann muss er auch Ableitungen von Kurven in seinem kartesischen Koordinatensystem umrechnen können auf die Kugelkoordinaten. Dazu muss seine Umrechnungsformel von kartesischen Koordinaten auf Kugelkoordinaten differenzierbar sein.

Diese Idee wird durch das Konzept der Mannigfaltigkeit verallgemeinert. Eine n-dimensionale *Mannigfaltigkeit* ist eine Menge M von Punkten, die lokal, also in der Umgebung eines Punktes, mit möglicherweise mehreren verschiedenen Koordinatensystemen versehen werden kann. Ein Koordinatensystem ist eine umkehrbare Abbildung einer offenen Teilmenge $U \subset M$ in den Raum $\mathbb{R}^n$. Die Komponenten dieser Abbildung heissen die *Koordinaten*.

Definition 7.3. *Eine Karte auf M ist eine umkehrbare Abbildung $\varphi\colon U \to \mathbb{R}^n$ (siehe auch Abbildung 7.3). Ein differenzierbarer Atlas ist eine Familie von Karten φ_α derart, dass die Definitionsgebiete U_α die ganze Menge M überdecken, und dass die Kartenwechsel Abbildungen*

$$\varphi_{\beta\alpha} = \varphi_\beta \circ \varphi_\alpha^{-1}\colon \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$$

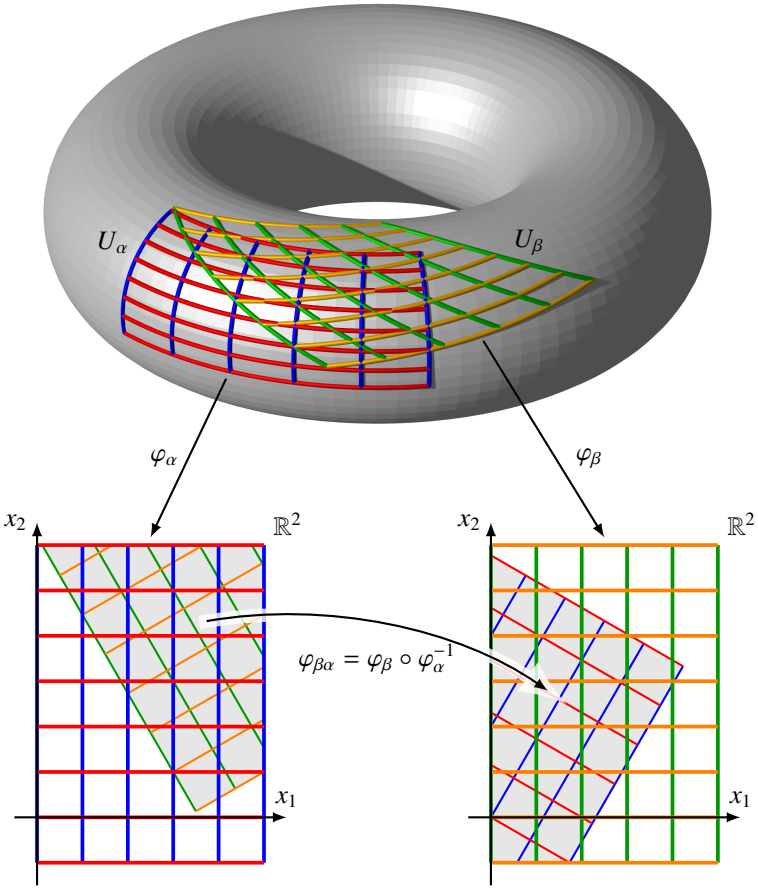


Abbildung 7.3: Karten $\varphi_\alpha \colon U_\alpha \to \mathbb{R}^2$ und $\varphi_\beta \colon U_\beta \to \mathbb{R}^2$ auf einem Torus. Auf dem Überschneidungsgebiet $\varphi_\alpha^{-1}(U_\alpha \cap U_\beta)$ ist der Kartenwechsel $\varphi_\beta \circ \varphi_\alpha^{-1}$ wohldefiniert und muss differnzierbar sein, wenn eine differenzierbare Mannigfaltigkeit entstehen soll.

als Abbildung von offenen Teilmengen von $\mathbb{R}^n$ differenzierbar ist. Eine n-dimensionale differenzierbare Mannigfaltigkeit ist eine Menge M mit einem differenzierbaren Atlas.

Karten und Atlanten regeln also nur, wie sich verschiedene lokale Koordinatensysteme ineinander umrechnen lassen.

Beispiel. $M = \mathbb{R}^n$ ist eine differenzierbare Mannigfaltigkeit denn die identische Abbildung $M \to \mathbb{R}^n$ ist eine Karte und ein Atlas von M. ◯

Beispiel. Die Kreislinie in in der Ebene ist eine 1-dimensionale Mannigfaltigkeit. Natürlich kann sie nicht mit einer einzigen Karte beschrieben werden, da es keine umkehrbaren Abbildungen zwischen $\mathbb{R}$ und der Kreislinie gibt. Die Projektionen auf die einzelnen Koordinaten liefern die folgenden vier Karten:

$$\varphi_1 \colon U_{x>0}\{(x,y) \mid x^2 + y^2 = 1 \wedge x > 0\} \to \mathbb{R} : (x,y) \mapsto y$$

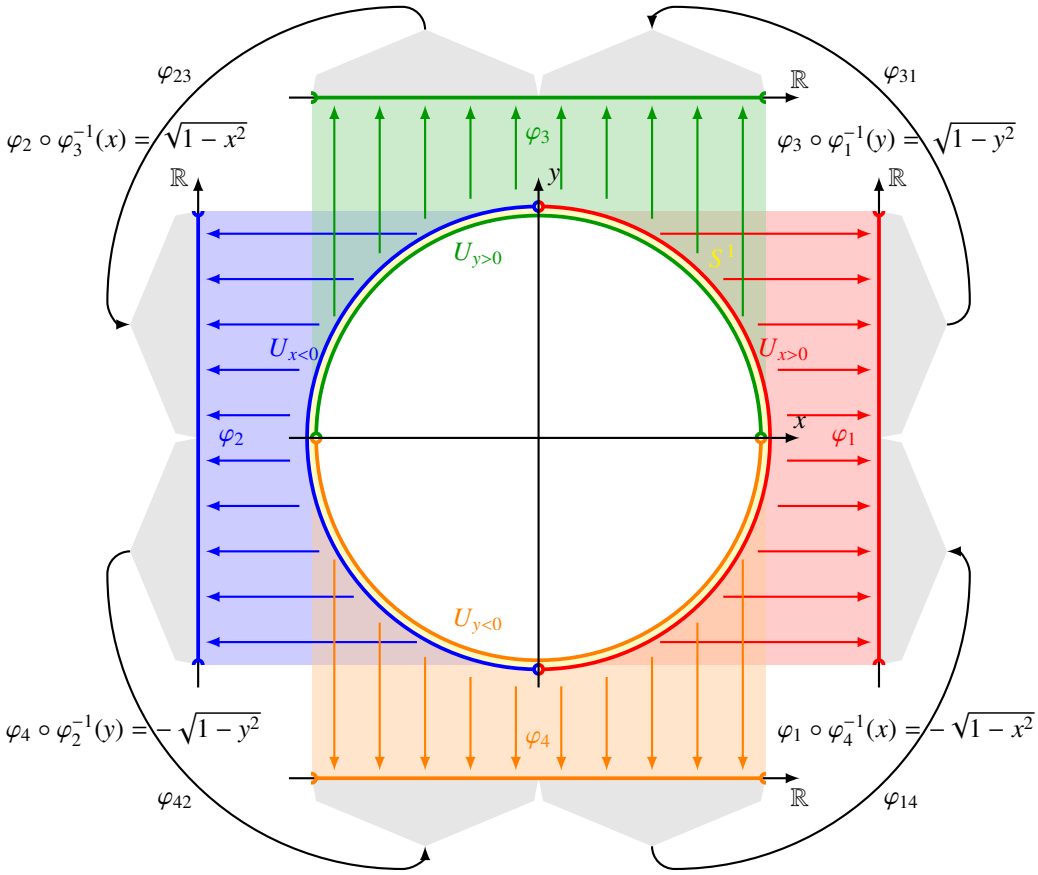


Abbildung 7.4: Karten für die Kreislinie $S^1 \subset \mathbb{R}^2$.

$$\varphi_2 \colon U_{x<0}\{(x,y) \mid x^2 + y^2 = 1 \wedge x < 0\} \to \mathbb{R} : (x,y) \mapsto y$$
$$\varphi_3 \colon U_{y>0}\{(x,y) \mid x^2 + y^2 = 1 \wedge y > 0\} \to \mathbb{R} : (x,y) \mapsto x$$
$$\varphi_4 \colon U_{y<0}\{(x,y) \mid x^2 + y^2 = 1 \wedge y < 0\} \to \mathbb{R} : (x,y) \mapsto x$$

Die Werte der Kartenabbildungen sind genau die x- und y-Koordinaten auf der in den Raum $\mathbb{R}^2$ eingebetteten Kreislinie.

Für φ_1 und φ_2 sind die Definitionsgebiete disjunkt, hier gibt es also keine Notwendigkeit, Koordinatenumrechnungen vornehmen zu können. Dasselbe gilt für φ_3 und φ_4.

Die nichtleeren Schnittmengen der verschiedenen Kartengebiete beschreiben jeweils die Punkte der Kreislinie in einem Quadranten. Die Umrechnung zwischen den Koordinaten und ihre Ableitung ist je nach Quadrant durch

1. Quadrant	$\varphi_{31} = \varphi_3 \circ \varphi_1^{-1} : y \mapsto$	$\sqrt{1-y^2}$	$D\varphi_{31} = -\dfrac{y}{\sqrt{1-y^2}}$
2. Quadrant	$\varphi_{24} = \varphi_3 \circ \varphi_1^{-1} : x \mapsto$	$\sqrt{1-x^2}$	$D\varphi_{24} = -\dfrac{x}{\sqrt{1-x^2}}$

3. Quadrant	$\varphi_{42} = \varphi_3 \circ \varphi_1^{-1} : y \mapsto -\sqrt{1-y^2}$	$D\varphi_{42} = \dfrac{y}{\sqrt{1-y^2}}$
4. Quadrant	$\varphi_{14} = \varphi_3 \circ \varphi_1^{-1} : x \mapsto -\sqrt{1-x^2}$	$D\varphi_{14} = \dfrac{x}{\sqrt{1-x^2}}$

gegeben. Diese Abbildungen sind im offenen Intervall $(-1, 1)$ differenzierbar, Schwierigkeiten mit der Ableitungen ergeben sich nur an den Stellen $x = \pm 1$ und $y = \pm 1$, die in einem Überschneidungsgebiet von Karten nicht vorkommen können. Somit bilden die vier Karten einen differenzierbaren Atlas für die Kreislinie (Abbildung 7.4). ○

Beispiel. Ganz analog zum vorangegangenen Beispiel über die Kreisline lässt sich für eine n-dimensionale Sphäre

$$S^n = \{(x_1, \ldots, x_{n+1}) \mid x_0^2 + \cdots + x_n^2 = 1\}$$

immer ein Atlas aus 2^{n+1} Karten mit den Koordinatenabbildungen

$$\varphi_{i,\pm} : U_{i,\pm} = \{p \in S^n \mid \pm x_i > 0\} \to \mathbb{R}^n : p \mapsto (x_1, \ldots, \hat{x}_i, \ldots, x_{n+1})$$

konstruieren, der S^n zu einer n-dimensionalen Mannigfaltigkeit macht. ○

Tangentialraum

Mit Hilfe einer Karte $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$ kann das Geschehen in einer Mannigfaltigkeit in den vertrauten n-dimensionalen Raum $\mathbb{B}^n$ transportiert werden. Eine Kurve $\gamma : \mathbb{R} \to M$, die so parametrisiert sein soll, dass $\gamma(t) \in U_\alpha$ für t in einer Umgebung I von 0 ist, wird von der Karte in eine Kurve $\gamma_\alpha = \varphi_\alpha \circ \gamma : I \to \mathbb{R}^n$ abgebildet, deren Tangentialvektor wieder ein Vektor in $\mathbb{R}^n$ ist.

Eine zweite Karte φ_β führt auf eine andere Kurve mit der Parametrisierung $\gamma_\beta = \varphi_\beta \circ \gamma : I \to \mathbb{R}^n$ und einem anderen Tangentialvektor. Die beiden Tangentialvektoren können aber mit der Ableitung der Koordinatenwechsel-Abbildung $\varphi_{\beta\alpha} = \varphi_\beta \circ \varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \mathbb{R}^n$ ineinander umgerechnet werden. Aus

$$\gamma_\beta = \varphi_\beta \circ \gamma = (\varphi_\beta \circ \varphi_\alpha^{-1}) \circ \varphi_\alpha \circ \gamma = \varphi_{\beta\alpha} \circ \varphi_\alpha \circ \gamma = \varphi_{\beta\alpha} \circ \gamma_\alpha$$

folgt durch Ableitung nach dem Kurvenparameter t, dass

$$\frac{d}{dt}\gamma_\beta(t) = D\varphi_{\beta\alpha} \cdot \frac{d}{dt}\gamma_\alpha(t).$$

Die Ableitung $D\varphi_{\beta\alpha}$ von $\varphi_{\beta\alpha}$ an der Stelle $\gamma_\alpha(t)$ berechnet also aus dem Tangentialvektor einer Kurve in der Karte φ_α den Tangentialvektor der Kurve in der Karte φ_β.

Die Forderung nach Differenzierbarkeit der Kartenwechselabbildungen $\varphi_{\beta\alpha}$ stellt also nur sicher, dass die Beschreibung eines Systemes mit Differentialgleichungen in verschiedenen Koordinatensystemen auf die gleichen Lösungskurven in der Mannigfaltigkeit führt. Insbesondere ist die Verwendung von Karten ist also nur ein Werkzeug, mit dem die Unmöglichkeit einer globalen Beschreibung einer Mannigfaltigkeit M mit einem einzigen globalen Koordinatensystem ohne Singularitäten umgangen werden kann.

Beispiel. Das Beispiel des Kreises in Abbildung 7.4 zeigt, dass die Tangentialvektoren je nach Karte sehr verschieden aussehen können. Der Tangentialvektor der Kurve $\gamma(t) = (x(t), y(t))$ im Punkt $\gamma(t)$ ist $\dot{y}(t)$ in den Karten φ_1 und φ_2 und $\dot{x}(t)$ in den Karten φ_3 und φ_4.

Die spezielle Kurve $\gamma(t) = (\cos t, \sin t)$ hat in einem Punkt $t \in (0, \frac{\pi}{2})$. in der Karte φ_1 den Tangentialvektor $\dot{y}(t) = \cos t$, in der Karte φ_3 aber den Tangentialvektor $\dot{x} = -\sin t$. Die Ableitung des Kartenwechsels in diesem Punkt ist die 1×1-Matrix

$$D\varphi_{31}(\gamma(t)) = -\frac{y(t)}{\sqrt{1-y(t)^2}} = -\frac{\sin t}{\sqrt{1-\sin^2 t}} = -\frac{\sin t}{\cos t} = -\tan t.$$

Die Koordinatenumrechnung ist gegeben durch

$$\dot{x}(t) = D\varphi_{31}(\gamma(t))\dot{y}(t)$$

wird für die spezielle Kurve $\gamma(t) = (\cos t, \sin t)$ wird dies zu

$$D\varphi_{31}(\gamma(t)) \cdot \dot{y}(t) = -\tan t \cdot \cos t = -\frac{\sin t}{\cos t} \cdot \cos t = -\sin t = \dot{x}(t). \qquad \bigcirc$$

Betrachtet man die Kreislinie als Kurve in $\mathbb{R}^2$, dann ist der Tangentialvektor durch $\dot{\gamma}(t) = (\dot{x}(t), \dot{y}(t))$ gegeben. Da die Karten Projektionen auf die x- bzw. y-Achsen sind, entsteht der Tangentialvektor in der Karte durch Projektion von $(\dot{x}(t), \dot{y}(t))$ auf die entsprechende Komponente.

Die Tangentialvektoren in zwei verschiedenen Punkten der Kurve können im Allgemeinen nicht miteinander verglichen werden. Darüber hinweg hilft auch die Tatsache nicht, dass die Kreislinie in den Vektorraum $\mathbb{R}^2$ eingebettet sind, wo sich Vektoren durch Translation miteinander vergleichen lassen. Ein nichtverschwindender Tangentialvektor im Punkt $(1, 0)$ hat, betrachtet als Vektor in $\mathbb{R}^2$ verschwindende x-Komponente, für Tangentialvektoren im Inneren eines Quadranten ist dies nicht der Fall.

Eine Möglichkeit, einen Tangentialvektor in $(1, 0)$ mit einem Tangentialvektor im Punkt $(\cos t, \sin t)$ zu vergleichen, besteht darin, den Vektor um den Winkel t zu drehen. Dies ist möglich, weil die Kreislinie eine kontinuierliche Symmetrie, nämlich die Drehung um den Winkel t hat, die es erlaubt, den Punkt $(1, 0)$ in den Punkt $(\cos t, \sin t)$ abzubilden. Erst diese Symmetrie ermöglicht den Vergleich. Dieser Ansatz ist für alle Matrizen erfolgreich, wie wir später sehen werden.

Ein weiterer Ansatz, Tangentialvektoren zu vergleichen, ist die Idee, einen sogenannten Zusammenhang zu definieren, eine Vorschrift, wie Tangentialvektoren infinitesimal entlang von Kurven in der Mannigfaltigkeit transportiert werden können. Auf einer sogenannten *Riemannschen Mannigfaltigkeit* ist zusätzlich zur Mannigfaltigkeitsstruktur die Längenmessung definiert. Sie kann dazu verwendet werden, den Transport von Vektoren entlang einer Kurve so zu definieren, dass dabei Längen und Winkel erhalten bleiben. Dieser Ansatz ist die Basis der Theorie der Krümmung sogenannter Riemannscher Mannigfaltigkeiten.

7.1.4 Der Satz von Noether

7.2 Lie-Gruppen

Die in bisherigen Beispielen untersuchten Matrizengruppen zeichnen sich durch zusätzliche Eigenschaften aus. Die Gruppe

$$\mathrm{GL}_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}) \mid \det A \neq 0\}$$

besteht aus den Matrizen, deren Determinante nicht 0 ist. Da die Menge der Matrizen mit $\det A = 0$ eine abgeschlossene Menge in $M_n(\mathbb{R}) \simeq \mathbb{R}^{n^2}$ ist, ist $\mathrm{GL}_n(\mathbb{R})$ eine offene Teilmenge in $\mathbb{R}^{n^2}$, sie besitzt also automatisch die Struktur einer n^2-Mannigfaltigkeit. Dies gilt jedoch auch für alle anderen Matrizengruppen, die in diesem Abschnitt genauer untersucht werden sollen.

7.2.1 Mannigfaltigkeitsstruktur der Matrizengruppen

Eine Matrizengruppe wird automatsich zu einer Mannigfaltigkeit, wenn es gelingt, eine Karte für eine Umgebung des neutralen Elements zu finden. Dazu muss gezeigt werden, dass sich aus einer solchen Karte für jedes andere Gruppenelement eine Karte für eine Umgebung ableiten lässt. Sei also $\varphi_e\colon U_e \to \mathbb{R}^N$ eine Karte für die Umgebung $U_e \subset G$ von $e \in G$. Für $g \in G$ ist dann die Abbildung

$$\varphi_g\colon U_g = gU_e \to \mathbb{R} : h \mapsto \varphi_e(g^{-1}h)$$

eine Karte für die Umgebung U_g des Gruppenelementes g. schreibt man l_g für die Abbildung $h \mapsto gh$, dann kann man die Kartenabbildung auch $\varphi_g = \varphi_e \circ l_{g^{-1}}$ schreiben.

Kartenwechsel

Die Kartenwechsel-Abbildungen für zwei Karten φ_{g_1} und φ_{g_2} ist die Abbildung

$$\varphi_{g_1,g_2} = \varphi_{g_1} \circ \varphi_{g_2}^{-1} = \varphi_e \circ l_{g_1^{-1}} \circ (\varphi_e \circ l_{g_2^{-1}})^{-1} = \varphi_e \circ l_{g_1^{-1}} \circ l_{g_2^{-1}}^{-1}\varphi_e^{-1} = \varphi_e \circ l_{g_1^{-1}} \circ l_{g_2}\varphi_e^{-1} = \varphi_e \circ l_{g_1^{-1}g_2}\varphi_e^{-1}$$

mit der Ableitung

$$D\varphi_e \circ Dl_{g_1^{-1}g_2} D\varphi_e^{-1} = D\varphi_e \circ Dl_{g_1^{-1}g_2}(D\varphi_e)^{-1}.$$

Die Abbildung $l_{g_1^{-1}g_2}$ ist aber nur die Multiplikation mit einer Matrix, also eine lineare Abbildung, so dass der Kartenwechsel nichts anderes ist als die Darstellung der Matrix der Linksmultiplikation $l_{g_1^{-1}g_2}$ im Koordinatensystem der Karte U_e ist. Differenzierbarkeit der Kartenwechsel ist damit sichergestellt, die Matrizengruppen sind automatisch differenzierbare Mannigfaltigkeiten.

Die Konstruktion aller Karten aus einer einzigen Karte für eine Umgebung des neutralen Elements zeigt auch, dass es für die Matrizengruppen reicht, wenn man die Elemente in einer Umgebung des neutralen Elementes parametrisieren kann. Dies ist jedoch nicht nur für die Matrizengruppen möglich. Wenn eine Gruppe gleichzeitig eine differenzierbare Mannigfaltigkeit ist, dann können Karten über die ganze Gruppe transportiert werden, wenn die Multiplikation mit Gruppenelementen eine differenzierbare Abbildung ist. Solche Gruppen heissen auch Lie-Gruppen gemäss der folgenden Definition.

Definition 7.4. *Eine* Lie-Gruppe *ist eine Gruppe, die gleichzeitig eine differenzierbare Mannigfaltigkeit ist derart, dass die Abbildungen*

$$G \times G \to G : (g_1, g_2) \mapsto g_1 g_2$$
$$G \to G : g \mapsto g^{-1}$$

differenzierbare Abbildungen zwischen Mannigfaltigkeiten sind.

Die Abstraktheit dieser Definition täuscht etwas über die Tatsache hinweg, dass sich mit Hilfe der Darstellungstheorie jede beliebige Lie-Gruppe als Untermannigfaltigkeit einer Matrizengruppe verstehen lässt. Das Studium der Matrizengruppen erlaubt uns daher ohne grosse Einschränkungen ein Verständnis für die Theorie der Lie-Gruppen zu entwickeln.

Tangentialvektoren und die Exponentialabbildung

Die Matrizengruppen sind alle in der n^2-dimensionalen Mannigfaltigkeit $\mathrm{GL}_n(\mathbb{R})$ enthalten. Diffferenzierbare Kurven $\gamma(t)$ in $\mathrm{GL}_n(\mathbb{R})$ haben daher in jedem Punkt Tangentialvektoren, die als Matrizen

in $M_n(\mathbb{R})$ betrachtet werden können. Wenn $\gamma(t)$ die Matrixelemente $\gamma_{ij}(t)$ hat, dann ist der Tangentialvektor im Punkt $\gamma(t)$ durch

$$\frac{d}{dt}\gamma(t) = \begin{pmatrix} \dot\gamma_{11}(t) & \dots & \dot\gamma_{1n}(t) \\ \vdots & \ddots & \vdots \\ \dot\gamma_{n1}(t) & \dots & \dot\gamma_{nn}(t) \end{pmatrix}$$

gegeben.

Im Allgemeinen kann man Tangentialvektoren in verschiedenen Punkten einer Mannigfaltigkeit nicht miteinander vergleichen. Die Multiplikation l_g, die den Punkt e in den Punkt g verschiebt, transportiert auch die Tangentialvektoren im Punkt e in Tangentialvektoren im Punkt g.

Aufgabe 7.5. *Gibt es eine Kurve $\gamma(t) \in \mathbb{GL}_n(\mathbb{R})$ mit $\gamma(0) = e$ derart, dass der Tangentialvektor im Punkt $\gamma(t)$ für $t > 0$ derselbe ist wie der Tangentialvektor im Punkt e, transportiert durch Matrixmultiplikation mit $\gamma(t)$?*

Eine solche Kurve muss die Differentialgleichung

$$\frac{d}{dt}\gamma(t) = \gamma(t) \cdot A \tag{7.4}$$

erfüllen, wobei $A \in M_n(\mathbb{R})$ der gegebene Tangentialvektor in $e = I$ ist.

Die Matrixexponentialfunktion

$$e^{At} = 1 + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \frac{A^4t^4}{4!} + \dots$$

liefert eine Einparametergruppe $\mathbb{R} \to \mathrm{GL}_n(\mathbb{R})$ mit der Ableitung

$$\frac{d}{dt}e^{At} = \lim_{h\to 0} \frac{e^{A(t+h)} - e^{At}}{h} = \lim_{h\to 0} e^{At}\frac{e^{Ah} - I}{h} = e^{At}A.$$

Sie ist also Lösung der Differentialgleichung (7.4).

7.2.2 Drehungen in der Ebene

Die Drehungen der Ebene sind die orientierungserhaltenden Symmetrien des Einheitskreises, der in Abbildung 7.4 als Mannigfaltigkeit erkannt wurde. Sie bilden eine Lie-Gruppe, die auf verschiedene Arten als Matrix beschrieben werden kann.

Die Untergruppe $\mathrm{SO}(2) \subset \mathrm{GL}_2(\mathbb{R})$

Drehungen der Ebene können in einer orthonormierten Basis durch Matrizen der Form

$$D_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

dargestellt werden. Wir bezeichnen die Menge der Drehmatrizen in der Ebene mit $\mathrm{SO}(2) \subset \mathrm{GL}_2(\mathbb{R})$. Die Abbildung

$$D_\bullet \colon \mathbb{R} \to \mathrm{SO}(2) : \alpha \mapsto D_\alpha$$

hat die Eigenschaften

$$D_{\alpha+\beta} = D_\alpha D_\beta$$

$$\begin{aligned} D_0 &= I \\ D_{2k\pi} &= I \qquad \forall k \in \mathbb{Z}. \end{aligned}$$

Daraus folgt zum Beispiel, dass $D_\bullet$ eine 2π-periodische Funktion ist. $D_\bullet$ bildet die Menge der Winkel $[0, 2\pi)$ bijektiv auf die Menge der Drehmatrizen in der Ebene ab.

Für jedes Intervall $(a, b) \subset \mathbb{R}$ mit Länge $b - a < 2\pi$ ist die Abbildung $\alpha \mapsto D_\alpha$ umkehrbar, die Umkehrung kann als Karte verwendet werden. Zwei verschiedene Karten $\alpha_1 \colon U_1 \to \mathbb{R}$ und $\alpha_2 \colon U_2 \to \mathbb{R}$ bilden die Elemente $g \in U_1 \cap U_2$ in Winkel $\alpha_1(g)$ und $\alpha_2(g)$ ab, für die $D_{\alpha_1(g)} = D_{\alpha_2(g)}$ gilt. Dies ist gleichbedeutend damit, dass $\alpha_1(g) = \alpha_2(g) + 2\pi k$ mit $k \in \mathbb{Z}$. In einem Intervall in $U_1 \cap U_2$ muss k konstant sein. Die Kartenwechselabblidung ist also nur die Addition eines Vielfachen von 2π, mit der identischen Abbildung als Ableitung. Diese Karten führen also auf besonders einfache Kartenwechselabbildungen.

Die Untergruppe $S^1 \subset \mathbb{C}$

Ein alternatives Bild für die Drehungen der Ebene kann man in der komplexen Ebene $\mathbb{C}$ erhalten. Die Multiplikation mit der komplexen Zahl $e^{i\alpha}$ beschreibt eine Drehung der komplexen Ebene um den Winkel α. Die Zahlen der Form $e^{i\alpha}$ haben den Betrag 1 und die Abbildung

$$f \colon \mathbb{R} \to \mathbb{C} : \alpha \mapsto e^{i\alpha}$$

hat die Eigenschaften

$$\begin{aligned} f(\alpha + \beta) &= f(\alpha) f(\beta) \\ f(0) &= 1 \\ f(2\pi k) &= 1 \qquad \forall k \in \mathbb{Z}, \end{aligned}$$

die zu den Eigenschaften der Abbildung $\alpha \mapsto D_\alpha$ analog sind.

Jede komplexe Zahl z vom Betrag 1 kann geschrieben werden in der Form $z = e^{i\alpha}$, die Abbildung f ist also eine Parametrisierung des Einheitskreises in der Ebene. Wir bezeichen $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ die komplexen Zahlen vom Betrag 1. S^1 ist eine Gruppe bezüglich der Multiplikation, da für jede Zahl $z, w \in S^1$ gilt $|z^{-1}| = 1$ und $|zw| = 1$ und damit $z^{-1} \in S^1$ und $zw \in S^1$.

Zu einer komplexen Zahl $z \in S^1$ gibt es einen bis auf Vielfache von 2π eindeutigen Winkel $\alpha(z)$ derart, dass $e^{i\alpha(z)} = z$. Damit kann man jetzt die Abbildung

$$\varphi \colon S^1 \to \mathrm{SO}(2) : z \mapsto D_{\alpha(z)}$$

konstruieren. Da D_α 2π-periodisch ist, geben um Vielfache von 2π verschiedene Wahlen von $\alpha(z)$ die gleiche Matrix $D_{\alpha(z)}$, die Abbildung φ ist daher wohldefiniert. φ erfüllt ausserdem die Bedingungen

$$\begin{aligned} \varphi(z_1 z_2) = D_{\alpha(z_1 z_2)} = D_{\alpha(z_1)+\alpha(z_2)} = D_{\alpha(z_1)} D_{\alpha(z_2)} = \varphi(z_1)\varphi(z_2) \\ \varphi(1) = D_{\alpha(1)} = D_0 = I \end{aligned}$$

Die Abbildung φ ist ein Homomorphismus der Gruppe S^1 in die Gruppe SO(2). Die Menge der Drehmatrizen in der Ebene kann also mit dem Einheitskreis in der komplexen Ebene identifiziert werden.

Tangentialvektoren von SO(2)

Da die Gruppe SO(2) eine eindimensionale Gruppe ist, kann jede Kurve $\gamma(t)$ durch den Drehwinkel $\alpha(t)$ mit $\gamma(t) = D_{\alpha(t)}$ beschrieben werden. Die Ableitung in $M_2(\mathbb{R})$ ist

$$\begin{aligned}\frac{d}{dt}\gamma(t) &= \frac{d}{d\alpha}\begin{pmatrix}\cos\alpha(t) & -\sin\alpha(t)\\ \sin\alpha(t) & \cos\alpha(t)\end{pmatrix}\cdot\frac{d\alpha}{dt}\\ &= \begin{pmatrix}-\sin\alpha(t) & -\cos\alpha(t)\\ \cos\alpha(t) & -\sin\alpha(t)\end{pmatrix}\cdot\dot\alpha(t)\\ &= \begin{pmatrix}\cos\alpha(t) & -\sin\alpha(t)\\ \sin\alpha(t) & \cos\alpha(t)\end{pmatrix}\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\cdot\dot\alpha(t) = D_{\alpha(t)}J\cdot\dot\alpha(t).\end{aligned}$$

Alle Tangentialvektoren von SO(2) im Punkt D_α entstehen aus J durch Drehung mit der Matrix D_α und Skalierung mit $\dot\alpha(t)$.

7.2.3 Isometrien von $\mathbb{R}^n$

Skalarprodukt

Lineare Abbildungen des Raumes $\mathbb{R}^n$ können durch $n\times n$-Matrizen beschrieben werden. Die Matrizen, die das Standardskalarprodukt $\mathbb{R}^n$ erhalten, bilden eine Gruppe, die in diesem Abschnitt genauer untersucht werden soll. Eine Matrix $A\in M_n(\mathbb{R})$ ändert das Skalarprodukt, wenn für jedes beliebige Paar x, y von Vektoren gilt $\langle Ax, Ay\rangle = \langle x, y\rangle$. Das Standardskalarprodukt kann mit dem Matrixprodukt ausgedrückt werden:

$$\langle Ax, Ay\rangle = (Ax)^t Ay = x^t A^t Ay = x^t y = \langle x, y\rangle$$

für jedes Paar von Vektoren $x, y\in\mathbb{R}$.

Mit dem Skalarprodukt kann man auch die Matrixelemente einer Matrix einer Abbildung f in der Standardbasis bestimmen. Das Skalarprodukt $\langle e_i, v\rangle$ ist die Länge der Projektion des Vektors v auf die Richtung e_i. Die Komponenten von Ae_j sind daher $a_{ij} = \langle e_i, f(e_j)\rangle$. Die Matrix A der Abbildung f hat also die Matrixelemente $a_{ij} = e_i^t Ae_j$.

Die orthogonale Gruppe O(n)

Die Matrixelemente von A^tA sind $\langle A^tAe_i, e_j\rangle = \langle e_i, e_j\rangle = \delta_{ij}$ sind diejenigen der Einheitsmatrix, die Matrix A erfüllt $AA^t = I$ oder $A^{-1} = A^t$. Dies sind die *orthogonalen* Matrizen. Die Menge O(n) der isometrischen Abbildungen besteht daher aus den Matrizen

$$\mathrm{O}(n) = \{A\in M_n(\mathbb{R}) \mid AA^t = I\}.$$

Die Matrixgleichung $AA^t = I$ liefert $n(n+1)/2$ unabhängige Bedingungen, die die orthogonalen Matrizen innerhalb der n^2-dimensionalen Menge $M_n(\mathbb{R})$ auszeichnen. Die Menge O(n) der orthogonalen Matrizen hat daher die Dimension

$$n^2 - \frac{n(n+1)}{2} = \frac{2n^2 - n^2 - n}{2} = \frac{n(n-1)}{2}.$$

Im Spezialfall $n = 2$ ist die Gruppe $O(2)$ eindimensional.

Tangentialvektoren

Die orthogonalen Matrizen bilden eine abgeschlossene Untermannigfaltigkeit von $\mathrm{GL}_n(\mathbb{R})$, nicht jede Matrix $M_n(\mathbb{R})$ kann also ein Tangentialvektor von $O(n)$ sein. Um herauszufinden, welche Matrizen als Tangentialvektoren in Frage kommen, betrachten wir eine Kurve $\gamma\colon \mathbb{R} \to O(n)$ von orthogonalen Matrizen mit $\gamma(0) = I$. Orthogonal bedeutet

$$\begin{aligned} & 0 = \frac{d}{dt} I = \frac{d}{dt}(\gamma(t)^t \gamma(t)) = \dot\gamma(t)^t \gamma(t)) + \gamma(t)^t \dot\gamma(t)) \\ \Rightarrow \quad & 0 = \dot\gamma(0)^t \cdot I + I \cdot \dot\gamma(0) = \dot\gamma(0)^t + \dot\gamma(0) = A^t + A = 0 \\ \Rightarrow \quad & A^t = -A \end{aligned}$$

Die Tangentialvektoren von $\mathrm{O}(n)$ sind also genau die antisymmetrischen Matrizen.

Für $n = 2$ sind alle antisymmetrischen Matrizen Vielfache der Matrix J, wie in Abschnitt 7.2.2 gezeigt wurde.

Für jedes Paar $i < j$ ist die Matrix A_{ij} mit den Matrixelementen $(A_{ij})_{ij} = -1$ und $(A_{ij})_{ji} = 1$ antisymmetrisch. Für $n = 2$ ist $A_{12} = J$. Die $n(n-1)/2$ Matrizen A_{ij} bilden eine Basis des $n(n-1)/2$-dimensionale Tangentialraumes von $\mathrm{O}(n)$.

Tangentialvektoren in einem anderen Punkt $g \in \mathrm{O}(n)$ haben die Form gA, wobei A eine antisymmetrische Matrix ist. Diese Matrizen sind nur noch in speziellen Fällen antisymmetrisch, zum Beispiel im Punkt $-I \in \mathrm{O}(n)$.

Die Gruppe $\mathrm{SO}(n)$

Die Gruppe $\mathrm{O}(n)$ enhält auch Isometrien, die die Orientierung des Raumes umkehren, wie zum Beispiel Spiegelungen. Wegen $\det(AA^t) = \det A \det A^t = (\det A)^2 = 1$ kann die Determinante einer orthogonalen Matrix nur ± 1 sein. Orientierungserhaltende Isometrien haben Determinante 1.

Die Gruppe

$$\mathrm{SO}(n) = \{A \in \mathrm{O}(n) \mid \ \det A = 1\}$$

heisst die *spezielle orthogonale Gruppe*. Die Dimension der Gruppe $\mathrm{O}(n)$ ist $n(n-1)/2$.

Die Gruppe $\mathrm{SO}(3)$

Die Gruppe $\mathrm{SO}(3)$ der Drehungen des dreidimensionalen Raumes hat die Dimension $3(3-1)/2 = 3$. Eine Drehung wird festgelegt durch die Richtung der Drehachse und den Drehwinkel. Die Richtung der Drehachse ist ein Einheitsvektor, also ein Punkt auf der zweidimensionalen Kugel. Der Drehwinkel ist der dritte Parameter.

Drehungen mit kleinen Drehwinkeln können zusammengesetzt werden aus den Matrizen

$$\begin{aligned} D_{x,\alpha} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}, & D_{y,\beta} &= \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}, & D_{z,\gamma} &= \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= e^{A_{23}t} & &= e^{-A_{13}t} & &= e^{A_{21}t} \end{aligned}$$

die Drehungen um die Koordinatenachsen um den Winkel α beschreiben. Auch die Winkel α, β und γ können als die drei Koordinaten der Mannigkfaltigkeit $\mathrm{SO}(3)$ angesehen werden.

7.2.4 Volumenerhaltende Abbildungen und die Gruppe $\mathrm{SL}_n(\mathbb{R})$

Die Elemente der Gruppe $SO(n)$ erhalten Längen, Winkel und die Orientierung, also auch das Volumen. Es gibt aber volumenerhaltende Abbildungen, die Längen oder Winkel nicht notwendigerweise erhalten. Matrizen $A \in M_n(\mathbb{R})$, die das Volumen erhalten, haben die Determinante $\det A = 1$. Wegen $\det(AB) = \det A \det B$ ist das Produkt zweier Matrizen mit Determinante 1 wieder eine solche, sie bilden daher eine Gruppe.

Definition 7.6. *Die volumenerhaltenden Abbildungen bilden die Gruppe*

$$\mathrm{SL}_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}) \mid \det(A) = 1\}$$

sie heisst die spezielle lineare Gruppe.

Wir wollen jetzt die Tangentialvektoren von $\mathrm{SL}_n(\mathbb{R})$ bestimmen. Dazu sei $A(t)$ eine Kurve in $\mathrm{SL}_n(\mathbb{R})$ mit $A(0) = I$. Für alle $t \in \mathbb{R}$ ist $\det A(t) = 1$, daher ist die Ableitung

$$\frac{d}{dt} \det A(t) = 0 \quad \text{an der Stelle } t = 0.$$

Für $n = 2$ ist

$$A(t) = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}) \quad \Rightarrow \quad \frac{d}{dt} \det A(t)\Big|_{t=0} = \dot{a}(0)d(0) + a(0)\dot{d}(0) - \dot{b}(0)c(0) - b(0)\dot{c}(0)$$
$$= \dot{a}(0) + \dot{d}(0)$$
$$= \mathrm{Spur}\, \frac{dA}{dt}.$$

Dies gilt nicht nur im Falle $n = 2$, sondern ganz allgemein für beliebige $n \times n$-Matrizen.

Satz 7.7. *Ist $A(t)$ eine differenzierbare Kurve in $\mathrm{SL}_n(\mathbb{B})$ mit $A(0) = I$, dann ist* $\mathrm{Spur}\, \dot{A}(0) = 0$.

Beweis. Die Entwicklung der Determinante von A nach der ersten Spalte ist

$$\det A(t) = \sum_{i=1}^{n} (-1)^{i+1} a_{i1}(t) \det A_{i1}(t).$$

Die Ableitung nach t ist

$$\frac{d}{dt} \det A(t) = \sum_{i=1}^{n} (-1)^{i+1} \dot{a}_{i1}(t) \det A_{i1}(t). + \sum_{i=1}^{n} (-1)^{i+1} a_{i1}(t) \frac{d}{dt} \det A_{i1}(t).$$

An der Stelle $t = 0$ enthält $\det A_{i1}(0)$ für $i \neq 1$ eine Nullzeile, der einzige nichtverschwindende Term in der ersten Summe ist daher der erste. In der zweiten Summe ist das einzige nicht verschwindende $a_{i1}(0)$ jenes für $i = 1$, somit ist die Ableitung von $\det A(t)$

$$\frac{d}{dt} \det A(t) = \dot{a}_{11}(t) \det A_{11}(t). + \frac{d}{dt} \det A_{11}(t) = \dot{a}_{11}(0) + \frac{d}{dt} \det A_{11}(t). \tag{7.5}$$

Die Beziehung (7.5) kann für einen Beweis mit vollständiger Induktion verwendet werden.

Die Induktionsverankerung für $n = 1$ besagt, dass $\det A(t) = a_{11}(t)$ genau dann konstant = 1 ist, wenn $\dot{a}_{11}(0) = \operatorname{Spur} A(0)$ ist. Unter der Induktionsannahme, dass für eine $(n-1) \times (n-1)$-Matrix $\tilde{A}(t)$ mit $\tilde{A}(0) = I$ die Ableitung der Determinante

$$\frac{d}{dt}\tilde{A}(0) = \operatorname{Spur} \dot{\tilde{A}}(0)$$

ist, folgt jetzt mit (7.5), dass

$$\frac{d}{dt}A(0) = \dot{a}_{11}(0) + \frac{d}{dt}\det A_{11}(t)\Big|_{t=0} = \dot{a}_{11}(0) + \operatorname{Spur} \dot{A}_{11}(0) = \operatorname{Spur} \dot{A}(0).$$

Damit folgt jetzt die Behauptung für alle n. □

Beispiel. Die Tangentialvektoren von $\mathrm{SL}_2(\mathbb{R})$ sind die spurlosen Matrizen

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \Rightarrow \quad \operatorname{Spur} A = a + d = 0 \quad \Rightarrow \quad A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}.$$

Der Tangentialraum ist also dreidimensional. Als Basis könnte man die folgenden Vektoren verwenden:

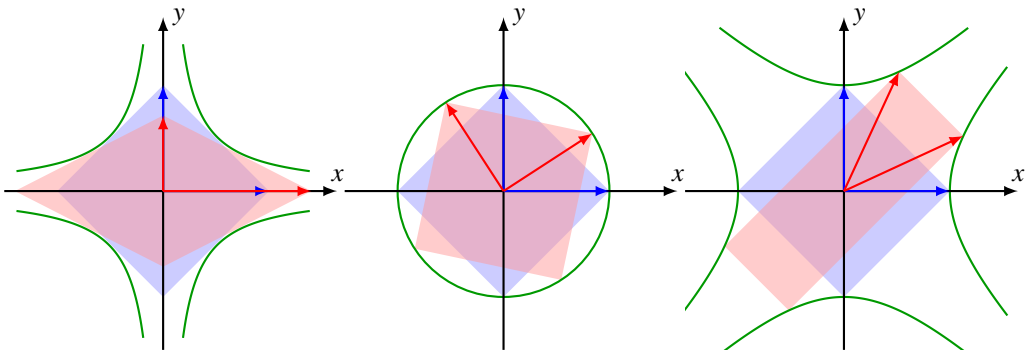
$$\begin{aligned}
A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad &\Rightarrow \qquad e^{At} = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \\
B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad &\Rightarrow \qquad e^{Bt} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \\
C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad &\Rightarrow \qquad e^{Ct} = I + Ct + \frac{C^2t^2}{2!} + \frac{C^3t^3}{3!} + \frac{C^4t^4}{4!} + \dots \\
&\qquad\qquad = I\Big(1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \dots\Big) + C\Big(t + \frac{t^3}{3!} + \frac{t^5}{5!} + \dots\Big) \\
&\qquad\qquad = I\cosh t + C\sinh t = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix},
\end{aligned}$$

wobei in der Auswertung der Potenzreihe für e^{Ct} verwendet wurde, dass $C^2 = I$.

Die Matrizen e^{At} Streckungen der einen Koordinatenachse und Stauchungen der anderen derart, dass das Volumen erhalten bleibt. Die Matrizen e^{Bt} sind Drehmatrizen, die Längen und Winkel und damit erst recht den Flächeninhalt erhalten. Die Matrizen der Form e^{Ct} haben die Vektoren $(1, \pm 1)$ als Eigenvektoren:

$$\begin{aligned}
\begin{pmatrix} 1 \\ 1 \end{pmatrix} \mapsto e^{Ct}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = (\cosh t + \sinh t)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \Big(\frac{e^t + e^{-t}}{2} + \frac{e^t - e^{-t}}{2}\Big)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = e^t\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
\begin{pmatrix} 1 \\ -1 \end{pmatrix} \mapsto e^{Ct}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = (\cosh t - \sinh t)\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \Big(\frac{e^t + e^{-t}}{2} - \frac{e^t - e^{-t}}{2}\Big)\begin{pmatrix} 1 \\ -1 \end{pmatrix} = e^{-t}\begin{pmatrix} 1 \\ -1 \end{pmatrix}
\end{aligned}$$

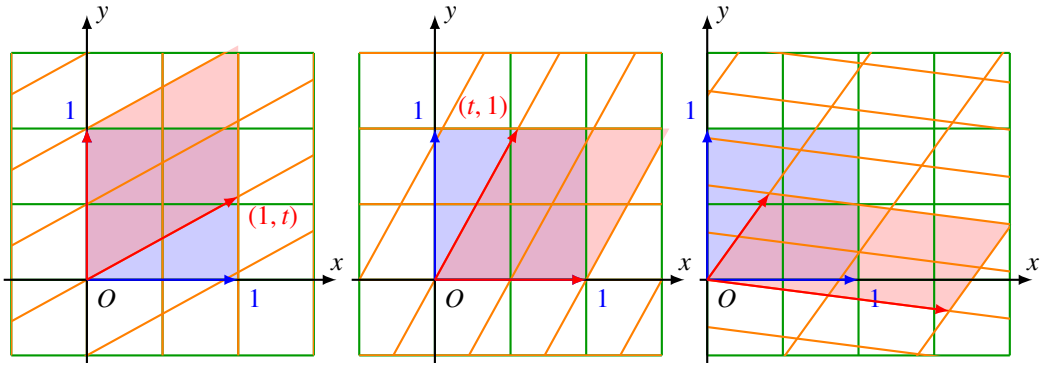
Die Matrizen e^{Ct} strecken die Richtung $(1, 1)$ um e^t und die dazu orthogonale Richtung $(1, -1)$ um den Faktor e^{-t}. Dies ist die gegenüber e^{At} um 45° verdrehte Situation, auch diese Matrizen sind flächenerhaltend. ○



$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$e^{At} = \begin{pmatrix} e^{t} & 0 \\ 0 & e^{-t} \end{pmatrix} \qquad e^{Bt} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \qquad e^{Ct} = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}$$

Abbildung 7.5: Tangentialvektoren und die davon erzeugen Einparameteruntergruppen für die Lie-Gruppe $\mathrm{SL}_2(\mathbb{R})$ der flächenerhaltenden linearen Abbildungen von $\mathbb{R}^2$. In allen drei Fällen wird ein blauer Rhombus mit den Ecken in den Standardbasisvektoren von einer Matrix der Einparameteruntergruppe zu zum roten Viereck verzerrt, der Flächeninhalt bleibt aber erhalten. In den beiden Fällen B und C stellen die grünen Kurven die Bahnen der Bilder der Standardbasisvektoren dar.



$$M = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \qquad N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 0.5 & 0.4 \\ -0.2 & -0.5 \end{pmatrix}$$

$$e^{Mt} = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \qquad e^{Nt} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \qquad e^{D\cdot 1} = \begin{pmatrix} 1.6005 & 0.4114 \\ -0.2057 & 0.5719 \end{pmatrix}$$

Abbildung 7.6: Weitere Matrizen mit Spur 0 und ihre Wirkung Die inken beiden Beispiele M und N sind nilpotente Matrizen, die zugehörigen Einparameter-Untergruppen beschreiben Schwerungen.

7.2.5 Die Gruppe SU(2)

Die Menge der Matrizen

$$\mathrm{SU}(2) = \left\{ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \;\middle|\; a, b, c, d \in \mathbb{C}, \det(A) = 1, AA^* = I \right\}$$

heisst die *spezielle unitäre Gruppe*. Wegen $\det(AB) = \det(A)\det(B) = 1$ und $(AB)^*AB = B^*A^*AB = B^*B = I$ ist SU(2) eine Untergruppe von $\mathrm{GL}_2(\mathbb{C})$. Die Bedingungen $\det A = 1$ und $AA^* = I$ schränken die möglichen Werte von a und b weiter ein. Aus

$$A^* = \begin{pmatrix} \overline{a} & \overline{c} \\ \overline{b} & \overline{d} \end{pmatrix}$$

und den Bedingungen führen die Gleichungen

$$\begin{aligned}
a\overline{a} + b\overline{b} = 1 \quad &\Rightarrow \quad |a|^2 + |b|^2 = 1 \\
a\overline{c} + b\overline{d} = 0 \quad &\Rightarrow \quad \frac{a}{b} = -\frac{\overline{d}}{\overline{c}} \\
c\overline{a} + d\overline{b} = 0 \quad &\Rightarrow \quad \frac{c}{d} = -\frac{\overline{b}}{\overline{a}} \\
c\overline{c} + d\overline{d} = 1 \quad &\Rightarrow \quad |c|^2 + |d|^2 = 1 \\
ad - bc = 1 &
\end{aligned}$$

Aus der zweiten Gleichung kann man ableiten, dass es eine Zahl $t \in \mathbb{C}$ gibt derart, dass $c = -t\overline{b}$ und $d = t\overline{a}$. Damit wird die Bedingung an die Determinante zu

$$1 = ad - bc = at\overline{a} - b(-t\overline{b}) = t(|a|^2 + |b|^2) = t,$$

also muss die Matrix A die Form haben

$$A = \begin{pmatrix} a & b \\ -\overline{b} & \overline{a} \end{pmatrix} \qquad \text{mit} \quad |a|^2 + |b|^2 = 1.$$

Schreibt man $a = a_1 + ia_2$ und $b = b_1 + ib_2$ mit rellen a_i und b_i, dann besteht $SU(2)$ aus den Matrizen der Form

$$A = \begin{pmatrix} a_1 + ia_2 & b_1 + ib_2 \\ -b_1 + ib_2 & a_1 - ia_2 \end{pmatrix}$$

mit der zusätzlichen Bedingung

$$|a|^2 + |b|^2 = a_1^2 + a_2^2 + b_1^2 + b_2^2 = 1.$$

Die Matrizen von SU(2) stehen daher in einer eins-zu-eins-Beziehung zu den Vektoren $(a_1, a_2, b_1, b_2) \in \mathbb{R}^4$ eines vierdimensionalen reellen Vektorraums mit Länge 1. Geometrisch betrachtet ist also SU(2) eine dreidmensionalen Kugel, die in einem vierdimensionalen Raum eingebettet ist.

7.3 Lie-Algebren

Im vorangegangenen Abschnitt wurde gezeigt, dass alle beschriebenen Matrizengruppen als Untermannigfaltigkeiten im n^2-dimensionalen Vektorraum $M_n(\mathbb{R}9$ betrachtet werden können. Die Gruppen haben damit nicht nur die algebraische Struktur einer Matrixgruppe, sie haben auch die geometrische Struktur einer Mannigfaltigkeit. Insbesondere ist es sinnvoll, von Ableitungen zu sprechen.

Eindimensionale Untergruppen einer Gruppe können auch als Kurven innerhalb der Gruppe angesehen werden. In diesem Abschnitt soll gezeigt werden, wie man zu jeder eindimensionalen Untergruppe einen Vektor in $M_n(\mathbb{R})$ finden kann derart, dass der Vektor als Tangentialvektor an diese Kurve gelten kann. Aus einer Abbildung zwischen der Gruppe und diesen Tagentialvektoren erhält man dann auch eine algebraische Struktur auf diesen Tangentialvektoren, die sogenannte Lie-Algebra. Sie ist charakteristisch für die Gruppe. Insbesondere werden wir sehen, wie die Gruppen SO(3) und SU(2) die gleich Lie-Algebra haben und dass die Lie-Algebra von SO(3) mit dem Vektorprodukt in $\mathbb{R}^3$ übereinstimmt.

7.3.1 Lie-Algebra einer Matrizengruppe

Zu jedem Tangentialvektor A im Punkt I einer Matrizengruppe gibt es eine Einparameteruntergruppe, die mit Hilfe der Exponentialfunktion e^{At} konstruiert werden kann. Für die folgende Konstruktion arbeiten wir in der Gruppe $\mathrm{GL}_n(\mathbb{R})$, in der jede Matrix auch ein Tangentialvektor ist. Wir werden daraus die Lie-Klammer ableiten und später verifizieren, dass diese auch für die Tangentialvektoren der Gruppen $\mathrm{SO}(n)$ oder $\mathrm{SL}_n(\mathbb{R})$ funktioniert.

