Anwendungen der Linearen Algebra

Philip Herrmann

7. Oktober 2014

Inhaltsverzeichnis

Kapitel 1

Einführung

Dieses Buch entsteht aus den Materialien meines Begleitangebots zu den Vorlesungen Lineare Algebra und analytische Geometrie I& II an der Universität Hamburg. Dieses Begleitangebot war Teil eines Projekt zur Verbesserung der Studieneingangsphase und zielte daher darauf ab, verschiedene grundlegende Probleme der Studienanfänger in diese Phase abzufedern. Die Kernziele des damaligen Begleitangebotes, sowie auch dieses Buches, sind:

Motivation. Die Bereitschaft sich auf eine neue mathematische Sprache und die damit einher gehende Abstraktion und Präzision einzulassen, soll durch das Aufzeigen von spannenden Anwendungen erhöht werden. Zum Beispiel ist es möglich an die in der Schule eingeführte Vorstellung vom Abstand eines Punktes zu einer Geraden anzuknüpfen und basierend auf dem allgemeinen Vektorraumbegriff eine Vielzahl moderner Anwendungen zu erschließen.

Math awareness. Die im Rahmen dieser Veranstaltung vorgestellten Anwendungen stammen überwiegend aus unserer Umwelt, sind 'außermathematische' Anwendungen. Die daraus abzuleitende Relevanz der Methoden und Konzepte der linearen Algebra, aber auch der Mathematik im Allgemeinen, soll ein Fundament für die Abschätzung der Bedeutung der Mathematik in der Gesellschaft bilden. Sobald die hinter den präsentierten Anwendungen stehenden mathematischen Konzepte (Approximation, Codierung, Optimierung,...) deutlich werden, wird das Auge des Betrachters dahingehend geschärft, dass nun selbstständig viele weitere Anwendungen dieser Konzepte erkannt werden können.

Orientierung und Übersicht. Die Anwendungen in dieser Vortragsreihe haben zwar den Fokus stets auf den jeweils gerade aktuellen Themen der Vorlesung, dennoch ist an vielen Stellen etwas mehr Mathematik nötig, um die Anwendung im Detail zu verstehen. Zum Teil kann dieses 'mehr' durch aus der Schulmathematik bekannte Inhalte grob abgedeckt werden, häufig muss aber auch die Tür zu weiteren Gebieten (z.B. Operations Research, Codierungstheorie, Kryptographie, Graphentheorie, Dynamische Systeme,...) der Mathematik geöffnet und der dahinter liegende Raum skizziert werden. Ohne jeweils wirklich tief in unbekanntes Gebiet vorzustoßen, hilft dieser Ausblick hoffentlich dabei schon sehr früh im Studium eine ungefähre Vorstellung von der Größe und von Wesenszügen der Mathematik zu bekommen. Es ist beabsichtigt häufig eine chronologische Dimension mit einzublenden und das Wechselspiel zwischen der Entstehung von Anwendungen und der Entstehung von Mathematik zu beleuch-

ten. Dies scheint mir einerseits schon in sich selbst eine wichtige Aufgabe einer solchen Veranstaltung, andererseits hilft die gewonnene Übersicht den Studierenden hoffentlich auch bei der Ausgestaltung ihres Studienverlaufs.

Verfestigung und Verständnis Die verschiedenen Anwendungen machen es erforderlich, die Vorlesungsinhalte immer wieder aus einem veränderten Blickwinkel zu rekapitulieren. Zum Beispiel werden Datenansammlungen oder Prozesse in Matrizen verwandelt und mit Methoden behandelt, die für das Studium linearer Abbildungen erarbeitet wurden. Solche Blickwinkelveränderungen ermöglichen einen Verständniszugewinn, da es einem vielleicht erst beim zweiten oder dritten Blickwinkel gelingen mag einen Zugang und ein Gefühl für die Sache zu finden. Außerdem wird bereits behandelter Vorlesungsstoff in einer leicht veränderten Form 'wiederholt' und dadurch in seiner Essenz vermutlich verfestigt.

Der Aufbau des Buches orientiert sich an einer stereotyptischen zweisemestrigen Vorlesung Linearen Algebra und Analytische Geometrie. Die einzelnen Kapitel tragen Überschriften, wie sie auch unter den Abschnittsüberschriften einer solchen Vorlesung vorkommen dürften. Die einzelnen Unterkapitel sind dann allerdings mit ihren wahren Inhalten überschrieben.