Lie-Klammer

Zu zwei verschiedenen Tagentialvektoren $A \in M_n(\mathbb{R})$ und $B \in M_n(\mathbb{R})$ gibt es zwei verschiedene Einparameteruntergruppen e^{At} und e^{Bt}. Wenn die Matrizen A und B oder die Einparameteruntergruppen e^{At} und e^{Bt} vertauschbar sind, dann stimmen $e^{At}e^{Bt}$ und $e^{Bt}e^{At}$ nicht überein. Die zugehörigen Potenzreihen sind:

$$
\begin{aligned}
e^{At} &= I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \dots \\
e^{Bt} &= I + Bt + \frac{B^2t^2}{2!} + \frac{B^3t^3}{3!} + \dots \\
e^{At}e^{Bt} &= \left(I + At + \frac{A^2t^2}{2!} + \dots\right)\left(I + Bt + \frac{B^2t^2}{2!} + \dots\right) \\
&= I + (A+B)t + \left(\frac{A^2}{2!} + AB + \frac{B^2}{2!}\right)t^2 + \dots \\
e^{Bt}e^{At} &= \left(I + Bt + \frac{B^2t^2}{2!} + \dots\right)\left(I + At + \frac{A^2t^2}{2!} + \dots\right) \\
&= I + (B+A)t + \left(\frac{B^2}{2!} + BA + \frac{A^2}{2!}\right)t^2 + \dots
\end{aligned}
$$

Die beiden Kurven $e^{At}e^{Bt}$ und $e^{Bt}e^{At}$ haben zwar den gleichen Tangentialvektor für $t = 0$, sie unterscheiden sich aber untereinander, und sie unterscheiden sich von der Einparameteruntergruppe von $A + B$

$$
e^{(A+B)t} = I + (A+B)t + \frac{t^2}{2}(A^2 + AB + BA + B^2) + \dots
$$

Für die Unterschiede finden wir

$$\begin{aligned}
e^{At}e^{Bt} - e^{(A+B)t} &= \left(AB - \frac{AB+BA}{2}\right)t^2 + \ldots = (AB-BA)\frac{t^2}{2} + \ldots = [A,B]\frac{t^2}{2} + \ldots \\
e^{Bt}e^{At} - e^{(A+B)t} &= \left(BA - \frac{AB+BA}{2}\right)t^2 + \ldots = (BA-AB)\frac{t^2}{2} + \ldots = -[A,B]\frac{t^2}{2} \\
e^{At}e^{Bt} - e^{Bt}e^{At} &= (AB-BA)t^2 + \ldots = \quad [A,B]t^2 + \ldots
\end{aligned}$$

wobei mit $[A,B] = AB - BA$ abgekürzt wird.

Definition 7.8. *Der Kommutator zweier Matrizen $A, B \in M_n(\mathbb{R})$ ist die Matrix $[A,B] = AB - BA$.*

Der Kommutator ist bilinear und antisymmetrisch, da

$$\begin{aligned}
[\lambda A + \mu B, C] &= \lambda AC + \mu BC - \lambda CA - \mu CB = \lambda[A,C] + \mu[B,C] \\
[A, \lambda B + \mu C] &= \lambda AB + \mu AC - \lambda BA - \mu CA = \lambda[A,B] + \mu[A,C] \\
[A,B] &= AB - BA = -(BA - AB) = -[B,A].
\end{aligned}$$

Aus der letzten Bedingung folgt insbesodnere $[A,A] = 0$

Der Kommutator $[A,B]$ misst in niedrigster Ordnung den Unterschied zwischen den e^{At} und e^{Bt}. Der Kommutator der Tangentialvektoren A und B bildet also die Nichtkommutativität der Matrizen e^{At} und e^{Bt} ab.

Die Jacobi-Identität

Der Kommutator hat die folgende zusätzliche algebraische Eigenschaft:

$$\begin{aligned}
[A,[B,C]] + [B,[C,A]] + [C,[A,B]] &= [A, BC-CB] + [B, CA-AC] + [C, AB-BA] \\
&= \quad ABC - ACB - BCA + CBA \\
&\quad + BCA - BAC - CAB + ACB \\
&\quad + CAB - CBA - ABC + BAC \\
&= 0.
\end{aligned}$$

Diese Eigenschaft findet man auch bei anderen Strukturen, zum Beispiel bei Vektorfeldern, die man als Differentialoperatoren auf Funktionen betrachten kann. Man kann dann einen Kommutator $[X,Y]$ für zwei Vektorfelder X und Y definieren. Dieser Kommutator von Vektorfeldern erfüllt ebenfalls die gleiche Identität.

Definition 7.9. *Ein bilineares Produkt $[\ ,\ ]\colon V \times V \to V$ auf dem Vektorraum erfüllt die* Jacobi-Identität*, wenn*

$$[u,[v,w]] + [v,[w,u]] + [w,[u,v]] = 0$$

ist für beliebige Vektoren $u, v, w \in V$.

Lie-Algebra

Die Tangentialvektoren einer Lie-Gruppe tragen also mit dem Kommutator eine zusätzliche Struktur, nämlich die Struktur einer Lie-Algebra.

Definition 7.10. *Ein Vektorraum V mit einem bilinearen, Produkt*

$$[\ ,\ ]\colon V \times V \ \to V : (u, v) \mapsto [u, v],$$

welches zusätzlich die Jacobi-Identität 7.9 erfüllt, heisst eine Lie-Algebra*.*

Die Lie-Algebra einer Lie-Gruppe G wird mit LG bezeichnet. LG besteht aus den Tangentialvektoren im Punkt I. Die Exponentialabbildung $\exp\colon LG \to G : A \mapsto e^A$ ist eine differenzierbare Abbildung von LG in die Gruppe G. Insbesondere kann die Inverse der Exponentialabbildung als eine Karte in einer Umgebung von I verwendet werden.

Für die Lie-Algebren der Matrizengruppen, die früher definiert worden sind, verwenden wir die als Notationskonvention, dass der Name der Lie-Algebra der mit kleinen Buchstaben geschrieben Name der Lie-Gruppe ist. Die Lie-Algebra von $\mathrm{SO}(n)$ ist also $L\,\mathrm{SO}(n) = \mathrm{os}(n)$, die Lie-Algebra von $\mathrm{SL}_n(\mathbb{R})$ ist $L\,\mathrm{SL}_n(\mathbb{R}) = \mathrm{sl}_n(\mathbb{R})$.

7.3.2 Die Lie-Algebra von SO(3)

Zur Gruppe SO(3) der Drehmatrizen gehört die Lie-Algebra so(3) der antisymmetrischen 3×3-Matrizen. Solche Matrizen haben die Form

$$\Omega = \begin{pmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{pmatrix}$$

Der Vektorraum so(3) ist also dreidimensional.

Die Wirkung von $I + t\Omega$ auf einem Vektor x ist

$$(I + t\Omega)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & t\omega_3 & -t\omega_2 \\ -t\omega_3 & 1 & t\omega_1 \\ t\omega_2 & -t\omega_1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 - t(-\omega_3 x_2 + \omega_2 x_3) \\ x_2 - t(\omega_3 x_1 - \omega_1 x_3) \\ x_3 - t(-\omega_2 x_1 + \omega_1 x_2) \end{pmatrix} = x - t\begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \times x = x + tx \times \omega.$$

Die Matrix Ω ist als die infinitesimale Version einer Drehung um die Achse ω.

Wir können die Analogie zwischen Matrizen in so(3) und Vektoren in $\mathbb{R}^3$ noch etwas weiter treiben. Zu jedem Vektor in $\mathbb{R}^3$ konstruieren wir eine Matrix in so(3) mit Hilfe der Abbildung

$$\mathbb{R}^3 \to \mathrm{so}(3) : \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \mapsto \begin{pmatrix} 0 & v_3 & -v_1 \\ -v_3 & 0 & v_2 \\ v_1 & -v_2 & 0 \end{pmatrix}.$$

Der Kommutator von zwei so aus Vektoren $\vec{u}$ und $\vec{v}$ konstruierten Matrizen U und V ist:

$$\begin{aligned}
[U, V] &= UV - VU \\
&= \begin{pmatrix} 0 & u_3 & -u_1 \\ -u_3 & 0 & u_2 \\ u_1 & -u_2 & 0 \end{pmatrix}\begin{pmatrix} 0 & v_3 & -v_1 \\ -v_3 & 0 & v_2 \\ v_1 & -v_2 & 0 \end{pmatrix} - \begin{pmatrix} 0 & v_3 & -v_1 \\ -v_3 & 0 & v_2 \\ v_1 & -v_2 & 0 \end{pmatrix}\begin{pmatrix} 0 & u_3 & -u_1 \\ -u_3 & 0 & u_2 \\ u_1 & -u_2 & 0 \end{pmatrix} \\
&= \begin{pmatrix} u_3v_3 + u_1v_1 - u_3v_3 - u_1v_1 & u_1v_2 - u_2v_1 & u_3v_2 - u_2v_3 \\ u_2v_1 - u_1v_2 & -u_3v_3 - u_2v_2 + u_3v_3 + u_2v_2 & u_3v_1 - u_1v_3 \\ u_2v_3 - u_3v_2 & u_1v_3 - u_3v_1 & -u_1v_1 - u_2v_2u_1v_1 + u_2v_2 \end{pmatrix} \\
&= \begin{pmatrix} 0 & u_1v_2 - u_2v_1 & -(u_2v_3 - u_3v_2) \\ -(u_1v_2 - u_2v_1) & 0 & u_3v_1 - u_1v_3 \\ u_2v_3 - u_3v_2 & -(u_3v_1 - u_1v_3) & 0 \end{pmatrix}
\end{aligned}$$

Die Matrix $[U, V]$ gehört zum Vektor $\vec{u} \times \vec{v}$. Damit können wir aus der Jacobi-Identität jetzt folgern, dass

$$\vec{u} \times (\vec{v} \times w) + \vec{v} \times (\vec{w} \times u) + \vec{w} \times (\vec{u} \times v) = 0$$

für drei beliebige Vektoren $\vec{u}$, $\vec{v}$ und $\vec{w}$ ist. Dies bedeutet, dass der dreidimensionale Vektorraum $\mathbb{R}^3$ mit dem Vektorprodukt zu einer Lie-Algebra wird. In der Tat verwenden einige Bücher statt der vertrauten Notation $\vec{u} \times \vec{v}$ für das Vektorprodukt die aus der Theorie der Lie-Algebren entlehnte Notation $[\vec{u}, \vec{v}]$, zum Beispiel das Lehrbuch der Theoretischen Physik [3] von Landau und Lifschitz.

Die Lie-Algebren sind vollständig klassifiziert worden, es gibt keine nicht trivialen zweidimensionalen Lie-Algebren. Unser dreidimensionaler Raum ist also auch in dieser Hinsicht speziell: es ist der kleinste Vektorraum, in dem eine nichttriviale Lie-Algebra-Struktur möglich ist.

Die antisymmetrischen Matrizen

$$\omega_{23} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \omega_{31} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \omega_{12} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

haben die Kommutatoren

$$\begin{aligned} [\omega_{23}, \omega_{31}] &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} = \omega_{12} \\ [\omega_{31}, \omega_{12}] &= \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \omega_{23} \\ [\omega_{12}, \omega_{23}] &= \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \omega_{31} \end{aligned} \tag{7.6}$$

7.3.3 Die Lie-Algebra von $\mathrm{SL}_n(\mathbb{R})$

Die Lie-Algebra von $\mathrm{SL}_n(\mathbb{R})$ besteht aus den spurlosen Matrizen in $M_n(\mathbb{R})$. Der Kommutator solcher Matrizen erfüllt

$$\mathrm{Spur}([A, B]) = \mathrm{Spur}(AB - BA) = \mathrm{Spur}(AB) - \mathrm{Spur}(BA) = 0,$$

somit ist

$$\mathrm{sl}_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}) \mid \mathrm{Spur}(A) = 0\}$$

mit dem Kommutator eine Lie-Algebra.

7.3.4 Die Lie-Algebra von $\mathrm{U}(n)$

Die Lie-Gruppe

$$U(n) = \{A \in M_n(\mathbb{C} \mid AA^* = I\}$$

heisst die unitäre Gruppe, sie besteht aus den Matrizen, die das sesquilineare Standardskalarprodukt auf dem komplexen Vektorraum $\mathbb{C}^n$ invariant lassen. Sei eine $\gamma(t)$ ein differenzierbare Kurve in $\mathrm{U}(n)$ derart, dass $\gamma(0) = I$. Die Ableitung der Identität $AA^* = I$ führt dann auf

$$0 = \frac{d}{dt}\gamma(t)\gamma(t)^*\Big|_{t=0} = \dot{\gamma}(0)\gamma(0)^* + \gamma(0)\dot{\gamma}(0)^* = \dot{\gamma}(0) + \dot{\gamma}(0)^* \quad \Rightarrow \quad \dot{\gamma}(0) = -\dot{\gamma}(0)^*.A \quad = -A^*$$

Die Lie-Algebra u(n) besteht daher aus den antihermiteschen Matrizen.

Wir sollten noch verifizieren, dass der Kommutator zweier antihermiteschen Matrizen wieder anithermitesch ist:

$$[A,B]^* = (AB - BA)^* = B^*A^* - A^*B^* = BA - AB = -[B,A].$$

Eine antihermitesche Matrix erfüllt $a_{ij} = -\overline{a}_{ji}$, für die Diagonalelemente folgt daher $a_{ii} = -\overline{a}_{ii}$ oder $\overline{a}_{ii} = -a_{ii}$. Der Realteil von a_{ii} ist

$$\Re a_{ii} = \frac{a_{ii} + \overline{a}_{ii}}{2} = 0,$$

die Diagonalelemente einer antihermiteschen Matrix sind daher rein imaginär.

7.3.5 Die Lie-Algebra von SU(2)

Die Lie-Algebra su(n) besteht aus den spurlosen antihermiteschen Matrizen. Sie erfüllen daher die folgenden Bedingungen:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{mit} \qquad \begin{cases} a + d = 0 \\ b^* = -c \end{cases} \quad \Rightarrow \quad a = is = -d$$

Damit hat A die Form

$$\begin{aligned} A = \begin{pmatrix} is & u+iv \\ -u+iv & -is \end{pmatrix} &= s\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} + u\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + v\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \\ &= \underbrace{iv\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{=\sigma_1} + \underbrace{iu\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}}_{=\sigma_2} + \underbrace{is\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}}_{=\sigma_3} \end{aligned}$$

Diese Matrizen heissen die *Pauli-Matrizen*, sie haben die Kommutatoren

$$\begin{aligned} [\sigma_1,\sigma_2] &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} - \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 2\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = 2i\sigma_3, \\ [\sigma_2,\sigma_3] &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = 2\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = 2i\sigma_1. \\ [\sigma_1,\sigma_3] &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 2i\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = 2i\sigma_2, \end{aligned}$$

Bis auf eine Skalierung stimmt dies überein mit den Kommutatorprodukten der Matrizen ω_{23}, ω_{31} und ω_{12} in (7.6). Die Matrizen $-\frac{1}{2}i\sigma_j$ haben die Kommutatorprodukte

$$\begin{aligned} [-\tfrac{1}{2}i\sigma_1, -\tfrac{1}{2}i\sigma_2] &= -\tfrac{1}{4}[\sigma_1,\sigma_2] = -\tfrac{1}{4}\cdot 2i\sigma_3 = -\tfrac{1}{2}i\sigma_3 \\ [-\tfrac{1}{2}i\sigma_2, -\tfrac{1}{2}i\sigma_3] &= -\tfrac{1}{4}[\sigma_2,\sigma_3] = -\tfrac{1}{4}\cdot 2i\sigma_1 = -\tfrac{1}{2}i\sigma_1 \\ [-\tfrac{1}{2}i\sigma_3, -\tfrac{1}{2}i\sigma_1] &= -\tfrac{1}{4}[\sigma_3,\sigma_1] = -\tfrac{1}{4}\cdot 2i\sigma_2 = -\tfrac{1}{2}i\sigma_2 \end{aligned}$$

Die lineare Abbildung, die

$$\omega_{23} \mapsto -\tfrac{1}{2}i\sigma_1$$

$$\omega_{31} \mapsto -\tfrac{1}{2} i\sigma_2$$
$$\omega_{12} \mapsto -\tfrac{1}{2} i\sigma_3$$

abbildet ist daher ein Isomorphismus der Lie-Algebra so(3) auf die Lie-Algebra su(2). Die Lie-Gruppen SO(3) und SU(2) haben also die gleiche Lie-Algebra.

Tatsächlich kann man Hilfe von Quaternionen die Matrix SU(2) als Einheitsquaternionen beschreiben und damit eine Darstellung der Drehmatrizen in SO(3) finden. Dies wird in Kapitel 18 dargestellt.

Übungsaufgaben

7.1. Die Elemente der Gruppe G der Translationen und Streckungen von $\mathbb{R}$ kann durch Paare $(\lambda, t) \in \mathbb{R}^+ \times \mathbb{R}$ beschrieben werden, wobei λ durch Streckung und t durch Translation wirkt:

$$(\lambda, t) \colon \mathbb{R} \to \mathbb{R} : x \mapsto \lambda x + t.$$

Dies ist allerdings noch keine Untergruppe einer Matrizengruppe. Dazu bettet man $\mathbb{R}$ mit Hilfe der Abbildung

$$\mathbb{R} \to \mathbb{R}^2 : x \mapsto \begin{pmatrix} x \\ 1 \end{pmatrix}$$

in $\mathbb{R}^2$ ein. Die Wirkung von (λ, t) ist dann

$$\begin{pmatrix} (\lambda, t) \cdot x \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda x + t \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

Die Wirkung des Paares (λ, t) kann also mit Hilfe einer 2×2-Matrix beschrieben werden. Die Abbildung

$$G \to \mathrm{GL}_2(\mathbb{R}) : (\lambda, t) \mapsto \begin{pmatrix} \lambda & t \\ 0 & 1 \end{pmatrix}$$

bettet die Gruppe G in $\mathrm{GL}_2(\mathbb{R})$ ein.

a) Berechnen Sie das Produkt $g_1 g_2$ zweier Elemente $g_j = (\lambda_j, t_j)$.

b) Bestimmen Sie das inverse Elemente von (λ, t) in G.

c) Der sogenannte Kommutator zweier Elemente ist $g_1 g_2 g_1^{-1} g_2^{-1}$, berechnen Sie den Kommutator für die Gruppenelemente von a).

d) Rechnen Sie nach, dass

$$s \mapsto \begin{pmatrix} e^s & 0 \\ 0 & 1 \end{pmatrix}, \qquad t \mapsto \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

Einparameteruntergruppen von $\mathrm{GL}_2(\mathbb{R})$ sind.

e) Berechnen Sie die Tangentialvektoren S und T dieser beiden Einparameteruntergruppen.

f) Berechnen Sie den Kommutator $[S, T]$

Lösung. a) Die beiden Gruppenelemente wirken auf x nach

$$(\lambda_1, t_1)(\lambda_2, t_2) \cdot x = (\lambda_1, t_1)(\lambda_2 x + t_2) = \lambda_1(\lambda_2 x + t_2) + t_1) = \lambda_1\lambda_2 x + (\lambda_1 t_2 + t_1),$$

also ist $g_1 g_2 = (\lambda_1\lambda_2, \lambda_1 t_2 + t_1)$.

b) Die Inverse von (λ, t) kann erhalten werden, indem man die Abbildung $x \mapsto y = \lambda x + t$ nach x auflöst:

$$y = \lambda x + t \qquad \Rightarrow \qquad \lambda^{-1}(y - t) = \lambda^{-1} y - \lambda^{-1} t.$$

Daraus liest man ab, dass $(\lambda, t)^{-1} = (\lambda^{-1}, -\lambda^{-1} t)$ ist.

c) Mit Hilfe der Identität $g_1 g_2 g_1^{-1} g_2^{-1} = g_1 g_2 (g_2 g_1)^{-1}$ kann man den Kommutator leichter berechnen

$$\begin{aligned}
g_1 g_2 &= (\lambda_1\lambda_2, t_1 + \lambda_1 t_2) \\
g_2 g_1 &= (\lambda_2\lambda_1, t_2 + \lambda_2 t_1) \\
(g_2 g_1)^{-1} &= (\lambda_1^{-1}\lambda_2^{-1}, -\lambda_2^{-1}\lambda_1^{-1}(t_2 + \lambda_2 t_1)) \\
g_1 g_2 g_1^{-1} g_2^{-1} &= (\lambda_1\lambda_2, t_1 + \lambda_1 t_2)(\lambda_1^{-1}\lambda_2^{-1}, -\lambda_2^{-1}\lambda_1^{-1}(t_2 + \lambda_2 t_1)) \\
&= (1, t_1 + \lambda_1 t_2 + \lambda_1\lambda_2(-\lambda_2^{-1}\lambda_1^{-1}(t_2 + \lambda_2 t_1))) \\
&= (1, t_1 + \lambda_1 t_2 - t_2 - \lambda_2 t_1) = (1, (1 - \lambda_2)(t_1 - t_2)).
\end{aligned}$$

Der Kommutator ist also das neutrale Element, wenn $\lambda_2 = 1$ ist.

d) Dies ist am einfachsten in der Matrixform nachzurechnen:

$$\begin{pmatrix} e^{s_1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} e^{s_2} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{s_1+s_2} & 0 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & t_1 + t_2 \\ 0 & 1 \end{pmatrix}$$

e) Die Tangentialvektoren werden erhalten durch ableiten der Matrixdarstellung nach dem Parameter

$$\begin{aligned}
S &= \frac{d}{ds}\begin{pmatrix} e^s & 0 \\ 0 & 1 \end{pmatrix}\Big|_{s=0} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \\
T &= \frac{d}{dt}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\Big|_{t=0} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}
\end{aligned}$$

f) Der Kommutator ist

$$[S, T] = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = T. \qquad \bigcirc$$

7.2. Eine Drehung eines Vektors $\vec{x}$ der Ebene $\mathbb{R}^2$ um den Winkel α gefolgt von einer Translation um $\vec{t}$ ist gegeben durch $D_\alpha \vec{x} + \vec{t}$. Darauf lässt sich jedoch die Theorie der Matrizengruppen nicht darauf anwenden, weil die Operation nicht die Form einer Matrixmultiplikation schreiben. Die Drehung und Translation kann in eine Matrix zusammengefasst werden, indem zunächst die Ebene mit

$$\mathbb{R}^2 \to \mathbb{R}^3 : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \qquad \text{oder in Vektorschreibweise} \qquad \vec{x} \mapsto \begin{pmatrix} \vec{x} \\ 1 \end{pmatrix}$$

in den dreidimensionalen Raum eingebettet wird. Die Drehung und Verschiebung kann damit in der Form

$$\begin{pmatrix} D_\alpha \vec{x} + \vec{t} \\ 1 \end{pmatrix} = \begin{pmatrix} D_\alpha & \vec{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{x} \\ 1 \end{pmatrix}$$

als Matrizenoperation geschrieben werden. Die Gruppe der Drehungen und Verschiebungen der Ebene ist daher die Gruppe

$$G = \left\{ A = \begin{pmatrix} D_\alpha & \vec{t} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha & t_x \\ \sin\alpha & \cos\alpha & t_y \\ 0 & 0 & 1 \end{pmatrix} \middle| \; \alpha \in \mathbb{R}, \vec{t} \in \mathbb{R}^2 \right\}$$

Wir kürzen die Elemente von G auch als $(\alpha, \vec{t})$ ab.

a) Verifizieren Sie, dass das Produkt zweier solcher Matrizen $(\alpha_1, \vec{t}_1)$ und $(\alpha_2, \vec{t}_2)$ wieder die selbe Form $(\alpha, \vec{t})$ hat und berechnen Sie α und $\vec{t}_j$.

b) Bestimmen Sie das inverse Element zu $(\alpha, \vec{t}) \in G$.

c) Die Elemente der Gruppe G sind parametrisiert durch den Winkel α und die Translationskomponenten t_x und t_y. Rechnen Sie nach, dass

$$\alpha \mapsto \begin{pmatrix} D_\alpha & 0 \\ 0 & 1 \end{pmatrix}, \quad t_x \mapsto \begin{pmatrix} I & \begin{pmatrix} t_x \\ 0 \end{pmatrix} \\ 0 & 1 \end{pmatrix}, \qquad t_y \mapsto \begin{pmatrix} I & \begin{pmatrix} 0 \\ t_y \end{pmatrix} \\ 0 & 1 \end{pmatrix}$$

Einparameteruntergruppen von G sind.

d) Berechnen Sie die Tangentialvektoren D, X und Y, die zu den Einparameteruntergruppen von c) gehören.

e) Berechnen Sie die Lie-Klammer für alle Paare von Tangentialvektoren.

Lösung. a) Die Wirkung beider Gruppenelemente auf dem Vektor $\vec{x}$ ist

$$\begin{aligned} \begin{pmatrix} D_{\alpha_1} & \vec{t}_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} D_{\alpha_2} & \vec{t}_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{x} \\ 1 \end{pmatrix} &= \begin{pmatrix} D_{\alpha_1} & \vec{t}_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} D_{\alpha_2}\vec{x} + \vec{t}_2 \\ 1 \end{pmatrix} = \begin{pmatrix} D_{\alpha_1}(D_{\alpha_2}\vec{x} + \vec{t}_2) + \vec{t}_1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} D_{\alpha_1} D_{\alpha_2}\vec{x} + D_{\alpha_1}\vec{t}_2 + \vec{t}_1 \\ 1 \end{pmatrix} = \begin{pmatrix} D_{\alpha_1+\alpha_2} & D_{\alpha_1}\vec{t}_2 + \vec{t}_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{x} \\ 1 \end{pmatrix}. \end{aligned}$$

Das Produkt in der Gruppe G kann daher

$$(\alpha_1, \vec{t}_1)(\alpha_2, \vec{t}_2) = (\alpha_1 + \alpha_2, \vec{t}_1 + D_{\alpha_1}\vec{t}_2)$$

geschrieben werden.

b) Die Inverse der Abbildung $\vec{x} \mapsto \vec{y} = D_\alpha \vec{x} + \vec{t}$ kann gefunden werden, indem man auf der rechten Seite nach $\vec{x}$ auflöst:

$$\begin{aligned} \vec{y} = D_\alpha \vec{x} + \vec{t} \qquad &\Rightarrow \qquad D_\alpha^{-1}(\vec{y} - \vec{t}) = \vec{x} \\ &\qquad\qquad \vec{x} = D_{-\alpha}\vec{y} + (-D_{-\alpha}\vec{t}) \end{aligned}$$

Die Inverse von $(\alpha, \vec{t})$ ist also $(-\alpha, -D_{-\alpha}\vec{t})$.

c) Da D_α eine Einparameteruntergruppe von SO(2) ist, ist $\alpha \mapsto (D_\alpha, 0)$ ebenfalls eine Einparameteruntergruppe. Für die beiden anderen gilt

$$\left(I, \begin{pmatrix} t_{x1} \\ 0 \end{pmatrix}\right)\left(I, \begin{pmatrix} t_{x2} \\ 0 \end{pmatrix}\right) = \left(I, \begin{pmatrix} t_{x1} + t_{x2} \\ 0 \end{pmatrix}\right) \quad \text{und} \quad \left(I, \begin{pmatrix} 0 \\ t_{y1} \end{pmatrix}\right)\left(I, \begin{pmatrix} 0 \\ t_{y2} \end{pmatrix}\right) = \left(I, \begin{pmatrix} 0 \\ t_{y1} + t_{y2} \end{pmatrix}\right),$$

also sind dies auch Einparameteruntergruppen.

d) Die Ableitungen sind

$$D = \frac{d}{d\alpha}\begin{pmatrix} D_\alpha & 0 \\ 0 & 1 \end{pmatrix}\Big|_{\alpha=0} = \begin{pmatrix} J & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$X = \frac{d}{dt_x}\begin{pmatrix} I & \begin{pmatrix} t_x \\ 0 \end{pmatrix} \\ 0 & 1 \end{pmatrix}\Big|_{t_x=0} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad Y = \frac{d}{dt_y}\begin{pmatrix} I & \begin{pmatrix} 0 \\ t_y \end{pmatrix} \\ 0 & 1 \end{pmatrix}\Big|_{t_y=0} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

e) Die Vertauschungsrelationen sind

$$[D, X] = DX - XD = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = Y$$

$$[D, Y] = DY - YD = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = -X$$

$$[X, Y] = XY - YX = 0 - 0 = 0$$ ○

Kapitel 8

Graphen

Ein Graph ist eine Menge von Knoten, die untereinander mit Kanten verbunden sind. Graphen können zum Beispiel verwendet werden um Netzwerke zu beschreiben, aber auch viele andere Datenstrukturen. Die Knoten können einzelne Objekte beschreiben, die Kanten beschreiben dann Beziehungen zwischen diesen Objekten. Graphen haben zwar nur eine eindimensionale Geometrie, sie können aber auch als erste Approximation dreidimensionaler Objekte dienen.

Die Bedeutung des Graphenkozeptes wird unterstrichen von der Vielzahl von Fragestellungen, die über Graphen gestellt, und der zugehöriten Lösungsalgorithmen, die zu ihrer Beantwortung gefunden worden sind. Die Komplexitätstheorie hat sogar gezeigt, dass sich jedes diskrete Problem in ein Graphenproblem umformulieren lässt.

Das Problem, einen Stundenplan zu finden, der sicherstellt, dass alle Studierenden jedes Fach besuchen können, für die sie sich angemeldet haben, lässt sich zum Beispiel wie folgt als ein Graphenproblem formulieren. Die Fächer betrachten wir als Knoten des Graphen. Für jedes Paar von Fächern ziehen wir eine Kante des Graphen, wenn sich mindestens ein Studierender für beide Fächer angemeldet hat. Die Kante drückt aus, dass die beiden Fächer nicht zur gleichen Zeit geplant werden dürfen. Das Problem, einen Stundenplan zu finden, besteht jetzt darin, für jedes Fach ein Zeitintervall zu finden, während dem es durchgeführt werden soll. Natürlich steht nur eine beschränkte Anzahl Zeitintervalle zur Verfügung und benachbarte Knoten dürfen nicht ins gleiche Zeitintervall geplant werden. Das zugehörige abstrakte Graphenproblem heisst das Färbeproblem: ist es möglich, mit einer beschränkten Anzahl von Farben die Knoten des Graphen so einzufärben, dass benachbarte Knoten niemals die gleiche Farbe haben.

In diesem Kapitel soll zunächst gezeigt werden, wie man Graphen mit Hilfe von Matrizen beschreiben kann (Abschnitt 8.1). Das Ziel dabei ist natürlich, die Hilfsmittel der Matrixalgebra zur Lösung von Graphenprobleme hinzuzuziehen. Die spektrale Graphentheorie in Abschnitt 8.2 verwendet die Eigenwerte und Eigenvektoren der zugehörigen Matrix, um Aussagen über den Graphen zu machen. In Abschnitt 8.4 wird gezeigt, wie spektralen Eigenschaften verwendet werden können, um eine Art von Wavelets auf einem Graphen zu definieren. Damit ensteht eine für gewisse Anwendungen besonders leistungsfähige Basis zur Beschreibung von Funktionen auf dem Graphen.

8.1 Beschreibung von Graphen mit Matrizen

Ein Graph ist eine Menge von Knoten, die untereinander mit Kanten verbunden sind. Graphen können zum Beispiel verwendet werden um Netzwerke zu beschreiben, aber auch viele andere Daten-

strukturen. Die Knoten können einzelne Objekte beschreiben, die Kanten beschreiben dann Beziehungen zwischen diesen Objekten.

8.1.1 Definition von Graphen

In der Einleitung zu diesem Abschnitt wurde bereits eine informelle Beschreibung des Konzeptes eines Graphen gegeben. Um zu einer Beschreibung mit Hilfe von Matrizen zu kommen, wird eine exakte Definition benötigt. Dabei werden sich einige Feinheiten zeigen, die für Anwendungen wichtig sind und sich in Unterschieden in der Definition der zugehörigen Matrix äussern.

Ungerichtete Graphen

Die Grundlage für alle Arten von Graphen ist eine Menge V von *Knoten*, auch *Vertices* genannt. Die Unterschiede zeigen sich in der Art und Weise, wie die Knoten mit sogenannten Kanten verbunden werden. Bei einen ungerichteten Graphen sind die beiden Endpunkte einer Kante gleichwertig, es gibt keine bevorzugte Reihenfolge oder Richtung der Kante. Eine Kante wird daher vollständig spezifiziert, wenn wir die Menge der Endpunkte kennen. Dies führt auf die folgende Definition eines ungerichteten Graphen.

Definition 8.1. *Ein* ungerichteter Graph *ist eine endliche Menge V von* Knoten *und eine Menge E von zweielementigen Teilmengen*

$$E \subset \{\, \{a, b\} \subset V \,|\, a \neq b\}.$$

Die Elemente von E heissen Kanten *(*edges*).*

Man beachte, dass es keine Kante gibt, die einen Knoten $a \in V$ mit sich selbst verbindet, da die zugehörige Menge $\{a, a\} = \{a\}$ nicht aus zwei verschiedenen Elementen besteht, wie die Definition 8.1 dies verlangt.

Ein elektrisches Netzwerk von ohmschen Widerständen kann mit Hilfe eines ungerichteten Graphen beschrieben werden. Ohmsche Widerstände hängen nicht von der Richtung des Stromflusses durch die Widerstände ab. Will man Spannungen und Ströme in einem solchen Netzwerk berechnen, ist auch das Fehlen von Schleifen, die von a zu a führen, kein Verlust. Die Endpunkte solcher Widerstände wären immer auf dem gleichen Potential. Folglich würde kein Strom fliessen und sie hätten keinen Einfluss auf das Verhalten des Netzwerkes. Sie können einfach weggelassen werden.

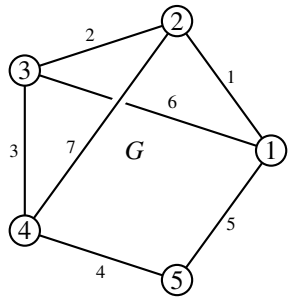
Gerichtete Graphen

In vielen Anwendungen sind die Endpunkte einer Kante nicht austauschbar. In einem Strassennetz sind Einbahnstrassen nicht in beiden Richtungen befahrbar. Anfangs- und Endpunkt einer Kante müssen in einem solche Graphen unterschieden werden. Eine zweielementige Menge ist daher nicht mehr eine geeignete Abstraktion für die Kante, ein (geordnetes) Paar von Vertizes passt besser.

Definition 8.2. *Ein* gerichteter Graph *ist eine endliche Menge V von Knoten und eine Menge $E \subset V \times V$ von gerichteten Kanten. Ausserdem gibt es zwei Abbildungen*

$$a\colon E \to V : (p, q) \mapsto a((p, q)) = p$$
$$e\colon E \to V : (p, q) \mapsto e((p, q)) = q.$$

Der Knoten $a(k)$ heisst der Anfangspunkt *der Kante $k \in E$, $e(k)$ heisst der* Endpunkt.



$$A(G) = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad D(G) = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

$$B(G) = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Abbildung 8.1: Adjazenz-, Inzidenz- und Gradmatrix eines ungerichteten Graphen mit 5 Knoten und 7 Kanten.

In einem gerichteten Graphen gehört also zu jeder Kante auch eine Richtung und die Unterscheidung von Anfangs- und Endpunkt einer Kante ist sinnvoll geworden. Ausderdem ist eine Kante (a, a) wohldefiniert, also eine Kante, die vom Knoten a wieder zu a zurückführt.

Man kann einen ungerichteten Graphen in einen gerichteten Graphen verwandeln, indem wir jede Kante $\{a, b\}$ durch zwei Kanten (a, b) und (b, a) ersetzen. Aus dem ungerichteten Graphen (V, E) mit Knotenmenge V und Kantenmenge E wird so der gerichtete Graph (V, E') mit der Kantenmenge

$$E' = \{(a, e) \mid \{a, e\} \in E\}.$$

Eine umgekehrte Zuordnung eines gerichteten zu einem ungerichteten Graphen ist nicht möglich, da eine "Schleife" (a, a) nicht in eine Kante des ungerichteten Graphen abgebildet werden kann.

In einem gerichteten Graphen kann man sinnvoll von gerichteten Pfaden sprechen. Ein *Pfad* γ in einem gerichteten Graphen (V, E) ist eine Folge $k_1, \ldots, k_r \in E$ von Kanten derart, dass $e(k_i) = a(k_{i+1})$ für $i = 1, \ldots, r-1$. Dies bedeutet, dass der Endpunkt jeder Kante mit dem Anfangspunkt der nachfolgenden Kante übereinstimmt. Die *Länge* des Pfades $\gamma = (k_1, \ldots, k_r)$ ist $|\gamma| = r$.

Adjazenzmatrix

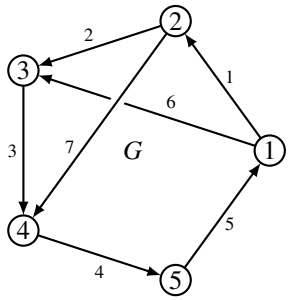
Eine naheliegende Beschreibung eines Graphen mit Hilfe einer Matrix kann man wie folgt erhalten. Zunächst werden die Knoten aus der Menge V durch die Zahlen $1, \ldots, n$ mit $n = |V|$ ersetzt. Diese Zahlen werden dann als Zeilen- uns Spaltenindizes interpretiert. Die zum Graphen gehörige sogenannte *Adjazenzmatrix* $A(G)$ enthält die Einträge

$$a_{ij} = \begin{cases} 1 & \{j, i\} \in E \\ 0 & \text{sonst.} \end{cases} \tag{8.1}$$

Die Matrix hat also genau dann einen von Null verschiedenen Eintrag in Zeile i und Spalte j, wenn die beiden Knoten i und j im Graphen verbunden sind. Die Adjazenzmatrix eines ungerichteten Graphen ist immer symmetrisch. Ein Beispiel ist in Abbildung 8.1 dargestellt.

Die Adjazenzmatrix kann auch für einen gerichteten Graphen definiert werden wie dies in in Abbildung 8.1 illustriert ist. Ihre Einträge sind in diesem Fall definiert mit Hilfe der gerichteten Kanten als

$$A(G)_{ij} = a_{ij} = \begin{cases} 1 & (j, i) \in E \\ 0 & \text{sonst.} \end{cases} \tag{8.2}$$



$$A(G) = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad D(G) = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$B(G) = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}$$

Abbildung 8.2: Adjazenz-, Inzidenz- und Gradmatrix eines gerichteten Graphen mit 5 Knoten und 7 Kanten.

Die Matrix $A(G)$ hat also genau dann einen nicht verschwindenden Matrixeintrag in Zeile i und Spalte j, wenn es eine Verbindung von Knoten j zu Knoten i gibt.

Adjazenzmatrix und die Anzahl der Pfade

Die Beschreibung des Graphen mit der Adjazenzmatrix $A = A(G)$ nach (8.2) ermöglicht bereits, eine interessante Aufgabe zu lösen.

Satz 8.3. *Der gerichtete Graph $G = ([n], E)$ werde beschrieben durch die Adjazenzmatrix $A = A(G)$. Dann gibt das Element in Zeile j und Spalte i von A^n die Anzahl der Wege der Länge n an, die von Knoten i zu Knoten j führen. Insbesondere kann man die Definition* (8.2) *formulieren als: In Zeile j und Spalte i der Matrix steht die Anzahl der Pfade der Länge* 1*, die i mit j verbinden.*

Beweis. Es ist klar, dass A^1 die genannte Eigenschaft hat. Wir beweisen, dass A^n Pfade der Länge n zählt, mit Hilfe von vollständiger Induktion. Zur Unterscheidung schreiben wir $A^{(n)}$ für die Matrix, die in Zeile j und Spalte i die Anzahl der Pfade der Länge n von i nach j enhält. Die zugehörigen Matrixelemente schreiben wir a^n_{ji} bzw. $a^{(n)}_{ji}$. Wir haben also zu zeigen, dass $A^n = A^{(n)}$.

Wir nehmen daher an, dass bereits bewiesen ist, dass das Element in Zeile j und Spalte i von A^{n-1} die Anzahl der Pfade der Länge $n-1$ zählt, dass also $A^{n-1} = A^{(n-1)}$. Dies ist die Induktionsannahme.

Wir bilden jetzt alle Pfade der Länge n von i nach k. Ein Pfad der Länge besteht aus einem Pfad der Länge $n-1$, der von i zu einem beliebigen Knoten j führt, gefolgt von einer einzelnen Kante, die von j nach k führt. Ob es eine solche Kante gibt, zeigt das Matrixelement a_{kj} an. Das Element in Zeile j und Spalte i der Matrix $A^{(n-1)}$ gibt die Anzahl der Wege von i nach j an. Es gibt also $a_{kj} \cdot a^{(n-1)}_{ji}$ Wege der Länge n, die von i nach k führen, aber als zweitletzten Knoten über den Knoten j führen. Die Gesamtzahl der Wege der Länge n von i nach k ist daher

$$a^{(n)}_{ki} = \sum_{j=1}^{n} a_{kj} a^{(n-1)}_{ji}.$$

In Matrixschreibweise bedeutet dies

$$A^{(n)} = A \cdot A^{(n-1)} = A \cdot A^{n-1} = A^n.$$

Beim zweiten Gleichheitszeichen haben wir die Induktionsannahme verwendet. □

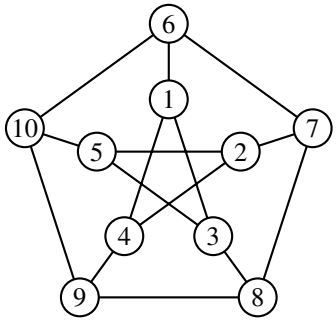


Abbildung 8.3: Peterson-Graph mit zehn Knoten.

Der Satz 8.3 ermöglicht auch, einen Algorithmus für den sogenannten Durchmesser eines Graphen zu formulieren.

Definition 8.4. *Der* Durchmesser *eines Graphen ist die kürzeste Länge d derart, dass es zwischen zwei beliebigen Knoten einen Pfad der Länge $\leq d$ gibt.*

Der Durchmesser d eines Graphen ist der kleinste Exponent derart, dass A^d keine ausserdiagonalen Einträge 0 hat. Die Diagonalelemente von A^n zählen die Anzahl der geschlossenen Pfade der Länge n, die durch einen Knoten führen. Diese können für den Durchmesser ignoriert werden. Man kann also Potenzen A^n berechnen bis keine Einträge 0 mehr vorhanden sind.

Beispiel. Der Peterson-Graph hat die Adjazenzmatrix

$$G = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Durch Nachrechnen kann man bestätigen, dass G^3 keine Ausserdiagonalelemente 0 enthält:

$$G^3 = \begin{pmatrix} 0 & 2 & 5 & 5 & 2 & 5 & 2 & 2 & 2 & 2 \\ 2 & 0 & 2 & 5 & 5 & 2 & 5 & 2 & 2 & 2 \\ 5 & 2 & 0 & 2 & 5 & 2 & 2 & 5 & 2 & 2 \\ 5 & 5 & 2 & 0 & 2 & 2 & 2 & 2 & 5 & 2 \\ 2 & 5 & 5 & 2 & 0 & 2 & 2 & 2 & 2 & 5 \\ 5 & 2 & 2 & 2 & 2 & 0 & 5 & 2 & 2 & 5 \\ 2 & 5 & 2 & 2 & 2 & 5 & 0 & 5 & 2 & 2 \\ 2 & 2 & 5 & 2 & 2 & 2 & 5 & 0 & 5 & 2 \\ 2 & 2 & 2 & 5 & 2 & 2 & 2 & 5 & 0 & 5 \\ 2 & 2 & 2 & 2 & 5 & 5 & 2 & 2 & 5 & 0 \end{pmatrix}$$

Daraus kann man jetzt ablesen, dass der Durchmesser des Petersongraphen $d = 5$ ist. Man kann aber auch mehr ablesen:

- Es gibt keine geschlossenen Pfade der Länge ≤ 3.
- Zwischen benachbarten Knoten gibt es jeweils 5 Pfade der Länge 3, zwischen nicht benachbarten Knoten gibt es genau 2 Pfade der Länge 3. ○

Das Beispiel illustriert, wie sich Zählaufgaben von Pfaden leicht mit dem Matrizenprodukt erledigen lassen. Trotzdem ist der Algorithmus nicht unbedingt effizient, da der Aufwand zur Berechnung des Matrizenproduktes relativ gross sein kann. Für den Peterson-Graphen können die gefundenen Aussagen über die Anzahl von Pfaden durch Ausnützung der Symmetrien des Graphen leichter direkt gefunden werden.

Beschriftete Graphen

Bei der Beschreibung eines elektrischen Netzwerkes mit Hilfe eines ungerichteten Graphen muss jeder Kante zusätzlich ein Widerstandswert zugeordnet werden. Dies ist, was eine Beschriftung einer Kante bewerkstelligt.

Definition 8.5. *Eine Beschriftung mit Elementen der Menge L eines gerichteten oder ungerichteten Graphen $G = (V, E)$ ist eine Abbildung $l\colon E \to L$.*

8.1.2 Inzidenzmatrix

Die Adjazenzmatrix kann zusätzliche Information, die möglicherweise mit den Kanten verbunden ist, nicht mehr darstellen. Dies tritt zum Beispiel in der Informatik bei der Beschreibung endlicher Automaten auf, wo zu jeder gerichteten Kante auch noch Buchstaben gehören, für die der Übergang entlang dieser Kante möglich ist.

Die *Inzidenzmatrix* löst dieses Problem. Dazu werden zunächst die Kanten numeriert $1, \ldots, m$ numeriert. Die Matrixeinträge

$$a_{ij} = \begin{cases} 1 & \text{Knoten } i \text{ ist ein Endpunkt von Kante } j \\ 0 & \text{sonst} \end{cases}$$

stellen die Beziehung zwischen Kanten und Knoten her.

Beschriftete Graphen

Die Inzidenzmatrix kann auch einen erweiterten Graphenbegriff abbilden, in dem zwischen zwei Kanten mehrere Verbindungen möglich sind. Graphen mit beschrifteten Kanten gehören dazu.

Definition 8.6. *Ein gerichteter Graph mit beschrifteten Kanten ist eine Menge V von Knoten und eine Menge von beschrifteten Kanten der Form*

$$E\{(a, b, l) \in V^2 \times L \mid \textit{Eine Kante mit Beschriftung } l \textit{ führt von } a \textit{ nach } b\}.$$

Die Menge L enthält die möglichen Beschriftungen der Kanten.

Für einen gerichteten Graphen wird in der Inzidenzmatrix für den Anfangspunkt einer Kante -1 eingetragen und für den Endpunkt $+1$.

Inzidenzmatrix und Adjazenzmatrix

Sei $B(G)$ die Inzidenzmatrix eines Graphen G. Die Spalten von $B(G)$ sind mit den Kanten des Graphen indiziert. Die Matrix $B(G)B(G)^t$ ist eine quadratische Matrix, deren Zeilen und Spalten mit den Knoten des Graphen indiziert sind. In dieser Matrix geht die Informatione über die individuellen Kanten wieder verloren. Sie hat für $i \neq j$ die Einträge

$$\begin{aligned}(B(G)B(G)^t)_{ij} &= \sum_{k \text{ Kante}} b_{ik}b_{jk} \\ &= \text{Anzahl der Kanten, die } i \text{ mit } j \text{ verbinden} \\ &= a_{ij}.\end{aligned}$$

Die Adjazenzmatrix eines Graphen lässt sich also aus der Inzidenzmatrix berechnen.

Gradmatrix

Die Diagonale von $B(G)B(G)^t$ enthält die Werte

$$(B(G)B(G)^t)_{ii} = \sum_{k \text{ Kante}} b_{ik}^2 = \text{Anzahl Kanten, die im Knoten } i \text{ enden}$$

Der *Grad* eines Knoten eines Graphen ist die Anzahl der Kanten, die in diesem Knoten enden. Die Diagonalmatrix die aus den Graden der Knoten besteht, heisst die Gradmatrix $D(G)$ des Graphen. Es gilt daher $B(G)B(G)^t = A(G) + D(G)$.

Gerichtete Graphen

Für einen gerichteten Graphen ändert sich an der Diagonalen der Matrix $B(G)B(G)^t$ nichts. Da es in einem gerichteten Graphen nur eine einzige Kante k gibt, die zwei Knoten i und j verbinden kann, muss das zugehörige Ausserdiagonalelement

$$a_{ij} = b_{ik}b_{jk} = -1$$

sein. Für einen gerichteten Graphen sind daher alle Ausserdiagonalelemente negativ und es gilt $B(G)B(G)^t = D(G) - A(G)$.

Anwendung: Netlist

Eine natürliche Anwendung eines gerichteten und beschrifteten Graphen ist eine eletronische Schaltung. Die Knoten des Graphen sind untereinander verbundene Leiter, sie werden auch *nets* genannt. Die beschrifteten Kanten sind die elektronischen Bauteile, die solche Nets miteinander verbinden. Die Inzidenzmatrix beschreibt, welche Anschlüsse eines Bauteils mit welchen Nets verbunden werden müssen. Die Informationen in der Inzidenzmatrix werden also in einer Applikation zum Schaltungsentwurf in ganz natürlicher Weise erhoben.

8.1.3 Die Adjazenzmatrix und Laplace-Matrix

Die Beschreibung mit der Matrix (8.2) “vergisst” den “Namen” der Kante, die eine Verbindung zwischen zwei Knoten herstellt. Damit ist sie keine geeignete Grundlage, um beschriftete Graphen

einer Matrixbeschreibung zuzuführen. Eine solche muss eine Matrix verwenden, die nicht nur das Vorhandensein einer Verbindung wiedergibt, sondern ausdrückt, welche Kante welche beiden Knoten miteinander verbindet. Dies führt zur sogenannten Adjazenzmatrix.

Definition 8.7. *Ist $G = (V, E)$ ein gerichteter Graph mit $n = |G|$ Vertizes und $m = |E|$ Kanten, dann ist die zugehörige* Adjazenzmatrix *$A = A(G)$ eine $n \times m$-Matrix. In der Spalte k wird der Anfangspunkt der Kante k mit -1, der Endpunkt mit $+1$ angezeigt, die übrigen Einträge sind 0. A hat also die Matrixelemente*

$$a_{ik} = \begin{cases} -1 & i = a(k) \\ +1 & i = e(k) \\ 0 & \text{sonst} \end{cases} \tag{8.3}$$

Der wesentliche Unterschied dieser Definition von der Matrix G liegt in der Bedeutung der Einträge. Für G drückt ein nicht verschwindendes Matrixelement das Vorhandensein einer Kante aus, in A ist es die Tatsache, dass in diesem Knoten eine Kante beginnt oder endet.

Es ist natürlich möglich, aus der Adjazenzmatrix auch die Link-Matrix zu rekonstruieren. Dazu muss für jedes Paar (j, i) von Knoten festgestellt werden, ob die Adjazenzmatrix eine entsprechende Verbindung enthält, also ob der Vektor

$$k_{ji} = e_i - e_j$$

als Spaltenvektor vorkommt, wobei die e_i die n-dimensionalen Standardbasisvektoren sind.

8.2 Spektrale Graphentheorie

Die Adjazenz-Matrix, die Grad-Matrix und damit natürlich auch die Laplace-Matrix codieren alle wesentliche Information eines ungerichteten Graphen. Sie operiert auf Vektoren, die für jeden Knoten des Graphen eine Komponente haben. Dies eröffnet die Möglichkeit, den Graphen über die linearalgebraischen Eigenschaften dieser Matrizen zu studieren. Dieser Abschnitt soll diese Idee an dem ziemlich übersichtlichen Beispiel der chromatischen Zahl eines Graphen illustrieren.

8.2.1 Chromatische Zahl und Unabhängigkeitszahl

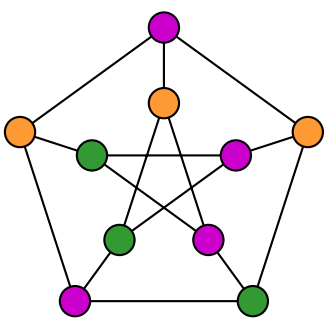
Der Grad eines Knotens ist ein mass dafür, wie stark ein Graph "vernetzt" ist. Je höher der Grad, desto mehr direkte Verbindungen zwischen Knoten gibt es. Noch etwas präziser können diese Idee die beiden mit Hilfe der chromatischen zahl und der Unabhängigkeitszahl erfasst werden.

Definition 8.8. *Die* chromatische Zahl $\operatorname{chr} G$ *eines Graphen G ist die minimale Anzahl von Farben, die Einfärben der Knoten eines Graphen nötig sind, sodass benachbarte Knoten verschiedene Farben haben.*

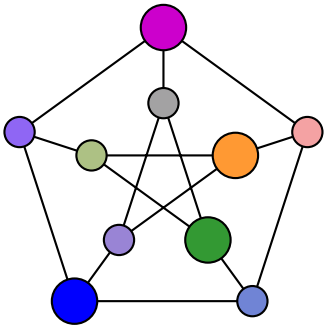
Definition 8.9. *Eine Menge von Knoten eines Graphen heisst* unabhängig*, wenn keine zwei Knoten im Graphen verbunden sind. Die* Unabhängigkeitszahl $\operatorname{ind} G$ *eines Graphen G ist die maximale Anzahl Knoten einer unabhängigen Menge.*

Zwischen der chromatischen Zahl und der Unabhängigkeitszahl eines Graphen muss es einen Zusammenhang geben. Je mehr Verbingungen es im Graphen gibt, desto grösser wird die chromatische Zahl. Gleichzeitig wird es schwieriger für Mengen von Knoten, unabhängig zu sein.

Satz 8.10. *Ist G ein Graph mit n Knoten, dann gilt* $\operatorname{chr} G \cdot \operatorname{ind} G \geq n$.



$\operatorname{chr} P = 3$



$\operatorname{ind} P = 4$

Abbildung 8.4: Chromatische Zahl und Unabhängigkeitszahl des Peterson-Graphen. Die chromatische Zahl ist 3, da der Graph sich mit drei Farben einfärben lässt (links). Die Unabhängigkeitszahl ist 4, die vier grösseren Knoten im rechten Graphen sind unabhängig. Die Farben der kleinen Knoten sind die additive Mischung der Farben der grossen Knoten, mit denen sie verbunden sind.

Beweis. Eine minimale Färbung des Graphen mit $\operatorname{chr} G$ Farben teilt die Knoten in $\operatorname{chr} G$ Mengen V_f von Knoten mit gleicher Farbe f ein. Da diese Mengen einfarbig sind, sind sie unabhängig, enthalten also höchstens so viele Knoten, wie die Unabhängigkeitszahl erlaubt, also $|V_f| \leq \operatorname{ind} G$. Da die Menge aller Knoten die Vereinigung der Mengen V_f ist, ist die Gesamtzahl der Knoten

$$V = \bigcup_{f \text{ eine Farbe}} V_f \qquad \Rightarrow \qquad n = \sum_{f \text{ eine Farbe}} |V_f|$$
$$\leq \sum_{f \text{ eine Farbe}} \operatorname{ind} G = (\text{Anzahl Farben}) \cdot \operatorname{ind} G = \operatorname{chr} G \cdot \operatorname{ind} G.$$

Damit ist $n \leq \operatorname{chr} G \cdot \operatorname{ind} G$ gezeigt. □

Beispiel. In einem vollständigen Graphen ist jeder Knoten mit jedem anderen verbunden. Jede Menge mit zwei oder mehr Knoten kann daher nicht unabhängig sein, die Unabhängigkeitszahl ist daher $\operatorname{ind} G = 1$. Andererseits ist für jeden Knoten eine eigene Farbe nötig, daher ist die chromatische Zahl $\operatorname{chr} G = n$. Die Ungleichung von Satz 8.10 ist erfüllt, sogar mit Gleichheit. Das Beispiel zeigt, dass die Ungleichung nicht ohne zusätzliche Annahmen verbessert werden kann. ○

Beispiel. Der Peterson-Graph P von Abbildung 8.4 hat chromatische Zahl $\operatorname{chr} P = 3$ und unabhängigkeitszahl $\operatorname{ind} P = 4$. Die Ungleichung von Satz 8.10 ist erfüllt, sogar als Ungleichung: $\operatorname{chr} P \cdot \operatorname{ind} P = 3 \cdot 4 = 12 > 10 = n$. ○

Nach Definition ist Unabhängigkeitszahl ein Mass für die Grösse einer unabhängigen Menge von Punkten. Der Beweis von Satz 8.10 zeigt, dass man sich die chromatische Zahl als ein Mass dafür, wieviele solche anabhängige Mengen in einem Graphen untergebracht werden können.

8.2.2 Chromatische Zahl und maximaler Grad

Wenn kein Knoten mehr als d Nachbarn hat, dann reichen $d+1$ Farben immer, um diesen Knoten und seine Nachbarn einzufärben. Das heisst aber noch nicht, dass dann auch $d+1$ Farben zur Einfärbung des ganzen Graphen reichen. Genau dies garantiert jedoch der folgende Satz.

Definition 8.11. *Der maximale Grad* $\max_{v \in V} \deg(v)$ *wird mit d bezeichnet.*

Satz 8.12. *Ist G ein Graph mit maximalem Grad d, dann gilt* $\operatorname{chr} G \leq d + 1$.

Beweis. Wir führen den Beweis mit Hilfe von vollständiger Induktion nach der Anzahl Knoten eines Graphen. Ein Graph mit nur einem Knoten hat keine Kanten, der maximale Grad ist daher 0 und $d + 1 = 1$ Farbe reicht auch tatsächlich zur Einfärbung des einen Knotens.

Wir nehmen jetzt an, die Behaupt sei für Graphen mit $n - 1$ Knoten bereits bewiesen, ein Graph G' mit $n - 1$ Knoten und maximalem Grad d' erfüllt also die Ungleichung $\operatorname{chr} G' \leq d' + 1$.

Wir wählen jetzt einen beleibigen Knoten v des Graphen G und bilden den Graphen G', der aus G entsteht, indem man den Knoten v entfernt: $G' = G \setminus \{v\}$. Der maximale Grad d' von G' kann dabei nicht grösser werden, es ist also $d' \leq d$. Da G' genau $n - 1$ Knoten hat, lässt er sich mit höchstens $d' + 1 \leq d + 1$ Farben einfärben. Es muss jetzt also nur noch eine Farbe für den Knoten v gefunden werden. Da d der maximale Grad ist, hat v höchstens d Nachbarn, die höchstens d verschiedene Farben haben können. Von den $d + 1$ zur Verfügung stehenden Farben bleibt also mindestens eine übrig, mit der man den Knoten v einfärben kann. Damit ist der Induktionsschritt gelungen und somit der Satz bewiesen. □

Das Argument im Beweis von Satz 8.12 ist für alle Begriffe anwendbar, die sich bei der Bildung eines Untergraphen auf "monotone" Art ändern. Die chromatische Zahl eines Untergraphen ist höchstens so gross wie die des ganzen Graphen. Dann kann man eine Ungleichung für grosse Graphen schrittweise aus entsprechenden Ungleichungen für die kleineren Teilgraphen gewinnen. Ziel der folgenden Abschnitte ist zu zeigen, dass sich eine Grösse mit ähnlichen Eigenschaften aus dem Eigenwertspektrum der Adjazenzmatrix ablesen lässt. Daraus ergibt sich dann eine bessere Abschätzung der chromatischen Zahl eines Graphen.

8.2.3 Maximaler Eigenwert von $A(G)$ und maximaler Grad

Die Adjazenzmatrix $A(G)$ eines Graphen G mit n Knoten enthält unter anderem auch die Information über den Grad eines Knotens. Die Summe der Elemente einer Zeile oder einer Spalte ergibt einen Vektor, der die Grade der Knoten als Komponenten enthält. Ist U ein n-dimensionaler Vektor aus lauter Einsen, dann ist ist $A(G)U$ ein Spaltenvektor bestehend aus den Zeilensummen der Matrix $A(G)$ und $U^t A(G)$ ein Zeilenvektor bestehend aus den Spaltensummen. $A(G)U$ ist also der Vektor der Grade der Knoten.

Das Skalarprodukt von $A(G)U$ mit U ist die Summe der Grade. Somit ist

$$\frac{\langle A(G)U, U\rangle}{\langle U, U\rangle} = \frac{1}{\langle U, U\rangle} \sum_{v \in V} \deg(v) = \frac{1}{n}(d_1 + \cdots + d_n) \tag{8.4}$$

der mittlere Grad, der mit $\overline{d}$ bezeichnet werden soll.

Da $A(G)$ eine symmetrische Matrix ist, ist $A(G)$ diagonalisierbar, die Eigenwerte sind also alle reell. Es ist ausserdem bekannt, dass der Eigenvektor f zum grössten Eigenwert $\alpha_{\max}$ von $A(G)$ den Bruch

$$\frac{\langle A(G)f, f\rangle}{\langle f, f\rangle}$$

für Vektoren $f \neq 0$ maximiert. Aus (8.4) folgt damit, dass

$$\overline{d} \leq \alpha_{\max} \tag{8.5}$$

ist.

In Abschnitt 9.3 des nächsten Kapitels wird die Perron-Frobenius-Theorie positiver Matrizen vorgestellt, welche einer Reihe interessanter Aussagen über den betragsgrössten Eigenwert und den zugehörigen Eigenvektor macht. Die Adjazenz-Matrix ist eine nichtnegative Matrix und $\alpha_{\max}$ ist der grösste Eigenwert, also genau die Grösse, auf die die Sätze 9.29 und anwendbar sind. Dazu muss die Matrix allerdings primitiv sein, was gleichbedeutend ist damit, dass der Graph zusammenhängend ist. Im folgenden soll dies daher jeweils angenommen werden.

Satz 8.13. *Ist G ein zusammenhänger Graph mit n Knoten und maximalem Grad d, dann gilt*

$$\frac{1}{n}\sum_{v\in V}\deg(v) = \overline{d} \leq \alpha_{max} \leq d.$$

Beweis. Wir wissen aus (8.5) bereits, dass $\overline{d} \leq \alpha_{\max}$ gilt, es bleibt also nur noch $\alpha_{\max} \leq d$ zu beweisen.

Sei f der Eigenvektor zum Eigenwert $\alpha_{\max}$. Nach Satz ist f ein positiver Vektor mit der Eigenschaft $A(G)f = \alpha_{\max} f$. Der Eigenvektor f ist eine Funktion auf den Knoten des Graphen, die v-Komponente des Vektors f für einen Vertex $v \in V$ ist $f(v)$. Die Eigenvektoreigenschaft bedeutet $(A(G)f)(v) = \alpha_{\max} f(v)$. Die Adjazenzmatrix $A(G)$ enthält in Zeile v Einsen genau für diejenigen Knoten $u \in V$, die zu v benachbart sind. Schreiben wir $u \sim v$ für die Nachbarschaftsrelation, dann ist

$$(A(G)f)(v) = \sum_{u\sim v} f(u).$$

Die Summe der Komponenten $A(G)f$ kann man erhalten durch Multiplikation von $A(G)f$ mit einem Zeilenvektor U^t aus lauter Einsen, also

$$\begin{aligned}\sum_{v\in V}\sum_{u\sim v} f(v) = U^t A(G)f = (U^t A(G))f = \begin{pmatrix} d_1 & d_2 & \dots & d_n \end{pmatrix} f \\ = \sum_{v\in V}\deg(v)f(v) \leq \sum_{v\in V} df(v) = d\sum_{v\in V} f(v).\end{aligned} \tag{8.6}$$

Andererseits ist $A(G)f = \alpha_{\max} f$, die linke Seite von (8.6) ist daher

$$\sum_{v\in V}\sum_{u\sim v} f(v) = U^t A(G)f = \alpha_{\max} U^t f = \alpha_{\max}\sum_{v\in V} f(v). \tag{8.7}$$

Die Ungleichung (8.6) und die Gleichung (8.7) ergeben zusammen die Ungleichung

$$\alpha_{\max}\sum_{v\in V} f(v) \leq d\sum_{v\in V} f(v) \qquad \Rightarrow \qquad \alpha_{\max} \leq d,$$

da die Summe der Komponenten des positiven Vektors f nicht verschwinden kann. Damit ist die Ungleichung bewiesen. □

8.2.4 $\alpha_{\max}$ eines Untergraphen

Der grösste Eigenwert $\alpha_{\max}$ ist ein potentieller Anwärter für eine bessere Abschätzung der chromatischen Zahl. Bereits früher wurde bemerkt, dass dies auch bedeutet, dass man das Verhalten des grössten Eigenwerts bei einem Übergang zu einem Untergraphen verstehen muss.

Satz 8.14. *Sei G' ein echter Untergraph von G mit Adjazenzmatrix $A(G')$ und grösstem Eigenwert $\alpha'_{max} = \varrho(A(G'))$, dann ist $\alpha'_{max} \leq \alpha_{max}$.*

Beweis. Sei f' der positive Eigenvektor zum Eigenwert $\alpha'_{\max}$ der Matrix $A(G')$. f' ist definiert auf der Menge V' der Knoten von G'. Aus f' lässt sich ein Vektor g mit den Werten

$$g(v) = \begin{cases} f'(v) & v \in V' \\ 0 & \text{sonst} \end{cases}$$

konstruieren, der auf ganz V definiert ist.

Die Vektoren f' und g haben die gleichen Komponenten, also ist auch $\langle f', f' \rangle = \langle g, g \rangle$. Die Matrixelemente von $A(G')$ und $A(G)$ auf gemeinsamen Knoten $u, v \in V'$ erfüllen $A(G')_{uv} \leq A(G)_{uv}$, da jede Kante von G' auch in G ist. Daher gilt

$$\langle A(G')f', f' \rangle \leq \langle A(G)g, g \rangle,$$

woraus sich die Ungleichung

$$\alpha'_{\max} = \frac{\langle A(G')f', f' \rangle}{\langle f', f' \rangle} = \frac{\langle A(G)g, g \rangle}{\langle g, g \rangle} \leq \alpha_{\max}$$

ergibt, da $\alpha_{\max}$ das Maximum von $\langle A(G)h, h \rangle / \langle h, h \rangle$ für alle Vektoren $h \neq 0$ ist. □

8.2.5 Chromatische Zahl und $\alpha_{\max}$: Der Satz von Wilf

Die in Satz 8.14 beschriebene Eigenschaft von $\alpha_{\max}$ beim Übergang zu einem Untergraphen ermöglich jetzt, eine besser Abschätzung für die chromatische Zahl zu finden.

Satz 8.15 (Wilf)**.** *Sie G ein zusammenhängder Graph und α_{max} der grösste Eigenwert seiner Adjazenzmatrix. Dann gilt*

$$\operatorname{chr} G \leq \alpha_{max} + 1.$$

Beweis. Wie der Satz 8.12 kann auch der Satz von Wilf mit Hilfe von vollständiger Induktion über die Anzahl n der Knoten bewiesen werden.

Ein Graph mit nur einem Knoten hat die 0-Matrix als Adjazenzmatrix, der maximale Eigenwert ist $\alpha_{\max} = 0$, und tatsächlich reicht $\alpha_{\max} + 1 = 1$ Farbe, um den einen Knoten einzufärben.

Wir nehmen jetzt an, der Satz sei für Graphen mit $n - 1$ Knoten bereits beweisen. Wir müssen dann zeigen, dass der Satz dann auch für Graphen mit n Knoten gilt.

Sei $v \in V$ ein Knoten minimalen Grades und $G' = G \setminus v$ der Untergraph, der entsteht, wenn der Knoten v entfernt wird. Da G' genau $n - 1$ Knoten hat, gilt der Satz von Wilf für G' und daher kann G' mit höchstens

$$\operatorname{chr} G' \leq 1 + \alpha'_{\max}$$

Farben eingefärbt werden. Nach Satz 8.14 ist $\alpha'_{\max} \leq \alpha_{\max}$, Also kann G' mit höchstens $\alpha_{\max} + 1$ Farben eingefärbt werden.

Da v ein Knoten minimalen Grades ist, ist sein Grad $d(v) \leq \overline{d} \leq \alpha_{\max}$. Die Nachbarn von v haben also hächstens $\alpha_{\max}$ verschiedene Farben, mit einer weiteren Farbe lässt sich also auch G einfärben. Daraus folgt $\operatorname{chr} G \leq \alpha_{\max} + 1$. □

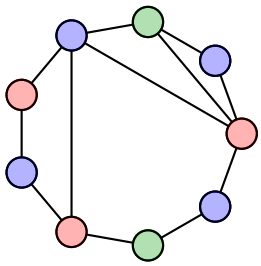


Abbildung 8.5: Beispiel für einen Graphen, für den der Satz 8.15 von Wilf die bessere Abschätzung für die chromatische Zahl eines Graphen gibt als der maximale Grad.

Beispiel. Der Graph in Abbildung 8.5 12 Kanten und 9 Knoten, daher ist $\overline{d} \leq \frac{24}{9}$. Der maximale Grad ist 4 und durch explizite Rechnung mit Hilfe zum Beispiel von Octave ergibt, dass $\alpha_{\max} \approx 2.9565$. Aus dem Satz von Wilf folgt, dass $\operatorname{chr} G \leq \alpha_{\max} + 1$, und daraus ergibt sich $\operatorname{chr} G \leq 3$. Tatsächlich ist die chromatische Zahl $\operatorname{chr} G = 3$, da der Graph mindestens ein Dreieck enthält. Der maximale Grad ist 4, somit gibt der Satz 8.12 die Schranke $\operatorname{chr} G \leq 4 + 1 = 5$ für die chromatische Zahl. Der Satz von Wilf ist also eine wesentliche Verbesserung, er liefert in diesem Fall den exakten Wert der chromatischen Zahl. ○

8.3 Wärmeleitung auf einem Graphen

Die Vektoren, auf denen die Laplace-Matrix operiert, können betrachtet werden als Funktionen, die jedem Knoten einen Wert zuordnen. Eine mögliche physikalische Interpretation davon ist die Temperaturverteilung auf dem Graphen. Die Kanten zwischen den Knoten erlauben der Wärmeenergie, von einem Knoten zu einem anderen zu fliessen. Je grösser die Temperaturdifferenz zwischen zwei Knoten ist, desto grösser ist der Wärmefluss und desto schneller ändert sich die Temperatur der beteiligten Knoten. Die zeitliche Änderung der Temperatur T_i im Knoten i ist proportional

$$\frac{dT_i}{dt} = \sum_{j \text{ Nachbar von } i} \kappa(T_j - T_i) = -\kappa\Big(d_i T_i - \sum_{j \text{ Nachbar von } i} T_j\Big)$$

Der Term auf der rechten Seite ist genau die Wirkung der Laplace-Matrix auf dem Vektor T der Temperaturen:

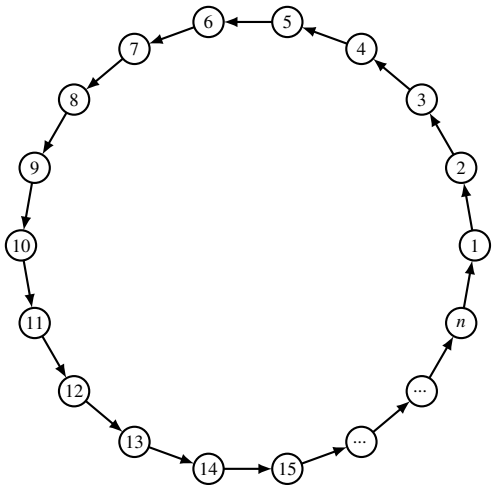
$$\frac{dT}{dt} = -\kappa L T. \tag{8.8}$$

Der Wärmefluss, der durch die Wärmeleitungsgleichung (8.8) beschrieben wird, codiert ebenfalls wesentliche Informationen über den Graphen. Je mehr Kanten es zwischen verschiedenen Teilen eines Graphen gibt, desto schneller findet der Wärmeaustausch zwischen diesen Teilen statt. Die Lösungen der Wärmeleitungsgleichung liefern also Informationen über den Graphen.

8.3.1 Eigenwerte und Eigenvektoren

Die Wärmeleitungsgleichung (8.8) ist eine lineare Differentialgleichung mit konstanten Koeffizienten, die mit der Matrixexponentialfunktion gelöst werden. Die Lösung ist

$$f(t) = e^{-\kappa L t} f(0).$$



$$L = \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & -1 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ -1 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}$$

Abbildung 8.6: Beispiel Graph zur Illustration der verschiedenen Basen auf einem Graphen.

Die Berechnung der Lösung mit der Matrixexponentialreihe ist ziemlich ineffizient, da grosse Matrizenprodukte berechnet werden müssen. Da die Matrix L symmetrisch ist, gibt es eine Basis aus orthonormierten Eigenvektoren und die Eigenwerte sind reell. Wir bezeichnen die Eigenvektoren mit $f_1, \dots, f_n$ und die zugehörigen Eigenwerte mit λ_i. Die Funktion $f_i(t) = e^{-\kappa\lambda_i t} f_i$ ist dann eine Lösung der Wärmeleitungsgleichung, denn die beiden Seiten

$$\begin{aligned} \frac{d}{dt} f_i(t) &= -\kappa\lambda_i e^{-\kappa\lambda_i t} f_i = -\kappa\lambda_i f_i(t) \\ -\kappa L f_i(t) &= -\kappa e^{-\kappa\lambda_i t} L f_i = -\kappa e^{-\kappa\lambda_i t} \lambda_i f_i = -\kappa\lambda_i f_i(t) \end{aligned}$$

von (8.8) stimmen überein.

Eine Lösung der Wärmeleitungsgleichung zu einer beliebigen Anfangstemperaturverteilung f kann durch Linearkombination aus den Lösungen $f_i(t)$ zusammengesetzt werden. Dazu ist nötig, f aus den Vektoren f_i linear zu kombinieren. Da aber die f_i orthonormiert sind, ist dies besonders einfach, die Koeffizienten sind die Skalarprodukte mit den Eigenvektoren:

$$f = \sum_{i=1}^{n} \langle f_i, f \rangle f_i.$$

Daraus kann man die allgmeine Lösungsformel

$$f(t) = \sum_{i=1}^{n} \langle f_i, f \rangle f_i(t) = \sum_{i=1}^{n} \langle f_i, f \rangle e^{-\kappa\lambda_i t} f_i \tag{8.9}$$

ableiten.

8.3.2 Beispiel: Ein zyklischer Graph

Wir illustrieren die im folgenden entwickelte Theorie an dem Beispielgraphen von Abbildung 8.6. Besonders interessant sind die folgenden Funktionen:

$$\left.\begin{aligned} s_m(k) &= \sin\frac{2\pi mk}{n} \\ c_m(k) &= \cos\frac{2\pi mk}{n} \end{aligned}\right\} \quad \Rightarrow \quad e_m(k) = e^{2\pi imk/n} = c_m(k) + is_m(k).$$

Das Skalarprodukt dieser Funktionen ist

$$\langle e_m, e_{m'}\rangle = \frac{1}{n}\sum_{k=1}^{n} \overline{e^{2\pi ikm/n}} e^{2\pi ikm'/n} = \frac{1}{n}\sum_{k=1}^{n} e^{\frac{2\pi i}{n}(m'-m)k} = \delta_{mm'}$$

Die Funktionen bilden daher eine Orthonormalbasis des Raums der Funktionen auf G. Wegen $\overline{e_m} = e_{-m}$ folgt, dass für gerade n die Funktionen

$$c_0, c_1, s_1, c_2, s_2, \ldots c_{\frac{n}{2}-1}, c_{\frac{n}{2}-1}, c_{\frac{n}{2}}$$

eine orthonormierte Basis.

Die Laplace-Matrix kann mit der folgenden Definition zu einer linearen Abbildung auf Funktionen auf dem Graphen gemacht werden. Sei $f\colon V \to \mathbb{R}$ und L die Laplace-Matrix mit Matrixelementen $l_{vv'}$ wobei $v, v' \in V$ ist. Dann definieren wir die Funktion Lf durch

$$(Lf)(v) = \sum_{v'\in V} l_{vv'} f(v').$$

8.3.3 Standardbasis und Eigenbasis

Die einfachste Basis, aus der siche Funktionen auf dem Graphen linear kombinieren lassen, ist die Standardbasis. Sie hat für jeden Knoten v des Graphen eine Basisfunktion mit den Werten

$$e_v\colon V \to \mathbb{R} : v' \mapsto \begin{cases} 1 & v = v' \\ 0 & \text{sonst.} \end{cases}$$

8.4 Wavelets auf Graphen

In Abschnitt 8.3.3 wurde gezeigt dass die Standardbasis den Zusammenhang zwischen den einzelnen Teilen des Graphen völlig ignoriert, während die Eigenbasis Wellen beschreibt, die mit vergleichbarer Amplitude sich über den ganzen Graphen entsprechen. Die Eigenbasis unterdrückt also die "Individualität" der einzelnen Knoten fast vollständig.

Wenn man einen Standardbasisvektor in einem Knoten i als Anfangstemperaturverteilung verwendet, erwartet man eine Lösung, die für kleine Zeiten t die Energie immer in der Nähe des Knotens i konzentriert hat. Weder die Standardbasis noch die Eigenbasis haben diese Eigenschaft.

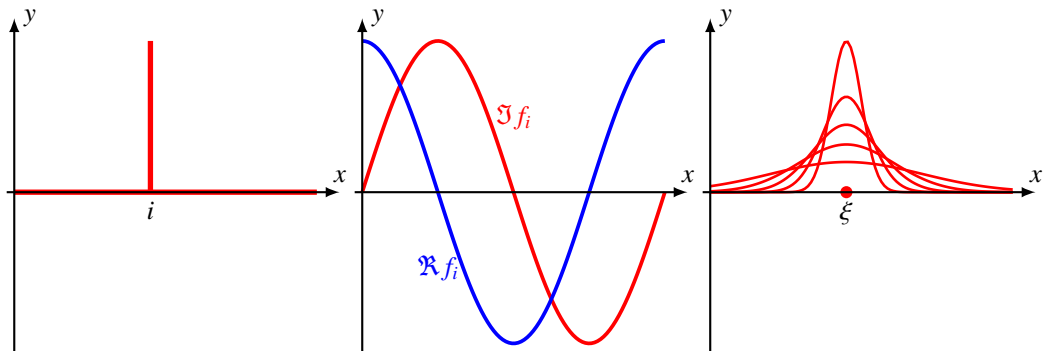


Abbildung 8.7: Vergleich der verschiedenen Funktionenfamilien, mit denen Lösungenfunktionen durch Linearkombination erzeugt werden können. In der Standarbasis (links) ist es am einfachsten, die Funktionswerte abzulesen, in der Eigenbasis (Mitte) kann die zeitliche Entwicklung besonders leicht berechnet werden. Dazuwischen liegen die Fundamentallösungen (rechts), die eine einigermassen übersichtliche Zeitentwicklung haben, die Berechnung der Temperatur an einer Stelle x zur Zeit t ist aber erst durch das Integral (8.10) gegeben.

8.4.1 Vergleich mit der Wärmeleitung auf $\mathbb{R}$

Ein ähnliches Phänomen findet man bei der Wärmeausbreitung gemäss der partiellen Differentialgleichung

$$\frac{\partial T}{\partial t} = -\kappa \frac{\partial^2 T}{\partial x^2}.$$

Die von Fourier erfundene Methode, die Fourier-Theorie, verwendet die Funktionen e^{ikx}, die Eigenvektoren der zweiten Ableitung $\partial^2/\partial x^2$ sind. Diese haben das gleiche Problem, der Betrag von e^{ikx} ist 1, die Entfernung von einem Punkt spielt überhaupt keine Rolle. Die Funktion

$$F(x,t) = \frac{1}{\sqrt{4\pi\kappa t}} e^{-x^2/4\kappa t}$$

ist eine Lösung der Wärmeleitungsgleichung mit einem Maximum an der Stelle 0. Sie heisst die Fundamentallösung der Wärmeleitungsgleichung. Durch Überlagerung von Translaten in eine Funktion

$$f(x,t) = \int_{-\infty}^{\infty} f(\xi) F(x-\xi, t)\, d\xi \tag{8.10}$$

kann man die allgemeine Lösung aus Fundamentallösungen zusammensetzen. Die Fundamentallösungen $f(x-\xi,t)$ sind für kleine Zeiten immer noch deutlich in einer Umgebung von ξ konzentriert.

8.4.2 Fundamentallösungen auf einem Graphen

Die Wärmeleitungsgleichung auf einem Graphen kann für einen Standardbasisvektor mit Hilfe der Lösungsformel (8.9) gefunden werden. Aus physikalischen Gründen ist aber offensichtlich, dass die

Wärmeenergie Fundamentallösungen $F_i(t)$ für kurze Zeiten t in der Nähe des Knoten i konzentriert ist. Dies ist aber aus der expliziten Formel

$$F_i(t) = \sum_{j=1}^{n} \langle f_j, e_i \rangle e^{-\kappa\lambda_i t} f_j = \sum_{j=1}^{n} \overline{f}_{ji} e^{-\kappa\lambda_i t}, \tag{8.11}$$

nicht unmittelbar erkennbar.

Man kann aber aus (8.11) ablesen, dass für zunehmende Zeit die hohen Frequenzen sehr schnell gedämpft werden. Die hohen Frequenzen erzeugen also den scharfen Peak für Zeiten nahe beim Knoten i, die zu kleineren λ_i beschreiben die Ausbreitung über grössere Distanzen. Die Fundamentallösung interpoliert also in einem gewissen Sinne zwischen den Extremen der Standardbasis und der Eigenbasis. Die "Interpolation" geht von der Differentialgleichung aus, sie ist nicht einfach nur ein Filter, der die verschiedenen Frequenzen auf die gleiche Art bearbeitet.

Gesucht ist eine Methode, eine Familie von Vektoren zu finden, aus der sich alle Vektoren linear kombinieren lassen, in der aber auch auf die für die Anwendung interessante Längenskala angepasste Funktionen gefunden werden können.

8.4.3 Wavelets auf einem Graphen

Die Fourier-Theorie analysiert Funktionen nach Frequenzen, wobei die zeitliche Position von interessanten Stellen der Funktion in der Phase der einzelnen Komponenten verschwindet. Die Lokalisierung geht also für viele praktische Zwecke verloren. Umgekehrt haben einzelne Ereignisse wie eine δ-Funktion keine charakteristische Frequenz, sie sind daher im Frequenzraum überhaupt nicht lokalisierbar. Die Darstellung im Frequenzraum und in der Zeit sind also extreme Darstellungen, entweder Frequenzlokalisierung oder zeitliche Lokalisierung ermöglichen, sich aber gegenseitig ausschliessen.

Dilatation

Eine Wavelet-Basis für die L^2-Funktionen auf $\mathbb{R}$ erlaubt eine Funktion auf $\mathbb{R}$ auf eine Art zu analysieren, die eine ungenaue zeitliche Lokalisierung bei entsprechend ungenauer Frequenzbestimmung ermöglicht. Ausserdem entstehen die Wavelet-Funktionen aus einer einzigen Funktion $\psi(t)$ durch Translation um b und Dilatation mit dem Faktor a:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) = T_b D_a \psi(t)$$

in der Notation von [4]. Auf einem Graphen ist so eine Konstruktion grundsätzlich nicht möglich, da es darauf weder eine Translations- noch eine Streckungsoperation gibt.

In der Theorie der diskreten Wavelet-Transformation ist es üblich, sich auf Zweierpotenzen als Streckungsfaktoren zu beschränken. Ein Gitter wird dadurch auf sich selbst abgebildet, aber auf einem Graphen gibt es keine Rechtfertigung für diese spezielle Wahl von Streckungsfaktoren mehr. Es stellt sich daher die Frage, ob man für eine beliebige Menge $T = \{t_1, t_2, \dots\}$ von Streckungsfaktoren eine Familie von Funktionen χ_j zu finden derart, dass man sich die χ_j in einem gewissen Sinn als aus χ_0 durch Dilatation entstanden vorstellen kann.

Die Dilatation kann natürlich nicht von einer echten Dilatation im Ortsraum herstammen, aber man kann wenigstens versuchen, die Dilatation im Frequenzraum nachzubilden. Für Funktionen in

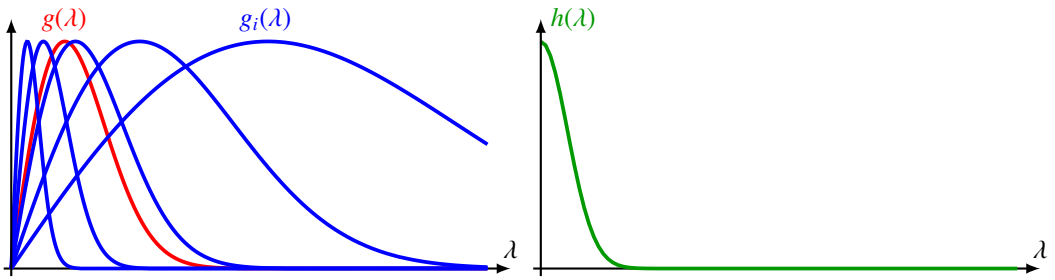


Abbildung 8.8: Lokalisierungsfunktion $g(\lambda)$ für die Dilatation (links). Die Dilatierten Funktionen $g_i = \tilde{D}_{1/a_i} g$ lokalisieren die Frequenzen jeweils um die Frequenzen a_i im Frequenzraum. Der Konstante Vektor ist vollständig delokalisiert, die Funktion h in der rechten Abbildung entfernt die hohen Frequenzen und liefert Funktionen, die in der Umgebung eines Knotens wie die Konstante Funktion aussehen.

$L^2(\mathbb{R})$ entspricht die Dilatation mit dem Faktor a im Ortsraum der Dilatation mit dem Faktor $1/a$ im Frequenzraum:

$$\widehat{D_a f}(\omega) = D_{1/a} \hat{f}(\omega).$$

[4, Satz 3.14]. Es bleibt aber das Problem, dass sich auch die Skalierung im Frequenzraum nicht durchführen lässt, da auch das Frequenzspektrum des Graphen nur eine Menge von reellen Zahlen ohne innere algebraische Struktur ist.

Mutterwavelets

Das Mutter-Wavelet einer Wavelet-Analyse zeichnet definiert, in welchem Mass sich Funktionen im Orts- und im Frequenzraum lokalisieren lassen. Die Standardbasis der Funktionen auf einem Graphen repräsentieren die perfekte örtliche Lokalisierung, Eigenbasis der Laplace-Matrix repräsentiert die perfekte Lokalisierung im Frequenzraum. Sei $g(\lambda) \geq 0$ eine Funktion im Frequenzraum, die für $\lambda \to 0$ und $\lambda \to \infty$ rasch abfällt mit einem Maximum irgendwo dazwischen (Abbildung 8.8). Sie kann als eine Lokalisierungsfunktion im Frequenzraum betrachtet werden.

Die Matrix $g(I)$ bildet entfernt aus einer Funktion die ganz hohen und die ganz tiefen Frequenz, lokalisiert also die Funktionen im Frequenzraum. Die Standardbasisvektoren werden dabei zu Funktionen, die nicht mehr nur auf einem Knoten von 0 verschieden sind, aber immer noch einigermassen auf dem Graphen lokalisiert sind. Natürlich sind vor allem die Werte auf den Eigenwerten $\lambda_0 < \lambda_1 \leq \cdots \leq \lambda_n$ der Laplace-Matrix von Interesse.

Die Matrix $g(I)$ kann mit Hilfe der Spektraltheorie berechnet werden, was im vorliegenden Fall naheliegend ist, weil ja die Eigenvektoren von der Laplace-Matrix bereits bekannt sind. Die Matrix χ^t bildet die Standardbasisvektoren in die Eigenbasis-Vektoren ab, also in eine Zerlegung im Frequenzraum ab, χ vermittelt die Umkehrabbildung. Mit der Spektraltheorie findet man für die Abbildung $g(I)$ die Matrix

$$g(I) = \chi \begin{pmatrix} g(\lambda_0) & 0 & \dots & 0 \\ 0 & g(\lambda_1) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g(\lambda_n) \end{pmatrix} \chi^t. \tag{8.12}$$

Dilatation

Die Dilatation um a im Ortsraum wird zu einer Dilatation um $1/a$ im Frequenzraum. Statt also nach einer echten Dilatation der Spaltenvektoren in $g(I)$ zu suchen, kann man sich darauf verlegen, Funktionen zu finden, deren Spektrum von einer Funktionen lokalisiert worden ist, die eine Dilatation von g ist. Man wählt daher eine ansteigende Folge $A = (a_1, \dots)$ von Streckungsfaktoren und betrachtet anstelle von g die dilatierten Funktionen $g_i = \tilde{D}_{1/a_i} g$. Die zugehörigen Wavelet-Funktionen auf dem Graphen können wieder mit der Formel (8.12) berechnet werden, man erhält

$$\tilde{D}_{1/a_i} g(I) = g_i(I) = \chi \begin{pmatrix} g(a_i\lambda_0) & 0 & \dots & 0 \\ 0 & g(a_i\lambda_1) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g(a_i\lambda_n) \end{pmatrix} \chi^t. \tag{8.13}$$

Die Spalten von $g_i(I)$ bilden wieder eine Menge von Funktionen, die eine gemäss g_i lokalisiertes Spektrum haben.

Vater-Wavelet

Wegen $g(0) = 0$ wird die konstante Funktion, die Eigenvektor zum Eigenwert $\lambda_0 = 0$ ist, von den Abbildungen $g_i(I)$ auf 0 abgebildet. Andererseits ist diese Funktion nicht lokalisiert, man möchte Sie also für die Analyse nicht unbedingt verwenden. Man wählt daher eine Funktion $h(\lambda)$ mit $h(0) = 1$ so, dass für $\lambda \to \infty$ der Wert $h(\lambda)$ genügend rasch gegen 0 geht. Die Matrix $h(I)$ bildet daher den konstanten Vektor nicht auf 0 ab, sondern lokalisiert ihn im Ortsraum. Wir erhalten daher in den Spalten von $h(I)$ Vektoren, die um die einzelnen Knoten lokalisiert sind.

Rekonstruktion

Die Operatoren $h(I)$ und $g_i(I)$ erzeugen analysieren eine Funktion nach den verschiedenen Frequenzen mit den Skalierungsfaktoren a_i, aber die Rekonstruktion ist noch nicht klar. Diese wäre einfacher, wenn die Operatoren zusammen die identische Abbildung ergäben, wenn also

$$h(I) + \sum_i g_i(I) = I$$

gelten würde. Nach der Spektraltheorie gilt das nur, wenn für alle Eigenwerte λ_k, $k = 1, \dots, n$

$$h(\lambda_k) + \sum_i g(a_i\lambda_k) = 1$$

gilt. Für beleibige Funktionen g und h kann man nicht davon ausgehen, aber man kann erwarten. Man muss daher zusätzlich verlangen, dass

$$h(\lambda_k) + \sum_i g(a_i\lambda_k) > 0$$

ist für alle Eigenwerte λ_k.

Frame

Die Menge von Vektoren, die in der vorangegangenen Konstruktion gefunden wurden, ist zu gross, um eine Basis zu sein. Vektoren lassen sich darin auf verschiedene Art darstellen. Wir verlangen aber auch keine eindeutige Darstellung, nur eine Darstellung, in der wir die "dominierenden" Komponenten in jeder Frequenzskala identifizieren können.

Definition 8.16. *Ein Frame des Vektorraumes $\mathbb{R}^n$ ist eine Menge $F = \{e_k \mid k = 1, \dots, N\}$ von Vektoren mit der Eigenschaft*

$$A\|v\|^2 \leq \sum_{k=1}^{N} |\langle v, e_k \rangle|^2 \leq B\|v\|^2 \tag{8.14}$$

Die Zahlen A und B heissen die Frame-Konstanten *des Frames.*

Die oben gefundenen Vektoren, die Spalten Vektoren von $h(I)$ und $g_i(I)$ bilden daher ein Frame. Die Frame-Konstanten kann man unmittelbar ausrechnen. Der mittlere Term von (8.14) ist

$$\|h(I)v\|^2 + \sum_i \|g_i(I)v\|^2,$$

die durch die Funktion

$$f(\lambda) = h(\lambda)^2 + \sum_i g_i(\lambda)^2$$

abgeschätzt werden kann. Die Frame-Konstanten sind daher

$$A = \min_k f(\lambda_k) \qquad \text{und} \qquad B = \max_k f(\lambda_k).$$

Die Konstruktion hat also ein Frame für die Funktionen auf dem Graphen etabliert, die viele Eigenschaften einer Multiskalenanalyse in diese wesentlich weniger symmetrische Situation rettet.

Kapitel 9

Wahrscheinlichkeitsmatrizen

Matrizen beschreiben lineare Abbildungen, also einen Prozess, der jedem Vektor einen neuen Vektor zuordnet. Es ist daher nicht abwegig zu erwarten, dass sich die Zeitentwicklung eines vom Zufall beeinflussten Systems, welches sich in mehreren verschiedenen Zuständen befinden kann, ebenfalls mit Hilfe von Matrizen beschreiben lässt. Eine solche Beschreiben ermöglicht leicht Verteilungen, Erwartungswerte und stationäre Zustände zu ermitteln.

Im Abschnitt 9.1 wird an Hand der Google Matrix bezeigt, wie ein anschauliches Beispiel in natürlicher Weise auf eine Matrix führt. Abschnitt 9.2 stellt dann die abstrakte mathematische Theorie der Markov-Ketten dar und behandelt einige wichtige Eigenschaften von Wahrscheinlichkeitsmatrizen. Es stellt sich heraus, dass thermodynamische Quantensysteme sehr gut mit solchen Matrizen beschrieben werden können, zum Beispiel kann man einfache Formen von Laser auf diese Art behandeln. Aus einem solchen System hat Parrondo ein System abgeleitet, welches ziemlich unerwartetes Verhalten an den Tag gelegt hat, welches mit Hilfe von Matrizen leicht zu analysieren ist. Dies wird in Abschnitt 9.4 dargestellt.

9.1 Google-Matrix

Das Internet besteht aus einer grossen Zahl von Websites, etwa 400 Millionen aktiven Websites, jede besteht aus vielen einzelnen Seiten. Es ist daher angemessen von $N \approx 10^9$ verschiedenen Seiten auszugehen. Eine natürliche Sprache umfasst dagegen nur einige 100000 bis Millionen von Wörtern. Ein durchschnittlicher Sprecher englischer Muttersprache verwendet nur etwa 50000 Wörter. Die Zahl der Wörter, die auf den N Seiten vorkommen können, ist also viel kleiner als die Zahl der zur Verfügung stehenden Wörter. Ein einzelnes Wort wird daher notwendigerweise auf einer grossen Zahl von Seiten vorkommen. Eine Suche nach einem bestimmten Wort wird also in der überwiegenden Zahl der Fälle derart viele Treffer zurückgeben, dass das Suchresultat nur dann nützlich sein kann, wenn eine zusätzliche Informationsquelle ermöglicht, die Treffer in eine sinnvolle Ordnung zu bringem.

Genau dieses Problem stellte sich den vielen traditionellen Suchmaschienen in der ersten grossen Boomphase des Internets. Traditionelle Informatione-Retrieval-Systeme operieren auf einem relativ kleinen Dokumentbestand und gehen davon aus, dass bereits wenige, spezifische Wörter nur in einem kleinen Teil des Dokumentbestandes vorkommen und damit eine übersichtliche Treffermenge ergeben. Die Einengung der Treffermenge dank der Suche nach spezifischer Menge bedeutet aber

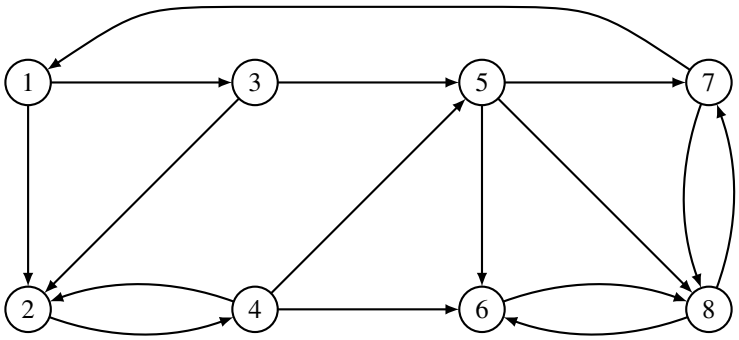


Abbildung 9.1: Modell-Internet als Beispiel für die Link-Matrix und die Google-Matrix.

auch, dass nach Synonymen oder alternative Formen eines Wortes separat gesucht werden muss, was die Übersichtlichkeit wieder zerstört.

9.1.1 Ein Modell für Webseitenbesucher

Das kombinierte Vorkommen von Wörtern oder Begriffen alleine kann also nicht ausreichen, um die Seiten zum Beispiel einem Fachgebiet zuzuordnen. Dazu muss eine externe Informationsquelle angezapft werden. Bei traditionellen Dokumenten liefert der Kontext, in dem ein Dokument erfasst wurde, solche ergänzenden Informationen. Eine Publikation in einem Fachjournal ordnet einen Text einem Fachgebiet zu. Im World-Wide-Web liefert die Link-Struktur diesen Kontext. Dokumente zu ähnlichen Themen werden bevorzugt untereinander verlinkt sein.

Gesucht ist jetzt also ein Modell, welches objektiv die Linkstruktur bewertet und daraus eine Rangordnung der passenden Wörter ableitet. Die Linkstruktur kann natürlich als gerichteter Graph betrachtet und mit Hilfe der Matrix (8.2) beschrieben werden. Dies trägt jedoch der Anzahl der Wahlmöglichkeiten nicht Rechnung. Eine Website mit nur einem Link auf die Seite j hat mehr Gewicht als eine Seite mit vielen Links, unter denen der Link auf die Seite j einer von vielen ist. Im Beispiel-Inter der Abbildung 9.1 signalisiert die Seite t mit nur einem Link auf die Seite 8 viel deutlicher, dass 8 eine wichtige Seite ist, also die die Seite 5 tut, die auch noch zwei andere Links enthält. Wir können diesen Unterschied berücksichtigen, indem wir zu einem Wahrscheinlichkeitsmodell übergehen, was wir im folgenden Abschnitt tun werden.

9.1.2 Wahrscheinlichkeitsinterpretation

Ein Internetbesucher kann eine grosse Zahl von Seiten besuchen. In diesem Abschnitt soll ein Modell entwickelt werden, welches die Wahrscheinlichkeit zu ermitteln gestattet, dass der Besucher auf einer bestimmten Seite landet.

Ereignisse und Wahrscheinlichkeiten

Wir bezeichnen mit S_i das Ereignis, dass sich der Besucher auf der Seite mit der Nummer i befindet, wobei $i = 1, \ldots, N$. Gesucht ist die Wahrscheinlichkeit $P(S_i)$. Ohne weitere Information müssten wir davon ausgehen, dass jede Seite etwa gleich wahrscheinlich ist, dass also $P(S_i) = 1/N$.

Wir wissen jedoch mehr. Wir wissen, dass der Besucher die verschiedenen Seiten zu einem guten Teil dadurch findet, dass er Links folgt. Die Wahrscheinlichkeit $P(S_i)$ verändert sich also,

wenn die Zahl der Links ansteigt, die auf die Seite i verweisen. Zur Beschreibung dieses Phänomens brauchen wir die zusätzliche Ereignisse S'_i, die mit Wahrscheinlichkeit $P(S'_i)$ eintreten, wenn sich der Besucher nach Navigation entlang eines Links auf der Seite i befindet.

Wir nehmen jetzt zusätzlich an, dass eine grosse Zahl von Besuchern über lange Zeit ungefähr nach den gleichen Dingen suchen und sich daher auf die gleiche Weise auf den verschiedenen Seiten verteilen und dass insbesondere die Verteilung stationär ist, dass also $P(S_i) = P(S'_i)$ gilt. Suchmaschinen wie Google gehen davon aus, dass alle Besucher ungefähr die gleichen Suchprioritäten haben, so dass es sich lohnt, die Suchresultate nach der Wahrscheinlichkeit $P(S_i)$ zu ordnen und dem Suchenden die wahrscheinlichsten Dokumente als erste zu zeigen.

Bedingte Wahrscheinlichkeit

Um einen Zusammenhang zwischen $P(S_i)$ und $P(S'_j)$ herzustellen, muss die Navigation entlang der Links modelliert werden. Die naheliegende Wahrscheinlichkeitsinterpretation ist die bedingte Wahrscheinlichkeit $P(S'_j|S_i)$ dass der Besucher auf der Seite j landet, nachdem er auf der Seite i die Linknavigation verwendet hat. Wenn es keinen Link zwischen den Seiten i und j gibt, dann ist diese Navigation natürlich nicht möglich und es folgt $P(S'_j|S_i) = 0$. Falls es einen Link gibt, ist $P(S'_j|S_i) \geq 0$.

A priori wissen wir nicht, wie wahrscheinlich es ist, dass der Besucher dem Link auf die Seite j folgt, normalerweise werden nicht alle Links mit gleicher Wahrscheinlichkeit verwendet. Wir nehmen daher zusätzlich an, dass alle Links gleich wahrscheinlich sind. Die Seite i enthält n_i Links, also ist die Wahrscheinlichkeit, auf einer von i aus verlinkten Seite j zu landen $P(S'_j|S_i) = 1/n_i$.

Totale Wahrscheinlichkeit

Der Satz von der totalen Wahrscheinlichkeit ermöglicht, einen Zusammenhang zwischen $P(S'_j)$ und $P(S_i)$ herzustellen. Es gilt

$$P(S'_j) = P(S'j|S_1)P(S_1) + P(S'j|S_2)P(S_2) + \cdots + P(S'j|S_N)P(S_N). \tag{9.1}$$

Dies kann in Matrix- und Vektorform übersichtlicher geschrieben werden. Dazu fassen wir die Wahrscheinlichkeiten $p'_j = P(S'_j)$ und $p_i = P(S_i)$ also Vektoren

$$p = \begin{pmatrix} P(S_1) \\ P(S_2) \\ \vdots \\ P(S_N) \end{pmatrix} \qquad \text{und} \qquad p' = \begin{pmatrix} P(S'_1) \\ P(S'_2) \\ \vdots \\ P(S'_N) \end{pmatrix}$$

zusammen. Die bedingten Wahrscheinlichkeiten $h_{ji} = P(S'_j|S_i)$ sind mit zwei Indizes beschrieben, sie bilden daher in natürlicher Weise eine Matrix

$$H = \begin{pmatrix} P(S'_1|S_1) & P(S'_1|S_2) & \dots & P(S'_1|S_N) \\ P(S'_2|S_1) & P(S'_2|S_2) & \dots & P(S'_2|S_N) \\ \vdots & \vdots & \ddots & \vdots \\ P(S'_N|S_1) & P(S'_N|S_2) & \dots & P(S'_N|S_N) \end{pmatrix}.$$

Die Formel (9.1) wird dann zur Formel für das Produkt Matrix mal Vektor:

$$(Hp)_j = \sum_{i=1}^{N} h_{ji}p_i = \sum_{i=1}^{N} P(S'_j|S_i)P(S_i) = p'_j \qquad \Rightarrow \qquad Hp = p'.$$

Die Matrix H modelliert also die Wahrscheinlichkeit der Navigation entlang eines Links.

Beispiel. Für das Beispiel-Internet von Abbildung 9.1 ist die zugehörige Matrix

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{2} & 0 \end{pmatrix}. \tag{9.2}$$

○

9.1.3 "Freier Wille"

Das Modell in Abschnitt (9.1.2) beschriebene Modell geht unter anderem davon aus, dass der Benutzer ausschliesslich die Navigation entlang der Links verwendet. Natürlich gibt es viele weitere Wege, auf denen ein Besucher auf einer bestimmten Seite landen kann. Er kann zum Beispiel einen Link auf eine Seite per Email zugesandt erhalten haben. Ein solcher Link ist nicht enthalten in einer öffentlich zugänglichen Seite des Internets und wird daher auch von der Matrix H nicht erfasst. Eine weitere wichtige Quelle von Links sind dynamisch erzeugte Links wie zum Beispiel die Suchresultate einer Suchmaschine. Hier entsteht die Möglichkeit, dass die erfolgreiche Suchmaschine, die ihre Suchresultate unter Zuhilfenahme der Matrix H sortiert, ihr eigenes Modell, auf dem ihr Erfolg basiert, torpediert.

Erweiterung der Link-Matrix

Wir bezeichnen das Ereignis, dass der Benutzer nicht die Link-Navigation verwendet mit F für "freier Wille", obwohl es so etwas natürlich nicht gibt. Die Wahrscheinlichkeit, auf der Seite S'_j zu landen, setzt sich jetzt aus den zwei Fällen F und $\overline{F}$ zusammen, für die erneut der Satz von der totalen Wahrscheinlichkeit den Zusammenhang

$$P(S'_j) = P(S'_j|\overline{F})P(\overline{F}) + P(S'_j|F)P(F)$$

Die Wahrscheinlichkeit $\alpha = P(F)$, mit der der Benutzer den "freiene Willen" bemüht, kann experimentell durch Studien ermittelt werden, die das Benutzerverhalten beobachten.

Die Wahrscheinlichkeit $P(S'_j|\overline{F})$ entsteht dadurch, dass der Benutzer der Linknavigation folgt, sie entspricht also der früher berechnenten Wahrscheinlichkeit

$$P(S'_j|\overline{F}) = \sum_{i=1}^{N} P(S'_j|S_i)P(S_i).$$

oder in Vektorform

$$(P(S'_j|\overline{F}))_{j=1,\dots,n} = Hp.$$

Über die spontane Besuchswahrscheinlichkeit $P(S'_j|F)$ wissen wir nichts. Eine erste Annahme könnte sein, dass jede Seite gleich wahrscheinlich ist, dass also $P(S'_j|F) = 1/N$. Alternativ könnte

man auch eine Wahrscheinlichkeitsverteilung $q_j = P(S'_j|F)$ experimentell zu ermitteln versuchen. Unter der Annahme, dass alle Seitenbesuche im Falle F auf Grund eines Sucheresultats einer Suchmaschine erfolgen, könnte die Suchmaschine den Vektor q aus ihrer eigenen Suchstatistik ermitteln.

Das erweiterte Modell kann also durch

$$P(S'_j) = \sum_{i=1}^{N} \alpha P(S'_j|S_i)P(S_i) + (1-\alpha)q_j \qquad \Rightarrow \qquad p' = \alpha H p + (1-\alpha)q \tag{9.3}$$

beschrieben werden.

Die Google-Matrix

Die Formel (9.1.3) erlaubt, die Wahrscheinlichkeit p' aus p und q zu berechnen. Für die Ermittlung der der stationären Verteilung war jedoch die Form $p = Hp$ besonders nützlich, weil sie das Problem in ein Eigenwertproblem mit einem bekanntem Eigenwert verwandelt. Wir streben daher an, die Formel (9.1.3) ebenfalls in die Form $p = Gp$ mit einer neuen Matrix G zu bringen.

Die Matrixform von zeigt, dass sich die gesuchte Matrix G zusammensetzt aus dem Summanden αH und einem weiteren Summanden A mit der Eigenschaft, dass $Ap = q$ für jeden beliebigen Wahrscheinlichkeitsvektor p. Da sich die Wahrscheinlichkeiten im Vektor p zu 1 summieren, gilt

$$\underbrace{\begin{pmatrix}1 & 1 & \dots & 1\end{pmatrix}}_{=U^t}\begin{pmatrix}P(S_1)\\P(S_2)\\\vdots\\P(S_N)\end{pmatrix} = P(S_1)+P(S_2)+\cdots+P(S_N) = 1.$$

Man erhält also die Wirkung der gewünschte Matrix A, indem man p erst mit dem Zeilenvektor U^t und das Resultat mit q multipliziert. Es gilt daher

$$Ap = qU^t p \qquad \Rightarrow \qquad A = qU^t.$$

Ausmultipliziert ist dies die Matrix

$$A = \begin{pmatrix}q_1 & q_1 & \dots & q_1\\q_2 & q_2 & \dots & q_2\\\vdots & \vdots & \ddots & \vdots\\q_N & q_N & \dots & q_N\end{pmatrix}.$$

Im Fall $q = \frac{1}{N}U$ kann dies zu

$$A = \frac{1}{N}UU^t = \frac{1}{N}\begin{pmatrix}1 & 1 & \dots & 1\\1 & 1 & \dots & 1\\\vdots & \vdots & \ddots & \vdots\\1 & 1 & \dots & 1\end{pmatrix}$$

vereinfacht werden.

Definition 9.1 (Google-Matrix)**.** *Die Matrix*

$$G = \alpha H + \frac{1-\alpha}{N}UU^t \qquad \textit{oder} \qquad G = \alpha H + (1-\alpha)qU^t \tag{9.4}$$

heisst die Google-Matrix.

Die Google-Matrix wurde von Sergei Brin und Larry Page in dem Artikel [2] als Basis der Suchmaschine Google beschrieben. Sie war die Basis für den Erfolg von Google und wird dem Prinzip nach auch heute noch zur Rangierung der Suchresultate verwendet. Dazu muss natürlich die Gleichung $p = Gp$ gelöst werden, was weiter unten in Abschnitt 9.1.4 diskutiert wird.

Natürlich ist die heutzutage verwendete Matrix mit Sicherheit komplizierter. In der vorgestellten Form unterstützt sie zum Beispiel auch das folgende Geschäftsmodell, welches in der Anfangszeit von Google eine Zeitlang erfolgreich war. Ein Anbieter betreibt zu diesem Zweck eine grosse Zahl von Websites, deren Seiten im Wesentlichen aus Suchbegriffen und Links untereinander und auf die Website des Kunden verweisen. Dadurch entsteht für die Google-Matrix der "Eindruck", dass sehr viele Websites gibt, die die Kundenwebsite als relevant für die Suchbegriffe ansehen. Die Kundenwebsite wird daher in den Suchresultaten weiter oben gezeigt. Das Problem rührt natürlich daher, dass alle Links als gleichermassen aussagekräftig betrachtet werden.

Die aktuell verwendete Variante der Google-Matrix ist natürlich ein Betriebsgeheimnis der Firma Google.

9.1.4 Wahrscheinlichkeitsverteilung

Die Google-Matrix G selbst interessiert weniger als die Wahrscheinlichkeitsverteilung p. Ziel dieses Abschnittes, ist den Vektor p zu berechnen.

Stationäre Verteilung

Die Einträge $P(S_i)$ des Vektors p geben die Wahrscheinlichkeit an, mit der sich ein Benutzer auf der Seite i befindet. Wir interpretieren diese Wahrscheinlichkeit auch als ein Mass für die Relevanz einer Seite.

Wir nehmen an, dass sich diese Wahscheinlichkeit nur langsam ändert. Diese Annahme trifft nicht zu für neue Nachrichten, die durchaus eine hohe Relevanz haben, für es aber noch nicht viele Links geben kann, die die Relevanz in der Google-Matrix erkennbar machen. Die Annahme bedeutet, dass sich die Verteilung p sehr viel langsamer ändert als der Navigationsprozess entlang der Links erfolgt. In erster Näherung ist es daher zulässig, nach einem Vektor p zu suchen, der sich unter Navigation nicht ändert, also nach einer *stationären* Lösung.

Für eine stationäre Wahrscheinlichkeitsverteilung gilt $p' = p$. Der Vektor p erfüllt daher die Gleichung

$$Gp = p. \tag{9.5}$$

p ist also ein Eigenvektor der Matrix G zum Eigenwert 1.

Für ein sehr kleines Netzwerk wie im oben dargestellten Beispiel ist es einfach, mit verbreiteten numerischen Algorithmen alle Eigenwerte und Eigenvektoren zu finden. Benötigt wird allerdings nur der Eigenvektor zum Eigenwert 1.

Beispiel. Ein Eigenvektor zum Eigenwert 1 der Matrix G, die aus der Matrix H von (9.2) und dem

Vektor $q = \frac{1}{8}u$ und $\alpha = 0.9$ gebildet wurde, ist

$$p_0 = \begin{pmatrix} 0.20100 \\ 0.25440 \\ 0.12163 \\ 0.26014 \\ 0.16394 \\ 0.45543 \\ 0.37739 \\ 0.66007 \end{pmatrix} \qquad \Rightarrow \qquad p = \frac{1}{\|p_0\|_1} p_0 = \begin{pmatrix} 0.080595 \\ 0.102004 \\ 0.048769 \\ 0.104305 \\ 0.065735 \\ 0.182609 \\ 0.151320 \\ 0.264664 \end{pmatrix}.$$

Der Vektor p_0 ist ein Einheitsvektor in der euklidischen Norm. Er kann daher nicht eine Wahrscheinlichkeitsverteilung sein, da sich die Elemente nicht zu 1 summieren. Die L^1-Norm $\|\cdot\|_1$ eines Vektors ist die Summe der Beträge aller Elemente eines Vektors. Indem man p_0 durch die Summe aller Einträge von p_0 teilt, erhält man die Wahrscheinlichkeitsverteilung p. ○

Potenzverfahren

Die üblichen Algorithmen wie der Francis-Algorithmus zur Bestimmung von Eigenwerten und Eigenvektoren ist für grosse Matrizen nicht praktikabel. Da aber 1 der betragsgrösste Eigenwert ist, kann sehr oft ein zugehöriger Eigenvektor mit der nachfolgend beschriebenen *Potenzmethode* gefunden werden.

Sei A eine $n \times n$-Matrix, der Einfachheit halber nehmen wir an, dass die Eigenwerte $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ absteigend geordnet sind, und dass $v_1, \ldots, v_n$ zugehörige linear unabhängige Eigenvektoren sind. Ein beliebiger Vektor v lässt sich in der Eigenbasis von A als

$$v = a_1 v_1 + \cdots + a_n v_n$$

ausdrücken. Wendet man darauf die Matrix A k-mal an, erhält man

$$A^k v = a_1 \lambda_1^k v_1 + a_2 \lambda_2^k v_2 + \cdots + a_n \lambda_2^k v_n.$$

Da λ_1 der betragsmässig grösste Eigenwert ist, wird der Vektor $A^k v$ ungefähr mit der k-ten Potenz anwachsen. Indem man durch λ_1^k teilt, erhält man

$$\frac{1}{\lambda_1^k} A^k v = a_1 v_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n.$$

Da alle Brüche Betrag < 1 haben, konvergiert die rechte Seite für $k \to \infty$ gegeben den ersten Summanden. Durch wiederholte Anwendung von A/λ_1 auf einen (fast) belieibigen Startvektor v erhält man also eine Folge von Vektoren, die gegen eine Eigenvektor zum Eigenwert λ_1 konvergiert.

Numerische Ungenauigkeiten können bewirken, dass die Iteration mit der Matrix A/λ_1 trotzdem nicht konvergiert. Man kann dies komponsieren, indem man nach jeder Iteration normiert. Da der gesuchte Eigenvektor eine Wahrscheinlichkeitsverteilung sein muss, muss die L^1-Norm 1 sein. Statt mit der euklidischen L^2-Norm zu normieren, normiert man daher besser mit der L^1-Norm. Damit ergibt sich das folgende Verfahren zur Bestimmung der Pagerank-Verteilung p für die Google-Matrix.

Satz 9.2. *Für die Google-Matrix p konvergiert die Folge*

$$p^{(0)} = u, \qquad p^{(k+1)} = \frac{G^{(k)}}{\|G^{(k)}\|_1}$$

gegen die stationäre Verteilung p mit $Gp = p$.

9.2 Diskrete Markov-Ketten und Wahrscheinlichkeitsmatrizen

Die einführend im Abschnitt 9.1 vorgestellte Google-Matrix ist nur ein Beispiel für ein Modell eines stochastischen Prozesses, der mit Hilfe von Matrizen modelliert werden kann. In diesem Abschnitt soll diese Art von Prozessen etwas formalisiert werden.

9.2.1 Markov-Eigenschaft

Ein stochastischer Prozess ist eine Familie von Zustandsvariablen X_t mit Werten in einer Menge $\mathcal{S}$ von Zuständen. Der Parameter t wird üblicherweise als die Zeit interpretiert, er kann beliebige reelle Werte oder diskrete Werte annahmen, im letzten Fall spricht man von einem Prozess mit diskreter Zeit.

Das Ereignis $\{X_t = x\}$ wird gelesen als "zur Zeit t befindet sich der Prozess im Zustand x". Mit $P(X_t = x)$ wir die Wahrscheinlichkeit bezeichnet, dass sich der Prozess zur Zeit t im Zustand x befindet. Die Funktion $t \mapsto X_t$ beschreiben also den zeitlichen Ablauf der vom Prozess durchlaufenen Zustände. Dies ermöglicht, Fragen nach dem Einfluss früherer Zustände, also des Eintretens eines Ereignisses $\{X_{t_0} = x\}$ auf das Eintreten eines Zustands $s \in \mathcal{S}$ zu einem späteren Zeitpunkt $t_1 > t_0$ zu studieren. Das Ereignis $\{X_t = x\}$ kann man sich als abhängig von der Vorgeschichte vorstellen. Die Vorgeschichte besteht dabei aus dem Eintreten gewisser Ereignisse

$$\{X_0 = x_0\}, \{X_1 = x_1\}, \{X_2 = x_2\}, \ldots, \{X_n = x_n\}$$

zu früheren Zeiten $t_0 < t_1 < \cdots < t_n < t$. Die bedingte Wahrscheinlichkeit

$$P(X_t = x | X_{t_n} = x_n \wedge X_{t_{n-1}} = x_{n-1} \wedge \cdots \wedge X_{t_1} = x_1 \wedge X_{t_0} = x_0) \tag{9.6}$$

ist die Wahrscheinlichkeit dafür, dass der Prozess zur Zeit t den Zustand x erreicht, wenn er zu den Zeitpunkten $t_0, t_1, \ldots, t_n$ die Zustände $x_0, x_1, \ldots, x_n$ durchlaufen hat.

Gedächtnislosigkeit

In vielen Fällen ist nur der letzte durchlaufene Zustand wichtig. Die Zustände in den Zeitpunkten $t_0 < \cdots < t_{n-1}$ haben dann keinen Einfluss auf die Wahrscheinlichkeit. Auf die bedingte Wahrscheinlichkeit (9.6) sollten also die Ereignisse $\{X_{t_0} = x_0\}$ bis $\{X_{t_{n-1}} = x_{n-1}\}$ keinen Einfluss haben.

Definition 9.3. *Ein stochastischer Prozess erfüllt die Markov-Eigenschaft, wenn für jede Folge von früheren Zeitpunkten $t_0 < t_1 < \cdots < t_n < t$ und Zuständen $x_0, \ldots, x_n, x \in \mathcal{S}$ die Wahrscheinlichkeit* (9.6) *nicht von der Vorgeschichte abhängt, also*

$$P(X_t = x | X_{t_n} = x_n \wedge X_{t_{n-1}} = x_{n-1} \wedge \cdots \wedge X_{t_1} = x_1 \wedge X_{t_0} = x_0) = P(X_t = x | X_{t_n} = x_n).$$

Die Wahrscheinlichkeiten $P(X_t = x | X_s = y)$ mit $t > s$ bestimmen das zeitliche Verhalten der Wahrscheinlichkeiten vollständig. Wir schreiben daher auch

$$p_{xy}(t, s) = P(X_t = x | X_s = y)$$

für die sogenannte *transiente Übergangswahrscheinlichkeit*. Für eine endliche Menge von Zuständen, können die transienten Übergangswahrscheinlichkeiten auch als zeitabhängige quadratische Matrix $P(s, t)$ geschrieben werden, deren Einträge

$$(P(s, t))_{xy} = p_{xy}(t, s)$$

mit den Zuständen $x, y \in \mathcal{S}$ indiziert sind.

Die Chapman-Kolmogorov-Gleichung

Man beachte, dass in der Definition der Markov-Eigenschaft keine Voraussetzungen darüber gemacht werden, wie nahe am Zeitpunkt t der letzte Zeitpunkt t_n der Vorgeschichte liegt. Die transienten Übergangswahrscheinlichkeiten $p_{xy}(s,t)$ werden aber im allgemeinen davon abhängen, wie weit in der Vergangenheit der Zeitpunkt $s < t$ liegt. Für eine näheren Zeitpunkt τ mit $s < \tau < t$ muss es daher einen Zusammenhang zwischen den transienten Übergangswahrscheinlichkeiten $p_{xy}(s,\tau)$, $p_{xy}(\tau,t)$ und $p_{xy}(s,t)$ geben.

Satz 9.4 (Chapman-Kolmogorov)**.** *Hat der Prozess die Markov-Eigenschaft und ist $s < \tau < t$, dann gilt*

$$p_{xy}(t,s) = \sum_{z\in\mathcal{S}} p_{xz}(t,\tau)p_{zy}(\tau,s),$$

was in Matrixform auch als

$$P(t,s) = P(t,\tau)P(\tau,s)$$

geschrieben werden kann.

Auch hier spielt es keine Rolle, wie nahe an t der Zwischenzeitpunkt τ liegt. Die Formel von Chapman-Kolmogoroff kann natürlich für zusätzliche Zwischenpunkte $s < t_1 < t_2 < \cdots < t_n < t$ formuliert werden. In Matrix-Notation gilt

$$P(t,s) = P(t,t_n)P(t_n,t_{n-1})\dots P(t_2,t_1)P(t_1,s),$$

was ausgeschrieben zu

$$p_{xy}(t,s) = \sum_{x_1,\dots,x_n\in\mathcal{S}} p_{xx_n}(t,t_n)p_{x_nx_{n-1}}(t_n,t_{n-1})\dots p_{x_2x_1}(t_2,t_1)p_{x_1y}(t_1,s)$$

wird. Jeder Summand auf der rechten Seite beschreibt einen Weg des Prozesses derart, dass er zu den Zwischenzeitpunkten bestimmte Zwischenzustände durchläuft.

Definition 9.5. *Die Wahrscheinlichkeit, dass der stochastische Prozess zwischen Zeitpunkten t_0 und t_n die Zwischenzustände x_i zu Zeiten t_i durchläuft ist das Produkt*

$$\sum_{x_1,\dots,x_n\in\mathcal{S}} p_{x_{n+1}x_n}(t_{n+1},t_n)p_{x_nx_{n-1}}(t_n,t_{n-1})\dots p_{x_2x_1}(t_2,t_1)p_{x_1x_0}(t_1,s) = \prod_{i=0}^{n} p_{x_{i+1}x_i}(t_{i+1}t_i)$$

heisst die Pfadwahrscheinlichkeit *für genannten Pfad.*

9.2.2 Diskrete Markov-Kette

Die Markov-Eigenschaft besagt, dass man keine Information verliert, wenn man die Vorgeschichte eines Zeitpunktes ignoriert. Insbesondere kann man eine Menge von geeigneten diskreten Zeitpunkten wählen, ohne viel Information über den Prozess zu verlieren. Eine *diskrete Markov-Kette* ist eine stochastischer Prozess, für den die Menge der Zeitpunkte t diskret ist. Es ist üblich, für die Zeitpunkte ganze oder natürliche Zahlen zu verwenden.

Definition 9.6. *Eine diskrete Markov-Kette ist ein stochastischer Prozess $(X_t)_{t\in\mathbb{N}}$ mit Werten in $\mathcal{S}$, der die Markov-Eigenschaft*

$$P(X_{n+1} = x_{n+1}|X_n = x_n \wedge \dots X_0 = x_0) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

hat.

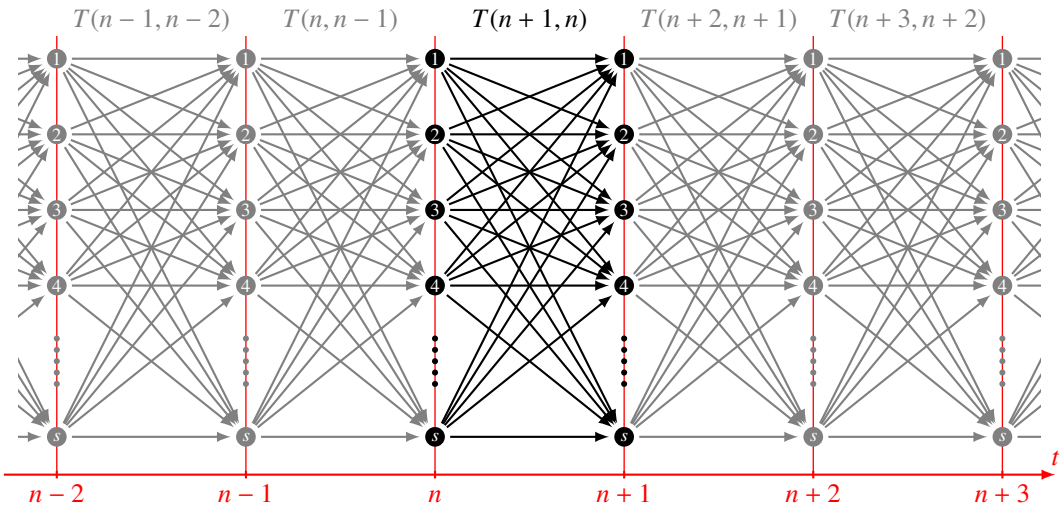


Abbildung 9.2: Diskrete Markovkette mit Zuständen $\mathcal{S} = \{1, 2, 3, \ldots, s\}$ und Übergangsmatrizen $T(n+1, n)$.

Die transienten Übergangswahrscheinlichkeiten zwischen aufeinanderfolgenden Zeitpunkten stellen jetzt die vollständige Information über die zeitliche Entwicklung dar (Abbildung 9.2). Aus der Matrix

$$T(n+1,n) = \begin{pmatrix} p_{11}(n+1,n) & \ldots & p_{1s}(n+1,n) \\ \vdots & \ddots & \vdots \\ p_{11}(n+1,n) & \ldots & p_{1s}(n+1,n) \end{pmatrix},$$

auch die 1-Schritt Übergangswahrscheinlichkeit genannt, kann man jetzt auch die Matrix der Überganswahrscheinlichkeiten für mehrere Schritte

$$T(n+m,n) = T(n+m,n+m-1)T(n+m-1,n+m-2)\ldots T(n+1,n)$$

mit der Chapman-Komogorov-Formel bestimmen. Die Markov-Eigenschaft stellt also sicher, dass man nur die 1-Schritt-Übergangswahrscheinlichkeiten kennen muss.

Eine Matrix T kann als Matrix der Übergangswahrscheinlichkeiten verwendet werden, wenn sie zwei Bedingungen erfüllt:

1. Die Einträge von T müssen als Wahrscheinlichkeiten interpretiert werden können, sie müssen also alle zwischen 0 und 1 sein: $0 \leq t_{ij} \leq 1$ für $i, j \in \mathcal{S}$

2. Die Matrix muss alle möglichen Fälle erfassen. Dazu ist notwendig, dass sich die Wahrscheinlichkeiten aller Übergänge aus einem Zustand j zu 1 summieren, also

$$\sum_{i \in \mathcal{S}} p_{ij} = 1.$$

Die Summe der Elemente einer Spalte

Beispiel. Die Permutationsmatrix einer Permutation $\sigma \in S_n$ (Abschnitt) ist eine Matrix mit Einträgen 0 und 1, so dass die erste Bedingung erfüllt ist. In jeder Zeile oder Spalte kommt genau eine 1 vor, so dass auch die zweite Bedingung erfüllt ist. Eine Permutationsmatrix beschreibt einen stochastischen Prozess, dessen Übergänge deterministisch sind. ○

Zustandswahrscheinlichkeiten

Die Wahrscheinlichkeit, mit der sich der Prozess zum Zeitpunkt n im Zustand $i \in \mathcal{S}$ befindet, wird

$$p_i(n) = P(X_i = n)$$

geschrieben, die auch in einem Vektor $p(n)$ zusammengefasst werden können. Die Matrix der Übergangswahrscheinlichkeiten erlaubt, die Verteilung $p(n+1)$ aus der Verteilung $p(n)$ zu berechnen. Nach dem Satz von der totalen Wahrscheinlichkeit ist nämlich

$$P(X_{n+1} = x) = \sum_{y \in \mathcal{S}} P(X_{n+1} = x | X_n = y) P(X_n = y) \qquad \text{oder} \qquad p^{(n+1)} = T(n+1, n) p^{(n)}$$

in Matrixform. Die Zeitentwicklung kann also durch Multiplikation mit der Übergangsmatrix berechnet werden.

Zeitunabhängige Übergangswahrscheinlichkeiten

Besonderes einfach wird die Situation, wenn die Übergangsmatrix $T(n+1, n)$ nicht von der Zeit abhängt. In diesem Fall ist $T(n+1, n) = T$ für alle n. Eine solche Markov-Kette heisst *homogen*. Die Mehrschritt-Übergangswahrscheinlichkeiten sind dann gegeben durch die Matrix-Potenzen $T(n+m, n) = T^m$. Im Folgenden gehen wir immer von einer homogenen Markov-Kette aus.

Stationäre Verteilung

Im Beispiel der Google-Matrix erwarten wir intuitiv, dass sich mit der Zeit eine Verteilung einstellt, die sich über die Zeit nicht mehr ändert. Ein solche Verteilung heisst stationär.

Definition 9.7. *Eine Verteilungsvektor p heisst* stationär *für die homogene Markov-Kette mit Übergangsmatrix T, wenn $Tp = p$.*

Eine stationäre Verteilung ist offenbar ein Eigenvektor der Matrix T zum Eigenwert 1. Gefunden werden kann er als Lösung des Gleichungssystems $Tp = p$. Dazu muss die Matrix $T - E$ singulär sein. Die Summe einer Spalte von T ist aber immer ein, da E in jeder Spalte genau eine 1 enthält, ist die Summe der Einträge einer Spalte von $T - E$ folglich 0. Die Summe aller Zeilen von $T - E$ ist also 0, die Matrix $T - E$ ist singulär. Dies garantiert aber noch nicht, dass alle Einträge in diesem Eigenvektor auch tatsächlich nichtnegativ sind. Die Perron-Frobienus-Theorie von Abschnitt 9.3 beweist, dass sich immer ein Eigenvektor mit nichtnegativen Einträgen finden lässt.

Es ist aber nicht garantiert, dass eine stationäre Verteilung auch eindeutig bestimmt ist. Dieser Fall tritt immer ein, wenn die geometrische Vielfachheit des Eigenwerts 1 grösser ist als 1. In Abschnitt 9.3.1 werden Bedingungen an eine Matrix T untersucht, die garantieren, dass der Eigenraum zum Eigenvektor 1 einedeutig bestimmt ist.

Beispiel. Als Beispiel dafür betrachten wir eine Permutation $\sigma \in S_n$ und die zugehörige Permutationsmatrix P, wie sie in Abschnitt beschrieben worden ist. Wir verwenden die Zyklenzerlegung (Abschnitt 6.1.2) $[n] = \{Z_1, Z_2, \dots\}$ der Permutation σ, ist ist also $\sigma(Z_i) = Z_i$ für alle Zyklen.

Jede Verteilung p, die auf jedem Zyklus konstant ist, ist eine stationäre Verteilung. Ist nämlich $i \in Z_k$, dann ist natürlich auch $\sigma(i) \in Z_k$, und damit ist $p_{\sigma(i)} = p_i$.

Für jede Wahl von nichtnegativen Zahlen z_i für $i = 1, \ldots, k$ mit der Eigenschaft $z_1 + \cdots + z_k = 1$ kann man eine stationäre Verteilung $p(z)$ konstruieren, indem man

$$p_i(z) = \frac{z_i}{|Z_r|} \qquad \text{wenn} \quad i \in Z_r$$

setzt. Die Konstruktion stellt sicher, dass sich die Komponenten zu 1 summieren. Wir können aus dem Beispiel auch ableiten, dass die geometrische Vielfachheit des Eigenvektors 1 mindestens so gross ist wie die Anzahl der Zyklen der Permutation σ. ○

Irreduzible Markov-Ketten

Die Zyklen-Zerlegung einer Permutation bilden voneinander isolierte Mengen von Zuständen, es gibt keine Möglichkeit eines Übergangs zu einem anderen Zyklus. Die Zyklen können daher unabhängig voneinander studiert werden. Diese Idee kann auf allgemeine Markov-Ketten verallgemeinert werden.

Definition 9.8. *Zwei Zustände $i, j \in \mathcal{S}$ kommunizieren, wenn die Übergangswahrscheinlichkeiten $T_{ij}(n) \neq 0$ und $T_{ij}(n) \neq 0$ sind für n gross genug.*

Die Zustände, die zu verschiedenen Zyklen einer Permutation gehören, kommunizieren nicht. Gerade deshalb waren auch die verschiedenen stationären Verteilungen möglich. Eine eindeutige stationäre Verteilung können wir also nur erwarten, wenn alle Zustände miteinander kommunizieren.

Definition 9.9. *Eine homogene Markov-Kette heisst* irreduzibel*, alle Zustände miteinander kommunizieren.*

Die Bedingung der Irreduzibilität ist gleichbedeutend damit, dass für genügend grosses n alle Matrixelemente von T^n positiv sind. Solche Matrizen nennt man positiv, in Abschnitt 9.3 wird gezeigt, dass positive Matrizen immer eine eindeutige stationäre Verteilung haben. In Abbildung 9.3 ist eine reduzible Markov-Kette dargestellt, die Zustandsmenge zerfällt in zwei Teilmengen von Zuständen, die nicht miteinander kommunizieren. Ein irreduzible Markov-Kette liegt vor, wenn sich ähnlich wie in Abbildung 9.2 jeder Zustand von jedem anderen aus erreichen lässt.

Wenn sich der Vektorraum $\mathbb{R}^n$ in zwei unter T invariante Unterräme zerlegen lässt, dann hat nach Wahl von Basen in den Unterräumen die Matrix T die Form

$$\left(\begin{array}{c|c} T_1 & 0 \\ \hline 0 & T_2 \end{array}\right).$$

Insbesondere kann man stationäre Verteilungen von T_1 und T_2 unabhängig voneinander suchen. Ist p_i eine stationäre Verteilung von T_i, dann ist

$$T\begin{pmatrix} g_1 p_1 \\ g_2 p_2 \end{pmatrix} = \begin{pmatrix} g_1 T_1 p_1 \\ g_2 T_2 p_2 \end{pmatrix} = \begin{pmatrix} g_1 p_1 \\ g_2 p_2 \end{pmatrix}, \qquad \text{für } g_i \in \mathbb{R}.$$

Durch Wahl der Gewichte $g_i \geq 0$ mit $g_1 + g_2 = 1$ lassen sich so die stationären Verteilungen für T aus den stationären Verteilungen der T_i ermitteln. Das Problem, die stationären Verteilungen von T zu finden, ist auf die Untermatrizen T_i reduziert worden.

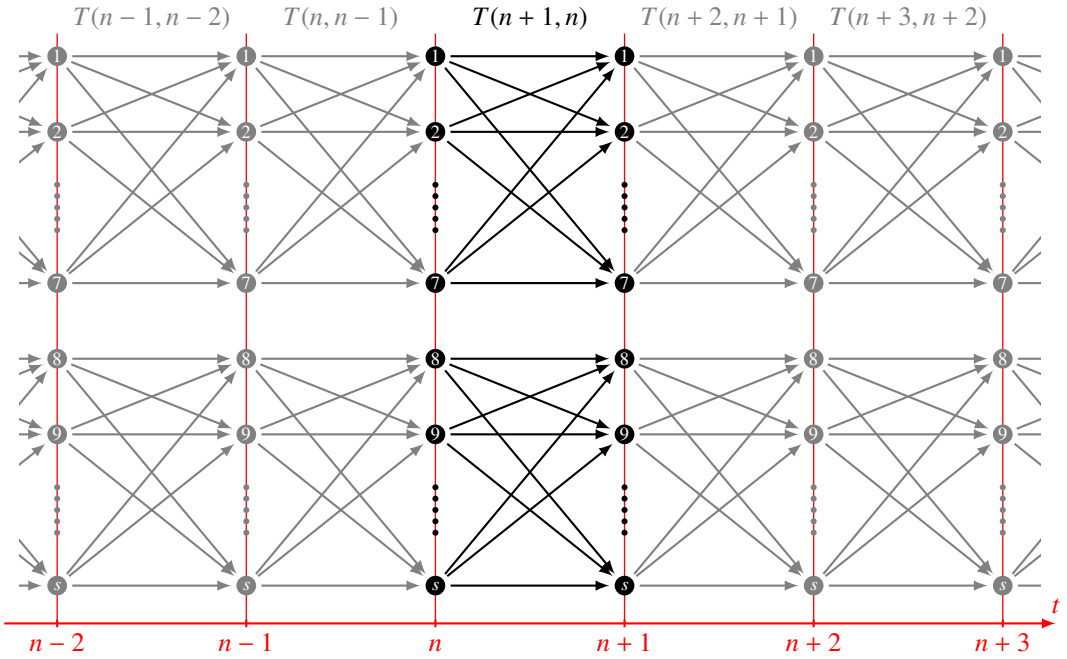


Abbildung 9.3: Diese Markov-Kette zerfällt in verschiedene irreduzible Markov-Ketten, dere Zustandsmengen nicht miteinander kommunizieren. Solche Markov-Ketten können unabhängig voneinander studiert werden.

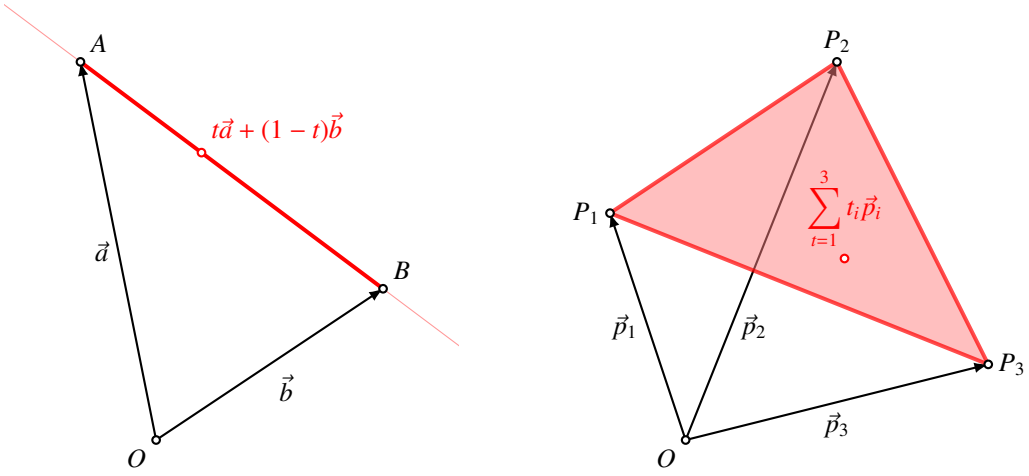


Abbildung 9.4: Die Konvexe Kombination von Vektoren $\vec{p}_1,\ldots,\vec{p}_n$ ist eine Summe der Form $\sum_{i=1}^{n} t_i\vec{p}_i$ wobei die $t_i \geq 0$ sind mit $\sum_{i=1}^{n} t_i = 1$. Für zwei Punkte bilden die konvexen Kombinationen die Verbindungsstrecke zwischen den Punkten, für drei Punkte in drei Dimensionen spannen die konvexen Kombinationen ein Dreieck auf.

Die konvexe Menge der stationären Verteilungen

Die stationären Verteilungen

$$\mathrm{Stat}(T) = \{p \in \mathbb{R}^n_+ \mid Tp = p \text{ und } \|p\|_1 = 1\}$$

bilden was man eine konvexe Menge nennt. Sind nämlich p und q stationäre Verteilungen, dann gilt zunächst $Tp = p$ und $Tq = q$. Wegen der Linearität gilt aber auch $T(tp+(1-t)q) = tTp+(1-t)Tq = tp + (1-t)q$. Jede Verteilung auf der "Verbindungsstrecke" zwischen den beiden Verteilungen ist auch wieder stationär.

Definition 9.10. *Eine* konvexe Kombination *von Vektoren $v_1, \ldots, v_k \in \mathbb{R}^\varkappa$ ist ein Vektor der Form*

$$v = t_1 v_1 + \cdots + t_k v_k \qquad \textit{mit} \quad t_i \geq 0 \textit{ und } t_1 + \cdots + t_n = 1.$$

Eine Teilmenge $M \subset \mathbb{R}^n$ heisst konvex, wenn zu zwei Vektoren $x, y \in M$ auch jede konvexe Kombination von x und y wieder in M ist.

Die konvexen Kombinationen der Vektoren sind Linearkombination mit nichtnegativen Koeffizienten. Sie bilden im Allgemeinen einen $(k-1)$-Simplex in $\mathbb{R}^n$. Für zwei Punkte x und y bilden die konvexen Kombination $tx + (1-t)y$ für $t \in [0,1]$ die Verbindungsstrecke der beiden Vektoren. Eine Menge ist also konvex, wenn sie mit zwei Punkten immer auch ihre Verbindungsstrecke enthält

Grenzverteilung

Im Beispiel der Google-Matrix wurde ein iterativer Algorithmus zur Berechnung des Pagerank verwendet. Es stellt sich daher die Frage, ob diese Methode für andere homogene Markov-Ketten auch funkioniert. Man beginnt also mit einer beliebigen Verteilung $p(0)$ und wendet die Übergangsmatrix T wiederholt an. Es entsteht somit eine Folge $p(n) = T^n p(0)$.

Definition 9.11. *Falls die Folge $p(n) = T^n p(0)$ konvergiert, heisst der Grenzwert*

$$p(\infty) = \lim_{n\to\infty} p(n)$$

eine Grenzverteilung *von T.*

Falls eine Grenzverteilung existiert, dann ist sie eine stationäre Verteilung. Für eine stationäre Verteilung $p(0)$ ist die Folge $p(n)$ eine konstante Folge, sie konvergiert also gegen $p(0)$. Stationäre Verteilungen sind also automatisch Grenzverteilungen. Falls der Raum der stationären Verteilungen mehrdimensional sind, dann ist auch die Grenzverteilung nicht eindeutig bestimmt, selbst wenn sie existiert. Aber nicht einmal die Existenz einer Grenzverteilung ist garantiert, wie das folgende Beispiel zeigt.

Beispiel. Sei T die Permutationsmatrix der zyklischen Verteilung von drei Elementen in S_3, also die Matrix

$$T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Die konstante Verteilung $\frac{1}{3}U$ ist offensichtlich eine stationäre Verteilung. In Abschnitt 9.3 wird gezeigt, dass es die einzige ist. Sei jetzt $p(0)$ eine beliebiger Vektor in $\mathbb{R}^3$ mit nichtnegativen Einträgen,

die sich zu 1 summieren. Dann bilden die Vektoren $p(n) = T^n p(0)$ einen Dreierzyklus

$$p(0) = p(3) = p(6) = \cdots = \begin{pmatrix} p_1(0) \\ p_2(0) \\ p_3(0) \end{pmatrix},$$
$$p(1) = p(4) = p(7) = \cdots = \begin{pmatrix} p_2(0) \\ p_3(0) \\ p_1(0) \end{pmatrix},$$
$$p(2) = p(5) = p(8) = \cdots = \begin{pmatrix} p_3(0) \\ p_1(0) \\ p_2(0) \end{pmatrix}.$$

Die Folge $p(n)$ kann also nur dann konvergieren, wenn die drei Komponenten gleich sind. ◯

Erwartungswert und Varianz

Wenn sich im Laufe der Zeit eine Grenzverteilung einstellen soll, dann muss es auch möglich sein, Erwartungswert und Varianz dieser Verteilung zu berechnen. Dazu muss jedem Zustand ein Zahlenwert zugeordnet werden. Sei also $g : \mathcal{S} \to R$ eine Funktion, die einem Zustand eine reelle Zahl zuordnet. Aus der Zufallsvariable X_n des Zustands zur Zeit n wird daraus die Zufallsvariable $Y_n = g(X_n)$ des Wertes zur Zeit n. Die Abbildung g kann auch als Vektor mit der Komponenten g_i für $i \in \mathcal{S}$ betrachtet werden, wir verwenden für diesen Vektor wieder die Schreibweise g.

Für die Verteilung $p(n)$ kann man jetzt auch Erwartungswert und Varianz berechnen. Der Erwartungswert ist

$$E(Y) = \sum_{i \in \mathcal{S}} g_i p_i(n) = g^t p(n).$$

Für die Varianz muss g_i durch g_i^2 ersetzt werden. Dies kann am einfachsten mit dem Hadamard-Produkt geschrieben werden:

$$E(Y^2) = \sum_{i \in \mathcal{S}} g_i p_i(n) = (g \odot g)^t p(n)$$
$$E(Y^k) = (g^{\odot k})^t p(n),$$

wobei wir die Hadamard-Potenz $A^{\odot k}$ einer Matrix A rekursiv durch

$$A^{\odot 0} = E \qquad \text{und} \qquad A^{\odot k} = A \odot A^{\odot(k-1)}$$

definieren.

Erwartungswert von Werten auf Übergängen

In Abschnitt 9.4 wird ein Spiel vorgestellt, in dem der Gewinn davon abhängt, welcher Übergang stattfindet, nicht welcher Zustand erreicht wird. Es git daher eine Matrix G von Gewinnen, der Eintrag g_{ij} ist der Gewinn, der bei einem Übergang von Zustand j in den Zustand i ausgezahlt wird. Mit dieser Matrix lassen sich jetzt viele verschiedene Fragen beantworten:

Frage 9.12. *Mit welchem Gewinn kann man in Runde n des Spiels rechnen, wenn $p(n-1)$ die Verteilung zur Zeit $n-1$ ist?*

Der Erwartungswert ist

$$E(Y) = \sum_{i,j \in \mathcal{S}} g_{ji} t_{ji} p_i(n-1)$$

oder in Matrixform

$$= U^t (G \odot T) p(n-1).$$

Frage 9.13. *Mit welchen Gewinnen kann man rechnen, wenn der Prozess sich zu Beginn einer Spielrunde im Zustand i befindet?*

Dies ist der Spezialfall der Frage 9.12 für die Verteilung $p_j(n-1) = \delta_{ij}$. Der Erwartungswert ist die Summe der Spalte j der Matrix $G \odot T$. Man kann das Produkt $U^t(G \odot T)$ also auch als eine Zeilenvektor von Gewinnerwartungen unter der Vorbedingung $X_{n-1} = j$ betrachten.

$$\Big(E(Y|X_{n-1} = 1) \quad \dots \quad E(Y|X_{n-1} = n)\Big) = U^t(G \odot T).$$

Indem man G durch $G^{\odot k}$ ersetzt, kann man beliebige höhere Momente berechnen.

9.2.3 Absorbierende Zustände

Eine Grenzverteilung beschreibt die relative Häufigkeit, mit der der Prozess in den verschiedenen Zuständen vorbeikommt. In einem Spiel, in dem der Spieler ruiniert werden kann, gibt es aus dem Ruin-Zustand keinen Weg zurück. Der Spieler bleibt in diesem Zustand.

Definition 9.14. *Ein Zustand i einer homogenen Markov-Kette mit Übergangsmatrix T heisst* absorbierend*, wenn $T_{ii} = 1$ ist. Eine Markov-Kette mit mindestens einem absorbierenden Zustand heisst* absorbierende Markov-Kette*. Nicht absorbierende Zustände heissen* transient

Eine Markov-Kette kann mehrere absorbierende Zustände haben, wie in Abbildung 9.5 dargestellt. Indem man die absorbierenden Zustände zuerst auflistet, bekommt die Übergangsmatrix die Form

$$T = \left(\begin{array}{c|c} E & R \\ \hline 0 & Q \end{array}\right).$$

Die Matrix R beschreibt die Wahrscheinlichkeiten, mit denen man ausgehend von einem transienten Zustand in einem bestimmten absorbierenden Zustand landet. Die Matrix Q beschreibt die Übergänge, bevor dies passiert. Die Potenzen von T sind

$$T^2 = \left(\begin{array}{c|c} E & R+RQ \\ \hline 0 & Q^2 \end{array}\right), \quad T^3 = \left(\begin{array}{c|c} E & R+RQ+RQ^2 \\ \hline 0 & Q^3 \end{array}\right), \ \dots, \ T^k = \left(\begin{array}{c|c} E & R\sum_{l=0}^{k-1} Q^l \\ \hline 0 & Q^k \end{array}\right).$$

Da man früher oder später in einem absorbierenden Zustand landet, muss $\lim_{k\to\infty} Q^k = 0$ sein. Die Summe in der rechten oberen Teilmatrix kann man als geometrische Reihe summieren, man erhält die Matrix

$$\sum_{l=0}^{k-1} Q^l = (E-Q)^{-1}(E-Q^k),$$

die für $k \to \infty$ gegen

$$N = \lim_{k\to\infty} \sum_{l=0}^{k-1} Q^l = (E-Q)^{-1}$$

konvergiert. Die Matrix N heisst die *Fundamentalmatrix* der absorbierenden Markov-Kette.

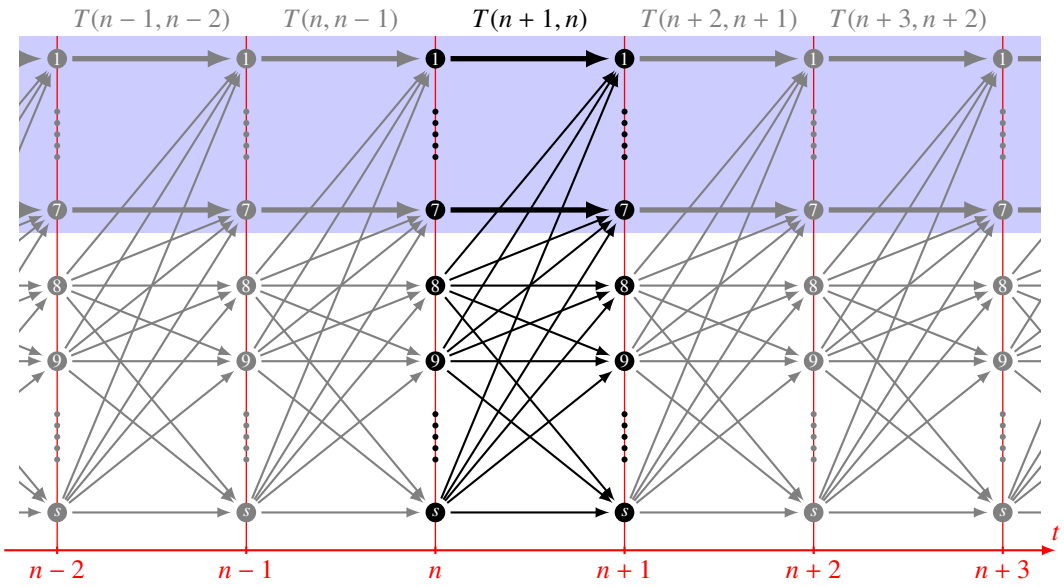


Abbildung 9.5: Markov-Kette mit absorbierenden Zuständen (blau hinterlegt). Erreicht die Markov-Kette einen absorbierenden Zustand, dann verbleibt sie für alle zukünftigen Zustände in diesem Zustand.

Absorbtionszeit

Wie lange dauert es im Mittel, bis der Prozess in einem Absorptionszustand i stecken bleibt? Die Fundamentalmatrix N der Markov-Kette beantwortet diese Frage. Wenn der Prozess genau im Schritt k zum ersten Mal Zustand i ankommt, dann ist $E(k)$ die mittlere Wartezeit. Der Prozess verbringt also zunächst $k-1$ Schritte in transienten Zuständen, bevor er in einen absorbierenden Zustand wechselt.

Wir brauchen die Wahrscheinlichkeit für einen Entwicklung des Zustandes ausgehend vom Zustand j, die nach $k-1$ Schritten im Zustand l landet, von wo er in den absorbierenden Zustand wechselt. Diese Wahrscheinlichkeit ist

$$P(X_k = i \wedge X_{k-1} = l \wedge X_0 = j) = \sum_{i_1,\ldots,i_{k-2}} r_{il} q_{li_{k-2}} q_{i_{k-2}i_{k-3}} \cdots q_{i_2 i_1} q_{i_1 j}$$

Von den Pfaden, die zur Zeit $k-1$ im Zustand l ankommen gibt es aber auch einige, die nicht absorbiert werden. Für die Berechnung der Wartezeit möchten wir nur die Wahrscheinlichkeit innerhalb der Menge der Pfade, die auch tatsächlich absorbiert werden, das ist die bedingte Wahrscheinlichkeit

$$\begin{aligned} P(X_k = i \wedge X_{k-1} = l \wedge X_0 = j | X_k = i) &= \frac{P(X_k = i \wedge X_{k-1} = l \wedge X_0 = j)}{P(X_k = i)} \\ &= \sum_{i_1,\ldots,i_{k-2}} q_{li_{k-2}} q_{i_{k-2}i_{k-3}} \cdots q_{i_2 i_1} q_{i_1 j}. \end{aligned} \tag{9.7}$$

Auf der rechten Seite steht das Matrixelement (l, j) von Q^{k-1}.

Für die Berechnung der erwarteten Zeit ist müssen wir die Wahrscheinlichkeit mit k multiplizie-

ren und summieren:

$$\begin{aligned} E(k) &= \sum_{k=0}^{\infty} k(q_{lj}^{(k)} - q_{lj}^{(k-1)}) \\ &= \cdots + (k+1)(q_{lj}^{(k)} - q_{lj}^{(k+1)}) + k(q_{lj}^{(k-1)} - q_{lj}^{(k)}) + \ldots \\ &= \cdots + q_{lj}^{(k-1)} + \cdots = \sum_{k} q_{lj}^{(k)}. \end{aligned} \tag{9.8}$$

In zwei benachbarten Termen in (9.8) heben sich die Summanden $kq_{lj}^{(k)}$ weg, man spricht von einer teleskopischen Reihe. Die verbleibenden Terme sind genau die Matrixelemente der Fundamentalmatrix N. Die Fundamentalmatrix enthält also im Eintrag (l, j) die Wartezeit bis zur Absorption über den Zustand l.

Wartezeit

Die mittlere Wartezeit bis zum Erreichen eines Zustands kann mit der Theorie zur Berechnung der Absorptionszeit berechnet werden. Dazu modifiziert man den Prozess dahingehend, dass der Zielzustand ein absorbierender Zustand wird. Der Einfachheit halber gehen wir davon aus, dass der Zustand 1 der Zielzustand ist. Wir ersetzen die Übergangsmatrix T durch die Matrix

$$\tilde{T} = \left(\begin{array}{c|ccc} 1 & t_{12} & \ldots & t_{1n} \\ \hline 0 & t_{22} & \ldots & t_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & t_{n2} & \ldots & t_{nn} \end{array}\right).$$

$\tilde{T}$ hat den Zustand 1 als absorbierenden Zustand. Die Q und R sind

$$\tilde{R} = \begin{pmatrix} t_{12} & \ldots & t_{1n} \end{pmatrix}, \quad \tilde{Q} = \begin{pmatrix} t_{22} & \ldots & t_{2n} \\ \vdots & \ddots & \vdots \\ t_{n2} & \ldots & t_{nn} \end{pmatrix}.$$

Die Wartezeit bis zum Erreichen des Zustands i ausgehend von einem Zustand n kann jetzt aus der Absorbtionszeit der Markov-Kette im Zustand 1 mit Hilfe der Fundamentalmatrix

$$\tilde{N} = (E - \tilde{Q})^{-1}$$

berechnet werden.

9.3 Positive Vektoren und Matrizen

Die Google-Matrix und die Matrizen, die wir in Markov-Ketten angetroffen haben, zeichnen sich dadurch aus, dass alle ihre Einträge positiv oder mindestens nicht negativ sind. Die Perron-Frobenius-Theorie, die in diesem Abschnitt entwickelt werden soll, zeigt, dass Positivität einer Matrix nützliche Konsequenzen für Eigenwerte und Eigenvektoren hat. Das wichtigste Resultat ist die Tatsache, dass postive Matrizen immer einen einzigen einfachen Eigenwert mit Betrag $\varrho(A)$ haben, was zum Beispiel die Konvergenz des Pagerank-Algorithmus garantiert. Dies wird im Satz von Perron-Frobenius in Abschnitt 9.3.3 erklärt.

9.3.1 Elementare Eigenschaften

In diesem Abschnitt betrachten wir ausschliesslich reelle Vektoren und Matrizen. Die Komponenten sind somit immer mit miteinander vergleichbar, daraus lässt sich auch eine Vergleichsrelation zwischen Vektoren ableiten.

Definition 9.15. *Ein Vektor $v \in \mathbb{R}^n$ heisst* positiv*, geschrieben $v > 0$, wenn alle seine Komponenten positiv sind: $v_i > 0 \forall i$. Ein Vektor $v \in \mathbb{R}^n$ heisst* nichtnegativ*, in Formeln $v \geq 0$, wenn alle seine Komponenten nicht negativ sind: $v_i \geq 0 \forall i$.*

Geometrisch kann man sich die Menge der positven Vektoren in zwei Dimensionen als die Punkte des ersten Quadranten oder in drei Dimensionen als die Vektoren im ersten Oktanten vorstellen.

Aus der Positivität eines Vektors lässt sich jetzt eine Vergleichsrelation für beliebige Vektoren ableiten. Mit der folgenden Definition wird erreicht, das mit Ungleichungen für Vektoren auf die gleiche Art und Weise gerechnet werden kann, wie man sich dies von Ungleichungen zwischen Zahlen gewohnt ist.

Definition 9.16. *Für zwei Vektoren $u, v \in \mathbb{R}^n$ ist genau dann $u > v$, wenn $u - v > 0$ ist. Ebenso ist $u \geq v$ genau dann, wenn $u - v \geq 0$.*

Ungleichungen zwischen Vektoren kann man daher auch so interpretieren, dass sie für jede Komponente einzeln gelten müssen. Die Definition funktionieren analog auch für Matrizen:

Definition 9.17. *Eine Matrix $A \in M_{m\times n}(\mathbb{R})$ heisst* positiv*, wenn alle ihre Einträge a_{ij} positiv sind: $a_{ij} > 0 \forall i, j$. Eine Matrix $A \in M_{m\times n}(\mathbb{R})$ heisst* nichtnegativ*, wenn alle ihre Einträge a_{ij} nichtnegativ sind: $a_{ij} \geq 0 \forall i, j$. Man schreibt $A > B$ bzw. $A \geq B$ wenn $A - B > 0$ bzw. $A - B \geq 0$.*

Die Permutationsmatrizen sind Beispiele von nichtnegativen Matrizen, deren Produkte wieder nichtnegativ sind. Dies ist aber ein sehr spezieller Fall, wie das folgende Beispiel zeigt.

Beispiel. Wir betrachten die Matrix

$$A = \begin{pmatrix} 0.9 & 0.1 & & & & \\ 0.1 & 0.8 & 0.1 & & & \\ & 0.1 & 0.8 & 0.1 & & \\ & & 0.1 & 0.8 & 0.1 & \\ & & & 0.1 & 0.8 & 0.1 \\ & & & & 0.1 & 0.9 \end{pmatrix} \tag{9.9}$$

Die Multiplikation eines Vektors mit dieser Matrix bewirkt, dass die Komponenten des Vektors auf benachbarte Komponenten "verschmiert" werden. Wendet man A wiederholt auf den ersten Standardbasisvektor $v_1 = e_1$ an, erhält man nacheinander die Vektoren $v_2 = Av_1$, $v_n = Av_{n-1}$. In Abbildung 9.6 sind die Komponenten als Säulen dargestellt. Man kann erkennen, dass für genügend grosse n alle Komponenten der Vektoren positiv werden.

Man kann auch direkt die Potenzen A^n ausrechen und sehen, dass

$$A^5 = \begin{pmatrix} 0.65658 & 0.27690 & 0.05925 & 0.00685 & 0.00041 & 0.00001 \\ 0.27690 & 0.43893 & 0.22450 & 0.05281 & 0.00645 & 0.00041 \\ 0.05925 & 0.22450 & 0.43249 & 0.22410 & 0.05281 & 0.00685 \\ 0.00685 & 0.05281 & 0.22410 & 0.43249 & 0.22450 & 0.05925 \\ 0.00041 & 0.00645 & 0.05281 & 0.22450 & 0.43893 & 0.27690 \\ 0.00001 & 0.00041 & 0.00685 & 0.05925 & 0.27690 & 0.65658 \end{pmatrix} > 0$$

und dass daher für alle $n \geq 5$ die Matrix A^n positiv ist. ○

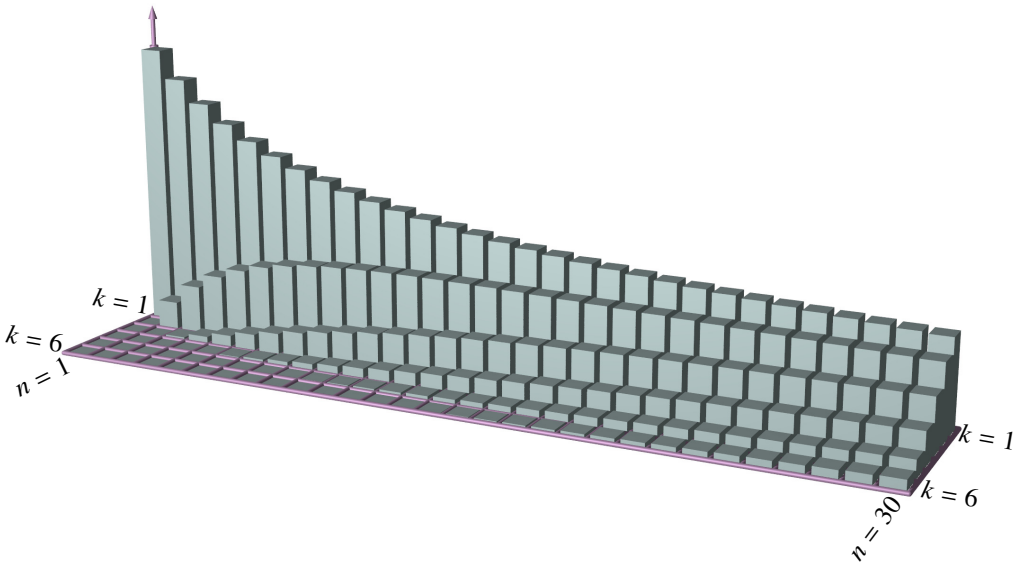


Abbildung 9.6: Die sechs Komponenten für $k = 1$ bis $k = 6$ der Vektoren $A^{n-1}e_1$ für die Matrix A in (9.9) sind als Säulen dargestellt. Sie zeigen, dass für genügend grosses n, alle Komponenten des Vektors $A^{n-1}e_1$ positiv werden.

Die Eigenschaft der Matrix A von (9.9), dass $A^n > 0$ für genügend grosses n ist bei Permutationsmatrizen nicht vorhanden. Die Zyklen-Zerlegung einer Permutationsmatrix zeigt, welche Unterräume von $\mathbb{R}^n$ die iterierten Bilder eines Standardbasisvektors aufspannen. Diese sind invariante Unterräume der Matrix. Das im Beispiel illustrierte Phänomen findet dann nur in invarianten Unterräumen statt.

Beispiel. Die Matrix

$$A = \begin{pmatrix} 0.9 & 0.1 & & & & & \\ 0.1 & 0.8 & 0.1 & & & & \\ & 0.1 & 0.9 & & & & \\ & & & 0.9 & 0.1 & & \\ & & & 0.1 & 0.8 & 0.1 & \\ & & & & 0.1 & 0.9 & \end{pmatrix} \tag{9.10}$$

besteht aus zwei 3×3-Blöcken. Die beiden Unterräume $V_1 = \langle e_1, e_2, e_3 \rangle$ und $V_2 = \langle e_4, e_5, e_6 \rangle$ sind daher invariante Unterräume von A und damit auch von A^n. Die Potenzen haben daher auch die gleich Blockstruktur. Insbesondere sind zwar die Blöcke von A^n für $n > 1$ positive Teilmatrizen, aber die Matrix A^n ist für alle n nicht positiv. ○

Definition 9.18. *Eine nichtnegative Matrix mit der Eigenschaft, dass $A^n > 0$ für ein genügend grosses n, heisst* primitiv.

Die Matrix A von (9.9) ist also primitiv, Permutationsmatrizen sind niemals primitiv. Die Matrix A von (9.10) ist nicht primitiv, aber die einzelnen Blöcke sind primitiv. Viele der Ausssagen über positive Matrizen lassen sich auf primitive nichtnegative Matrizen verallgemeinern.

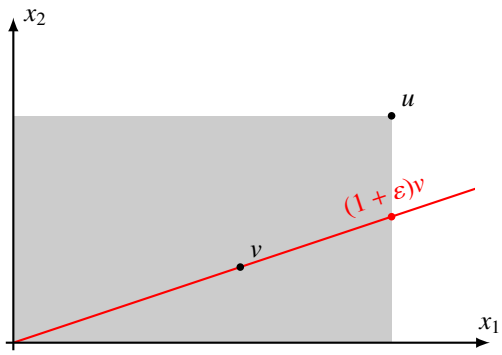


Abbildung 9.7: Die Vektoren $w \leq u$ liegen im grauen Rechteck. Zwei nichtnegative Vektoren u und v mit $u > v$ haben keine gleichen Komponenten. Daher kann man v mit einer Zahl $\vartheta = 1 + \varepsilon > 1$ strecken, so dass der gestreckte Vektor $(1 + \varepsilon)v$ gerade noch im grauen Rechteck liegt: $u \geq (1 + \varepsilon)v$. Streckung mit einem grösseren Faktor führt dagegen aus dem Rechteck hinaus.

Das Beispiel zeigt auch, dass der Begriff der primitiven Matrix eng mit der Idee verknüpft ist, die Problemstellung in invariante Unterräume aufzuteilen, in denen eine primitive Matrix vorliegt. Primitive Matrizen werden damit zu naheliegenden Bausteinen für die Problemlösung für nicht primitive Matrizen.

Eine interessante Eigenschaft positiver Vektoren oder Matrizen ist, dass die Positivität sich manchmal "upgraden" lässt, wie im folgenden Satz. Er zeigt, dass ein Vektor, der grösser ist als ein anderer, auch um einen definierten Faktor > 1 grösser ist. Dies wird geometrisch in Abbildung 9.7 illustriert.

Satz 9.19 (Trenntrick)**.** *Sind u und v nichtnegative Vektoren und $u > v$, dann gibt es eine positive Zahl $\varepsilon > 0$ derart, dass $u \geq (1 + \varepsilon)v$. Ausserdem kann ε so gewählt werden, dass $u \not\geq (1 + \mu)v$ für $\mu > \varepsilon$.*

Beweis. Wir betrachten die Zahl

$$\vartheta = \max_{v_i \neq 0} \frac{u_i}{v_i}.$$

Wegen $u > v$ sind die Quotienten auf der rechten Seite alle > 0. Da nur endlich viele Quotienten miteinander verglichen werden, ist daher auch $\vartheta > 1$. Es folgt $u \geq \vartheta v$. Wegen $\vartheta > 1$ ist $\varepsilon = \vartheta - 1 > 0$ und $u \geq (1 + \varepsilon)v$. □

Der Satz besagt also, dass es eine Komponente $v_i \neq 0$ gibt derart, dass $u_i = (1 + \varepsilon)v_i$. Diese Komponenten limitiert also, wie stark man v strecken kann, so dass er immer noch $\leq u$ ist. Natürlich folgt aus den der Voraussetzung $u > v$ auch, dass u ein positiver Vektor ist (Abbildung 9.7).

Satz 9.20 (Vergleichstrick)**.** *Sei A eine positive Matrix und seinen u und v Vektoren mit $u \geq v$ und $u \neq v$, dann ist $Au > Av$ (siehe auch Abbildung 9.8).*

Beweis. Wir schreiben $d = u - v$, nach Voraussetzung ist $d \neq 0$. Der Satz besagt dann, dass aus $d \geq 0$ folgt, dass $Ad > 0$, dies müssen wir beweisen.

Die Ungleichung $Ad > 0$ besagt, dass alle Komponenten von Ad positiv sind. Um dies nachzuweisen, berechnen wir

$$(Ad)_i = \sum_{j=1}^{n} a_{ij} d_j. \tag{9.11}$$

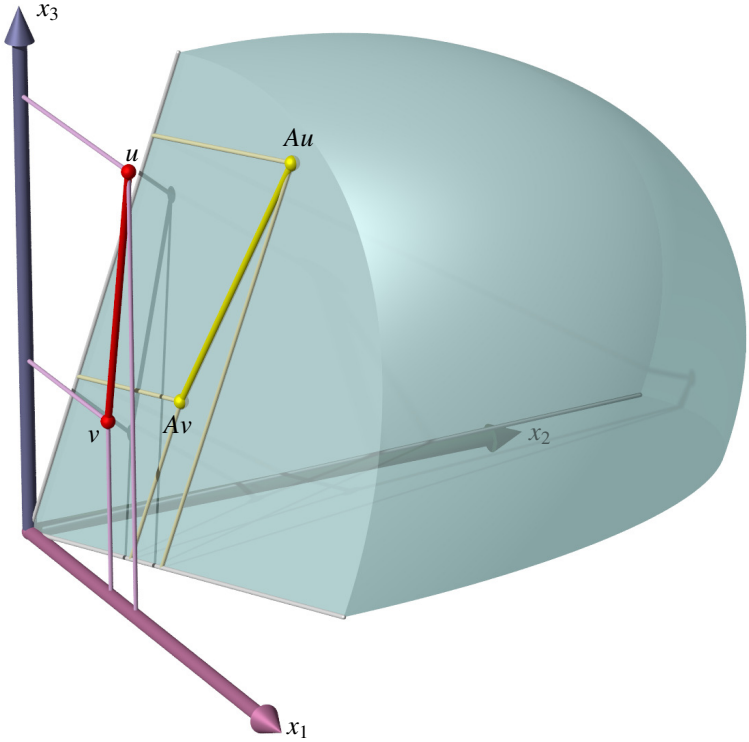


Abbildung 9.8: Eine positive Matrix A bildet nichtnegative Vektoren in positive Vektoren ab (Korollar 9.21). Zwei verschiedene Vektoren auf einer Seitenfläche erfüllen $u \geq v$, aber nicht $u > v$, da sie sich in der Koordinaten x_2 nicht unterscheiden. Die Bilder unter A unterscheiden sich dann auch in x_2, es gilt $Au > Av$ (siehe auch Satz 9.20)

Alle Terme $a_{ij} > 0$, weil A positiv ist, und mindestens eine der Komponenten $d_j > 0$, weil $d \neq 0$. Insbesondere sind alle Terme der Summe ≥ 0, woraus wir bereits schliessen können, dass $(Ad)_i \geq 0$ sein muss. Die Komponente $d_j > 0$ liefert einen positiven Beitrag $a_{ij}d_j > 0$ zur Summe (9.11), also ist $(Ad)_i > 0$. □

Der folgende Spezialfall folgt unmittelbar aus dem Satz 9.20.

Korollar 9.21. *Ist A eine positive Matrix und $u \geq 0$ mit $u \neq 0$, dann ist $Au > 0$.*

Eine positive Matrix macht also aus nicht verschwindenden und nicht negativen Vektoren positive Vektoren.

9.3.2 Die verallgemeinerte Dreiecksungleichung

Die Dreiecksungleichung besagt, dass für beliebige Vektoren $u, v \in \mathbb{R}^n$ gilt

$$|u + v| \leq |u| + |v|$$

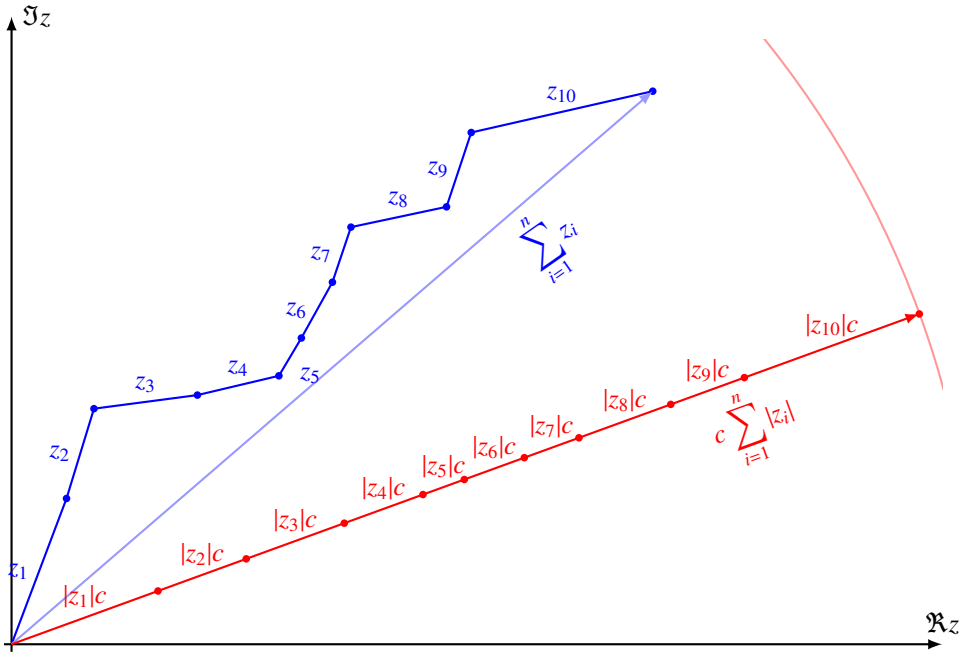


Abbildung 9.9: Die verallgemeinerte Dreiecksungleichung von Satz 9.22 besagt, dass die Länge einer Summe von Vektoren (blau) höchstens so gross ist wie die Summe der Längen, mit Gleichheit genau dann, wenn alle Vektoren die gleiche Richtung haben (rot). Hier dargestellt am Beispiel von Zahlen in der komplexen Zahlenebene. In dieser Form wird die verallgemeinerte Dreiecksungleichung in Satz 9.23

mit Gleichheit genau dann, wenn u und v linear abhängig sind. Wenn beide von 0 verschieden sind, dann gibt es eine positive Zahl t mit $u = tv$. Wir brauchen eine Verallgemeinerung für eine grössere Zahl von Summanden.

Satz 9.22 (Verallgemeinerte Dreiecksungleichung)**.** *Für n Vektoren $v_i \neq 0$ gilt*

$$|u_1 + \cdots + u_n| \leq |u_1| + \cdots + |u_n|$$

mit Gleichheit genau dann, wenn alle Vektoren nichtnegative Vielfache eines gemeinsamen Einheitsvektors c sind: $u_i = |u_i|c$ (siehe auch Abbildung 9.9).

Beweis. Die Aussage kann mit vollständiger Induktion bewiesen werden. Die Induktionsverankerung ist der Fall $n = 2$ gegeben durch die gewöhnliche Dreiecksungleichung.

Wir nehmen daher jetzt an, die Aussage sei für n bereits bewiesen, wir müssen sie dann für $n + 1$ beweisen. Die Summe von $n + 1$ Vektoren kann man $u = u_1 + \cdots + u_n$ und $v = u_{n+1}$ aufteilen. Es gilt dann

$$|u + v| = |u_1 + \cdots + u_n + u_{n+1}|$$

und

$$|u_1 + \cdots + u_n| = |u_1| + \cdots + |u_n|.$$

Aus der Induktionsannahme folgt dann, dass die Vektoren $u_1, \dots, u_n$ positive Vielfache eines Einheitsvektors u sind, $u_i = |u_i|c$. Es ist dann

$$u = u_1 + \cdots + u_n = \Big(\sum_{i=1}^{n} |u_i|\Big).$$

Aus der gewöhnlichen Dreiecksungleichung, angewendet auf u und v folgt jetzt, dass v ebenfalls ein nichtnegatives Vielfaches von c ist. Damit ist der Induktionsschritt vollzogen. □

Satz 9.23. *Seien $a_1, \dots, a_n$ positive Zahlen und $u_i \in \mathbb{C}$ derart, dass*

$$\Big|\sum_{i=1}^{n} a_i u_i\Big| = \sum_{i=1}^{n} a_i |u_i|,$$

dann gibt es eine komplexe Zahl c und einen nichtnegativen Vektor v derart, dass $u = cv$.

Der Satz besagt, dass die komplexen Komponenten u_i alle das gleiche Argument haben. Die motiviert den nachstehenden geometrischen Beweis des Satzes.

Beweis. Wer stellen uns die komplexen Zahlen u_i als Vektoren in der zweidimensionalen Gaussschen Ebene vor. Dann ist die Aussage nichts anderes als ein Spezialfall von Satz 9.22 für den zweidimensionalen reellen Vektorraum $\mathbb{C}$. □

9.3.3 Der Satz von Perron-Frobenius

Wir sind an den Eigenwerten und Eigenvektoren einer positiven oder primitiven Matrix interessiert. Nach Definition des Spektralradius $\varrho(A)$ muss es einen Eigenvektor zu einem Eigenwert λ mit Betrag $|\lambda| = \varrho(A)$ geben, aber a priori wissen wir nicht, ob es einen reellen Eigenwert vom Betrag $\varrho(A)$ gibt, und ob der Eigenvektor dazu reell ist.

In Abbildung 9.8 kann man sehen, dass eine positive Abbildung den positiven Oktanten in einen etwas engeren Kegel hinein abbildet. Iteriert man dies (Abbildung 9.10), wird die Bildmenge immer enger, bis sie nur ein sehr enger Kegel um die Richtung des Eigenvektors ist. Tatsächlich kann man aus dieser Idee auch einen topologischen Beweis des untenstehenden Satzes von Perron-Frobenius konstruieren. Er beruht darauf, dass eine Abbildung, die Distanzen verkleinert, einen Fixpunkt hat. Die Konstruktion einer geeigneten Metrik ist allerdings eher kompliziert, weshalb wir im Beweise der nachstehenden Aussagen den konventionellen Weg wählen.

Wir beginnen damit zu zeigen, dass für positive Matrizen A, nichtnegative Eigenvektoren zu Eigenwerten $\lambda \neq 0$ automatisch positiv sind. Ausserdem müssen die zugehörigen Eigenwerte sogar positiv sein.

Satz 9.24. *Sei A eine positive Matrix und u ein nichtnegativer Eigenvektor zum Eigenwert $\lambda \neq 0$. Dann ist u ein positiver Vektor und $\lambda > 0$.*

Beweis. Nach dem Korollar 9.21 folgt, dass $Au > 0$ ein positiver Vektor ist, es sind also alle Komponenten positiv. Der Vektor u ist aber auch ein Eigenvektor, es gilt also $\lambda u = Au$. Da alle Komponenten von Au positiv sind, müssen auch alle Komponenten von λu positiv sein. Das ist nur möglich, wenn $\lambda > 0$. □

Satz 9.25. *Sei A eine positive Matrix und v ein Eigenvektor von A zu einem Eigenwert λ mit Betrag $|\lambda| = \varrho(A)$, dann ist der Vektor u mit den Komponenten $u_i = |v_i|$ ein positiver Eigenvektor zu Eigenwert $\varrho(A)$.*

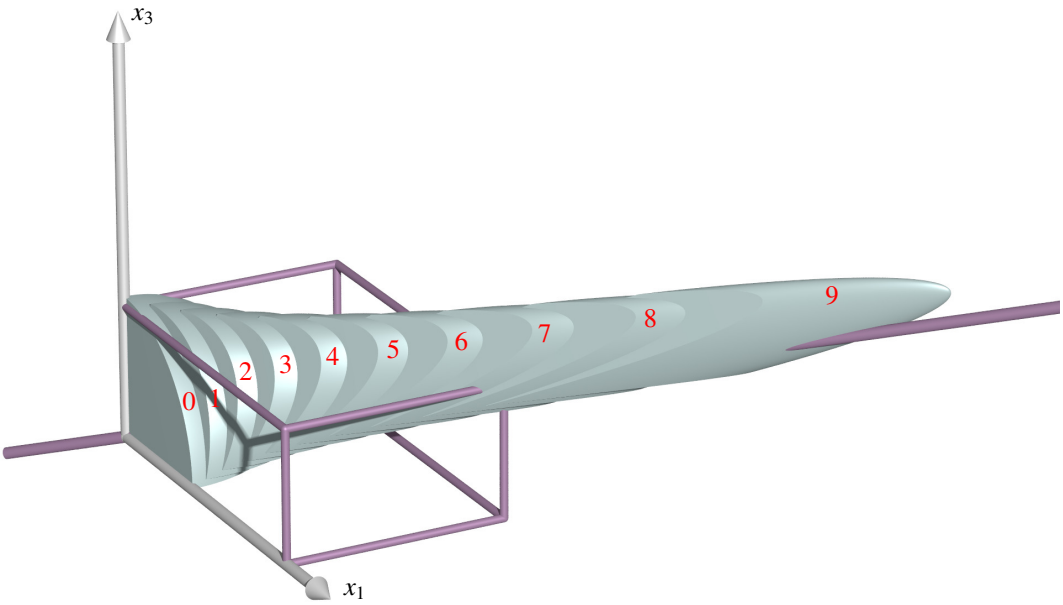


Abbildung 9.10: Die Iteration einer positiven Matrix bildet den positiven Oktanten in immer enger werdende Kegel ab, die die Richtung des gesuchten Eigenvektors gemeinsam haben.

Beweis. Es gilt natürlich auch, dass

$$(Au)_i = \sum_{j=1}^{n} a_{ij}u_j = \sum_{j=1}^{n} |a_{ij}v_j| \geq \left|\sum_{j=1}^{n} a_{ij}v_j\right| = |(Av)_i| = |\lambda v_i| = \varrho(A)|v_i| = \varrho(A)u_i,$$

oder $Au \geq \varrho(A)u$. Wir müssen zeigen, dass sogar $Au = \varrho(A)u$ gilt. Wir nehmen daher an, dass $Au \neq \varrho(A)u$ ist, und führen dies zu einem Widerspruch.

Da $\varrho(A)u$ ein nichtnegativer Vektor ist, können wir den Vergleichstrick Satz 9.20, auf die beiden Vektoren Au und $\varrho(A)u$ anwenden. Wir schliessen $A^2u > \varrho(A)Au$.

Mit dem Trenntrick Satz 9.19 können wir jetzt eine Zahl $\vartheta > 1$ finden derart, dass

$$A^2u \geq \vartheta\varrho(A)Au$$

ist. Durch weitere Anwendung von A findet man

$$A^3u \geq (\vartheta\varrho(A))^2Au$$
$$\vdots$$
$$A^{k+1}u \geq (\vartheta\varrho(A))^kAu$$

Daraus kann man jetzt die Norm abschätzen:

$$\begin{aligned} \|A^k\|\,|Au| \geq \|A^{k+1}u\| \geq (\vartheta\varrho(A))^k|Au| \quad &\Rightarrow \quad \|A^k\| \geq (\vartheta\varrho(A))^k \\ &\Rightarrow \quad \|A^k\|^{\frac{1}{k}} \geq \vartheta\varrho(A) \\ &\Rightarrow \quad \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} \geq \vartheta\varrho(A) \\ &\Rightarrow \quad \varrho(A) \geq \vartheta\varrho(A) \end{aligned}$$

Wegen $\vartheta > 1$ ist dies aber gar nicht möglich. Dieser Widerspruch zeigt, dass $u = v$ sein muss, insbesondere ist v ein nichtnegativer Eigenvektor. □

Satz 9.26. *Sei A eine positive Matrix und v ein Eigenvektor zu einem Eigenwert λ mit Betrag $|\lambda| = \varrho(A)$. Dann ist $\lambda = \varrho(A)$.*

Beweis. Nach Satz 9.25 ist der Vektor u mit den Komponenten $u_i = |v_i|$ ein positiver Eigenvektor zum Eigenwert $\varrho(A)$. Aus der Eigenvektorgleichung für u folgt

$$Au = \varrho(A)u \quad \Rightarrow \quad \sum_{j=1}^{n} a_{ij}|v_j| = \varrho(A)|v_i|. \tag{9.12}$$

Anderseits ist v ein Eigenvektor zum Eigenwert λ, also gilt

$$\sum_{j=1}^{n} a_{ij}v_j = \lambda v_i.$$

Der Betrag davon ist

$$\left|\sum_{j=1}^{n} a_{ij}v_j\right| = |\lambda v_i| = \varrho(A)|v_i| = \varrho|v_i|. \tag{9.13}$$

Die beiden Gleichungen (9.12) und (9.13) zusammen ergeben die Gleichung

$$\left|\sum_{j=1}^{n} a_{ij}v_j\right| = \sum_{j=1}^{n} a_{ij}|v_j|.$$

Nach der verallgemeinerten Dreiecksungleichung Satz 9.3.2 folgt jetzt, dass es eine komplexe Zahl c vom Betrag 1 gibt derart, dass $v_j = |v_j|c = u_jc$. Insbesondere ist $v = cu$ und damit ist

$$\lambda v = Av = Acu = cAu = c\varrho(A)u = \varrho(A)v,$$

woraus $\lambda = \varrho(A)$ folgt. □

Satz 9.27. *Der Eigenraum einer positiven Matrix A zum Eigenwert $\varrho(A)$ ist eindimensional.*

Beweis. Sei u der bereits gefundene Eigenvektor von A zum Eigenwert $\varrho(A)$ und sei v ein weiterer, linear unabhängiger Eigenvektor zum gleichen Eigenwert. Da A und $\varrho(A)$ reell sind, sind auch die Vektoren $\Re v$ und $\Im v$ aus den Realteilen $\Re v_i$ oder den Imaginärteilen $\Im v_i$ Eigenvektoren. Beide Vektoren sind reelle Vektoren und mindestens einer davon ist mit u linear unabhängig. Wir dürfen daher annehmen, dass v ein linear unabhängiger Eigenvektor zum Eigenwert $\varrho(A)$ ist.

Weil wir wissen, dass u ein positiver Vektor ist, gibt es einen grösstmöglichen Faktor $c > 0$ derart, dass $u \geq cv$ oder $u \geq cv$. Insbesondere verschwindet mindestens eine Komponente von $u - cv$. Da u und v Eigenvektoren zum Eigenwert $\varrho(A)$ sind, ist

$$A(u - cv) = \varrho(A)(u - cv).$$

Der Vektor auf der rechten Seite hat mindestens eine verschwindende Komponente. Der Vektor auf der linken Seite ist nach Vergleichstrick Satz 9.20

$$A(u - cv) > 0,$$

alle seine Komponenten sind > 0. Dieser Widerspruch zeigt, dass die Annahme, es gäbe einen von u linear unabhängigen Eigenvektor zum Eigenwert $\varrho(A)$ nicht haltbar ist. □

Satz 9.28. *Der verallgemeinerte Eigenraum zum Eigenwert $\varrho(A)$ einer positiven Matrix A ist eindimensional. Ist u der Eigenvektor von A zum Eigenwert $\varrho(A)$ nach Satz 9.27 und p^t der entsprechende Eigenvektor A^t, dann ist*

$$\mathbb{R}^n = \langle u \rangle \oplus \{x \in \mathbb{R}^n \mid px = 0\} = \langle u \rangle \oplus \ker p$$

eine Zerlegung in invariante Unterräume von A.

Beweis. Die beiden Vektoren x und p sind beide positiv, daher ist das Produkt $pu \neq 0$. Insbesondere ist $u \notin \ker p$

Es ist klar, dass $A\langle u \rangle = \langle Au \rangle = \langle u \rangle$ ein invarianter Unterraum ist. Für einen Vektor $x \in \mathbb{R}^n$ mit $px = 0$ erfüllt das Bild Ax

$$p(Ax) = (pA)x = (A^t p^t)^t x = \varrho(A)(p^t)^t x = \varrho(A)px = 0,$$

somit ist $A \ker p \subset \ker p$. Beide Räume sind also invariante Unteräume.

$\ker p$ ist $(n-1)$-dimensional, $\langle u \rangle$ ist eindimensional und u ist nicht in $\ker p$ enthalten. Folglich spannen $\langle u \rangle$ und $\ker p$ den ganzen Raum auf.

Gäbe es einen weitern linear unabhängigen Vektor im verallgemeinerten Eigenraum von $\mathcal{E}_{\varrho(A)}$, dann müsste es auch einen solchen Vektor in $\ker p$ geben. Da $\ker p$ invariant ist, müsste es also auch einen weiteren Eigenvektor u_2 zum Eigenwert $\varrho(A)$ in $\ker p$ geben. Die beiden Vektoren u und u_1 sind dann beide Eigenvektoren, was nach Satz 9.27 nicht möglich ist. □

Die in den Sätzen 9.25 bis 9.28 gefundenen Resultate können wir folgt zusammengefasst werden:

Satz 9.29 (Perron-Frobenius)**.** *Sei A eine positive Matrix mit Spektralradius $\varrho(A)$. Dann gibt es einen positiven Eigenvektor zum Eigenwert $\varrho(A)$, mit geometrischer und algebraischer Vielfachheit* 1.

Beispiel. In der Google-Matrix mit freiem Willen nach (9.4) enthält den Term $((1-\alpha)/N)UU^t$. Die Matrix UU^t ist eine Matrix aus lauter Einsen, der Term ist also für $\alpha < 1$ eine positive Matrix. Die Google-Matrix ist daher eine positive Matrix. Nach dem Satz von Perron-Frobenius ist die Grenzverteilung eindeutig bestimmt. ○

Der Satz 9.29 von Perron-Frobenius kann auf primitive Matrizen verallgemeinert werden.

Satz 9.30. *Sei A ein primitive, nichtnegative Matrix. Dann ist $\varrho(A)$ der einzige Eigenwert vom Betrag $\varrho(A)$ und er hat geometrische und algebraische Vielfachheit* 1.

Beweis. Nach Voraussetzung gibt es ein n derart, dass $A^n > 0$. Für A^n gelten die Resultate von Satz 9.29.

XXX TODO □

9.4 Das Paradoxon von Parrondo

Das Paradoxon von Parrondo ist ein der Intuition widersprechendes Beispiel für eine Kombination von Spielen mit negativer Gewinnerwartung, deren Kombination zu einem Spiel mit positiver Gewinnerwartung führt. Die Theorie der Markov-Ketten und der zugehörigen Matrizen ermöglicht eine sehr einfache Analyse.

9.4.1 Die beiden Teilspiele

Das Spiel *A*

Das Spiel A besteht darin, eine Münze zu werfen. Je nach Ausgang gewinnt oder verliert der Spieler eine Einheit. Sei X die Zufallsvariable, die den gewonnen Betrag beschreibt. Für eine faire Münze ist die Gewinnerwartung in diesem Spiel natürlich $E(X) = 0$. Wenn die Wahrscheinlichkeit für einen Gewinn $\frac{1}{2} + e$ ist, dann muss die Wahrscheinlichkeit für einen Verlust $\frac{1}{2} - e$ sein, und die Gewinnerwartung ist $E(X) = 1 \cdot P(X = 1) + (-1) \cdot P(X = -1) = \frac{1}{2} + e + (-1)\left(\frac{1}{2} - e\right) = 2e$. Die Gewinnerwartung ist also genau dann negativ, wenn $e < 0$ ist.

Das Spiel *B*

Das zweite Spiel B ist etwas komplizierter, da der Spielablauf vom aktuellen Kapital K des Spielers abhängt. Wieder gewinnt oder verliert der Spieler eine Einheit, die Gewinnwahrscheinlichkeit hängt aber vom Dreierrest des Kapitals ab. Sei Y die Zufallsvariable, die den Gewinn beschreibt. Ist K durch drei teilbar, ist die Gewinnwahrscheinlichkeit $\frac{1}{10}$, andernfalls ist sie $\frac{3}{4}$. Formell ist

$$\begin{aligned} P(Y = 1|K \text{ durch 3 teilbar}) &= \frac{1}{10} \\ P(Y = 1|K \text{ nicht durch 3 teilbar}) &= \frac{3}{4} \end{aligned} \tag{9.14}$$

Insbesondere ist die Wahrscheinlichkeit für einen Gewinn in zwei der Fälle recht gross, in einem Fall aber sehr klein.

Übergangsmatrix im Spiel *B*

Für den Verlauf des Spiels spielt nur der Dreierrest des Kapitals eine Rolle. Es gibt daher drei mögliche Zustände 0, 1 und 2. In einem Spielzug finde ein Übergang in einen anderen Zustand statt, der Eintrag b_{ij} ist die Wahrscheinlichkeit

$$b_{ij} = P(K \equiv i|K \equiv j),$$

dass ein Übergang vom Zustand j in den Zustand i stattfindet. Die Matrix ist

$$B = \begin{pmatrix} 0 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{10} & 0 & \frac{1}{4} \\ \frac{9}{10} & \frac{3}{4} & 0 \end{pmatrix}.$$

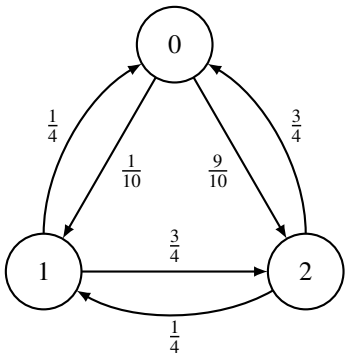


Abbildung 9.11: Zustandsdiagramm für das Spiel B, Zustände sind die Dreierreste des Kapitals.

Gewinnerwartung in einem Einzelspiel B

Die Gewinnerwartung einer einzelnen Runde des Spiels B hängt natürlich ebenfalls vom Ausgangskapital ab. Mit den Wahrscheinlichkeiten von (9.14) findet man die Gewinnerwartung

$$\begin{aligned} E(Y|K \text{ durch 3 teilbar}) &= 1 \cdot P(Y = 1|K \equiv 0 \mod 3) + (-1) \cdot P(Y = -1|K \equiv 0 \mod 3) \\ &= \frac{1}{10} - \frac{9}{10} = -\frac{8}{10} \\ E(Y|K \text{ nicht durch 3 teilbar}) &= 1 \cdot P(Y = 1|K \not\equiv 0 \mod 3) + (-1) \cdot P(Y = -1|K \not\equiv 0 \mod 3) \\ &= \frac{3}{4} - \frac{1}{4} = \frac{1}{2}. \end{aligned} \tag{9.15}$$

Falls K durch drei teilbar ist, muss der Spieler also mit einem grossen Verlust rechnen, andernfalls mit einem moderaten Gewinn.

Ohne weiteres Wissen über das Anfangskapital ist es zulässig anzunehmen, dass die drei möglichen Reste die gleiche Wahrscheinlichkeit haben. Die Gewinnerwartung in diesem Fall ist dann

$$\begin{aligned} E(Y) &= E(Y|K \text{ durch 3 teilbar}) \cdot \frac{1}{3} + E(Y|K \text{ nicht durch 3 teilbar}) \cdot \frac{2}{3} \\ &= -\frac{8}{10} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} = -\frac{8}{30} + \frac{10}{30} = \frac{2}{30} = \frac{1}{15}. \end{aligned} \tag{9.16}$$

Unter der Annahme, dass alle Reste die gleiche Wahrscheinlichkeit haben, ist das Spiel also ein Gewinnspiel.

Die Berechnung der Gewinnerwartung in einem Einzelspiel kann man wie folgt formalisieren. Die Matrix B gibt die Übergangswahrscheinlichkeiten zwischen verschiedenen Zuständen. Die Matrix

$$G = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$$

gibt die Gewinne an, die bei einem Übergang anfallen. Die Matrixelemente $g_{ij}b_{ij}$ des Hadamard-Produktes $G \odot B$ von G mit B enthält in den Spalten die Gewinnerwartungen für die einzelnen

Übergänge aus einem Zustand. Die Summe der Elemente der Spalte j enthält die Gewinnerwartung

$$E(Y|K \equiv j) = \sum_{i=0}^{2} g_{ij} b_{ij}$$

für einen Übergang aus dem Zustand j. Man kann dies auch als einen Zeilenvektor schreiben, der durch Multiplikation der Matrix $G \odot B$ mit dem Zeilenvektor $U^t = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$ entsteht:

$$\begin{pmatrix} E(Y|K \equiv 0) & E(Y|K \equiv 1) & E(Y|K \equiv 2) \end{pmatrix} = U^t G \odot B.$$

Die Gewinnerwartung ist dann das Produkt

$$E(Y) = \sum_{i=0}^{2} E(Y|K \equiv i) p_i = U^t (G \odot B) p.$$

Tatsächlich ist

$$G \odot B = \begin{pmatrix} 0 & -\frac{1}{4} & \frac{3}{4} \\ \frac{1}{10} & 0 & -\frac{1}{4} \\ -\frac{9}{10} & \frac{3}{4} & 0 \end{pmatrix} \quad \text{und} \quad U^t G \odot B = \begin{pmatrix} -\frac{8}{10} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Dies stimmt mit den Erwartungswerten in (9.15) überein. Die gesamte Geinnerwartung ist dann

$$(G \odot B) \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} -\frac{8}{10} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \frac{1}{3} U = \frac{1}{3} \left(-\frac{8}{10} + \frac{1}{2} + \frac{1}{2} \right) = \frac{1}{3} \cdot \frac{2}{10} = \frac{1}{15}, \tag{9.17}$$

dies stimmt mit (9.16) überrein.

Das wiederholte Spiel B

Natürlich spielt man das Spiel nicht nur einmal, sondern man wiederholt es. Es ist verlockend anzunehmen, dass die Dreierreste 0, 1 und 2 des Kapitals immer noch gleich wahrscheinlich sind. Dies braucht jedoch nicht so zu sein. Wir prüfen die Hypothese daher, indem wir die Wahrscheinlichkeit für die verschiedenen Dreierreste des Kapitals in einem interierten Spiels ausrechnen.

Das Spiel kennt die Dreierreste als die drei für das Spiel ausschlaggebenden Zuständen. Das Zustandsdiagramm 9.11 zeigt die möglichen Übergänge und ihre Wahrscheinlichkeiten, die zugehörige Matrix ist

$$B = \begin{pmatrix} 0 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{10} & 0 & \frac{1}{4} \\ \frac{9}{10} & \frac{3}{4} & 0 \end{pmatrix}$$

Die Matrix B ist nicht negativ und man kann nachrechnen, dass $B^2 > 0$ ist. Damit ist die Perron-Frobenius-Theorie von Abschnitt 9.3 anwendbar.

Ein Eigenvektor zum Eigenwert 1 kann mit Hilfe des Gauss-Algorithmus gefunden werden:

$$\begin{bmatrix} -1 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{10} & -1 & \frac{1}{4} \\ \frac{9}{10} & \frac{3}{4} & -1 \end{bmatrix} \to \begin{bmatrix} 1 & -\frac{1}{4} & -\frac{3}{4} \\ 0 & -\frac{39}{40} & \frac{13}{40} \\ 0 & \frac{39}{40} & -\frac{13}{40} \end{bmatrix} \to \begin{bmatrix} 1 & -\frac{1}{4} & -\frac{3}{4} \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & -\frac{5}{6} \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix}$$

Daraus liest man einen möglichen Lösungsvektor mit den Komponenten 5, 2 und 6 ab. Wir suchen aber einen Eigenvektor, der als Wahrscheinlichkeitsverteilung dienen kann. Dazu müssen sich die Komponente zu 1 summieren, was man durch normieren in der l^1-Norm erreichen kann:

$$p = \begin{pmatrix} P(K \equiv 0) \\ P(K \equiv 1) \\ P(K \equiv 2) \end{pmatrix} = \frac{1}{5+2+6} \begin{pmatrix} 5 \\ 2 \\ 6 \end{pmatrix} = \frac{1}{13} \begin{pmatrix} 5 \\ 2 \\ 6 \end{pmatrix} \approx \begin{pmatrix} 0.3846 \\ 0.1538 \\ 0.4615 \end{pmatrix}. \tag{9.18}$$

Die Hypothese, dass die drei Reste gleich wahrscheinlich sind, ist also nicht zutreffend.

Die Perron-Frobenius-Theorie sagt, dass sich die Verteilung 9.18 nach einiger Zeit einstellt. Wir können jetzt auch die Gewinnerwartung in einer einzelnen Runde des Spiels ausgehend von dieser Verteilung der Reste des Kapitals berechnen. Dazu brauchen wir zunächst die Wahrscheinlichkeiten für Gewinn oder Verlust, die wir mit dem Satz über die totale Wahrscheinlichkeit nach

$$\begin{aligned} P(Y=+1) &= P(Y=+1|K\equiv 0)\cdot P(K\equiv 0) + P(Y=+1|K\equiv 1)\cdot P(K\equiv 1) + P(Y=+1|K\equiv 2)\cdot P(K\equiv 2) \\ &= \frac{1}{10}\cdot\frac{5}{13} + \frac{3}{4}\cdot\frac{2}{13} + \frac{3}{4}\cdot\frac{6}{13} \\ &= \frac{1}{13}\left(\frac{1}{2} + \frac{3}{2} + \frac{9}{2}\right) = \frac{13}{26} = \frac{1}{2} \\ P(Y=-1) &= P(Y=-1|K\equiv 0)\cdot P(K\equiv 0) + P(Y=-1|K\equiv 1)\cdot P(K\equiv 1) + P(Y=-1|K\equiv 2)\cdot P(K\equiv 2) \\ &= \frac{9}{10}\cdot\frac{5}{13} + \frac{1}{4}\cdot\frac{2}{13} + \frac{1}{4}\cdot\frac{6}{13} \\ &= \frac{1}{13}\left(\frac{9}{2} + \frac{1}{2} + \frac{3}{2}\right) = \frac{1}{2} \end{aligned}$$

berechnen können. Gewinn und Verlust sind also gleich wahrscheinlich, das Spiel B ist also ebenfalls fair.

Auch diese Gewinnwahrscheinlichkeit kann etwas formeller mit dem Hadamard-Produkt berechnet werden:

$$U^t(G \odot B)p = \begin{pmatrix} -\frac{8}{10} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \frac{1}{13} \begin{pmatrix} 5 \\ 2 \\ 6 \end{pmatrix} = -\frac{8}{10}\cdot\frac{5}{13} + \frac{1}{2}\cdot\frac{2}{13} + \frac{1}{2}\cdot\frac{6}{13} = \frac{1}{26}(-8+2+6) = 0,$$

wie erwartet.

Das modifizierte Spiel $\tilde{B}$

Wir modifizieren jetzt das Spiel B derart, dass die Wahrscheinlichkeiten für Gewinn um ε verringert werden und die Wahrscheinlichkeiten für Verlust um ε vergrössert werden. Die Übergangsmatrix des modifzierten Spiels $\tilde{B}$ ist

$$\tilde{B} = \begin{pmatrix} 0 & \frac{1}{4}+\varepsilon & \frac{3}{4}-\varepsilon \\ \frac{1}{10}-\varepsilon & 0 & \frac{1}{4}+\varepsilon \\ \frac{9}{10}+\varepsilon & \frac{3}{4}-\varepsilon & 0 \end{pmatrix} = B + \varepsilon \underbrace{\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}}_{F}$$

Wir wissen bereits, dass der Vektor p von (9.18) als stationäre Verteilung Eigenvektor zum Eigenwert B ist, wir versuchen jetzt in erster Näherung die modifizierte stationäre Verteilung $p_\varepsilon = p + \varepsilon p_1$ des modifizierten Spiels zu bestimmen.

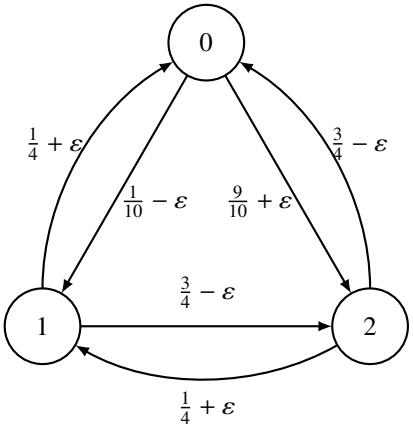


Abbildung 9.12: Zustandsdiagramm für das modifizerte Spiel $\tilde{B}$, Zustände sind die Dreierreste des Kapitals. Gegenüber dem Spiel B (Abbildung 9.11) sind die Wahrscheinlichkeiten für Verlust um ε vergrössert und die Wahrscheinlichkeiten für Gewinn um ε verkleinert worden.

Gewinnerwartung im modifizierten Einzelspiel

Die Gewinnerwartung aus den verschiedenen Ausgangszuständen kann mit Hilfe des Hadamard-Produktes berechnet werden. Wir berechnen dazu zunächst

$$G \odot \tilde{B} = G \odot (B + \varepsilon F) = G \odot B + \varepsilon G \odot F \quad \text{mit} \quad G \odot F = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Nach der früher dafür gefundenen Formel ist

$$\begin{aligned} \Big(E(Y|K \equiv 0) \quad E(Y|K \equiv 1) \quad E(Y|K \equiv 2)\Big) &= U^t(G \odot \tilde{B}) \\ &= U^t(G \odot B) + \varepsilon U^t(G \odot F) \\ &= \Big(-\tfrac{8}{10} \quad \tfrac{1}{2} \quad \tfrac{1}{2}\Big) + 2\varepsilon U^t \\ &= \Big(-\tfrac{8}{10} + 2\varepsilon \quad \tfrac{1}{2} + 2\varepsilon \quad \tfrac{1}{2} + 2\varepsilon\Big). \end{aligned}$$

Unter der Annahme gleicher Wahrscheinlichkeiten für die Ausgangszustände, erhält man die Gewinnerwartung

$$\begin{aligned} E(Y) &= U^t(G \odot \tilde{B}) \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \\ &= U^t(G \odot B)\frac{1}{3}U + \varepsilon U^t(G \odot F)\frac{1}{3}U \\ &= \frac{1}{15} + 2\varepsilon \end{aligned}$$

unter Verwendung der in (9.17) berechneten Gewinnerwartung für das Spiel B.

Iteration des modifizierten Spiels

Der Gaussalgorithmus liefert nach einiger Rechnung, die man am besten mit einem Computeralgebrasystem durchführt,

$$\left(\begin{array}{ccc} -1 & \frac{1}{4}+\varepsilon & \frac{3}{4}-\varepsilon \\ \frac{1}{10}-\varepsilon & -1 & \frac{1}{4}+\varepsilon \\ \frac{9}{10}+\varepsilon & \frac{3}{4}-\varepsilon & -1 \end{array}\right) \to \left(\begin{array}{ccc} 1 & 0 & -\frac{65-40\varepsilon+80\varepsilon^2}{78+12\varepsilon+80\varepsilon^2} \\ 0 & 0 & -\frac{26+12\varepsilon+80\varepsilon^2}{78+12\varepsilon+80\varepsilon^2} \\ 0 & 0 & 0 \end{array}\right),$$

woraus man die Lösung

$$p = \begin{pmatrix} 65-40\varepsilon+80\varepsilon^2 \\ 26+12\varepsilon+80\varepsilon^2 \\ 78+12\varepsilon+80\varepsilon^2 \end{pmatrix}$$

ablesen kann. Allerdings ist dies keine Wahrscheinlichkeitsverteilung, wir müssen dazu wieder normieren. Die Summe der Komponenten ist

$$\|p\|_1 = 169 - 16\varepsilon + 240\varepsilon^2.$$

Damit bekommen wir für die Lösung bis zur ersten Ordnung

$$p_\varepsilon = \frac{1}{169-16\varepsilon+240\varepsilon^2}\begin{pmatrix} 65-40\varepsilon+80\varepsilon^2 \\ 26+12\varepsilon+80\varepsilon^2 \\ 78+12\varepsilon+80\varepsilon^2 \end{pmatrix} = \frac{1}{13}\begin{pmatrix} 5 \\ 2 \\ 6 \end{pmatrix} + \frac{\varepsilon}{2197}\begin{pmatrix} -440 \\ 188 \\ 252 \end{pmatrix} + O(\varepsilon^2).$$

Man beachte, dass der konstante Vektor der ursprüngliche Vektor p für das Spiel B ist. Der lineare Term ist ein Vektor, dessen Komponenten sich zu 1 summieren, in erster Ordnung ist also die l^1-Norm des Vektors wieder $\|p_\varepsilon\|_1 = 0 + O(\varepsilon^2)$.

Mit den bekannten Wahrscheinlichkeiten kann man jetzt die Gewinnerwartung in einem einzeln Spiel ausgehend von der Verteilung p_ε berechnen. Dazu braucht man das Hadamard-Produkt

$$G \odot \tilde{B} = G = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} \odot \begin{pmatrix} 0 & \frac{1}{4}+\varepsilon & \frac{3}{4}-\varepsilon \\ \frac{1}{10}-\varepsilon & 0 & \frac{1}{4}+\varepsilon \\ \frac{9}{10}+\varepsilon & \frac{3}{4}-\varepsilon & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{4}-\varepsilon & \frac{3}{4}-\varepsilon \\ \frac{1}{10}-\varepsilon & 0 & -\frac{1}{4}-\varepsilon \\ -\frac{9}{10}-\varepsilon & \frac{3}{4}-\varepsilon & 0 \end{pmatrix}$$

Wie früher ist der erwartete Gewinn

$$\begin{aligned} E(Y) &= U^t(G \odot \tilde{B})p_\varepsilon \\ &= \left(-\tfrac{3}{10}-2\varepsilon \quad \tfrac{1}{2}-2\varepsilon \quad \tfrac{1}{2}-2\varepsilon\right) p_\varepsilon \\ &= -\varepsilon \cdot \frac{294-48\varepsilon+480\varepsilon^2}{169-16\varepsilon+240\varepsilon^2} = -\frac{294}{169}\varepsilon + O(\varepsilon^2). \end{aligned}$$

Insbesondere ist also die Gewinnerwartung negativ für nicht zu grosse $\varepsilon > 0$. Das Spiel ist also ein Verlustspiel.

9.4.2 Kombination der Spiele

Jetzt werden die beiden Spiele A und B zu einem neuen Spiel kombiniert. Für das Spiel A haben wir bis jetzt keine Übergansmatrix aufgestellt, da das Kapital darin keine Rolle spielt. Um die beiden Spiele kombinieren zu können brauchen wir aber die Übergansmatrix für die drei Zustände $K \equiv 0, 1, 2$. Sie ist

$$A = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

Das Spiel C

In jeder Durchführung des Spiels wird mit einem Münzwurf entschieden, ob Spiel A oder Spiel B gespielt werden soll. Mit je Wahrscheinlichkeit $\frac{1}{2}$ werden also die Übergansmatrizen A oder B verwendet:

$$P(K \equiv i | K \equiv j) = A \cdot P(\text{Münzwurf Kopf}) + B \cdot P(\text{Münzwurf Kopf}) = \frac{1}{2}(A+B) = \begin{pmatrix} 0 & \frac{3}{8} & \frac{5}{8} \\ \frac{3}{10} & 0 & \frac{2}{8} \\ \frac{7}{10} & \frac{5}{8} & 0 \end{pmatrix}$$

Die Gewinnerwartung in einem Einzelspiel ist

$$\begin{aligned} E(Y) &= U^t (G \odot C) \frac{1}{3} U \\ &= U^t \begin{pmatrix} 0 & -\frac{3}{8} & \frac{5}{8} \\ \frac{3}{10} & 0 & -\frac{3}{8} \\ -\frac{7}{10} & \frac{5}{8} & 0 \end{pmatrix} \frac{1}{3} U \\ &= \begin{pmatrix} -\frac{2}{5} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \frac{1}{3} U = \frac{1}{3}\left(-\frac{2}{5} + \frac{1}{4} + \frac{1}{4}\right) = -\frac{1}{30} \end{aligned}$$

Das Einzelspiel ist also ein Verlustspiel.

Das iterierte Spiel C

Für das iterierte Spiel muss man wieder den Eigenvektor von C zum Eigenwert 1 finden, die Rechnung mit dem Gauss-Algorithmus liefert

$$p = \frac{1}{709} \begin{pmatrix} 245 \\ 180 \\ 284 \end{pmatrix}.$$

Damit kann man jetzt die Gewinnwahrscheinlichkeit im iterierten Spiel berechnen, es ist

$$\begin{aligned} E(Y) &= U^t (G \odot C) p \\ &= \begin{pmatrix} -\frac{2}{5} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \frac{1}{709} \begin{pmatrix} 245 \\ 180 \\ 84 \end{pmatrix} \\ &= \frac{-2 \cdot 49 + 45 + 71}{709} = \frac{18}{709}, \end{aligned}$$

Das iteriert Spiel B ist also ein Gewinnspiel! Obwohl die Spiele A und B für sich alleine in der iterierten Form keine Gewinnspiele sind, ist das kombinierte Spiel, wo man zufällig die beiden Spiel verbindet immer ein Gewinnspiel.

Man kann statt des Spiels B auch das modifizierte Spiel $\tilde{B}$ verwenden, welches für kleine $\varepsilon > 0$ ein Verlustspiel ist. Die Analyse lässt sich in der gleichen Weise durchführen und liefert wieder, dass für nicht zu grosses ε das kombinierte Spiel ein Gewinnspiel ist.

Kapitel 10

Anwendungen in Kryptographie und Codierungstheorie

Die algebraische Theorie der endlichen Körper hat sich als besonders nützliche herausgestellt in der Krypographie. Die Eigenschaften dieser Körper sind reichhaltig genug, um kryptographsch widerstandsfähige Algorithmen zu liefern, die auch in ihrer Stärke beliebig skaliert werden können. Gleichzeitig liefert die Algebra auch eine effiziente Implementierung. In diesem Abschnitt soll dies an einigen Beispielen gezeigt werden.

10.1 Arithmetik für die Kryptographie

Die Algorithmen der mathematischen Kryptographie basieren auf den Rechenoperationen in grossen, aber endlichen Körpern. Für die Division liefert der euklidische Algorithmus eine Methode, der in so vielen Schritten die Inverse findet, wie Dividend und Divisor Binärstellen haben. Dies ist weitgehend optimal.

Die Division ist umkehrbar, in der Kryptographie strebt man aber an, Funktionen zu konstruieren, die nur mit grossem Aufwand umkehrbar sind. Eine solche Funktion ist das Potenzieren in einem endlichen Körper. Die Berechnung von Potenzen durch wiederholte Multiplikation ist jedoch prohibitiv aufwendig, daher ist ein schneller Potenzierungsalgorithmus nötig, der in Abschnitt 10.1.1 beschrieben wird. Bei der Verschlüsselung grosser Datenmengen wie zum Beispiel bei der Verschlüsselung ganzer Harddisks mit Hilfe des AES-Algorithmus kommt es auf die Geschwindigkeit auch der elementarsten Operationen in den endlichen Körpern an. Solche Methoden werden in den Abschnitten 10.1.2 und 10.1.3 besprochen.

10.1.1 Potenzieren

Wir gehen davon aus, dass wir einen schnellen Algorithmus zur Berechnung des Produktes zweier Elemente a, b in einer beliebigen Gruppe G haben. Die Gruppe G kann die Multiplikation der ganzen oder reellen Zahlen sein, dies wird zum Beispiel in Implementation der Potenzfunktion verwendet. Für kryptographische Anwendungen ist G die multiplikative Gruppe eines endlichen Körpers oder eine elliptische Kurve.

Zur Berechnung von a^k sind bei einer naiven Durchführung des Algorithmus $k-1$ Multiplikationen nötig, immer sofort gefolgt von einer Reduktion $\bmod p$ um sicherzustellen, dass die Resultate nicht zu gross werden. Ist l die Anzahl der Binärstellen von k, dann benötigt dieser naive Algorithmus $O(2^l)$ Multiplikationen, die Laufzeit wächst also exponentiell mit der Bitlänge von k an. Der nachfolgend beschriebene Algorithmus reduziert die Laufzeit auf die $O(l)$.

Zunächst schreiben wir den Exponenten k in binärer Form als

$$k = k_l 2^l + k_{l-1} 2^{l-1} + \ldots k_2 2^2 + k_1 2^1 k_0 2^0.$$

Die Potenz a^k kann dann geschrieben werden als

$$a^k = a^{k_l 2^l} \cdot a^{k_{l-1} 2^{l-1}} \cdot \cdots \cdot a^{k_2 2^2} \cdot a^{k_1 2^1} \cdot a^{k_0 2^0}$$

Nur diejenigen Faktoren tragen etwas bei, für die $k_i \neq 0$ ist. Die Potenz kann man daher auch schreiben als

$$a^k = \prod_{k_i \neq 0} a^{2^i}.$$

Es sind also nur so viele Faktoren zu berücksichtigen, wie k Binärstellen 1 hat.

Die einzelnen Faktoren a^{2^i} können durch wiederholtes Quadrieren erhalten werden:

$$a^{2^i} = a^{2 \cdot 2^{i-1}} = (a^{2^{i-1}})^2,$$

also durch maximal $l-1$ Multiplikationen. Wenn k keine Ganzzahl ist sondern binäre Nachkommastellen hat, also

$$k = k_l 2^l + \cdots + k_1 2^1 + k_0 2^0 + k_{-1} 2^{-1} + k_{-2} 2^{-2} + \ldots,$$

dann können die Potenzen $a^{2^{-i}}$ durch wiederholtes Wurzelziehen

$$a^{2^{-i}} = a^{\frac{1}{2} \cdot 2^{-i+1}} = \sqrt{a^{2^{-i+1}}}$$

gefunden werden. Die Berechnung der Quadratwurzel lässt sich in Hardware effizient implementieren.

Algorithmus 10.1. *Der folgende Algorithmus berechnet a^k in $O(\log_2(k))$ Multiplikationen*

1. *Initialisiere $p = 1$ und $q = a$*
2. *Falls k ungerade ist, setze $p := p \cdot q$*
3. *Setze $q := q^2$ und $k := k/2$, wobei die ganzzahlige Division durch 2 am effizientesten als Rechtsshift implementiert werden kann.*
4. *Falls $k > 0$, fahre weiter bei 2.*

Beispiel. Die Berechnung von 1.1^{17} mit diesem Algorithmus ergibt

1. $p = 1$, $q = 1.1$
2. k ist ungerade: $p := 1.1$
3. $q := q^2 = 1.21$, $k := 8$
4. k ist gerade

5. $q := q^2 = 1.4641$, $k := 4$
6. k ist gerade
7. $q := q^2 = 2.14358881$, $k := 2$
8. k ist gerade
9. $q := q^2 = 4.5949729863572161$, $k := 1$
10. k ist ungerade: $p := 1.1 \cdot p = 5.05447028499293771$
11. $k := 0$

Multiplikationen sind nur nötig in den Schritten 3, 5, 7, 9, 10, es werden also genau 5 Multiplikationen ausgeführt. ○

10.1.2 Rechenoperationen in $\mathbb{F}_p$

Die Multiplikation macht aus zwei Faktoren a und b ein Resultat mit Bitlänge $\log_2 a + \log_2 b$, die Bitlänge wird also typischerweise verdoppelt. In $\mathbb{F}_p$ muss anschliessend das Resultat $\bmod p$ reduziert werden, so dass die Bitlänge wieder höchstens $\log_2 p$ ist. In folgenden soll gezeigt werden, dass dieser Speicheraufwand für eine Binärimplementation deutlich reduziert werden kann, wenn die Reihenfolge der Operationen modifiziert wird.

Für die Multiplikation von $41 \cdot 47$ rechnet man im Binärsystem

$$
\begin{array}{r}
101001 \cdot 101111 \\
\hline
\qquad 101111 \\
\quad 101111 \\
101111 \\
\hline
11110000111 \\
\hline
\end{array}
$$

In $\mathbb{F}_{53}$ muss im Anschluss Modulo $p = 53$ reduziert werden.

Der Speicheraufwand entsteht zunächst dadurch, dass durch die Multiplikation mit 2 die Summanden immer länger werden. Man kann den die Sumanden kurz halten, indem man jedesmal, wenn der Summand nach der Multiplikation mit 2 grösser als p geworden ist, p subtrahiert (Abbildung 10.1). Ebenso kann bei nach jeder Addition das bereits reduzierten zweiten Faktors wieder reduziert werden. Die Anzahl der nötigen Reduktionsoperationen wird durch diese frühzeitig durchgeführten Reduktionen nicht teurer als bei der Durchführung des Divisionsalgorithmus.

Es ist also möglich, mit gleichem Aufwand an Operationen aber mit halbe Speicherplatzbedarf die Multiplikationen in $\mathbb{F}_p$ durchzuführen. Die Platzeinsparung ist besonders bei Implementationen in Hardware hilfreich, wo on-die Speicherplatz teuer sein kann.

10.1.3 Rechenoperationen in $\mathbb{F}_{2^l}$

Von besonderem praktischem Interesse sind die endlichen Körper $\mathbb{F}_{2^l}$. Die arithmetischen Operationen in diesen Körpern lassen sich besonders effizient in Hardware realisieren.

Multiplikation mit 2	Reduktion?	reduziert	Summanden	Summe	reduziert
101111		101111	101111	101111	101111
101111	1011110	101001			
101111	0111010	011101			
101111	0001010	000101	000101	110100	110100
101111	0010100	001010			
101111	0101000	010100	010100	1001000	10011 = 19

Abbildung 10.1: Multiplikation von $41 = 101001_2$ mit $47 = 101111_2$, Reduktion nach jeder Multiplikation mit 2: falls das Resultat $> p$ ist, wie in den rot markierten Zeilen $p = 53 = 110101_2$ durchgeführt. Bei der Bildung der Summe wird ebenfalls in jedem Schritt falls nötig reduziert, angezeigt durch die roten Zahlen in der zweitletzten Spalte. Die Anzahl der Subtraktionen, die für die Reduktionen nötig sind, ist von der selben Grössenordnung wie bei der Durchführung des Divisionsalgorithmus.

Zahldarstellung

Ein endlicher Körper $\mathbb{F}_{2^l}$ ist definiert durch ein irreduzibles Polynom in $\mathbb{F}_2[X]$ vom Grad 2^l

$$m(X) = X^l + m_{l-1}X^{l-1} + m_{l-2}X^{l-2} + \cdots + m_2X^2 + m_1X + m_0$$

gegeben. Ein Element in $\mathbb{F}_2[X]/(m)$ kann dargestellt werden durch ein Polynom vom Grad $l-1$, also durch

$$a = a_{l-1}X^{l-1} + a_{l-2}X^{l-2} + \cdots + a_2X^2 + a_1X + a_0.$$

In einer Maschine kann eine Zahl also als eine Bitfolge der Länge l dargestellt werden.

Addition

Die Addition in $\mathbb{F}_2$ ist in Hardware besonders leicht zu realisieren. Die Addition ist die XOR-Operation, die Multiplikation ist die UND-Verknüfung. Ausserdem stimmen in $\mathbb{F}_2$ Addition und Subtraktion überein.

Die Addition zweier Polynome erfolgt komponentenweise. Die Addition von zwei Elemente von $\mathbb{F}_{2^l}$ kann also durch die bitweise XOR-Verknüpfung der Darstellungen der Summanden erfolgen. Diese Operation ist in einem einzigen Maschinenzyklus realisierbar. Die Subtraktion, die für die Reduktionsoperation module $m(X)$ nötig ist, ist mit der Addition identisch.

Multiplikation

Die Multiplikation zweier Polynome benötigt zunächst die Multiplikation mit X, wodurch der Grad des Polynoms ansteigt und möglicherweise so gross wird, dass eine Reduktionsoperation modulo $m(X)$ nötig wird. Die Reduktion wird immer dann nötig, wenn der Koeffizient von X^l nicht 0 ist. Der Koeffizient kann dann zum Verschwinden gebracht werden, indem $m(X)$ addiert wird.

In Abbildung 10.2 wird gezeigt, wie die Reduktion erfolgt, wenn die Multiplikation mit X, also der Shift nach links, einen Überlauf ergibt. Das Minimalpolynom $m(X) = X^8 + X^4 + X^3 + X + 1$ bedeutet, dass in $\mathbb{F}_{2^l}$ $X^8 = X^4 + X^3 + X + 1$ gilt, so dass man das Überlaufbit durch $X^4 + X^3 + X + 1$ ersetzen und addieren kann.

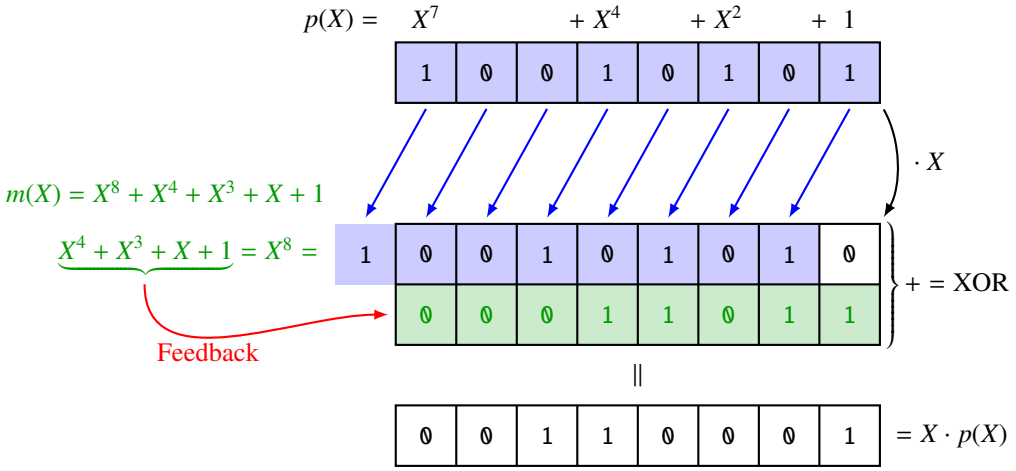


Abbildung 10.2: Implementation der Multiplikation mit X in einem endlichen Körper $\mathbb{F}_{2^l}$ mit dem Minimalpolynom $m(X) = X^8 + X^4 + X^3 + X^+1$ als Feedback-Schieberegister.

Ein Produktes $p(X) \cdot q(X)$, wobei $p(X)$ und $q(X)$ Repräsentaten von Elementen $\mathbb{F}_{2^l}$ sind, kann jetzt wie folgt berechnet werden. Mit dem Schieberegister werden die Vielfachen $X^k \cdot p(X)$ für $k = 0, \ldots, l-1$ berechnet. Diejenigen Vielfachen, für die der Koeffizient von X^k in $q(X)$ von 0 verschieden ist werden aufsummiert und ergeben das Produkt. Der Prozess in Abbildung 10.3 dargestellt.

10.2 Kryptographie und endliche Körper

10.2.1 Potenzen in $\mathbb{F}_p$ und diskreter Logarithmus

Für kryptographische Anwendungen wird eine einfach zu berechnende Funktion benötigt, die ohne zusätzliches Wissen, üblicherweise der Schlüssel genannt, nicht ohne weiteres umkehrbar ist. Die arithmetischen Operationen in einem endlichen Körper sind mit geringem Aufwand durchführbar. Für die "schwierigste" Operation, die Division, steht der euklidische Algorithmus zur Verfügung.

Die nächstschwierigere Operation ist die Potenzfunktion. Für $g \in \mathbb{k}$ und $a \in \mathbb{N}$ ist die Potenz $g^a \in \mathbb{k}$ natürlich durch die wiederholte Multiplikation definiert. In der Praxis werden aber g und a Zahlen mit vielen Binärstellen sein, die die wiederholte Multiplikation ist daher sicher nicht effizient, das Kriterium der einfachen Berechenbarkeit scheint also nicht erfüllt. Der folgende Algorithmus berechnet die Potenz in $O(\log_2 a)$ Multiplikationen.

Algorithmus 10.2 (Divide-and-conquer). *Sei $a = a_0 + a_1 2^1 + a_2 2^2 + \cdots + a_k 2^k$ die Binärdarstellung der Zahl a.*

1. *setze $f = g$, $x = 1$, $i = 0$*
2. *solange $i \geq k$ ist, führe aus*
   (a) *falls $a_i = 1$ setze $x := x \cdot f$*
   (b) *$i := i + 1$ und $f := f \cdot f$*

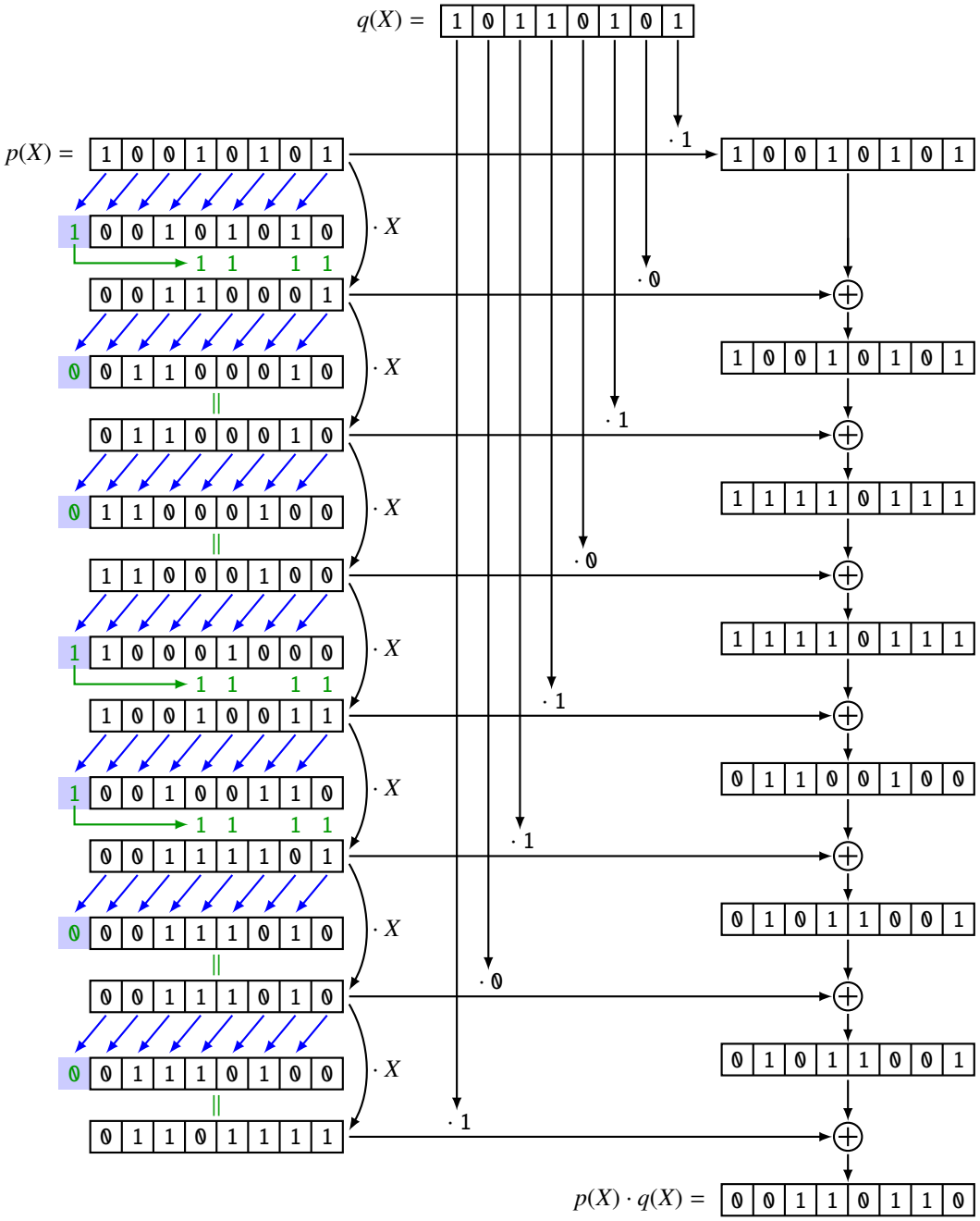


Abbildung 10.3: Multiplikation zweier Elemente von $\mathbb{F}_{2^l}$. Mit Hilfe des Schieberegisters am linken Rand werden die Produkte $X \cdot p(X)$, $X^2 \cdot p(X), \ldots, X^7 \cdot p(X)$ nach der in Abbildung 10.2 dargestellten Methode berechnet. Am rechten Rand werden diejenigen $X^k \cdot p(X)$ aufaddiert, für die der X^k-Koeffizient von $q(X)$ von 0 verschieden ist.

Die Potenz $x = g^a$ kann so in $O(\log_2 a)$ Multiplikationen berechnet werden.

Beweis. Die Initalisierung in Schritt 1 stellt sicher, dass x den Wert g^0 hat. Schritt 2b stellt sicher, dass die Variable f immer den Wert g^{2^i} hat. Im Schritt 2a wird zu x die Potenz $g^{a_i 2^i}$ hinzumultipliziert. Am Ende des Algorithmus hat daher x den Wert

$$x = g^{a_0 2^0} \cdot g^{a_1 2^1} \cdot g^{a_2 2^2} \cdot \ldots \cdot 2^{a_k 2^k} = g^{a_0 + a_1 2 + a_2 2^2 + \cdots + a_k 2^k} = g^a.$$

Die Schleife wird $\lfloor 1 + \log_2 ab \rfloor$ mal durchlaufen. In jedem Fall wird auf jeden Fall die Multiplikation in Schritt 2b durchgeführt und im schlimmsten Fall auch noch die Multiplikation in Schritt 2a. Es werden also nicht mehr als $2\lfloor 1 + \log_2 a \rfloor = O(\log_2 a)$ Multiplikationen durchgeführt. □

Beispiel. Man berechne die Potenz 7^{2021} in $\mathbb{F}_p$. Die Binärdarstellung von 2021 ist $2021_{10} = \texttt{11111100101}_2$. Wir stellen die nötigen Operationen des Algorithmus 10.2 in der folgenden Tabelle

i	f	a_i	x
0	7	1	7
1	49	0	7
2	1110	1	24
3	486	0	24
4	1234	0	24
5	667	1	516
6	785	1	977
7	418	1	430
8	439	1	284
9	362	1	819
10	653	1	333

Daraus liest man ab, dass $7^{2021} = 333 \in \mathbb{F}_{1291}$. ○

Die Tabelle suggeriert, dass die Potenzen von g "wild", also scheinbar ohne System in $\mathbb{F}_p$ herumspringen. Dies deutet an, dass die Umkehrung der Exponentialfunktion in $\mathbb{F}_p$ schwierig ist. Die Umkehrfunktion der Exponentialfunktion, die Umkehrfunktion von $x \mapsto g^x$ in $\mathbb{F}_p$ heisst der *diskrete Logarithmus*. Tatsächlich ist der diskrete Logarithmus ähnlich schwierig zu bestimmen wie das Faktorisieren von Zahlen, die das Produkt grosser Primafaktoren ähnlicher Grössenordnung wie p sind. Die Funktion $x \mapsto g^x$ ist die gesuchte, schwierig zu invertierende Funktion.

Auf dern ersten Blick scheint der Algorithmus 10.2 den Nachteil zu haben, dass erst die Binärdarstellung der Zahl a ermittelt werden muss. In einem Computer ist dies aber normalerweise kein Problem, da a im Computer ohnehin binär dargestellt ist. Die Binärziffern werden in der Reihenfolge vom niederwertigsten zum höchstwertigen Bit benötigt. Die folgende Modifikation des Algorithmus ermittelt laufend auch die Binärstellen von a. Die dazu notwendigen Operationen sind im Binärsystem besonders effizient implementierbar, die Division durch 2 ist ein Bitshift, der Rest ist einfach das niederwertigste Bit der Zahl.

Algorithmus 10.3. *1. Setze $f = g$, $x = 1$, $i = 0$*

2. Solange $a > 0$ ist, führe aus

(a) Verwende den euklidischen Algorithmus um r und b zu bestimmen mit $a = 2b + r$

(b) Falls $r = 1$ setze $x := x \cdot f$

(c) $i := i + 1$, $a = b$ und $f := f \cdot f$

Die Potenz $x = g^a$ kann so in $O(\log_2 a)$ Multiplikationen berechnet werden.

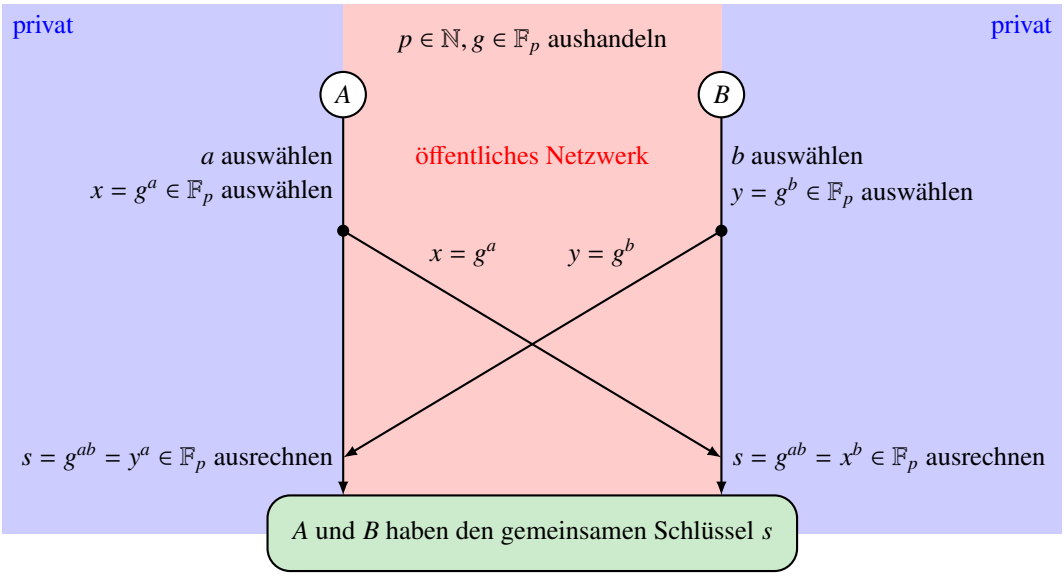


Abbildung 10.4: Schlüsselaustausch nach Diffie-Hellman. Die Kommunikationspartner A und B einigen sich öffentlich auf $p \in \mathbb{N}$ und $g \in \mathbb{F}_p$. A wählt dann einen privaten Schlüssel $a \in \mathbb{N}$ und B wählt $b \in \mathbb{N}$, sie tauschen dann $x = g^a$ und $y = g^b$ aus. A erhält den gemeinsamen Schlüssel aus y^a, B erhält ihn aus x^b.

10.2.2 Diffie-Hellman-Schlüsseltausch

Eine Grundaufgabe der Verschlüsselung im Internet ist, dass zwei Kommunikationspartner einen gemeinsamen Schlüssel für die Verschlüsselung der Daten aushandeln können müssen. Es muss davon ausgegangen werden, dass die Kommunikation abgehört wird. Trotzdem soll es für einen Lauscher nicht möglich sein, den ausgehandelten Schlüssel zu ermitteln.

Die beiden Partner A und B einigen sich zunächst auf eine Zahl g, die öffentlich bekannt sein darf. Weiter erzeugen sie eine zufällige Zahl a und b, die sie geheim halten. Das Verfahren soll aus diesen beiden Zahlen einen Schlüssel erzeugen, den beide Partner berechnen können, ohne dass sie a oder b übermitteln müssen. Die beiden Zahlen werden daher auch die privaten Schlüssel genannt.

Die Idee von Diffie und Hellman ist jetzt, die Werte $x = g^a$ und $y = g^b$ zu übertragen. In $\mathbb{R}$ würden dadurch natürlich dem Lauscher auch a offenbart, er könnte einfach $a = \log_g x$ berechnen. Ebenso kann auch b als $b = \log_g y$ erhalten werden, die beiden privaten Schlüssel wären also nicht mehr privat. Statt der Potenzfunktion in $\mathbb{R}$ muss also eine Funktion verwendet werden, die nicht so leicht umgekehrt werden kann. Die Potenzfunktion in $\mathbb{F}_p$ erfüllt genau diese Eigenschaft. Die Kommunikationspartner einigen sich also auch noch auf die (grosse) Primzahl p und übermitteln $x = g^a \in \mathbb{F}_p$ und $y = g^b \in \mathbb{F}_p$.

Aus x und y muss jetzt der gemeinsame Schlüssel abgeleitet werden. A kennt $y = g^b$ und a, B kennt $x = g^a$ und b. Beide können die Zahl $s = g^{ab} \in \mathbb{F}_p$ berechnen. A macht das, indem er $y^a = (g^b)^a = g^{ab}$ rechnet, B rechnet $x^b = (g^a)^b = g^{ab}$, beide natürlich in $\mathbb{F}_p$. Der Lauscher kann aber g^{ab} nicht ermitteln, dazu müsste er a oder b ermitteln können. Die Zahl $s = g^{ab}$ kann also als gemeinsamer Schlüssel verwendet werden.

10.2.3 Elliptische Kurven

Das Diffie-Hellman-Verfahren basiert auf der Schwierigkeit, in einem Körper $\mathbb{F}_p$ die Gleichung $a^x = b$ nach x aufzulösen. Die Addition in $\mathbb{F}_p$ wird dazu nicht benötigt. Es reicht, eine Menge mit einer Multiplikation zu haben, in der das die Gleichung $a^x = b$ schwierig zu lösen ist. Ein Gruppe wäre also durchaus ausreichend.

Ein Kandidat für eine solche Gruppe könnte der Einheitskreis $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ in der komplexen Ebene sein. Wählt man eine Zahl $g = e^{i\alpha}$, wobei α ein irrationales Vielfaches von π ist, dann sind alle Potenzen g^n für natürliche Exponenten voneinander verschieden. Wäre nämlich $g^{n_1} = g^{n_2}$, dann wäre $e^{i\alpha(n_1-n_2)} = 1$ und somit müsste $\alpha = 2k\pi/(n_1 - n_2)$ sein. Damit wäre aber α ein rationales Vielfaches von π, im Widerspruch zur Voraussetzung. Die Abbildung $n \mapsto g^n \in S^1$ ist auf den ersten Blick etwa ähnlich undurchschaubar wie die Abbildung $n \mapsto g^n \in \mathbb{F}_p$. Es gibt zwar die komplexe Logarithmusfunktion, mit der man n bestimmen kann, dazu muss man aber den Wert von g^n mit beliebiger Genauigkeit kennen, denn die Werte von g^n können beliebig nahe beieinander liegen.

Der Einheitskreis ist die Lösungsmenge der Gleichung $x^2 + y^2 = 1$ für reelle Koordinaten x und y, doch Rundungsunsicherheiten verunmöglichen den Einsatz in einem Verfahren ähnlich dem Diffie-Hellman-Verfahren. Dieses Problem kann gelöst werden, indem für die Variablen Werte aus einem endlichen Körper verwendet werden. Gesucht ist also eine Gleichung in zwei Variablen, deren Lösungsmenge in einem endlichen Körper eine Gruppenstruktur trägt. Die Lösungsmenge ist eine "Kurve" von Punkten mit Koordinaten in einem endlichen Körper.

In diesem Abschnitt wird gezeigt, dass sogenannte elliptische Kurven über endlichen Körpern genau die verlangen Eigenschaften haben.

Elliptische Kurven

Elliptische Kurven sind Lösungen einer Gleichung der Form

$$Y^2 + XY = X^3 + aX + b \tag{10.1}$$

mit Werten von X und Y in einem geeigneten Körper. Die Koeffizienten a und b müssen so gewählt werden, dass die Gleichung (10.1) genügend viele Lösungen hat. Über den komplexen Zahlen hat die Gleichung natürlich für jede Wahl von X drei Lösungen. Für einen endlichen Körper können wir dies im allgemeinen nicht erwarten, aber wenn wir genügend viele Wurzeln zu $\mathbb{F}$ hinzufügen können wir mindestens erreichen, dass die Lösungsmenge so viele Elemente hat, dass ein Versuch, die Gleichung $g^x = b$ mittels Durchprobierens zu lösen, zum Scheitern verurteil ist.

Definition 10.4. *Die* elliptische Kurve $E_{a,b}(\Bbbk)$ *über dem Körper* $\Bbbk$ *ist die Menge*

$$E_{a,b}(\Bbbk) = \{(X, Y) \in \Bbbk^2 \mid Y^2 + XY = X^3 + aX + b\},$$

für $a, b \in \Bbbk$.

Um die Anschauung zu vereinfachen, werden wir elliptische Kurven über dem Körper $\mathbb{R}$ visualisieren. Die daraus gewonnenen geometrischen Einsichten werden wir anschliessend algebraisch umsetzen. In den reellen Zahlen kann man die Gleichung (10.1) noch etwas vereinfachen. Indem man in (10.1) quadratisch ergänzt, bekommt man

$$Y^2 + XY + \frac{1}{4}X^2 = X^3 + \frac{1}{4}X^2 + aX + b$$

$$\Rightarrow \qquad v^2 = X^3 + \frac{1}{4}X^2 + aX + b, \tag{10.2}$$

indem man $v = Y + \frac{1}{2}X$ setzt. Man beachte, dass man diese Substition nur machen kann, wenn $\frac{1}{2}$ definiert ist. In $\mathbb{R}$ ist dies kein Problem, aber genau über den Körpern mit Charakteristik 2, die wir für die Computer-Implementation bevorzugen, ist dies nicht möglich. Es geht hier aber nur um die Visualisierung.

Auch die Form (10.2) lässt sich noch etwas vereinfachen. Setzt man $X = u - \frac{1}{12}$, dann verschwindet nach einiger Rechnung, die wir hier nicht durchführen wollen, der quadratische Term auf der rechten Seite. Die interessierenden Punkte sind Lösungen der einfacheren Gleichung

$$v^2 = u^3 + \left(a - \frac{1}{48}\right)u + b - \frac{a}{12} + \frac{1}{864} = u^3 + Au + B. \tag{10.3}$$

In dieser Form ist mit (u, v) immer auch $(u, -v)$ eine Lösung, die Kurve ist symmetrisch bezüglich der u-Achse. Ebenso kann man ablesen, dass nur diejenigen u-Werte möglich sind, für die das kubische Polynom $u^3 + Au + B$ auf der rechten Seite von (10.3) nicht negativ ist.

Sind u_1, u_2 und u_3 die Nullstellen des kubischen Polynoms auf der rechten Seite von (10.3), folgt

$$v^2 = (u - u_1)(u - u_2)(u - u_3) = u^3 - (u_1 + u_2 + u_3)u^2 + (u_1u_2 + u_1u_3 + u_2u_3)u - u_1u_2u_3.$$

Durch Koeffizientenvergleich sieht man, dass $u_1 + u_2 + u_3 = 0$ sein muss. Abbildung 10.5 zeigt eine elliptische Kurve in der Ebene.

Geometrische Definition der Gruppenoperation

In der speziellen Form 10.3 ist die elliptische Kurve symmetrisch unter Spiegelung an der u-Achse. Die Spiegelung ist eine Involution, zweimalige Ausführung führt auf den ursprünglichen Punkt zurück. Die Inverse in einer Gruppe hat diese Eigenschaft auch, es ist daher naheliegend, den gespiegelten Punkt als die Inverse eines Elementes zu nehmen.

Eine Gerade durch zwei Punkte der in Abbildung 10.5 dargestellten Kurve schneidet die Kurve ein drittes Mal. Die Gruppenoperation wird so definiert, dass drei Punkte der Kurve auf einer Geraden das Gruppenprodukt e haben. Da aus $g_1g_2g_3 = e$ folgt $g_3 = (g_1g_2)^{-1}$ oder $g_1g_2 = g_3^{-1}$, erhält man das Gruppenprodukt zweier Elemente auf der elliptischen Kurve indem erst den dritten Schnittpunkt ermittelt und diesen dann an der u-Achse spiegelt.

Die geometrische Konstruktion schlägt fehl, wenn $g_1 = g_2$ ist. In diesem Fall kann man die Tangente im Punkt g_1 an die Kurve verwenden. Dieser Fall tritt zum Beispiel auch in den drei Punkten $(u_1, 0)$, $(u_2, 0)$ und $(u_3, 0)$ ein.

Um das neutrale Element der Gruppe zu finden, können wir zwei Punkte g und g^{-1} miteinander verknüpfen. Die Gerade durch g und g^{-1} schneidet aber die Kurve kein drittes Mal. Ausserdem sind alle Geraden durch g und g^{-1} für verschiedene g parallel. Das neutrale Element entspricht also einem unendlich weit entfernten Punkt. Das neutrale Element entsteht immer dann als Produkt, wenn zwei Punkte die gleiche u-Koordinaten haben.

Gruppenoperation, algebraische Konstruktion

Nach den geometrischen Vorarbeiten zur Definition der Gruppenoperation kann können wir die Konstruktion jetzt algebraisch umsetzen.

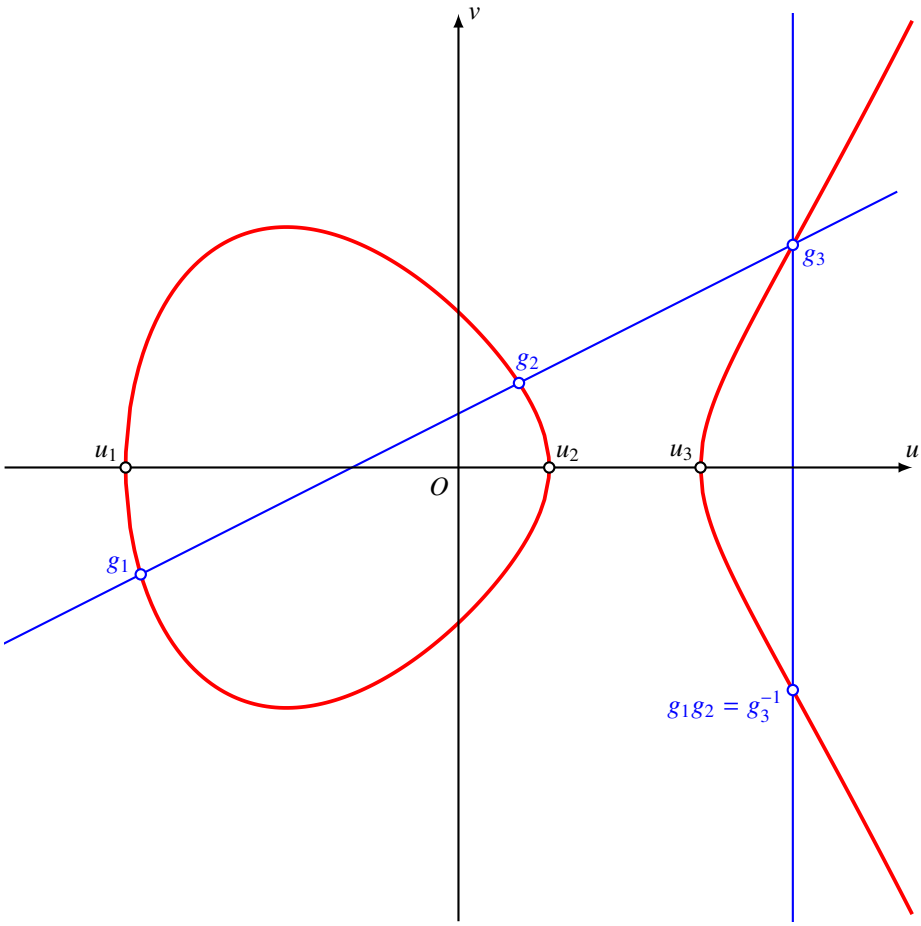


Abbildung 10.5: Elliptische Kurve in $\mathbb{R}$ in der Form $v^2 = u^3 + Au + B$ mit Nullstellen u_1, u_2 und u_3 des kubischen Polynoms auf der rechten Seite. Die blauen Punkte und Geraden illustrieren die Definition der Gruppenoperation in der elliptischen Kurve.

Zunächst überlegen wir uns wieder eine Involution, welche als Inverse dienen kann. Dazu beachten wir, dass die linke Seite der definierenden Gleichung

$$Y^2 + XY = X^3 - aX + b. \tag{10.4}$$

auch als $Y(Y + X)$ geschrieben werden kann. Die Abbildung $Y \mapsto -X - Y$ macht daraus

$$(-X - Y)(-X - Y + X) = (X + Y)Y,$$

dies ist also die gesuchte Involution.

Seien also $g_1 = (x_1, y_1)$ und $g_2 = (x_2, y_2)$ zwei verschiedene Lösungen der Gleichung (10.4) Als erstes brauchen wir eine Gleichung für die Gerade durch die beiden Punkte. Sei also $l(X, Y)$ eine Linearform derart, dass $l(g_1) = d$ und $l(g_2) = d$ für ein geeignetes $d \in \Bbbk$. Dann gilt auch für die Punkte

$$g(t) = tg_1 + (1 - t)g_2 \qquad \Rightarrow \qquad l(g(t)) = tl(g_1) + (1 - t)l(g_2) = tc + (1 - t)c = (t + 1 - t)c = c,$$

jeder Punkt der Geraden durch g_1 und g_2 lässt sich in dieser Form schreiben.

Setzt man jetzt $g(t)$ in die Gleichung ein, erhält man eine kubische Gleichung in t, von der wir bereits zwei Nullstellen kennen, nämlich 0 und 1. Die kubische Gleichung muss also durch t und $(t-1)$ teilbar sein. Diese Berechnung kann man einfach in einem Computeralgebrasystem durchführen. Das Polynom ist

$$p(t) =$$

Nach Division durch $t(t-1)$ erhält man als den Quotienten

$$q(t) = (y_2 - y_1)^2 + (y_2 - y_1)(x_2 - x_1) + t(x_2 - x_1)^3 - 2x_2^3 + 3x_1x_2^2 - x_1^3$$

und den Rest

$$r(t) = t(y_1^2 + x_1y_1 - x_1^3 - ax_1 - b) + (1-t)(y_2^2 + x_2y_2 - x_2^3 - ax_2 - b).$$

Die Klammerausdrücke verschwinden, da die sie gleichbedeutend damit sind, dass die Punkte Lösungen von (10.4) sind.

Für den dritten Punkt auf der Geraden muss t so gewählt werden, dass $q(t) = 0$ ist. Dies ist aber eine lineare Gleichung mit der Lösung

$$t = -\frac{(y_1 - y_2)^2 + (y_2 - y_1)(x_2 - x_1) - 2x_2^3 + 3x_1x_2^2 - x_1^3}{(x_2 - x_1)^3}.$$

Setzt man dies $g(t)$ ein, erhält man für die Koordinaten des dritten Punktes g_3 die Werte

$$x_3 = \frac{(y_2 - y_1)^2(x_2 - x_1) + (y_2 - y_1)(x_2 - x_1)^2 - (x_2^4 + x_1^4)}{(x_2 - x_1)^3} \tag{10.5}$$

$$y_3 = \frac{(y_2 - y_1)^3 + (x_2 - x_1)(y_2 - y_1)^2 - (x_2 - x_1)^3(y_2 - y_1) - (x_2 - x_1)^2(x_1y_2 - x_2y_1)}{(x_2 - x_1)^3} \tag{10.6}$$

Die Gleichungen (10.5) und (10.6) ermöglichen also, das Element $g_1g_2^{-1}$ zu berechnen. Interessant daran ist, dass in den Formeln die Konstanten a und b gar nicht vorkommen.

Es bleibt noch der wichtige Fall des Quadrierens in der Gruppe zu behandeln, also den Fall $g_1 = g_2$. In diese Fall sind die Formeln (10.5) und (10.6) ganz offensichtlich nicht anwendbar. Die geometrische Anschauung hat nahegelegt, die Tangent an die Kurve im Punkt g_1 zu nehmen. In $\mathbb{R}$ würde man dafür einen Grenzübergang $g_2 \to g_1$ machen, aber in einem endlichen Körper ist dies natürlich nicht möglich.

Wir schreiben die Gerade als Parameterdarstellung in der Form $t \mapsto g(t) = (x_1 + ut, y_1 + vt)$ für beliebige Parameter in $\Bbbk$. Die Werte u_1 und u_2 müssen so gewählt werden, dass $g(t)$ eine Tangente wird. Setzt man $g(t)$ in die Gleichung (10.4) ein, entsteht ein kubische Gleichung, die genau dann eine doppelte Nullstelle bei 0 hat, wenn u, v die Tangentenrichtung beschreiben. Einsetzen von $g(t)$ in (10.4) ergibt die Gleichung

$$0 = -u^3t^3 + (-3u^2x_1 + v^2 + uv)t^2 + (2vy_1 + uy_1 - 3ux_1^2 + vx_1 - au)t + (y_1^2 + x_1y_1 - x_1^3 - ax_1 - b) \tag{10.7}$$

Damit bei $t = 0$ eine doppelte Nullstelle mussen die letzten beiden Koeffizienten verschwinden, dies führt auf die Gleichungen

$$y_1^2 + x_1y_1 = x_1^3 + ax_1 + b \tag{10.8}$$

$$(2y_1 + x_1)v + (y_1 - 3x_1^2 - a)u = 0 \tag{10.9}$$

Die erste Gleichung (10.8) drückt aus, dass g_1 ein Punkt der Kurve ist, sie ist automatisch erfüllt.

Die zweite Gleichung (10.2.3) legt das Verhältnis von u und v, also die Tangentenrichtung fest. Eine mögliche Lösung ist

$$\begin{aligned} u &= x_1 + 2y_1 \\ v &= -y_1 + 3x_1^2 + a. \end{aligned} \tag{10.10}$$

Der Quotient ist ein lineares Polynom in t, die Nullstelle parametrisiert den Punkt, der $(g_1)^{-2}$ entspricht. Der zugehörige Wert von t ist

$$t = -\frac{3u^2x_1 - v^2 - uv}{u^3}. \tag{10.11}$$

Setzt man und (10.10) in $g(t)$ ein, erhält man sehr komplizierte Ausdrücke für den dritten Punkt. Wir verzichten darauf, diese Ausdrücke hier aufzuschreiben. In der Praxis wird man in einem Körper der Charakteristik 2 arbeiten. In diesem Körper werden alle geraden Koeffizienten zu 0, alle ungeraden Koeffizienten werden unabhängig vom Vorzeichen zu 1. Damit bekommt man die folgenden, sehr viel übersichtlicheren Ausdrücke für den dritten Punkt:

$$\begin{aligned} x &= -\frac{y_1^2 + x_1y_1 + x_1^4 + x_1^3 + ax_1 - a^2}{x_1^2} \\ y &= \frac{y_1^3 + (x_1^2 + x_1 + a)y_1^2 + (x_1^4 + a^2)y_1 + x_1^6 + ax_1^4 + ax_1^3 + a^2x_1^2 + a^2x_1 + a^3}{x_1^3} \end{aligned} \tag{10.12}$$

Damit haben wir einen vollständigen Formelsatz für die Berechnung der Gruppenoperation in der elliptischen Kurve mindestens für den praktisch relevanten Fall einer Kurve über einem Körper der Charakteristik 2.

Satz 10.5. *Die elliptische Kurve*

$$E_{a,b}(\mathbb{F}_{p^l}) = \{(X, Y) \in \mathbb{F}_{p^l} \mid Y^2 + XY = X^3 - aX - b\}$$

trägt eine Gruppenstruktur, die wie folgt definiert ist:

1. *Der Punkt* $(0, 0)$ *entspricht dem neutralen Element.*
2. *Das inverse Element von* (x, y) *ist* $(-x, -y - x)$.
3. *Für zwei verschiedene Punkte* g_1 *und* g_2 *kann* $g_3 = (g_1g_2)^{-1}$ *mit Hilfe der Formeln* (10.5) *und* (10.6) *gefunden werden.*
4. *Für einen Punkt* g_1 *kann* $g_3 = g_1^{-2}$ *in Charakteristik* 2 *mit Hilfe der Formeln* (10.12) *gefunden werden.*

Diese Operationen machen $E_{a,b}(\mathbb{F}_{p^l})$ *zu einer endlichen abelschen Gruppe.*

Beispiele

TODO: elliptische Kurven in IPsec: Oakley Gruppen

Diffie-Hellman in einer elliptischen Kurve

TODO: g^x in einer elliptischen Kurve

10.3 Advanced Encryption Standard – AES

Eine wichtige Forderung bei der Konzeption des damals neuen Advanced Encryption Standard war, dass darin keine "willkürlich" erscheinenden Operationen geben darf, bei denen der Verdacht entstehen könnte, dass sich dahinter noch offengelegtes Wissen über einen möglichen Angriff auf den Verschlüsselungsalgorithmus verbergen könnte. Dies war eine Schwäche des vor AES üblichen DES Verschlüsselungsalgorithmus. In seiner Definition kommt eine Reihe von Konstanten vor, über deren Herkunft nichts bekannt war. Die Gerüchteküche wollte wissen, dass die NSA die Konstanten aus dem ursprünglichen Vorschlag abgeändert habe, und dass dies geschehen sei, um den Algorithmus durch die NSA angreifbar zu machen.

Eine weiter Forderung war, dass die Sicherheit des neuen Verschlüsselungsstandards "skalierbar" sein soll, dass man also die Schlüssellänge mit der Zeit von 128 Bit auf 196 oder sogar 256 Bit steigern kann. Der Standard wird dadurch langlebiger und gleichzeitig entsteht die Möglichkeit, Sicherheit gegen Rechenleistung einzutauschen. Weniger leistungsfähige Systeme können den Algorithmus immer noch nutzen, entweder mit geringerer Verschlüsselungsrate oder geringerer Sicherheit.

In diesem Abschnitt soll gezeigt werde, wie sich die AES spezifizierten Operationen als mit der Arithmetik der endlichen Körper beschreiben lassen. Im Abschnitt 10.3.1 werden Bytes als Elemente in einem endlichen Körper $\mathbb{F}_{2^8}$ interpretiert. Damit kann dann die sogenannte S-Box konstruiert werden und es ist leicht zu verstehen, dass sie invertierbar ist. Aus den Byte-Operationen können dann Mischoperationen erzeugt werden, die Bytes untereinander verknüpfen, die aber auch wieder als Operationen in einem endlichen Körper verstanden werden können.

10.3.1 Byte-Operationen

Moderne Prozessoren operieren auf Wörtern, die Vielfache von Bytes sind. Byte-Operationen sind besonders effizient in Hardware zu realisieren. AES verwendet daher als Grundelemente Operationen auf Bytes, die als Elemente eines endlichen Körpers $\mathbb{F}_{2^8}$ interpretiert werden.

Bytes als Elemente von $\mathbb{F}_{2^8}$

Das Polynom $m(X) = X^8 + X^4 + X^3 + X + 1 \in \mathbb{F}_2[X]$ ist irreduzibel, somit ist $\mathbb{F}_{2^8} = \mathbb{F}_2[X]/(m)$ ein Körper. Die Elemente können dargestellt werden als Polynome, das Byte 63_{16} bekommt die Form

$$p(X) = p_7X^7 + p_6X^6 + \cdots + p_2X^2 + p_1X + p_0,$$

sie bestehen daher aus den 8 Bits $p_7, \ldots, p_0$.

Die Interpretation der Bytes als Elemente eines Körpers bedeutet, dass jede Multiplikation mit einem nicht verschwindenden Byte invertierbar ist. Ausserdem mischen diese Operationen die einzelnen Bits auf einigermassen undurchsichtige, aber umkehrbare Art durcheinander, wie dies für ein Verschlüsselungsverfahren wünschenswert ist.

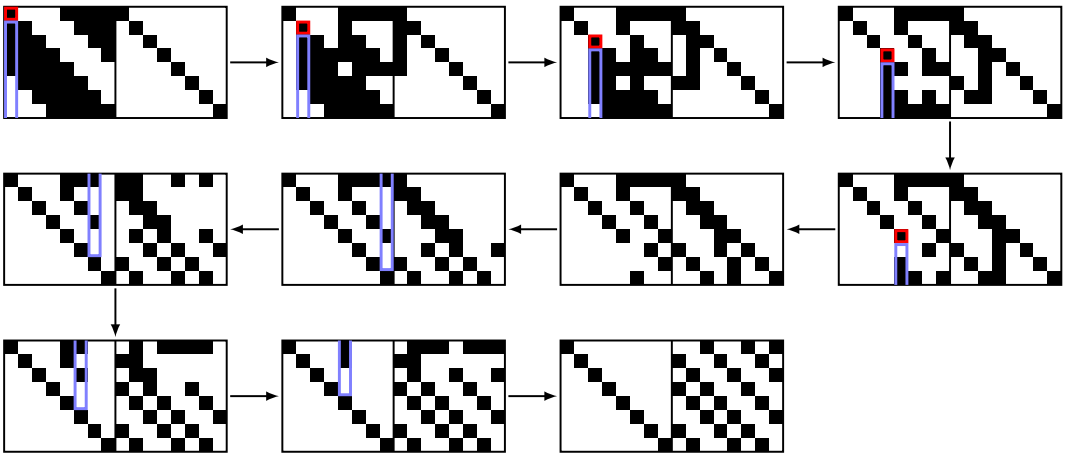


Abbildung 10.6: Berechnung der Inversen der Matrix A in der S-Box des AES-Algorithmus mit dem Gauss-Algorithmus

S-Box

Für die Operation der S-Box wird wie folgt zusammengesetzt. Zunächst wird ein Byte x durch das zugehörige multiplikative inverse Element

$$x \mapsto \bar{x} = \begin{cases} x^{-1} & \text{für } x \in \mathbb{F}_{2^8}^* \\ 0 & \text{für } x = 0 \end{cases}$$

ersetzt.

Im zweiten Schritt betrachten wir $\mathbb{F}_{2^8}$ als einen 8-dimensionalen Vektorraum über $\mathbb{F}_2$. Einem Polynom $p(X) = p_7X^7 + \cdots + p_1X + p_0$ wird der Spaltenvektor mit den Komponenten p_0 bis p_7 zugeordnet.

Eine lineare Transformation in diesem Vektorraum kann durch eine 8×8-Matrix in $M_8(\mathbb{F}_2)$ betrachtet werden. In der S-Box wird die Matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

verwendet. Mit dem Gauss-Algorithmus, schematisch dargestellt in Abbildung 10.6, kann man die Inverse bestimmen, die Multiplikation mit A ist also eine invertierbare Abbildung.

Der letzte Schritt ist dann wieder eine Addition von $q(X) = X^7 + X^6 + X + 1 \in \mathbb{F}_{2^8}$, durch Subtraktion von $q(X)$ invertiert werden kann. Die S-Box-Operation kann also bektoriell geschrieben werden also

$$S(x) = A\overline{x} + q.$$

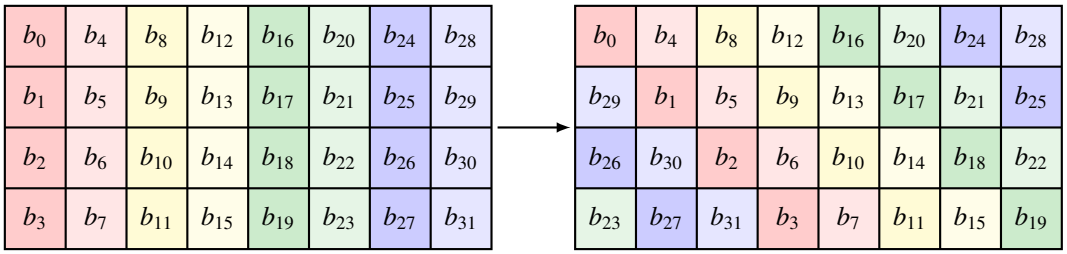


Abbildung 10.7: Zeilenshift in einem Block von 256 bits

Die Implementation ist möglicherweise mit einer Tabelle am schnellsten, es sind ja nur 256 Bytes im Definitionsbereich der S-Box-Abbildung und ebenso nur 256 möglich Werte.

10.3.2 Block-Operationen

Die zu verschlüsselnden Daten werden in in Blöcke aufgeteilt, deren Länge Vielfache von 32 bit sind. Die kleinste Blockgrösse ist 128 Bit, die grösste ist 256 Bit. Die Bytes eines Blockes werden dann in einem Rechteck angeordnet als

$$\begin{array}{|cccccccc|}\hline b_0 & b_4 & b_8 & b_{12} & b_{16} & b_{20} & b_{24} & b_{28} \\ b_1 & b_5 & b_9 & b_{13} & b_{17} & b_{21} & b_{25} & b_{29} \\ b_2 & b_6 & b_{10} & b_{14} & b_{18} & b_{22} & b_{26} & b_{30} \\ b_3 & b_7 & b_{11} & b_{15} & b_{19} & b_{23} & b_{27} & b_{31} \\ \hline\end{array} \tag{10.13}$$

für eine Blocklänge von 256 Bits.

Zeilenshift

Die Verschlüsselung muss sicherstellen, dass die Bytes des Blockes untereinander gut gemischt werden. Die bisher beschriebenen Operationen operieren immer nur auf einzelnen Bytes während die im nächsten Abschnitt beschriebene Spalten-Mischoperation nur auf Spalten wird. Die Zeilenmischoperation permutiert die Zeilen in den vier Zeilen eines Blocks zyklisch, die erste Zeile bleibt an Ort, die zweite Zeile wird um ein Byte rotiert, die dritte um zwei und die letzte um 3 Bytes, wie in Abbildung **??** dargestellt. Diese Operation könnte mit einer Permutationsmatrix beschrieben werden, dies wäre jedoch keine effiziente Implementation. Der Zeilenschift hat ansonsten keine elegante algebraische Beschreibung.

Spalten mischen

Jede Spalte von (10.13) kann als Vektor des vierdimensionalen Vektorraumes $\mathbb{F}_{2^8}^4$. Die Zeilenmischoperation wendet ein lineare Abbildung auf jeden Spaltenvektor von (10.13). Die Koeffizienten der Matrix sind Elemente von $\mathbb{F}_{2^8}$. Die Matrix ist

$$C = \begin{pmatrix} 02_{16} & 03_{16} & 01_{16} & 01_{16} \\ 01_{16} & 02_{16} & 03_{16} & 01_{16} \\ 01_{16} & 01_{16} & 02_{16} & 03_{16} \\ 03_{16} & 01_{16} & 01_{16} & 02_{16} \end{pmatrix}.$$

Um nachzuprüfen, dass die Matrix C invertierbar ist, könnte man den Gauss-Algorithmus verwenden und damit die Inverse berechnen. Dazu müsste man die multiplikativen Inversen kennen, was etwas mühsam ist. Man kann aber aber auch die Determinante bestimmen, dazu braucht man nur multiplizieren zu können, was in diesem Fall sehr leicht möglich ist, weil kein Überlauf entsteht. Dabei hilft es zu beachten, dass die Multiplikation mit 02_{16} nur eine Einbit-Shiftoperation nach links ist. Nur die Multiplikation $03_{16} \cdot 03_{16} = 05_{16}$ gibt etwas mehr zu überlegen. Mit geeigneten Zeilen-Operationen kann man die Berechnung der Determinante von C mit dem Entwicklungssatz etwas vereinfachen. Man erhält

$$\begin{aligned}
\det(C) &= \begin{vmatrix} 02_{16} & 03_{16} & 01_{16} & 01_{16} \\ 01_{16} & 02_{16} & 03_{16} & 01_{16} \\ 00_{16} & 03_{16} & 01_{16} & 02_{16} \\ 00_{16} & 00_{16} & 03_{16} & 02_{16} \end{vmatrix} \\
&= 02_{16} \begin{vmatrix} 02_{16} & 03_{16} & 01_{16} \\ 03_{16} & 01_{16} & 02_{16} \\ 00_{16} & 03_{16} & 02_{16} \end{vmatrix} + 01_{16} \begin{vmatrix} 03_{16} & 01_{16} & 01_{16} \\ 03_{16} & 01_{16} & 02_{16} \\ 00_{16} & 03_{16} & 02_{16} \end{vmatrix} \\
&= 02_{16} \begin{vmatrix} 02_{16} & 03_{16} & 01_{16} \\ 01_{16} & 02_{16} & 03_{16} \\ 00_{16} & 03_{16} & 02_{16} \end{vmatrix} + 01_{16} \begin{vmatrix} 03_{16} & 01_{16} & 01_{16} \\ 00_{16} & 00_{16} & 01_{16} \\ 00_{16} & 03_{16} & 02_{16} \end{vmatrix} \\
&= 02_{16} \left(02_{16} \begin{vmatrix} 02_{16} & 03_{16} \\ 03_{16} & 02_{16} \end{vmatrix} + 01_{16} \begin{vmatrix} 03_{16} & 01_{16} \\ 03_{16} & 02_{16} \end{vmatrix} \right) + 01_{16} \begin{vmatrix} 03_{16} & 01_{16} & 01_{16} \\ 00_{16} & 03_{16} & 02_{16} \\ 00_{16} & 00_{16} & 01_{16} \end{vmatrix} \\
&= 02_{16}(02_{16}(04_{16} + 05_{16}) + (06_{16} + 03_{16})) + 03_{16}03_{16} \\
&= 02_{16}(02_{16} + 05_{16}) + 05_{16} = 0e_{16} + 05_{16} = 0a_{16} \neq 0.
\end{aligned}$$

Damit ist gezeigt, dass die Matrix C invertierbar auf den Spaltenvektoren wirkt. Die Inverse der Matrix kann einmal berechnet und anschliessend für die Entschlüsselung verwendet werden.

Alternativ kann man die Multiplikation mit der Matrix C auch interpretieren als eine Polynommultiplikation. Dazu interpretiert man die Spalten des Blocks als Polynom vom Grad 3 mit Koeffizienten in $\mathbb{F}_{2^8}$. Durch Reduktion mit dem irreduziblen Polynom $n(Z) = Z^4 + 1 \in \mathbb{F}_{2^8}[X]$ entsteht aus dem Polynomring wieder ein Körper. Die Wirkung der Matrix C ist dann nichts anderes als Multiplikation mit dem Polynom

$$c(Z) = 03_{16}Z^3 + Z^2 + Z^1 + 02_{16},$$

die natürlich ebenfalls umkehrbar ist.

10.3.3 Schlüssel

Die von den Byte- und Blockoperationen mischen die einzelnen Bits der Daten zwar ganz schön durcheinander, aber es wird noch kein Schlüsselmaterial eingearbeitet, welches den Prozess einzigartig macht.

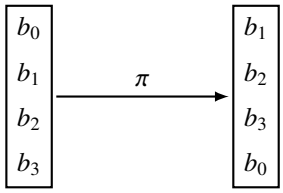
Schlüsseladdition

Nach jeder Spaltenmischoperation wird ein Rundenschlüssel zum Blockhinzuaddiert. Beim ersten Mal wird dazu einfach das Schlüsselmaterial verwendet. Für die folgenden Runden muss aus diesem Schlüssel neues Material, die sogenannten Rundenschlüssel, gewonnen werden.

Rundenschlüssel

Die Erzeugung der Rundenschlüssel ist in Abbildung 10.8 schematisch dargestellt. Die Blöcke beschreiben wieder Spaltenvektoren im vierdimensionalen Raum $\mathbb{F}_{2^8}^4$. Die Blöcke K_0 bis K_7 stellen den ursprünglichen Schlüssel dar. Die Erzeugung eines neuen Blocks Schlüsselmatrial beginnt damit, dass der letzte Vektor des vorangegangenblocks drei Operationen unterworfen werden.

- Die Operation π vertauscht die Bytes des Vektors zyklisch:



- Die S-Operation wendet die S-Box auf alle Bytes eines Vektors an.
- Die r_i Operation addiert in Runde eine Konstante r_i zur 0-Komponente.

Die Konstante r_i ist wieder ein einzelnes Byte und es ist daher naheliegend, diese Bytes mit Hilfe der Arithmetik in $\mathbb{F}_{2^8}$ zu erzeugen. Man kann daher r_i definieren als $(\mathtt{02}_{16})^{i-1} \in \mathbb{F}_{2^8}$.

10.3.4 Runden

Der AES-Verschlüsselungsalgorithmus besteht jetzt darin, die bisher definierten Operationen wiederholt anzuwenden. Eine einzelne Runde besteht dabei aus folgenden Schritten:

1. Wende die S-Box auf alle Bytes des Blocks an.
2. Führe den Zeilenshift durch.
3. Mische die Spalten (wird in der letzten Runde)
4. Erzeuge den nächsten Rundenschlüssel
5. Addiere den Rundenschlüssel

Der AES-Verschlüsselungsalgorithmus beginnt damit, dass der Schlüssel zum Datenblock addiert wird. Anschliessend werden je nach Blocklänge verschiedene Anzahlen von Runden durchgeführt, 10 Runden für 128 bit, 12 Runden für 192 bit und 14 Runden für 256 bit.

Übungsaufgaben

10.1. *A* und *B* einigen sich darauf, das Diffie-Hellman-Verfahren für $p = 2027$ durchzuführen und mit $g = 3$ zu arbeiten. *A* verwenden $a = 49$ als privaten Schlüssel und erhält von *B* den öffentlichen Schlüssel $y = 1772$. Welchen gemeinsamen Schlüssel verwenden *A* und *B*?

Lösung. Der zu verwendende gemeinsame Schlüssel ist $g^{ab} = (g^b)^a = y^a \in \mathbb{F}_{2027}$. Diese Potenz kann man mit dem Divide-and-Conquer-Algorithmus effizient berechnen. Die Binärdarstellung des privaten Schlüssels von *A* ist $a = 49_{10} = \mathtt{110001}_2$. Der Algorithmus verläuft wie folgt:

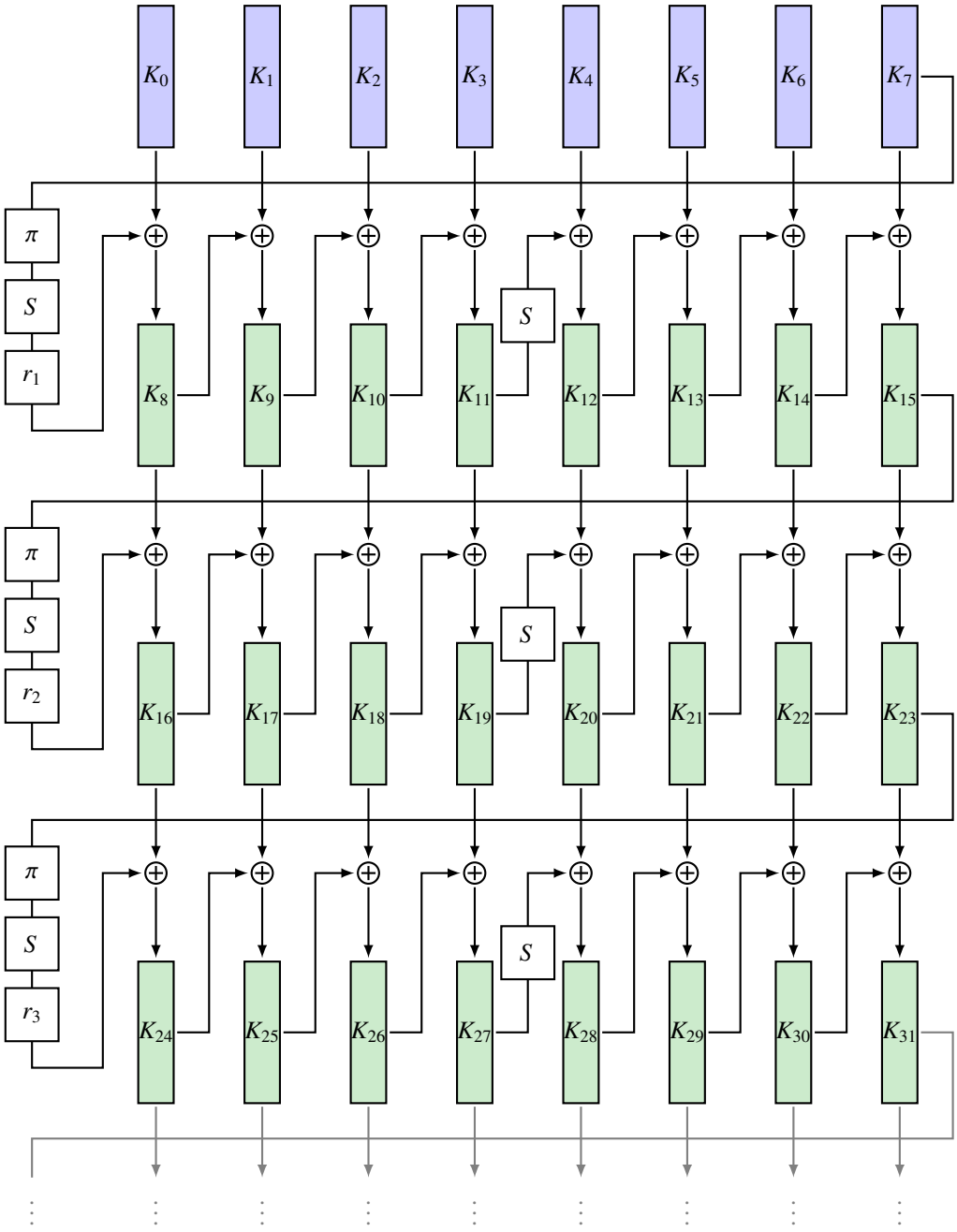


Abbildung 10.8: Erzeugung der erweiterten Schlüsseldaten aus dem Schlüssel $K_0, \ldots, K_7$ für Schlüssellänge 256 bit. Die mit S beschrifteten Blöcke wenden die S-Box auf jedes einzelne Byte an. π ist die zyklische Vertauschung der Bytes eines Wortes. Die Operation r_i ist eine Addition einer Konstanten, die in jeder Runde anders ist.

i	g^{2^i}	a_i	x
0	3	1	3
1	9	0	3
2	81	0	3
3	480	0	3
4	1349	1	2020
5	1582	1	1088

Der gemeinsame Schlüssel ist daher $s = 1088$. ○

Kapitel 11

Homologie

Mit der Inzidenzmatrix war es möglich, einen Graphen zu beschreiben und verschiedene interessante Eigenschaften desselben zu berechnen. Damit können aber nur eindimensionale Strukturen analysiert werden: Es ist zum Beispiel nicht möglich, ein Dreieck vom Rand eines Dreiecks zu unterscheiden 11.1. Die Randkurve ist in einem Dreieck zusammenziehbar, aber sobald man das innere des Dreiecks entfernt, ist die Randkurve nicht mehr zusammenziehbar. Dreieck und der Rand des Dreiecks sind also grundsätzlich verschieden.

Die Inzidenzmatrix ordnet jeder Kante ihre beiden Endpunkte zu. Die Homologietheorie verallgemeinert diese Idee. Der sogenannte Randoperator ordnet jedem Dreieck, Tetraeder oder allgemein jedem Simplex seinen Rand zu. Damit wird es möglich, das Dreieck vom Rand des Dreiecks zu unterschieden.

11.1 Simplexe und simpliziale Komplexe

Die Idee, das Dreieck und seinen Rand zu unterscheiden verlangt, dass wir zunächst Dreiecke und deren höherdimensionale Verallgemeinerungen, die sogenannten Simplizes entwickeln müssen.

11.1.1 Simplexe und Rand

Rand eines Dreiecks

Die Inzidenz-Matrix eines Graphen hat einer Kante die beiden Endpunkte mit verschiedenen Vorzeichen zugeordnet. Dieses Idee soll jetzt verallgemeinert werden. Der Rand des Dreiecks $\triangle$ in Abbildung 11.1 besteht aus den Kanten P_0P_1, P_1P_2 und P_0P_2. Für eine algebraische Definition müssen die Kanten offenbar eine Orientierung haben, die ist aber garantiert, da wir den Anfangs- und Endpunkten einer Kante verschiedene Vorzeichen gegeben haben. Dem Dreieck $\triangle$ werden dann die drei Kanten k_{01}, k_{02} und k_{12} zuogeordnet, aber mit zusätzlichen Vorzeichen, die die Orientierung festhalten. Durchläuft man den Rand von $\triangle$ in der Reihenfolge $P_0P_1P_2$, dann müssen die Kanten k_{12} und k_{02} ein negatives Vorzeichen erhalten.

Wir können diese Zuordnung wieder mit einer Matrix ausdrücken.

$$\begin{array}{ll} k_{01}: & \\ k_{02}: & \partial = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \\ k_{12}: & \end{array}$$

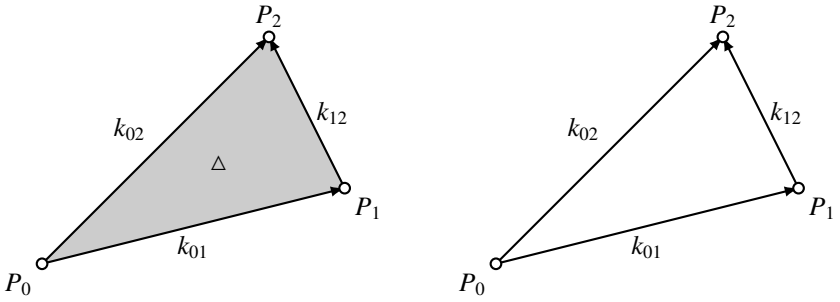


Abbildung 11.1: Ein Dreieck $\triangle$ (rechts) und der Rand des Dreicks (links) sind mit den Methoden der Graphentheorie nicht unterschiedbar. Als topologische Räume sind das Dreieck und sein Rand aber ganz klar unterschiedbar: In einem Dreieck ist jeder geschlossene Pfad in einen Punkt zusammenziehbar, aber die Randkurve ist nicht mehrzusammenziehbar, sobald man das innere des Dreiecks entfernt.

Simplizes

Punkte, Kanten und Dreiecke sind die einfachsten Fälle sogenannter Simplizes. Wir formulieren die Definition dieser Objekte auf eine Weise, die uns ermöglichen soll, sie auf beliebige Dimension zu verallgemeinern.

Die Strecke, die die Punkte P und Q miteinander verbindet, kann beschrieben werden durch eine Parametrisierung der Form

$$s_1 : t \mapsto t\vec{p} + (1-t)\vec{q} = t_0\vec{p} + t_1\vec{q}, \tag{11.1}$$

wobei die beiden positiven reellen Zahlen $t_0, t_1 \in \mathbb{R}$ die Bedingung $t_0 + t_1 = 1$ erfüllen. Für ein eindimensionales Objekt brauchen wir also zwei Punkte und zwei positive Parameter, die sich zu 1 summieren. Die Mengen $\triangle_1 = \{(t_0, t_1) | t_i \geq 0, t_0 + t_1 = 1\}$ kann also ganz allgemein als Parameterraum zur Beschreibung eindimensionalen Objektes mit den Endpunkten dienen. Eine Strecke ist also eine Abbildung der Form

$$s_1 : \triangle_1 \to \mathbb{R}^N : (t_0, t_1) \mapsto t_0\vec{p} + t_1\vec{q}, \tag{11.2}$$

und der Rand besteht aus den Punkten $s_1(0)$ und $s_1(1)$, wobei der Anfangspunkt $s_1(0)$ mit einem negativen Vorzeichen versehen wird.

Für höhere Dimensionen brauchen wir auf analoge Weise erst wieder einen geeigneten Parameterraum. Die Menge

$$\triangle_n = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} \,|\, t_i \geq 0, t_0 + t_1 + \cdots + t_n = 1\}$$

beschreibt zum Beispiel für $n = 2$ ein Dreieck und für $n = 3$ ein Tetraeder.

Gegeben $n+1$-Punkte $P_0, \dots, P_n$ mit Ortsvektoren $\vec{p}_0, \dots, \vec{p}_n$ können wir eine Abbildung

$$s_n : \triangle_n \to \mathbb{R}^N : (t_0, \dots, t_n) \mapsto t_0\vec{p}_0 + t_1\vec{p}_1 + \cdots + t_n\vec{p}_n \tag{11.3}$$

Eine solche Abbildung verallgemeinert also den Begriff einer Strecke auf höhere Dimensionen.

Definition 11.1. *Ein n-dimensionales* Simplex *oder n-*Simplex *ist eine stetige Abbildung $s_n : \triangle_n \to X$.*

Die Ecken des n-Simplex $\triangle_n$ sind die Standardbasisvektoren in $\mathbb{R}^{n+1}$. Mit e_k bezeichnen wird die Ecke, deren Koordinaten $t_i = 0$ sind für $k \neq i$, ausser der Koordinaten t_k, die den Wert $t_k = 1$ hat.

Rechnen mit Simplizes

Damit wir leichter mit Simplizes rechnen können, betrachten wir jedes Simplex als einen Basisvektor eines abstrakten Vektorraumes. Zu einem n-Simplex gehören Vektorräume C_l für jede Dimension $l = 0$ bis $l = n$. Der Vektorraum C_0 besteht aus Linearkombinationen

$$C_0 = \{x_0 P_0 + \cdots + x_n P_n \,|x_i \in \mathbb{R}\},$$

C_0 ist ein n-dimensionaler Raum. Der Vektorraum C_1 besteht aus Linearkombinationen der Kanten

$$C_1 = \left\{ \sum_{i<j} x_{ij} k_{ij} \,\middle|\, x_{ij} \in \mathbb{R} \right\},$$

wobei k_{ij} die Kante von der Ecke i zur Ecke j ist.

In Dimension l bezeichnen wir mit C_l den Vektorraum bestehend aus den Linearkombinationen

$$C_l = \left\{ \sum_{i_1<\cdots<i_l} x_{i_1\ldots i_l} s_{i_1\ldots i_l} \,\middle|\, s_{i_1\ldots i_l} \in \mathbb{R} \right\},$$

wobei $s_{i_1\ldots i_l}$ das Simplex mit den Ecken $i_1, \ldots, i_l$ ist.

Für $n = 1$ gibt ist C_1 ein eindimensionaler Vektorraum und C_0 ist zweidimensional. Die Randabbildung, die einer Kante den Rand zuordnet, ist

$$\partial \colon C_1 \to C_0 : s_{01} \mapsto 1 \cdot s_0 + (-1) \cdot s_1$$

und hat in den oben beschriebenden Basen die Matrix

$$\partial = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Rand eines Simplex

Einem Simplex muss auch der Rand zugeordnet werden können. Setzt man in $\triangle_2$ den Parameter $t_k = 0$, dann erhält man die Kante, die der Ecke mit Nummer k gegenüberliegt. Für jedes k gibt es also eine Abbildung

$$i_k \colon \triangle_{n-1} \to \triangle_n : (t_0, \ldots, t_n) \mapsto (t_0, \ldots, t_{k-1}, 0, t_k, \ldots, t_n),$$

in die Kante gegenüber der Ecke e_k. Dies ist auch die Art, wie Kanten des Dreiecks $\triangle$ in Abbildung 11.1 orientiert wurden.

Für den Rand des 2-Simplexes mussten die Kanten mit alternierenden Vorzeichen zugeordnet werden. Damit wird erreicht, dass jeder Punkt sowohl Endpunkt einer Kante und ausserdem Anfangspunkt der nächsten Kannte ist. Diese Eigenschaft soll auch in höheren Dimensionen erhalten bleiben. Die vier Dreiecke, die den Rand eines 3-Simplex ausmachen, müssen so orientiert werden, dass jede Kante in beiden Richtungen durchlaufen wird.

Definition 11.2. *Der Randoperator ordnet die Kanten eines n-Simplex mit alternierenden Vorzeichen zu, die Matrix ist*

11.1.2 Triangulation

11.2 Kettenkomplexe

11.2.1 Randoperator von Simplexen

11.2.2 Kettenkomplexe und Morphismen

11.3 Homologie

11.3.1 Homologie eines Kettenkomplexes

11.3.2 Induzierte Abbildung

11.3.3 Homologie eines simplizialen Komplexes

11.4 Exaktheit und die Mayer-Vietoris-Folge

Die Berechnung der Homologie-Gruppen ist zwar im Wesentlichen ein kombinatorisches Problem, trotzdem ist eher aufwändig. Oft weiss man, wie sich toplogische Räume aus einfacheren Räumen zusammensetzen lassen. Eine Mannigkfaltigkeit zum Beispiel wird durch die Karten definiert, also zusammenziehbare Teilmengen von $\mathbb{R}^n$, die die Mannigkfaltigkeit überdecken. Das Ziel dieses Abschnittes ist, Regeln zusammenzustellen, mit denen man die Homologie eines solchen zusammengesetzten Raumes aus der Homologie der einzelnen Teile und aus den "Verklebungsabbildungen", die die Teile verbinden, zu berechnen.

11.4.1 Kurze exakte Folgen von Kettenkomplexen

11.4.2 Schlangenlemma und lange exakte Folgen

11.4.3 Mayer-Vietoris-Folge

11.5 Fixpunkte

Zu jeder Abbildung $f\colon X \to X$ eines topologischen Raumes in sich selbst gehört die zugehörige lineare Abbildung $f_*\colon H_*(X) \to H_*(X)$ der Homologiegruppen. Diese linearen Abbildungen sind im Allgemeinen viel einfacher zu analysieren. Zum Beispiel soll in Abschnitt 11.5.1 die Lefshetz-Spurformel abgeleitet werden, die eine Aussagen darüber ermöglicht, ob eine Abbildung einen Fixpunkt haben kann. In Abschnitt 11.5.2 wird gezeigt wie man damit den Browerschen Fixpunktsatz beweisen kann, der besagt, dass jede Abbildung eines Einheitsballs in sich selbst immer einen Fixpunkt hat.

11.5.1 Lefshetz-Spurformel

11.5.2 Brower-Fixpunktsatz

Literatur

[1] Hans-Dieter Ebbinghaus et al. *Zahlen*. Bd. 1. Springer-Verlag, 1983. ISBN: 3-540-12666-X.

[2] Sergey Brin und Lawrence Page. “The anatomy of a large-scale hypertextual Web search engine”. In: *Computer Networks and ISDN Systems* 30.1 (1998). Proceedings of the Seventh International World Wide Web Conference, S. 107 –117. ISSN: 0169-7552. DOI: https://doi.org/10.1016/S0169-7552(98)00110-X. URL: http://www.sciencedirect.com/science/article/pii/S016975529800110X.

[3] L. D. Landau und E. M. Lifschitz. *Mechanik*. Bd. 1. Lehrbuch der theoretischen Physik. Akademie-Verlag, 1981.

[4] Andreas Müller u. a. *Mathematisches Seminar Wavelets*. 2019.

[5] Andreas Müller. *Source Code Repository*. 2020. URL: https://github.com/AndreasFMueller/SeminarNumerik.git.

Teil II

Anwendungen und weiterführende Themen

Übersicht

Im zweiten Teil kommen die Teilnehmer des Seminars selbst zu Wort. Die im ersten Teil dargelegten mathematischen Methoden und grundlegenden Modelle werden dabei verfeinert, verallgemeinert und auch numerisch überprüft.

Kapitel 12

Thema

Pascal Andreas Schmid und Robine Luchsinger

12.1 Versuchsreihe

Um zwei der vorgestellten Suchalgorithmen zu vergleichen, wurden zwei Versuchsreihen erstellt. Dazu wurden in einem ersten Schritt zufällige Netzwerke generiert und anschliessend der *Dijkstra*-, sowie der A^*-Algorithmus auf das Netzwerk angewandt. Dieser Vorgang wurde für die zufällig generierten Netzwerke mit einer Knotenzahl von 10, 20 50, 100, 200, 500 und 1000 je zehnmal repetiert. Die Anzahl der Knoten im abgesuchten Netzwerk wirkt sich direkt auf die Rechenzeit aus. Der *Dijkstra*-Algorithmus weist eine Zeitkomplexität von $O(E \log V)$ auf, wobei E die Anzahl Kanten (engl. *edges*) und V die Anzahl Knoten (engl. *vertices*) darstellt. Für den $A*$-Algorithmus ist die Zeitkomplexität einerseits abhängig von der verwendeten Heuristik, andererseits aber auch vom vorliegenden Netzwerk selbst. Aus diesem Grund lässt sich keine defintive Angabe zu O machen.

Die beiden Versuchsreihen unterscheiden sich zudem dahingehend, dass der Start- und Zielknoten bei der ersten Versuchsreihe im Netzwerk diametral gegenüber liegen. Dadurch gehen viele Knoten verloren, welcher *Dijkstra* als uninformierter Suchalgorithmus absuchen würde. In der zweiten Veruschsreihe werden hingegen Start- un Zielpunkt zufällig im Netzwerk ausgewählt. Es wird deshalb erwwartet, dass die Unterschiede in der Rechenzeit der beiden Algorithmen in der zweiten Versuchsreihe deutlich ausgeprägter sind.

12.1.1 Einfluss der Knotenzahl auf die Rechenzeit

In 12.5 ist ersichtlich, dass der Unterschied in der Rechenzeit zwischen *Dijkstra* und $A*$ erst aber einer Knotenzahl von ca. $n = 500$ merklich ansteigt. Dieses etwas überraschende Resultat ist darauf zurückzuführen, dass bei steigender Knotenzahl die Abweichung des effektiven kürzesten Pfades von der Distanz der Luftlinie abnimmt. Die Effektivität von $A*$ mit euklidischer Heuristik ist wiederum grösser, wenn die Abweichung des kürzesten Pfads von der Luftlinie minimal ist. Bei Betrachtung von 12.6 wird dies ersichtlich, wobei die relative Abweichung erstaunlicherweise bei einer Knotenzahl von $n = 100$ maximal ist und nach $n = 500$ nur noch marginal abnimmt.

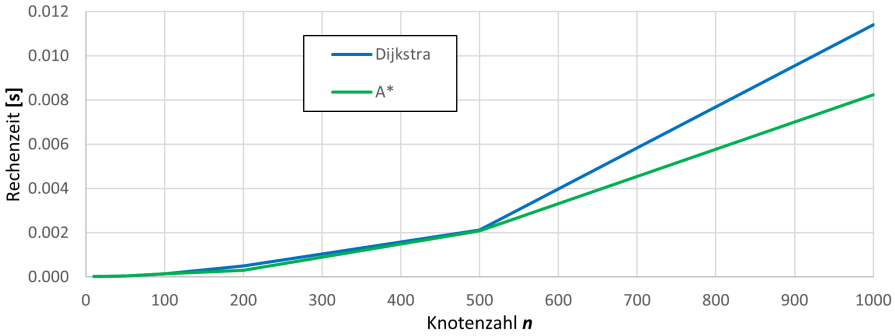


Abbildung 12.1: Gemessene Rechenzeiten der ersten Versuchsreihe in Abhängigkeit der Knotenzahl.

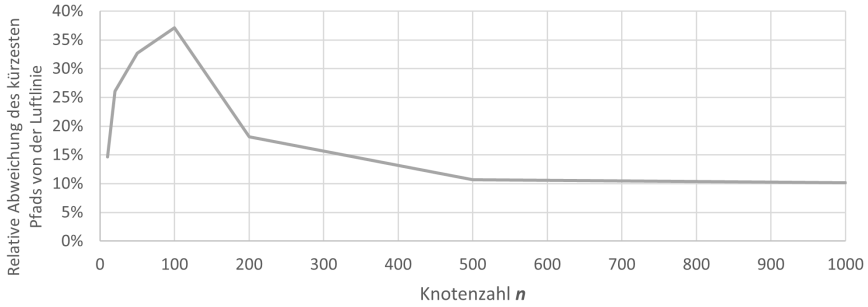


Abbildung 12.2: Relative Abweichung des kürzesten Pfads von der Luftlinie.

12.1.2 Einfluss der Position der Start- und Zielknoten auf die Rechenzeit

Zum Vergleich der Resultate in 12.2.1 zeigt 12.7 die Rechenzeiten der zweiten Versuchsreihe, in welcher die Start- und Zielknoten zufällig im Netzwerk ausgewählt wurden. Einerseits ist eine reduzierte durchschnittliche Rechenzeit festzustellen, was schlicht daran liegt, dass die zufällige Wahl der Knoten dazu führt, dass diese tendenziell weniger weit auseinander liegen.
Des weiteren ist festzustellen, dass sich die Unterschiede der Rechenzeiten zwischen *Dijkstra* und *A** deutlich früher abzeichnen. Dieses Phänomen lässt sich leicht durch die zielgerichtete Suche des *A**-Algorithmus erklären.

In 12.8 ist ersichtlich, dass bei einem im Netzwerk liegenden Startknoten die zielgerichtete Suche von *A** deutlich ausgeprägter zum Zuge kommt, als wenn dieser am Rand des Netzwerks liegen würde.

12.2 Versuchsreihe

Um zwei der vorgestellten Suchalgorithmen zu vergleichen, wurden zwei Versuchsreihen erstellt. Dazu wurden in einem ersten Schritt zufällige Netzwerke generiert und anschliessend der *Dijkstra*-, sowie der A^*-Algorithmus auf das Netzwerk angewandt. Dieser Vorgang wurde für die zufällig generierten Netzwerke mit einer Knotenzahl von 10, 20 50, 100, 200, 500 und 1000 je zehnmal

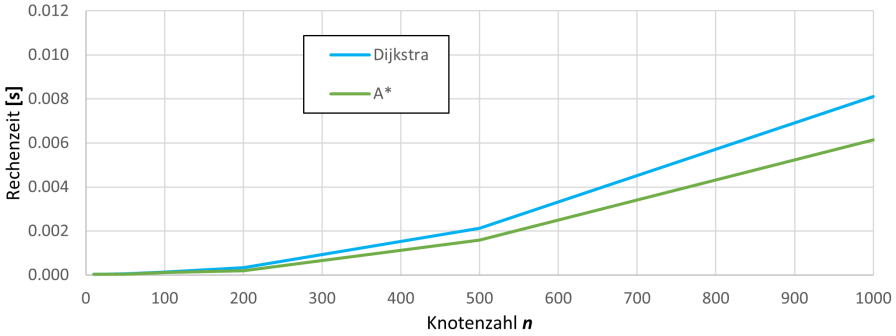


Abbildung 12.3: Gemessene Rechenzeiten der zweiten Versuchsreihe in Abhängigkeit der Knotenzahl.

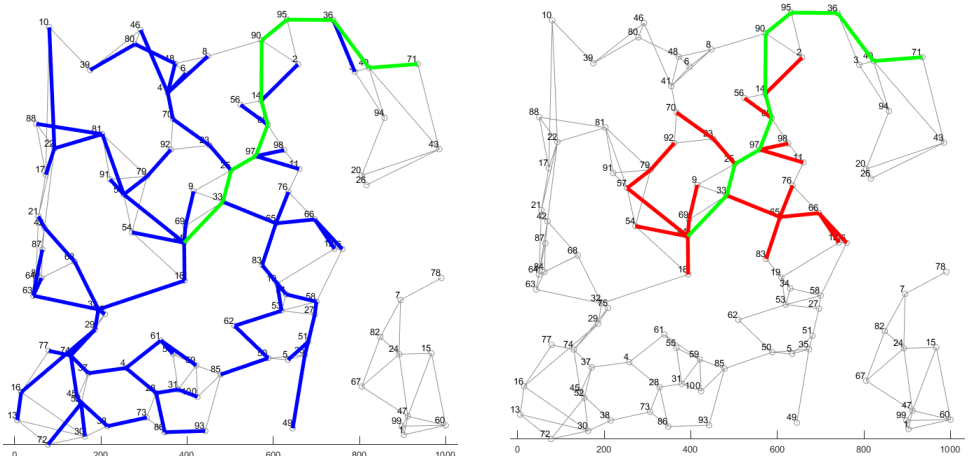


Abbildung 12.4: Suchpfad in grün mit *Dijkstra* (links), und *A** (rechts). Besuchte Knoten sind in blau, resp. rot markiert.

repetiert. Die Anzahl der Knoten im abgesuchten Netzwerk wirkt sich direkt auf die Rechenzeit aus. Der *Dijkstra*-Algorithmus weist eine Zeitkomplexität von $O(E \log V)$ auf, wobei E die Anzahl Kanten (engl. *edges*) und V die Anzahl Knoten (engl. *vertices*) darstellt. Für den *A**-Algorithmus ist die Zeitkomplexität einerseits abhängig von der verwendeten Heuristik, andererseits aber auch vom vorliegenden Netzwerk selbst. Aus diesem Grund lässt sich keine defintive Angabe zu O machen.

Die beiden Versuchsreihen unterscheiden sich zudem dahingehend, dass der Start- und Zielknoten bei der ersten Versuchsreihe im Netzwerk diametral gegenüber liegen. Dadurch gehen viele Knoten verloren, welcher *Dijkstra* als uninformierter Suchalgorithmus absuchen würde. In der zweiten Veruschsreihe werden hingegen Start- un Zielpunkt zufällig im Netzwerk ausgewählt. Es wird deshalb erwwartet, dass die Unterschiede in der Rechenzeit der beiden Algorithmen in der zweiten Versuchsreihe deutlich ausgeprägter sind.

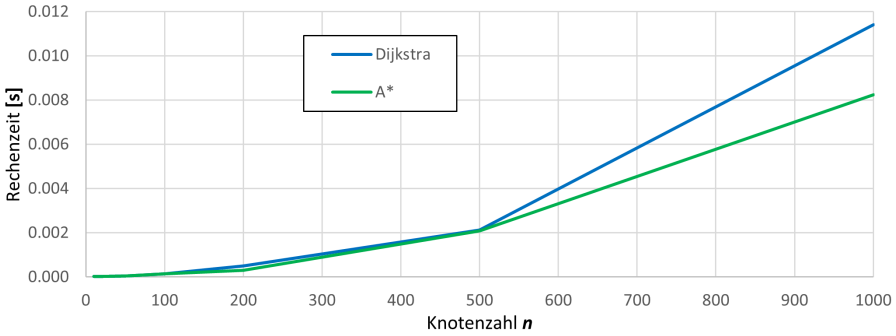


Abbildung 12.5: Gemessene Rechenzeiten der ersten Versuchsreihe in Abhängigkeit der Knotenzahl.

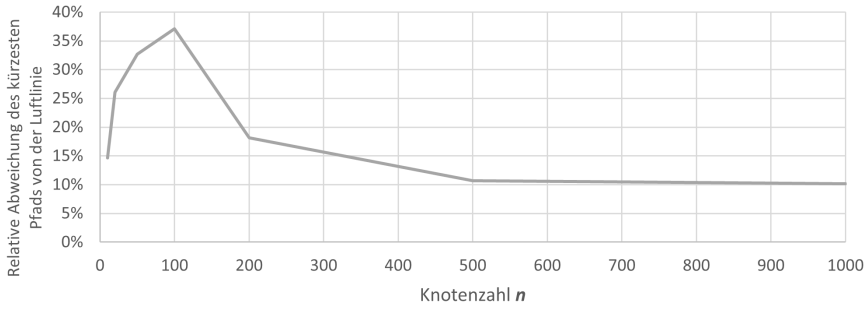


Abbildung 12.6: Relative Abweichung des kürzesten Pfads von der Luftlinie.

12.2.1 Einfluss der Knotenzahl auf die Rechenzeit

In 12.5 ist ersichtlich, dass der Unterschied in der Rechenzeit zwischen *Dijkstra* und *A** erst aber einer Knotenzahl von ca. $n = 500$ merklich ansteigt. Dieses etwas überraschende Resultat ist darauf zurückzuführen, dass bei steigender Knotenzahl die Abweichung des effektiven kürzesten Pfades von der Distanz der Luftlinie abnimmt. Die Effektivität von *A** mit euklidischer Heuristik ist wiederum grösser, wenn die Abweichung des kürzesten Pfads von der Luftlinie minimal ist. Bei Betrachtung von 12.6 wird dies ersichtlich, wobei die relative Abweichung erstaunlicherweise bei einer Knotenzahl von $n = 100$ maximal ist und nach $n = 500$ nur noch marginal abnimmt.

12.2.2 Einfluss der Position der Start- und Zielknoten auf die Rechenzeit

Zum Vergleich der Resultate in 12.2.1 zeigt 12.7 die Rechenzeiten der zweiten Versuchsreihe, in welcher die Start- und Zielknoten zufällig im Netzwerk ausgewählt wurden. Einerseits ist eine reduzierte durchschnittliche Rechenzeit festzustellen, was schlicht daran liegt, dass die zufällige Wahl der Knoten dazu führt, dass diese tendenziell weniger weit auseinander liegen.
Des weiteren ist festzustellen, dass sich die Unterschiede der Rechenzeiten zwischen *Dijkstra* und *A** deutlich früher abzeichnen. Dieses Phänomen lässt sich leicht durch die zielgerichtete Suche des *A**-Algorithmus erklären.

In 12.8 ist ersichtlich, dass bei einem im Netzwerk liegenden Startknoten die zielgerichtete Suche

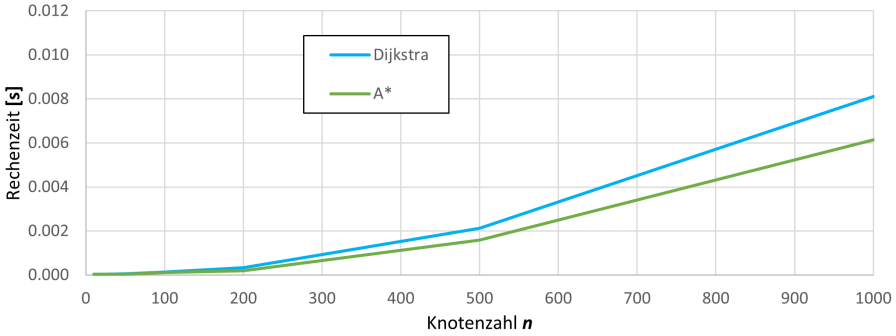


Abbildung 12.7: Gemessene Rechenzeiten der zweiten Versuchsreihe in Abhängigkeit der Knotenzahl.

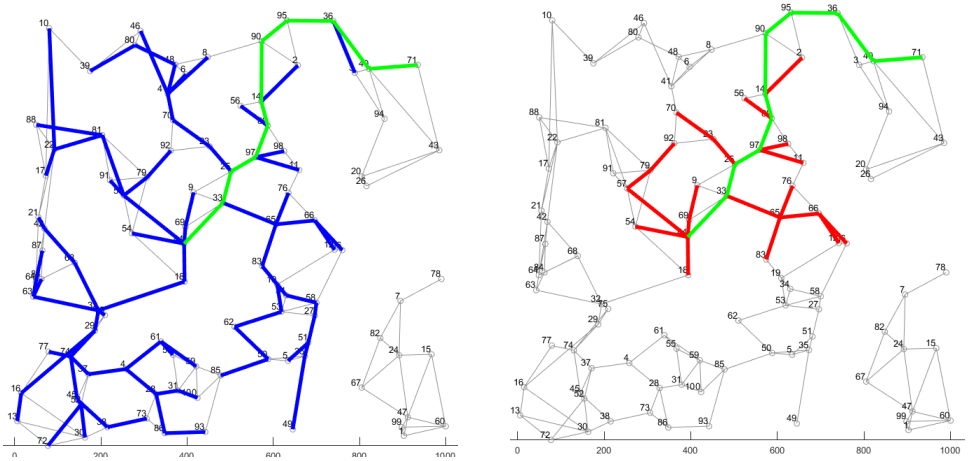


Abbildung 12.8: Suchpfad in grün mit *Dijkstra* (links), und *A** (rechts). Besuchte Knoten sind in blau, resp. rot markiert.

von *A** deutlich ausgeprägter zum Zuge kommt, als wenn dieser am Rand des Netzwerks liegen würde.

12.3 Ausblick

12.3.1 Optimierungsprobleme bei Graphen

Das Finden eines kürzesten Pfades, sprich die Minimierung der Summe der Kantengewichte, ist nur eines der Optimierungsprobleme, die sich im Bereich von Grafen aufstellen lassen. Verschiedene, ähnliche Problemstellungen lassen sich teilweise mit denselben Algorithmen lösen.
Im Bereich vom Computernetzwerken könnte zum Beispiel die Minimierung der Knotenzahl zur Datenübbertragung von Interesse sein. Dabei lässt sich dieses Problem einfach dadurch lösen, dass dem *Dijkstra*, oder dem *A**-Algorithmus anstelle der Graph-Matrix (mit Kantengewichten als Einträgen) die Adjazenz-Matrix als Argument übergeben wird. Der gefundene kürzeste Pfad enstpricht

der Anzahl benutzter Kanten, bzw. der Anzahl besuchter Knoten.

12.3.2 Wahl der Heuristik

Ein grundlegendes Problem bei der Anwendung des $A*$ oder ähnlicher informierter Suchalgorithmen ist die Wahl der Heurstik. Bei einem physischen Verkehrsnetz kann bspw. die euklidische Distanz problems ermittelt werde. Bei einem regionalen Netzwerk ist die Annahme eines orthogonalen X-Y-Koordinatenetzes absolut ausreichend. Dies gilt z.B. auch für das Vernessungsnetz der Schweiz[1] Bei überregionalen Netzwerken (Beispiel: Flugverbindungen) ist hingegen eine Berechnung im dreidimensionalen Raum, oder vereinfacht als Projektion auf das Geoid notwendig. Anonsten ist der Ablauf bei der Ausführung des Algorithmus allerdings identisch.
In nicht-physischen Netzwerken stellt sich jedoch eine zweite Problematik. Da eine physische Distanz entweder nicht ermittelt werden kann, oder aber nicht ausschlaggebend ist, sind andere Netzwerk-Eigenschaften zur Beurteilung beizuziehen. Die Zuverlässigkeit ist dabei aber in den meisten Fällen nicht vergleichbar hoch, wie bei der euklidischen Heuristik. Oftmals werden deshalb bei derartigen Problem auch Algorithmen angewendet, die eine deutlich optimierte Zeitkomplexität aufweisen, dafür aber nicht mit Sicherheit den effizienstesten Pfad finden.

[1] Die aktuelle Schweizer Referenzsystem LV95 benutzt ein E/N-Koordinatennetz, wobei aufgrund zunehmender Abweichung vom Referenzellipsoid bei grosser Entfernung vom Nullpunkt ein Korrekturfaktor für die Höhe angebracht werden muss.

Kapitel 13

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

13.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

13.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{13.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

13.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

13.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

13.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 14

Crystal Mɐth

Tim Tönz, Naoki Pross

14.1 Einleitung

Es gibt viele möglichkeiten sich in Kristallen zu verlieren. Auch wen man nur die Mathematischen möglichkeiten in betracht zieht, hat man noch viel zu viele Möglichkeiten sich mit kristallen zu beschäftigen. In diesem Articel ist daher der Fokus "nuräuf die Symmetrie gelegt. Im Abschitt über Symmetrien werden wir sehen, wie eine Symmetrie eines Objektes weit 2.ter versuch: Die Kristallographie ist ein grosses Thema, Symmetrien auch. Für beide bestehen schon bewährte Mathematische Modelle und Definitionen. Die

14.2 Symmetrie

Das Wort Symmetrie ist sehr alt und hat sich seltsamerweise von seinem ursprünglichen griechischen Wort $\sigma\nu\mu\mu\varepsilon\tau\rho\iota\alpha$ [1] fast nicht verändert. In der Alltagssprache mag es ein locker definierter Begriff sein, aber in der Mathematik hat Symmetrie eine sehr präzise Bedeutung.

Definition 14.1 (Symmetrie)**.** *Ein mathematisches Objekt wird als symmetrisch bezeichnet, wenn es unter einer bestimmten Operation invariant ist.*

Wenn der Leser noch nicht mit der Gruppentheorie in Berührung gekommen ist, ist vielleicht nicht ganz klar, was eine Operation ist, aber die Definition sollte trotzdem Sinn machen. Die Formalisierung dieser Idee wird bald kommen, aber zunächst wollen wir eine Intuition aufbauen.

Die intuitivsten Beispiele kommen aus der Geometrie, daher werden wir mit einigen geometrischen Beispielen beginnen. Wie wir jedoch später sehen werden, ist das Konzept der Symmetrie eigentlich viel allgemeiner. In Abbildung 14.1 haben wir einige Formen, die offensichtlich symmetrisch sind. Zum Beispiel hat ein Quadrat viele Achsen, um die es gedreht werden kann, ohne sein Aussehen zu verändern. Regelmässige Polygone mit n Seiten sind gute Beispiele, um eine diskrete Rotationssymmetrie zu veranschaulichen, was bedeutet, dass eine Drehung um einen Punkt um

[1] *Simmetría*: ein gemeinsames Mass habend, gleichmässig, verhältnismässig

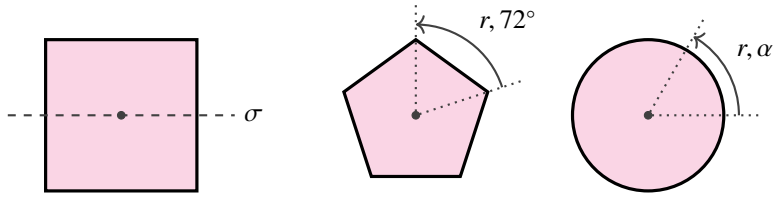


Abbildung 14.1: Beispiele für geometrisch symmetrische Formen.

einen bestimmten Winkel $360°/n$ sie unverändert lässt. Das letzte Beispiel auf der rechten Seite ist eine unendliche Rotationssymmetrie. Sie wird so genannt, weil es unendlich viele Werte für $\alpha \in \mathbb{R}$ gibt, die die Form unverändert lassen. Dies ist hoffentlich ausreichend, um die Bedeutung hinter der Notation zu verstehen, die nun eingeführt wird.

Definition 14.2 (Symmetriegruppe)**.** *Sei g eine Operation, die ein mathematisches Objekt unverändert lässt. Bei einer anderen Operation h definieren wir die Komposition $h \circ g$ als die Anwendung der Operationen nacheinander. Alle Operationen bilden unter Komposition eine Gruppe, die Symmetriegruppe genannt wird.*

Mit dem oben Gesagten können wir das n-Gon Beispiel formalisieren. Wenn wir r eine Drehung von $2\pi/n$ sein lassen, gibt es eine wohlbekannte Symmetriegruppe

$$C_n = \langle r \rangle = \left\{ \mathbb{1}, r, r^2, \ldots, r^{n-1} \right\} = \mathbb{Z}/n\mathbb{Z},$$

die Zyklische Gruppe heisst. Hier die Potenzen von r sind als wiederholte Komposition gemeint, d.h. $r^n = r \circ r \circ \cdots \circ r$. Die Schreibweise mit den spitzen Klammern wird als Erzeugendensystem bezeichnet. Das liegt daran, dass alle Elemente der Symmetriegruppe aus Kombinationen einer Teilmenge erzeugt werden, die als erzeugende Elemente bezeichnet werden. Die Reflexionssymmetriegruppe ist nicht so interessant, da sie nur $\{\mathbb{1}, \sigma\}$ enthält. Kombiniert man sie jedoch mit der Rotation, erhält man die so genannte Diedergruppe

$$D_n = \langle r, \sigma : r^{n-1} = \sigma^2 = (\sigma r)^2 = \mathbb{1} \rangle = \left\{ \mathbb{1}, r, \ldots, r^{n-1}, \sigma, \sigma r, \ldots, \sigma r^{n-1} \right\}.$$

Diesmal muss die Generator-Notation die Beziehungen zwischen den beiden Operationen beinhalten. Die ersten beiden sind leicht zu erkennen, für die letzte empfehlen wir, sie an einem 2D-Quadrat auszuprobieren.

Wir haben nun unseren Operationen Symbole gegeben, mit denen es tatsächlich möglich ist, eine nicht kommutative Algebra zu erstellen. Die naheliegende Frage ist dann, könnte es sein, dass wir bereits etwas haben, das dasselbe tut? Natürlich, ja. Dafür führen wir den Begriff der Darstellung ein.

Definition 14.3 (Darstellung einer Gruppe, Gruppenhomomorphismus)**.** *Seien G und H Gruppe mit unterschiedlicher Operation $\diamond$ bzw. $\star$. Ein Homomorphismus[2] ist eine Funktion $f : G \to H$, so dass für jedes $a, b \in G$ gilt $f(a \diamond b) = f(a) \star f(b)$. Man sagt, dass der Homomorphismus f G in H transformiert, oder dass H eine Darstellung von G ist.*

Beispiel. Die Elemente $r^k \in C_n$, wobei $0 < k < n$, stellen abstrakt eine Drehung von $2\pi k/n$ um den Ursprung dar. Die mit der Matrix

$$\Phi(r^k) = \begin{pmatrix} \cos(2\pi k/n) & -\sin(2\pi k/n) \\ \sin(2\pi k/n) & \cos(2\pi k/n) \end{pmatrix}$$

[2] Für eine ausführlichere Diskussion siehe §2.3.1 im Buch.

definierte Funktion von C_n nach $O(2)$ ist eine Darstellung von C_n. In diesem Fall ist die erste Gruppenoperation die Komposition und die zweite die Matrixmultiplikation. Man kann überprüfen, dass $\Phi(r^2 \circ r) = \Phi(r^2)\Phi(r)$. ○

Beispiel. Die Rotationssymmetrie des Kreises C_∞, mit einem unendlichen Kontinuum von Werten $\alpha \in \mathbb{R}$, entspricht perfekt dem komplexen Einheitskreis. Der Homomorphismus $\phi : C_\infty \to \mathbb{C}$ ist durch die Eulersche Formel $\phi(r) = e^{i\alpha}$ gegeben. ○

Die Symmetrien, die wir bis jetzt besprochen haben, haben immer mindestens einen Punkt unbesetzt gelassen. Im Fall der Rotation war es der Drehpunkt, bei der Spiegelung die Achse. Dies ist jedoch keine Voraussetzung für eine Symmetrie, da es Symmetrien gibt, die jeden Punkt zu einem anderen Punkt verschieben können. Ein aufmerksamer Leser wird bemerken, dass die unveränderten Punkte zum Eigenraum[3] der Matrixdarstellung der Symmetrieoperation gehören. Diesen Spezialfall, bei dem mindestens ein Punkt unverändert bleibt, nennt man Punktsymmetrie.

Definition 14.4 (Punktgruppe)**.** *Wenn jede Operation in einer Symmetriegruppe die Eigenschaft hat, mindestens einen Punkt unverändert zu lassen, sagt man, dass die Symmetriegruppe eine Punktgruppe ist.*

Um das Konzept zu illustrieren, werden wir den umgekehrten Fall diskutieren: eine Symmetrie, die keine Punktsymmetrie ist, die aber in der Physik sehr nützlich ist, nämlich die Translationssymmetrie. Von einem mathematischen Objekt U wird gesagt, dass es eine Translationssymmetrie $Q(x) = x + a$ hat, wenn es die Gleichung

$$U(x) = U(Q(x)) = U(x + a),$$

für ein gewisses a, erfüllt. Zum Beispiel besagt das erste Newtonsche Gesetz, dass ein Objekt, auf das keine Kraft einwirkt, eine zeitranslationsinvariante Geschwindigkeit hat, d.h. wenn $\vec{F} = \vec{0}$ dann $\vec{v}(t) = \vec{v}(t + \tau)$.

14.3 Kristalle

Unter dem Begriff Kristall sollte sich jeder ein Bild machen können. Wir werden uns aber nicht auf sein Äusseres fokussieren, sondern was ihn im Inneren ausmacht. Die Innereien eines Kristalles sind glücklicherweise relativ einfach definiert.

Definition 14.5 (Kristall)**.** *Ein Kristall besteht aus Atomen, welche sich in einem Muster arrangieren, welches sich in drei Dimensionen periodisch wiederholt.*

Ein Zweidimensionales Beispiel eines solchen Muster ist Abbildung **??**. Für die Überschaubarkeit haben wir ein simples Muster eines einzelnen XgrauenX Punktes gewählt in nur Zwei Dimensionen. Die eingezeichneten Vektoren a und b sind die kleinstmöglichen Schritte im Raum bis sich das Kristallgitter wiederholt. Dadurch können von einem einzelnen XGrauenX Gitterpunkt in **??** können mit einer ganzzahligen Linearkombination von a und b alle anderen Gitterpunkte des Kristalles erreicht werden. Ein Kristallgitter kann eindeutig mit a und b und deren winkeln beschrieben werden weswegen a und b auch Gitterparameter genannt werden. Im Dreidimensionalen-Raum können alle Gitterpunkte mit derselben Idee und einem zusätzlichen Vektor also FRMEL FÜR TRANSLATIONSVEKTOR erreicht werden. Da sich das Ganze Kristallgitter wiederholt, wiederholen sich auch die Eigenschaften eines Gitterpunktes Periodisch mit eiem

[3] Zur Erinnerung $E_\lambda = \text{null}(\Phi - \lambda I)$, $\vec{v} \in E_\lambda \implies \Phi\vec{v} = \lambda\vec{v}$

14.4 Piezoelektrizität

Literatur

[1] Hans-Dieter Lang. *Elektrotechnik 2*. Fachhochschule Ostschweiz Rapperswil, Feb. 2020.

[2] Charles C. Pinter. *A Book of Abstract Algebra*. Dover Publications Inc.; 2. Edition, Jan. 2010. ISBN: 978-0-486-47417-5.

[3] Donald E. Sands. *Introduction to Crystallography*. Dover Publications Inc., 1993. ISBN: 978-0-486-67839-9.

Kapitel 15

Thema

Joshua Bär und Michael Steiner

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

15.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

15.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{15.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

15.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

15.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

15.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 16

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

16.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

16.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{16.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

16.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

16.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

16.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 17

McEliece-Kryptosystem

Reto Fritsche

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

17.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

17.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{17.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

17.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

17.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

17.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

17.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

17.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 18

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

18.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

18.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{18.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

18.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

18.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

18.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 19

Thema

Adrian Schuler und Thomas Reichlin

19.1 Einleitung

In diesem Kapitel geht es darum die Matrix im dreidimensionalen Spannungszustand genauer zu untersuchen. In der Geotechnik wendet man solche Matrizen an, um Spannungen im Boden zu berechnen. Mit diesen Grundlagen dimensioniert man beispielsweise Böschungen, Fundationen, Dämme und Tunnels. Ebenfalls benötigt man diese Matrix, um aus Versuchen Kennzahlen über den anstehenden Boden zu gewinnen. Besonderes Augenmerk liegt dabei auf dem Oedometer - Versuch.

Bei dieser Untersuchung der zugehörigen Berechnungen hat man es mit Vektoren, Matrizen und Tensoren zu tun. Um die mathematische Untersuchung vorzunehmen, beschäftigt man sich zuerst mit den spezifischen Gegebenheiten und Voraussetzungen. Ebenfalls gilt es ein paar wichtige Begriffe und deren mathematischen Zeichen einzuführen, damit sich den Berechnungen schlüssig folgen lässt.

In diesem Kapitel hat man es insbesondere mit Spannungen und Dehnungen zu tun. Mit einer Spannung ist hier jedoch keine elektrische Spannung gemeint, sondern eine Kraft geteilt durch Fläche.

19.2 Einführung wichtige Begriffe

l_0 = Ausgangslänge [m]

Δl = Längenänderung nach Kraftauftrag [m]

Δb = Längenänderung in Querrichtung nach Kraftauftrag [m]

ε = Dehnung [−]

σ = Spannung [kPa]

E = Elastizitätsmodul [kPa]

ν = Querdehnungszahl; Poissonzahl [−]

$$F = \text{Kraft [kN]}$$

$$A = \text{Fläche [m}^2\text{]}$$

$$t = \text{Tiefe [m]}$$

$$s = \text{Setzung, Absenkung [m]}$$

Beziehungen

$$\varepsilon = \frac{\Delta l}{l_0}$$

$$\varepsilon_q = \frac{\Delta b}{l_0} = \varepsilon \cdot \nu$$

$$\sigma = \frac{N}{A}$$

$$F = \int_A \sigma dA$$

$$\varepsilon' = \frac{1}{l_0}$$

19.3 Einführung wichtige Begriffe

Tensoren wurden als erstes in der Elastizitätstheorie eingesetzt. (Quelle Herr Müller) In der Elastizitätstheorie geht es darum viele verschiedene Komponenten zu beschreiben. Mit einer Matrix oder einem Vektor kann man dies nicht mehr bewerkstelligen. Wenn man den dreidimensionalen Spannungszustand abbilden möchte, müsste man mehrere Vektoren haben. Deshalb wurden 1840 von Rowan Hamilton Tensoren in die Mathematik eingeführt. Woldemar Voigt hat den Begriff in die moderne Bedeutung von Skalar, Matrix und Vektor verallgemeinert. Albert Einstein hat Tensoren zudem in der allgemeinen Relativitätstheorie benutzt. Tensor sind eine Stufe höher als Matrizen. Matrizen sind 2. Stufe. Da Tensoren eine Stufe höher sind, kann man auch Matrizen, Vektoren und Skalare als Tensoren bezeichnen. Der Nachteil von den Tensoren ist, dass man die gewohnten Rechenregeln, die man bei Vektoren oder Matrizen kennt, nicht darauf anwenden kann. Man ist deshalb bestrebt die Tensoren als Vektoren und Matrizen darzustellen, damit man die gewohnten Rechenregeln darauf anwenden kann. (Quelle Wikipedia) In der vorliegenden Arbeit sind bereits alle Tensoren als Matrizen 2. Stufe abgebildet. Trotzdem kann man diese Matrizen wie vorher beschrieben als Tensor bezeichnen. Da diese als Matrizen abgebildet sind, dürfen wir die bekannten Rechenregeln auf unsere Tensoren anwenden.

19.4 Spannungsausbreitung

Anhand untenstehendem Bild kann ein einfaches Beispiel betrachtet werden. Es gibt eine Flächenlast (Kraft), diese wird auf den Boden abgetragen. Diese Last muss dann vom Boden aufgenommen werden. Im Boden entsteht nebst der Eigenspannung eine weitere Spannung durch diese Last (Zusatzspannung). Diese Zusatzspannung σ ist abhängig von (x, y, t). Je nach dem, wo man sich im Boden befindet variert die Spannung. Mit der Tiefe wird die Zusatzspannung geringer. Die Ausbreitung der Zusatzspannung im Boden hat die Form einer Zwiebel. Durch Untersuchung der Spannung an verschiedenen Punkten im Boden, kann man eine Funktion abtragen. Dasselbe macht man auch

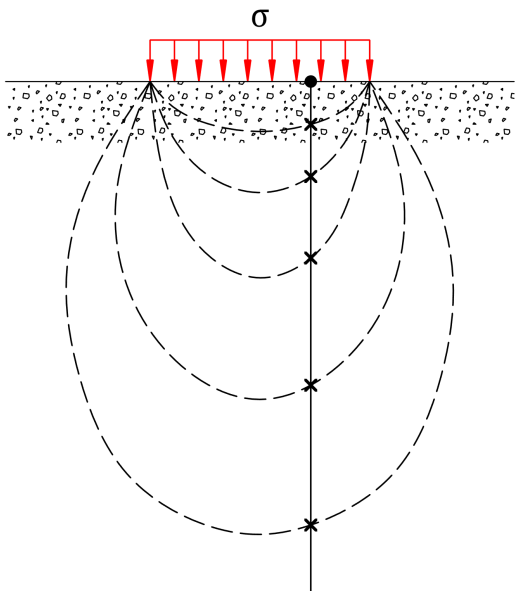


Abbildung 19.1: Ausbreitung der Spannung im Boden

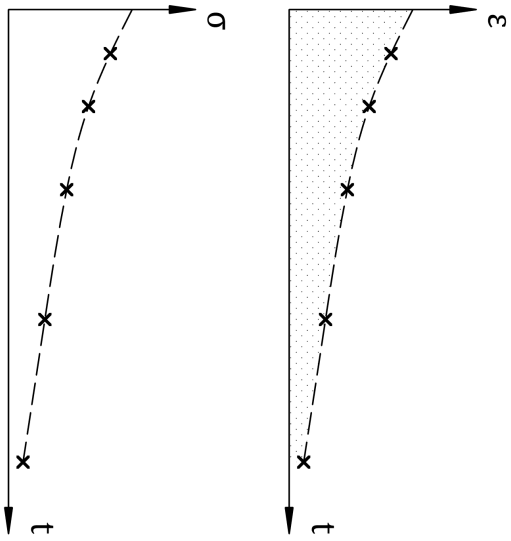


Abbildung 19.2: Funktionen Spannung und Dehnung

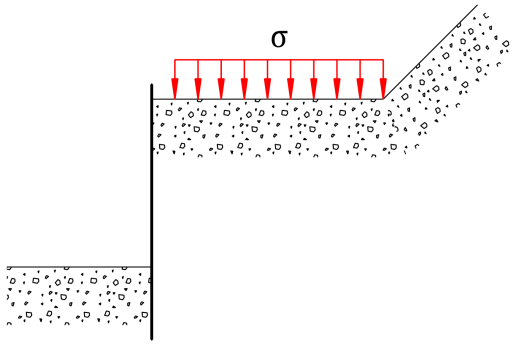


Abbildung 19.3: Beispiel Lastauftrag auf Boden

mit der Dehnung. Es zeigt sich, dass die Form der beiden Funktionen gleich ist. Dies erklärt sich dadurch, dass die Spannung und die Dehnung proportional zueinander sind.

Anhand eines etwas schwierigeren Beispiels sieht man, dass die Spannungsausbreitung nicht immer ganz einfach ist. Man hat hier eine Baugrube mit einem Baugrubenabschluss, wo ein Teil des Bodens abgetragen wurde. Was aber immer noch gilt ist, dass die Spannung σ von drei Variablen abhängig ist (x, y, t). Ansätze um die Spannungsausbreitung zu berechnen gibt es je nach Bodentyp verschiedene.

Die Spannungsausbreitung ist uns jedoch gegeben, es geht nicht darum, dies genauer zu untersuchen. Durch die Spannungsausbreitung und das Elastizitätsmodul kann man eine Dehnung berechnen. Anhand dieser Dehnung kann man mit einem Integral wiederum die Setzung berechnen.

$$\varepsilon = \frac{\sigma}{E}$$

$$s = \int_0^\infty \varepsilon \; dt$$

Die Setzung zu bestimmen ist in der Geotechnik sehr wichtig. Besonders ungleichmässige Setzungen können bei Bauwerken Probleme ergeben. Es gilt also die Bauwerke so zu dimensionieren, dass es verträgliche Setzungen gibt.

19.5 Proportionalität Spannung-Dehnung

Das Hook'sche Gesetz beschreibt die elastische Längenänderung von Festkörpern im Zusammenhang mit einer Krafteinwirkung. Die Längenänderung Δl ist proportional zur Krafteinwirkung F.

$$F \sim \Delta l$$

Man kann dies nur im Bereich vom linearen-elastischen Materialverhalten anwenden. Das heisst, dass alle Verformungen reversibel sind, sobald man die Kraft wegnimmt. Es findet somit keine dauernde Verformung statt. Da es sehr praktisch ist die Längenänderung nicht absolut auszudrücken haben wir ε. Die Dehnung ε beschreibt die relative Längenänderung. Die Dehnung ε ist wiederum proportional zu der aufgebrachten Spannung. Im Bauingenieurwesen hat man es oft mit grösseren Teilen oder grösseren Betrachtungsräumen zu tun. Da ist es nun natürlich sehr sinnvoll, wenn wir

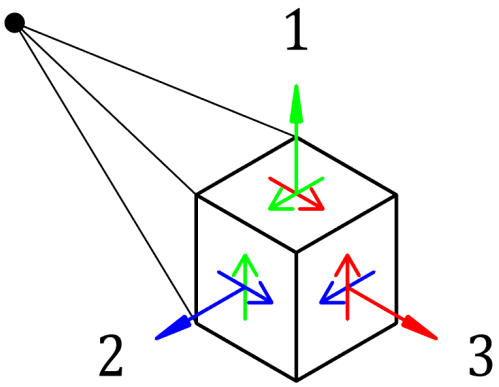


Abbildung 19.4: Infinitesimales Bodenteilchen

nicht mit absoluten Zahlen rechnen, sondern unabhängig von der Länge den Zustand mit Dehnung ε beschreiben können. Mithilfe vom E-Modul, (steht für Elastizitätsmodul) einer Proportionalitätskonstante, kann man das in eine Gleichung bringen, wie man hier sieht. Das E-Modul beschreibt, das Verhältnis von Kraftaufnahme eines Werkstoffes und dessen zusammenhängender Längenveränderung. (Quelle Wikipedia)

$$\sigma = E \cdot \varepsilon$$

$$E = \frac{\Delta\sigma}{\Delta\varepsilon} = const.$$

Aus diesem Verhältnis kann man das E-Modul berechnen. Je nach Material ist dies verschieden. Das E-Modul lässt sich nur im linearen-elastischen Materialverhalten anwenden. Für Bodenmaterial gibt es ein spezielles E-Modul. Dieses wird mit dem Oedometer-Versuch ermittelt. Es wird mit E_{OED} ausgedrückt. Dieser Versuch wird später noch beschrieben. Der Oedometer-Versuch ist abhängig von den diesem Kapitel zu untersuchenden Matrizen.

19.6 Dreiachsiger Spannungszustand

Wie im Kapitel Spannungsausbreitung beschrieben herrscht in jedem Punkt ein anderer Spannungszustand. Um die Spannung im Boden genauer untersuchen zu können, führt man einen infinitesimales Bodenteilchen ein. Das Bodenteilchen ist geometrisch gesehen ein Würfel. An diesem Bodenteilchen trägt man die Spannungen ein in alle Richtungen.

An diesem infinitesimalen Bodenteilchen hat man ein räumliches Koordinatensystem, die Achsen $(1, 2, 3)$. Die Achsen vom Koordinatensystem zeigen aus den 3 ersichtlichen Flächen heraus. Pro ersichtliche Fläche haben wir eine Normalspannung und zwei Schubspannungen. Im Gegensatz zum eindimensionalen Zustand entstehen bei einer Belastung des Bodenteilchens eine Vielzahl an Spannungen. Es entstehen diverse Normal- und Schubspannungen. Die Schubspannungen befinden sich an der Fläche, sie gehen rechtwinklig von den Achsen weg. Die Schubspannungen auf einer Fläche stehen im 90 Grad Winkel zueinander. Geschrieben werden diese mit σ, mit jeweils zwei Indizes. Die Indizes geben uns an, in welche Richtung die Spannungen zeigen. Der erste Index ist die Fläche auf welcher man sich befindet. Der zweite Index gibt an, in welche Richtung die Spannung

zeigt, dabei referenzieren die Indizes auch auf die Achsen $(1, 2, 3)$. Bei den Spannungen sind immer positive als auch negative Spannungen möglich. Es können also Druck- oder Zugspannungen sein.

Zunächst wird untenstehend der allgemeine Spannungszustand betrachtet.

Spannungstensor 2. Stufe i,j ∈ 1,2,3

$$\overline{\sigma} = \sigma_{ij} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} = \qquad \Rightarrow \qquad \vec{\sigma} = \begin{pmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{pmatrix}$$

Dehnungstensor 2. Stufe k,l ∈ 1,2,3

$$\overline{\varepsilon} = \varepsilon_{kl} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix} = \qquad \Rightarrow \qquad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{pmatrix}$$

Bei diesen zwei obenstehenden Formeln kann man sehen wie Matrizen zu einem Vektor umgewandelt wurden. Unter dem Kapitel Hadamard-Algebra kann man sehen, dass man dabei Zeile um Zeile in eine Spalte schreiben kann, sodass es einen Vektor ergibt.

Elastizitätstensor 4. Stufe i,j,k,l ∈ 1,2,3

$$\overline{\overline{C}} = C_{ijkl} = \begin{pmatrix} C_{1111} & C_{1112} & C_{1113} & C_{1121} & C_{1122} & C_{1123} & C_{1131} & C_{1132} & C_{1133} \\ C_{1211} & C_{1212} & C_{1213} & C_{1221} & C_{1222} & C_{1223} & C_{1231} & C_{1232} & C_{1233} \\ C_{1311} & C_{1312} & C_{1313} & C_{1321} & C_{1322} & C_{1323} & C_{1331} & C_{1332} & C_{1333} \\ C_{2111} & C_{2112} & C_{2113} & C_{2121} & C_{2122} & C_{2123} & C_{2131} & C_{2132} & C_{2133} \\ C_{2211} & C_{2212} & C_{1113} & C_{2221} & C_{2222} & C_{2223} & C_{2231} & C_{2232} & C_{2233} \\ C_{2311} & C_{2312} & C_{2313} & C_{2321} & C_{2322} & C_{2323} & C_{2331} & C_{2332} & C_{2333} \\ C_{3111} & C_{3112} & C_{3113} & C_{3121} & C_{3122} & C_{3123} & C_{3131} & C_{3132} & C_{3133} \\ C_{3211} & C_{3212} & C_{3213} & C_{3221} & C_{3222} & C_{3223} & C_{3231} & C_{3232} & C_{3233} \\ C_{3311} & C_{3312} & C_{3313} & C_{3321} & C_{3322} & C_{3323} & C_{3331} & C_{3332} & C_{3333} \end{pmatrix}$$

Dieser Elastizitätstensor muss eine quadratische Matrix mit 3^4 Einträgen ergeben, da die Basis mit den drei Richtungen $1, 2, 3$ und die Potenz mit den 4 Indizes mit je $1, 2, 3$ definiert sind. Dies gibt daher eine 9 x 9 Matrix, welche zudem symmetrisch ist.

Folglich gilt:

$$\overline{\overline{C}} = \overline{\overline{C}}^{\,T}$$

Allgemeine Spannungsgleichung (mit Vektoren und Tensor)

$$\vec{\sigma} = \overline{\overline{C}} \cdot \vec{\varepsilon}$$

$$\begin{pmatrix}\sigma_{11}\\ \sigma_{12}\\ \sigma_{13}\\ \sigma_{21}\\ \sigma_{22}\\ \sigma_{23}\\ \sigma_{31}\\ \sigma_{32}\\ \sigma_{33}\end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-2\nu & 0 & 0 & 0 & \nu & 0 & 0 & 0 & \nu \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ \nu & 0 & 0 & 0 & 1-2\nu & 0 & 0 & 0 & \nu \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ \nu & 0 & 0 & 0 & \nu & 0 & 0 & 0 & 1-2\nu \end{pmatrix} \begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{12}\\ \varepsilon_{13}\\ \varepsilon_{21}\\ \varepsilon_{22}\\ \varepsilon_{23}\\ \varepsilon_{31}\\ \varepsilon_{32}\\ \varepsilon_{33}\end{pmatrix}$$

Man kann das zudem auch als Indexnotation aufschreiben.

$$\sigma_{ij} == \sum_{k} = 1^3 \sum_{l} = 1^3 C_{ijkl} \cdot \varepsilon_{kl}$$

Um die Berechnung an einem Beispiel zu veranschaulichen:

$$\sigma_{22} = \frac{E \cdot \nu}{(1+\nu)(1-2\nu)} \cdot \varepsilon_{11} + \frac{E}{(1+\nu)} \cdot \varepsilon_{22} + \frac{E \cdot \nu}{(1+\nu)(1-2\nu)} \cdot \varepsilon_{33}$$

Anhand dem Tensor der allgemeinen Spannungsgleichung kann man zwar eine Symmetrie erkennen. Die verschiedenen Einträge wechseln sich aber mit einander ab und es gibt keine klaren Blöcke mit nur einem gleichen Eintrag. Man greift deshalb auf die Voigt'sche Notation zurück.

Zur Notation wird die Voigt'sche Notation benutzt. Das sieht wie folgt aus:

$$\overline{\sigma} = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix} = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ & \sigma_{22} & \sigma_{23}\\ sym & & \sigma_{33}\end{pmatrix} \Rightarrow \vec{\sigma} = \begin{pmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{23}\\ \sigma_{13}\\ \sigma_{12}\end{pmatrix}$$

In der Voigt'sche Notation hat man die Reihenfolge von der Ecke links oben, diagonal zur Ecke rechts unten. Danach ist noch σ_{23}, σ_{13} und σ_{12} aufzuschreiben um den Vektor zu erhalten.

Eine weitere Besonderheit ist die Symmetrie der Matrix. So entspricht σ_{23} dem Wert σ_{32} und σ_{13} dem Wert σ_{31}. Dies ist dadurch bedingt, dass die Kräfte in seitlicher Richtung im Boden die gleichen Werte annehmen. Man hat in dieser Berechnung ein isotropes Material. Im infinitesimalen Körper muss ein Gleichgewicht vorherrschen. Ist kein Gleichgewicht vorhanden, würde sich der Körper zu drehen beginnen. Es macht somit keinen Unterschied, ob man auf der Achse 2 in Richtung 3 geht, oder auf der Achse 3 in Richtung 2.

Da die Spannung proportional zur Dehnung ist, kann man die ganze Voigt'sche Notation auch mit der Dehnung ausdrücken. Auch hier wandelt man das ganze gemäss der Reihenfolge in einen Vektor um.

$$\overline{\varepsilon} = \begin{pmatrix}\varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13}\\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23}\\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33}\end{pmatrix} = \begin{pmatrix}\varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13}\\ & \varepsilon_{22} & \varepsilon_{23}\\ \text{sym} & & \varepsilon_{33}\end{pmatrix} \qquad \Rightarrow \qquad \vec{\varepsilon} = \begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{22}\\ \varepsilon_{33}\\ \varepsilon_{23}\\ \varepsilon_{13}\\ \varepsilon_{12}\end{pmatrix}$$

Mit der hergeleiteten Beziehung für die Spannungsgleichung anhand vom E-Modul, der allgemeinen linearen Spannungsgleichung kann man diese Beziehungen neu aufschreiben. Man benötigt dazu den zuvor berechneten Dehnungsvektor. Die Gleichung besagt:

$$\text{Spannungsvektor} = \text{Elastizitätstensor} \cdot \text{Dehnungsvektor}$$

$$\vec{\sigma} = \overline{\overline{C}} \cdot \vec{\varepsilon}$$

Die Vektoren haben je 6 Einträge. Um das ganze auszudrücken braucht es einen 6 x 6 Elastizitätstensor. Der Tensor hat sich also im Vergleich zum 9 x 9 Tensor verkleinert. Dies ist deshalb der Fall, da man in den Achsen 2 und 3 Symmetrien hat. Dadurch kann man die Einträge ($\varepsilon_{21} = \varepsilon_{12}; \varepsilon_{31} = \varepsilon_{13}; \varepsilon_{32} = \varepsilon_{23}$) zusammenfassen und drei Einträge verschwinden, da drei Dehnungen gleich sind. Das ganze sieht dann wie folgt aus:

$$\begin{pmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{23}\\ \sigma_{13}\\ \sigma_{12}\end{pmatrix} = \begin{pmatrix}C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16}\\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26}\\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36}\\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46}\\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56}\\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66}\end{pmatrix}\begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{22}\\ \varepsilon_{33}\\ \varepsilon_{23}\\ \varepsilon_{13}\\ \varepsilon_{12}\end{pmatrix}$$

Die Spannung σ_{11} besteht somit aus Anteilen von all diesen sechs Konstanten und den verschiedenen Dehnungen. Zuvor bei der Voigt'schen Notation hat man jedoch gesehen, dass die Tensoren symmetrisch sind. Folglich muss auch dieser Elastizitätstensor symmetrisch sein. Das sind folgendermassen aus:

$$\begin{pmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{23}\\ \sigma_{13}\\ \sigma_{12}\end{pmatrix} = \begin{pmatrix}C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16}\\ & C_{22} & C_{23} & C_{24} & C_{25} & C_{26}\\ & & C_{33} & C_{34} & C_{35} & C_{36}\\ & & & C_{44} & C_{45} & C_{46}\\ & & & & C_{55} & C_{56}\\ \text{sym} & & & & & C_{66}\end{pmatrix}\begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{22}\\ \varepsilon_{33}\\ \varepsilon_{23}\\ \varepsilon_{13}\\ \varepsilon_{12}\end{pmatrix}$$

Die Konstanten C kann man nun anders ausdrücken. Und zwar bewerkstelligt man dies mithilfe vom Hook'schen Gesetz.

$$\begin{pmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{23}\\ \sigma_{13}\\ \sigma_{12}\end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)}\begin{pmatrix}1-2\nu & \nu & \nu & 0 & 0 & 0\\ \nu & 1-2\nu & \nu & 0 & 0 & 0\\ \nu & \nu & 1-2\nu & 0 & 0 & 0\\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}\end{pmatrix}\begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{22}\\ \varepsilon_{33}\\ \varepsilon_{23}\\ \varepsilon_{13}\\ \varepsilon_{12}\end{pmatrix}$$

Mithilfe der Poissonzahl, welche uns die Querdehnung angibt, sprich wie viel sich der Körper in Querrichtung verformt und dem E-Modul kann man alle Konstanten ausdrücken. Bei einigen fällt auf, dass diese 0 werden. Der Tensor besagt also, dass diese jeweiligen Konstanten keinen Einfluss auf unsere Spannung haben. Man sieht nun auch ganz gut, dass sich im Vergleich bei der allgemeinen Darstellung der Spannungsgleichung, die Einträge verschoben haben. Man hat nun eine sehr vorteilhafte Anordnung der verschiedenen Blöcke im Tensor. Als Beispiel kann man sich σ_{33} anschauen. Es ist ersichtlich, dass die Konstante C_{31}, C_{32}, C_{33}, C_{35} und C_{36} keinen Einfluss auf σ_{33}

haben. Dies kann wie folgt erklärt werden. Auf Achse 3 geht σ_{33} in Richtung 3. Der Einfluss von C_{31}, Achse 3 in Richtung 1 hat keinen Einfluss auf σ_{33}.

Von $\overline{\overline{C}}$ bildet man nun die Inverse Matrix $\overline{\overline{C}}^{-1}$ stellt sich die ganze Gleichung um.

$$\vec{\varepsilon} = \overline{\overline{C}}^{-1} \cdot \vec{\sigma}$$

$$\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ \varepsilon_{23} \\ \varepsilon_{13} \\ \varepsilon_{12} \end{pmatrix} = \frac{1}{E} \begin{pmatrix} 1 & -\nu & -\nu & 0 & 0 & 0 \\ -\nu & 1 & -\nu & 0 & 0 & 0 \\ -\nu & -\nu & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2+2\nu & 0 & 0 \\ 0 & 0 & 0 & 0 & 2+2\nu & 0 \\ 0 & 0 & 0 & 0 & 0 & 2+2\nu \end{pmatrix} \begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{23} \\ \sigma_{13} \\ \sigma_{12} \end{pmatrix}$$

Die zwei Blöcke links unten und rechts oben sind immer noch vorhanden. Im Vergleich wo wir die Inverse noch nicht gemacht haben hat sich das nicht geändert. Um die Einflüsse der Parameter zu veranschaulichen schreibt man folgende Gleichung.

$$\varepsilon_{22} = \frac{1}{E}\sigma_{22} - \frac{\nu}{E}\sigma_{11} - \frac{\nu}{E}\sigma_{33}$$

ε_{22} beschreibt die Dehnung in Achse 2 und in Richtung 2. In erster Linie hängt ε_{22} von σ_{22} ab. Wenn die Poisson - Zahl grösser wird oder σ_{11} oder σ_{33}, dann wird dadurch die Dehnung ε_{22} kleiner. Das heisst, auf Kosten von Verformung in anderer Richtung als Achse 2 Richtung 2 erfolgt die Verformung an anderer Stelle. Wiederum hat die Schubspannung auf σ_{11} keinen Einfluss.

Nun kennt man die Beziehung der 6 Dehnungen mit den 6 Spannungen. In der Geotechnik wäre das aufgrund der vielen Komponenten sehr umständlich um damit Berechnungen zu machen. Es braucht daher eine Vereinfachung mit Invarianten, welche im nächsten Kapitel beschrieben sind.

19.7 Spannungsausbreitung

Trotz der Vereinfachung lässt sich mit den Invarianten die Realität adäquat abbilden. Als erste Bedingung stellt man folgendes Verhältnis auf:

$$\sigma_{22} = \sigma_{33}$$

Dies deshalb, da man von einem isotropen Bodenmaterial ausgeht. In Achse 22, Richtung 22 hat man den gleichen Boden wie in Achse 33 und Richtung 33. Das Verhalten bezüglich Kraftaufnahme, Dehnung Spannung ist somit dasselbe.

Man führt die zwei Werte p als hydrostatische Spannung und q als deviatorische Spannung ein. Die Berechnung von p und q sieht wie folgt aus:

$$p = \frac{\sigma_{11} + \sigma_{22} + \sigma_{33}}{3}$$

oder durch Vereinfachung, da $\sigma_{22} = \sigma_{33}$:

$$p = \frac{\sigma_{11} + 2\sigma_{33}}{3}$$

$$q = \sigma_{11} - \sigma_{33}$$

p ist das arithmetische Mittel von der Spannung im infinitesimalen Würfel. q ist die Differenz zwischen der Spannung in vertikaler Richtung und der Spannung in Richtung 2 und 3. Man kann p als Druckspannung und q als Schubspannung anschauen.

Aus der Formel vom vorherigen Kapitel konnten wir die Spannungen berechnen. Deshalb kann man nun p und q in die Gleichung einsetzen. Die Dehnungen werden mit neuen Variablen eingeführt. Die Deviatorische Dehnung kann mit einer Schubdehnung verglichen werden. Die hydrostatische Dehnung kann mit einer Kompressionsdehnung verglichen werden.

$$\overbrace{\sigma_{11} - \sigma_{33}}^{q} = \frac{3E}{2(1+\nu)} \overbrace{\frac{2}{3}(\varepsilon_{11} - \varepsilon_{33})}^{\varepsilon_\nu}$$

$$\overbrace{\frac{\sigma_{11} + 2\sigma_{33}}{3}}^{p} = \frac{E}{3(1-2\nu)} \overbrace{(\varepsilon_{11} - 2\varepsilon_{33})}^{\varepsilon_s}$$

ε_s = Hydrostatische Dehnung[−]

ε_ν = Deviatorische Dehnung[−]

Diese Komponenten kann man nun in die Vereinfachte Matrix einsetzen. Man hat dann eine Matrix multipliziert mit einem Vektor und erhält einen Vektor.

$$\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} \frac{3E}{2(1+\nu)} & 0 \\ 0 & \frac{E}{3(1-2\nu)} \end{pmatrix} \begin{pmatrix} \varepsilon_s \\ \varepsilon_\nu \end{pmatrix}$$

Mit dieser Formel lassen sich verschieden Parameter von Versuchen analysieren und berechnen. Ein solcher Versuch, den oft in der Geotechnik durchgeführt wird ist der Oedometer-Versuch. Im nächsten Kapitel wird die Anwendung der Matrix an diesem Versuch beschrieben.

19.8 Spannungsausbreitung

Beim Oedometer - Versucht hat man einen Stahlring mit einer Filterplatte am Boden. In diesen Stahlring wird eine Bodenprobe eingefüllt. Anschliessend wir mit einer Platte das Bodenmaterial mit einer ansteigenden Kraft belastet.

Die Probe wird sich so verdichten. Das Volumen nimmt ab. Der Stahlring verhindert ein seitliches ausbrechen oder entweichen der Bodenprobe. Die Dehnung auf der Seite beträgt somit 0. Mit dem Wert der Kraft und der Fläche lässt sich die Spannung berechnen. Anhand der Volumenabnahme errechnet man die Dehnung. Aus diesen Werten lässt sich wiederum das E-Modul bestimmen. Beim Oedometer Versuch ist das E-Modul als E_{OED} bezeichnet.

Das E_{OED} hat man speziell in der Geotechnik. Dies aufgrund der speziellen Situation wo man sich mit dem infinitesimalen Würfel befindet. Mit dem Stahlring, der verhindert das Material seitlich entweichen kann hat man ganz ähnliche Verhältnisse wie tief im Untergrund. Auch dort kann das Material bei einer Belastung nicht seitlich entweichen.

Wichtig ist nochmals zu betonen, dass alle diese beschriebenen Berechnungen ausschliesslich im linear-elastischen Materialverhalten funktionieren. So ist es auch beim Oedometer - Versuch. Den Versuch kann man auf einem σ und ε Diagramm abtragen.

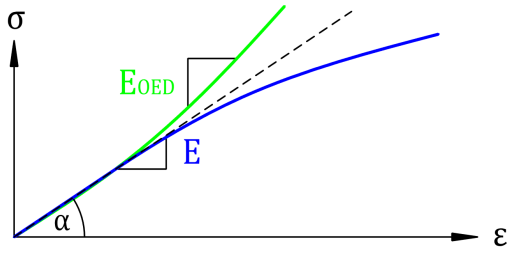


Abbildung 19.5: Diagramm Oedometer - Versuch

Bei einem Versuch mit anderem Baumaterial wie beispielsweise Holz nimmt die Dehnung im Laufe des Versuchs stärker zu, obwohl weniger Spannung abgetragen wird. Bei den meisten Böden ist dies anders. Durch die Komprimierung nimmt der Boden mehr Spannung auf, und verformt sich zugleich weniger stark.

Man kann die Dehnung in unsere vereinfachte Matrix einsetzen. Das E-Modul ersetzt man mit dem E_{OED}.

$$\overbrace{\sigma_{11} - \sigma_{33}}^{q} = \frac{3E}{2(1+\nu)} \overbrace{\frac{2}{3}(\varepsilon_{11} - 0)}^{\varepsilon_v}$$

$$\overbrace{\frac{\sigma_{11} + 2\sigma_{33}}{3}}^{p} = \frac{E}{3(1-2\nu)} \overbrace{(\varepsilon_{11} - 2 \cdot 0)}^{\varepsilon_s}$$

$$\begin{pmatrix} \sigma_{11} - \sigma_{33} \\ \sigma_{11} + 2\sigma_{33} \end{pmatrix} = \begin{bmatrix} \frac{E_{OED}}{(1+\nu)} & 0 \\ 0 & \frac{E_{OED}}{(1-2\nu)} \end{bmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{11} \end{pmatrix}$$

An einem geeigneten Punkt, wo man noch im linear-elastischen Materialverhalten ist, kann man nun das E_{OED} abtragen. Es wird nur ein Delta betrachtet um E_{OED} zu berechnen. Man darf die Dehnung nicht über den gesamten Verlauf betrachten um E_{OED} zu berechnen.

Mit diesem ermittelten E-Modul kann man nun weitere Berechnungen für die Geotechnik durchführen.

Kapitel 20

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

20.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [1]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

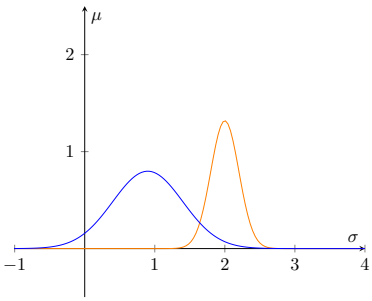


Abbildung 20.1: System

20.2 Kalman Filter

20.2.1 Geschichte

Das Kalman Filter wurde 1960 von Rudolf Emil Kalman entdeckt und direkt von der NASA für die Appollo Mission benutzt. Der Filter kommt mit wenig Rechenleistung aus und war somit dafür geeignet die Rakete bei der Navigation zu unterstützen. Das Filter schätzt den Zustand eines Systems anhand von Messungen und kann den nächsten Zustand errechnen. Typische Anwendungen des Kalman-Filters sind die Glättung von verrauschten Daten und die Schätzung von Parametern und kommt heutzutage in jedem Satellit, Navigationssystem, Smartphones und Videospielen vor.

20.2.2 Wahrscheinlichkeit

Das Kalman Filter versucht nichts anderes, als ein geeigneter Wert zwischen zwei Normalverteilungen zu schätzen. Die eine Kurve zeigt die errechnete Vorhersage des Zustands, bzw. deren Normal-Gauss-Verteilung. Die andere Kurve zeigt die verrauschte Messung des nächsten Zustand, bzw. deren Normal-Verteilung. Wie man in am Beispiel dieser zwei Gauss-Verteilungen sehen kann, ist sowohl der geschätzte Zustand als auch der gemessene Zustand nicht am selben Punkt.

Um eine genauere Schätzung des Zustandes zu machen, wird nun ein Wert zwischen den beiden Verteilungen gesucht. An diesem Punkt wird nun eine Eigenschaft ausgenutzt. Durch das Multiplizieren zweier Normalverteilungen entsteht eine neue Normalverteilung.

Wir haben eine Normalverteilung der Vorhersage:

$$y_1(x;\mu_1,\sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \quad e^{-\frac{(x-\mu_1)^2}{2{\sigma_1}^2}} \tag{20.1}$$

und für die Messung:

$$y_2(x;\mu_2,\sigma_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \quad e^{-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}}. \tag{20.2}$$

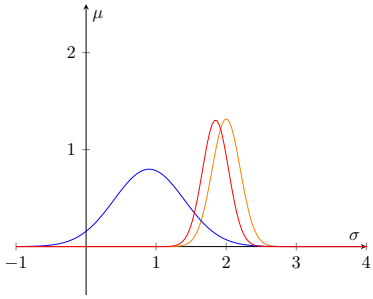


Abbildung 20.2: System

Diesen werden nun Multipliziert und durch deren Fläche geteilt um sie wieder zu Normieren:

$$y_f(x;\mu_f,\sigma_f) = \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2{\sigma_1}^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}}}{\int y_1 * y_2} \tag{20.3}$$

Dadurch gleicht sich die neue Kurve den anderen an. Interessant daran ist, dass die fusionierte Kurve sich der genauere Normal-Verteilung anpasst. ist σ_2 klein und σ_1 gross, so wird sich die fusionierte Kurve näher an $y_2(x;\mu_2,\sigma_2)$ begeben. Sie ist also Gewichtet und die best mögliche Schätzung.

Was in 2 Dimensionen erklärt wurde, funktioniert auch in mehreren Dimensionen. Dieses Prinzip mach sich der Kalman Filter zu nutze, und wird von uns für die Erdbeben Berechnung genutzt.

20.2.3 Anwendungsgrenzen

Nicht lineare Systeme

20.3 Aufbau

Um ein Erdbeben kenntlich zumachen werden in der Regel Seismographen mit vielen Sensoren verwendet. Ein Seismograph besteht im Grunde aus einer federgelagerten Masse. Wirkt eine Bodenerregung auf das Gerät ein, bleibt die gekoppelte Masse in der regel stehen und das Gehäuse schwingt mit.Relativbewegung des Bodens kann damit als Längenänderung im Zeitverlauf gemessen werden. In modernen Seismographen wird die Bodenbewegung in alle Richtungen gemessen, sowohl Horizontal als auch Vertikal. Wir konstruieren uns eine einfachere Version eines Seismographen, welcher rein mechanisch funktioniert. Zudem kann er nur in eine Dimension Messwerte aufnehmen. Würde das System ausgebaut werden, um alle Horizontalbewegungen aufzunehmen, würde der Verwendung des Kalman-Filters zu kompliziert werden. Für zwei Dimensionen (x,y) würde der Pythagoras für das System benötigt werden. Da sich der Pythagoras bekanntlich nicht linear verhält, kann kein lineares Kalman-Filter implementiert werden. Da das Kalman-Filter besonders effektiv und einfach für lineare Abläufe geeignet ist, würde eine Zweidimensionale Betrachtung den Rahmen dieser Arbeit sprengen. Für ein nicht-lineares System werden Extended Kalman-Filter benötigt, bei denen die System-Matrix (A) durch die Jacobi-Matrix des System ersetzt wird.

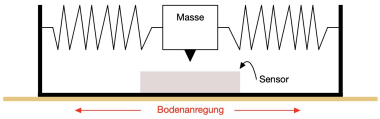


Abbildung 20.3: System

20.3.1 Optionen

Wollte man einen 2D Seismographen aufbauen, ohne den Pythagroas zu verwenden, kann dies mit der Annahme, das die Feder sehr lang sind erfolgen. Da sich bei langen Federn die Auslenkungen verkleiner...!!Noch nicht fertig!

20.4 Systemgleichung

Da das Kalman-Filter zum Schätzen des nächsten Zustand verwendet wird, wird eine Gleichung, welche das System beschreibt. Das Kalman-Filter benötigt eine Beschreibung der Systemdynamik. Im Fall unseres Seismographen, kann die Differentialgleichung zweiter Ordnung einer gedämpften Schwingung am harmonischen Oszillator verwendet werden. Diese lautet:

$$m\ddot{x} + 2k\dot{x} + Dx = f \tag{20.4}$$

mit den Konstanten m = Masse, k = Dämpfungskonstante und D = Federkonstante. Um diese nun in die Systemmatrix umzuwandeln, wird aus der Differentialgleichung zweiter Ordnung durch eine Substitution eine DGL erster Ordnung:

$$x_1 = x, \qquad x_2 = \dot{x}, \qquad x_3 = \ddot{x} \qquad | \quad \text{Substitution} \tag{20.5}$$

$$mx_3 + 2kx_2 + Dx_1 = f \qquad | \quad \text{DGL 1. Ordnung} \tag{20.6}$$

$$x_3 = -\frac{D}{m}x_1 - \frac{2k}{m}x_2 + \frac{f}{m} \qquad | \quad \text{nach} \quad x_3 \tag{20.7}$$

auch als Matrix-Vektor-Gleichung schreiben. Hierbei beschreibt die Matrix A die gesamte Systemdynamik in der Form, wie sie ein Kalman-Filter benötigt.

Um die lineare Differentialgleichung in das Kalman-Filter zu implementieren, muss dieses als Vektor-Gleichung umgewandelt werden. Dafür wird die Gleichung in die Zustände aufgeteilt. Die für uns relevanten Zustände sind die Position der Masse, die Geschwindigkeit der Masse und äussere Beschleunigung des ganzen System. Dabei muss unterschieden werden. um welche Beschleunigung es sich handelt. Das System beinhaltet sowohl eine Beschleunigung der Masse bzw. Feder (innere Beschleunigung), als auch eine Beschleunigung der ganzen Apparatur (äussere Beschleunigung). In unserem Fall wird die äusseren Beschleunigung gesucht, da diese der Erdbeben Anregung gleich kommt.

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{D}{m} & -\frac{2k}{m} & \frac{1}{m} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \tag{20.8}$$

Durch die Rücksubstituion ergibt sich:

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{D}{m} & -\frac{2k}{m} & \frac{1}{m} \end{pmatrix}\begin{pmatrix} x(t) \\ v(t) \\ f(t) \end{pmatrix}. \tag{20.9}$$

Da die Kraft unbekannt ist, wird die letzte Zeile später mit Nullen bestückt, denn genau diese Werte wollen wir.

20.5 Kalman Filter

Um den Kalman Filter zu starten, müssen gewisse Bedingungen definiert werden. In diesem Abschnitt werden die einzelnen Parameter/Matrizen erläutert und Erklärt, wofür sie nützlich sind.

20.5.1 Anfangsbedingungen

Anfangszustand x

Das Filter benötigt eine Anfangsbedingung. In unserem Fall ist es die Ruhelage, die Masse bewegt sich nicht. Zudem erföhrt die Apparatur keine äussere Kraft.

$$x_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{20.10}$$

Anfangsfehler / Kovarianzmatrix P

Da auch der Anfangszustand fehlerhaft sein kann, wird für den Filter einen Anfangsfehler eingeführt. Auf der Diagonalen werden die Varianzen eingesetzt, in den restlichen Felder stehen die Kovarianzen. In unserem Fall ist der Anfangszustand gut bekannt. Wir gehen davon aus, dass das System in Ruhe und in Abwesenheit eines Erdbeben startet, somit kann die Matrix mit Nullen bestückt werden. Somit ergibt sich für die Kovarianzmatrix

$$P_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{20.11}$$

Diese Matrix beschreibt die Unsicherheit des geschätzten Zustandes und wird sowohl für die Vorhersage als auch die Korrektur benötigt. Sie wird nach jeder Schätzung aktualisiert.. Für einen gut bekannten Zustandsvektor können kleine Werte eingesetzt werden, für ungenaue Anfangsbedingungen sollten grosse Werte (1 Million) verwendet werden. Grosse Werte ermöglichen dem Filter sich schnell einzupendeln.

Dynamikmatrix A

Die Dynamikmatrix bildet den Kern des Filters. Diese wurde weiter oben Bereits beschrieben. Dabei wollen wird die äussere Kraft des Systems ermitteln. Da nichts über die äussere Kraft bekannt ist, müssen wir annehmen das deren Ableitung 0 ist. Die System Vektor-Gleichung lautet daher:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{D}{m} & -\frac{2k}{m} & \frac{1}{m} \\ 0 & 0 & 0 \end{pmatrix} \tag{20.12}$$

Prozessrauschkovarianzmatrix Q

Die Prozessrauschmatrix teilt dem Filter mit, wie sich der Systemzustand verändert. Kalman-Filter berücksichtigen Unsicherheiten wie Messfehler und -rauschen. Bei unserem Modell könnte das beispielsweise ein Windstoss an die Masse sein. Für uns wäre dies:

$$Q = \begin{pmatrix} {\sigma_x}^2 & 0 & 0 \\ 0 & {\sigma_v}^2 & 0 \\ 0 & 0 & {\sigma_f}^2 \end{pmatrix} \tag{20.13}$$

Die Standabweichungen müssten Statistisch ermittelt werden, da der Fehler nicht vom Sensor kommt und somit nicht vom Hersteller gegeben ist. Das Bedeutet wiederum dass Q die Unsicherheit des Prozesses beschreibt, und die Messung.

Messmatrix H

Die Messmatrix gibt an, welcher Parameter gemessen werden soll. In unserem Fall ist es nur die Position der Massen.

$$H = (1, 0, 0)$$

Messrauschkovarianz R

Die Messrauschkovarianzmatrix beinhaltet, wie der Name es schon sagt, das Rauschen der Positionssensoren. In unserem Fall wird nur die Position der Masse gemessen. Da wir keine anderen Sensoren haben ist dies lediglich:

$$R = ({\sigma_x}^2). \tag{20.14}$$

Diese Messrauchen wird meistens vom Sensorhersteller angegeben. Für unsere Theoretische Apparatur wird hier ein kleiner Fehler eingesetzt.

20.5.2 Fiter Algorithmus

Nachdem alle Parameter aufgestellt sind, wird der Filter initialisiert und wird den Zustand der Feder vorherzusagen, die Messung zu präzisieren und laufend zu aktualisieren. Das Filter berechnet aufgrund der aktuellen Schätzung eine Vorhersage. Diese wird, sobald verfügbar, mit der Messung verglichen. Aus dieser Differenz und den Unsicherheiten des Prozesses (Q) und der Messung (R) wird der wahrscheinlichste, neue Zustand geschätzt.

Vorhersage

Im Filterschritt Vorhersage wird der nächste Zustand anhand des Anfangszustand und der Systemmatrix berechnet. Dies funktioniert ganz Trivial mit dem Rechenschritt:

$$x_{t+1} = A \cdot x_t. \tag{20.15}$$

Die Kovarianz P_{pred} wird ebenfalls neu berechnet, da die Unsicherheit im Vorhersage grösser wird als im Aktuellen. Da wir ein mehrdimensionales System haben, kommt noch die Messunsicherheit Q dazu, so dass die Unsicherheit des Anfangsfehlers P immer grösser wird. Dies funktioniert durch multiplizieren der Systemmatrix, deren Ableitung und mit dem aktualisierten Anfangsfehler. Dazu wird noch die Messunsicherheit addiert, somit entsteht die Gleichung

$$P_{pred} = APA^T + Q. \tag{20.16}$$

wird dieser Vorgang wiederholt, schaut der Filter wie genau die letzte Anpassung von P zur Messung stimmt. Ist der Unterschied klein, wird die Kovarianz P kleiner, ist der Unterschied gross, wird auch die Kovarianz grösser. Das Filter passt sich selber an und korrigiert sich bei grosser Abweichung.

Messen

Der Sensor wurde noch nicht benutz, doch genau der liefert Werte für den Filter. Die aktuellen Messwerte z werden die Innovation w mit dem Zustandsvektor x und der Messmatrix H zusammengerechnet. Hier bei wird lediglich die Messung mit dem Fehler behaftet, und die Messmatrix H

$$w = Z - (H \cdot x) \tag{20.17}$$

Die Innovation ist der Teil der Messung, die nicht durch die Systemdynamik erklärt werden kann. Innovation = Messung - Vorhersage. Dies ist Intuitiv logisch, eine Innovation von 0 bedeutet, dass die Messung nichts Neues hervorbrachte.

Im nächsten Schritt wir analysiert, mit welcher Kovarianz weiter gerechnet wird.

Korrigieren

Udpdate

20.6 Anfügen der Schwingung

Ein Erdbeben breitet sich im Boden wellenartig aus und bringt Objekte, wie zum Beispiel ein Gebäude, in Schwingung. Diese Schwingungen pflanzen sich im Gebäude mit gleicher Amplitude, Geschwindigkeit und Beschleunigung in horizontaler und vertikaler Bewegung fort. Wir möchten herauszufinden, wie gross die Massenbeschleunigung infolge eines Erdbeben ist. Mit Hilfe von fiktiven Sensoren, die eine Ortsveränderung des Gebäude messen, können wir mit Anwendung von Matrizen und dem Kalman-Filter die Beschleunigung berechnen.

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{20.18}$$

20.7 Erreger-Schwingung

Wir möchten mit einer gedämpften harmonischen Schwingung ein einfaches Erdbeben simulieren, die im Kalman Filter eingespeist wird. Die Gleichung lautet

$$x(t) = Ae^{t/2}sin(t). \tag{20.19}$$

Mit dieser Schwingung können wir ein einachsiger Seismograph simulieren, der eine Ortsverschiebung auf der x-Achse durchführt. Die Dämpfung der Schwingung ist relevant, da das System beim Schwingungsvorgang durch die Federkonstante und der Reibung, Energie verliert.

Die Ergebnisse dieser Schwingung setzen wir in die Messmatrix ein und können den Kalman-Filter starten.

20.8 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

20.8.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

20.9 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

20.9.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non pro-

vident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Literatur

[1] *BibTeX*. 6. Feb. 2020. URL: https://de.wikipedia.org/wiki/BibTeX.

Kapitel 21

Thema

Hans Muster

Ein paar Hinweise für die korrekte Formatierung des Textes

- Absätze werden gebildet, indem man eine Leerzeile einfügt. Die Verwendung von \\ ist nur in Tabellen und Arrays gestattet.
- Die explizite Platzierung von Bildern ist nicht erlaubt, entsprechende Optionen werden gelöscht. Verwenden Sie Labels und Verweise, um auf Bilder hinzuweisen.
- Beginnen Sie jeden Satz auf einer neuen Zeile. Damit ermöglichen Sie dem Versionsverwaltungssysteme, Änderungen in verschiedenen Sätzen von verschiedenen Autoren ohne Konflikt anzuwenden.
- Bilden Sie auch für Formeln kurze Zeilen, einerseits der besseren Übersicht wegen, aber auch um GIT die Arbeit zu erleichtern.

21.1 Teil 0

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua [**munkres:bibtex**]. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

21.2 Teil 1

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta

sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt

$$\int_a^b x^2\,dx = \left[\frac{1}{3}x^3\right]_a^b = \frac{b^3 - a^3}{3}. \tag{21.1}$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.2.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga (**??**).

Et harum quidem rerum facilis est et expedita distinctio **??**. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus **??**. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

21.3 Teil 2

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.3.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis

aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

21.4 Teil 3

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

21.4.1 De finibus bonorum et malorum

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat.

Index