Den Abschluss bildet ein Kaptiel über das Wesen der Mathematik. Um gleich zu sehr hochtrabenden Zielen auf Distanz zu gehen sei gesagt, dass hier lediglich eine unvollständige Übersicht über die mathematische Landkarte geboten werden soll, wie sie die Studierenden der Hamburger Vorlesung in einem Vortrag am Ende des zweiten Semesters von mir erhalten haben. Die ...

Bezug nehmen, auf Literatur, die insprierend für diese Buchidee war: ... Trend die lineare Algebra zusammen mit einigen Anwendungen zu präsentieren und mathematische Modellierung in der Lehramtsausbildung. Außerdem ... die tolle Darstellung von Rousseau et. al.

Evtl dafür entschuldigen, dass die Anwendungen an einigen Stellen kriegerisch erscheinen mögen. Durch Zitat dieses von V.I. Arnold: "All mathematics is divided into three parts..." rechtfertigen? Aber Zitat dann unbedingt relativieren.

Fahrplan für Inhalt dieses Buches als 2-SWS Begleitung zu einer Vorlesung über lineare Algebra erstellen.

Bielefeld, den 7. Oktober 2014
– Philip Herrmann

Kapitel 2

Grundlagen

Das Grundlagenkapitel einer Vorlesung über Lineare Algebra dient in der Regel als Einstiegskapitel in das Mathematikstudium. Einem kurzen Abriss von Aussagenlogik und mengentheoretischen Grundlagen, folgt eine Einführung in erste mathematische Strukturen, wie Gruppen, Ringe, Körper oder Vektorräume. Spezialfälle dieser Strukturen sind dem Studienanfänger aus der Schulzeit bereits vertraut und es ist höchst erfreulich, dass das Wiedersehen dieser Strukturen in einem abstrakteren Gewand gleich von tagtäglichen Anwendungen begleitet werden kann, die heute zweifellos auf dem bisherigen Höhepunkt ihrer weiter steigenden Bedeutung stehen: Kryptologie und Codierungstheorie. Tatsächlich verbergen sich hinter diesen zwei Namen jeweils ganze eigenständige mathematische Disziplinen, deren Ziele, Themen und Geschichte in grob zusammengefasster Form in den folgenden zwei Unterkapiteln eingeführt werden. Das Hauptaugenmerk soll allerdings darauf liegen, das Wirken zweier mathematischer Strukturen der ersten Semesterwochen in expliziten Anwendungen zu präsentieren. Im Kryptologie-Kapitel wird hierzu die modulare Arithmetik aufgegriffen, also das Rechnen in den endlichen Zahlbereichen $\mathbb{Z}/n$. Diese Zahlbereiche stellen für die Studierenden häufig die ersten nicht so geläufigen Repräsentanten der mathematischen Strukturen Gruppe, Ring und, anhängig von n, Körper dar.

Im Abschnitt über Codierungstheorie stehen Vektorräume im Mittelpunkt. Wir beschränken uns daher nach einer kleinen Einführung der generellen Ideen auf sogenannte lineare Codes und entdecken, wie in diesem Fall viele der grundlegenden Definitionen der ersten Kapitel einer beliebigen Vorlesung zur linearen Algebra ein codierungstheoretische Bedeutung erlangen. Auf dieser Grundlage ist es bereits möglich eine weitreichende Einführung in einige Codes zu erarbeiten. Natürlich soll auch hier schnell wieder die Anwendung im Vordergrund stehen.

2.1 Kryptologie

Kryptologie ist die Wissenschaft vom Verborgenen, der Geheimnisse, bestehend aus den zwei Disziplinen Kryptographie und Kryptoanalyse. Die Kryptographie beschäftigt sich mit dem Verbergen und Verschlüsseln von Kommunikationsinhalten, sogar manchmal mit dem Verbergen von Kommunikationsfluss überhaupt. Im Gegensatz dazu, geht es in der Kryptoanalyse um das Aufdecken und

Enttarnen von Nachrichten. In diesem Unterkapitel soll ein wenig von der spannenden Geschichte der Kryptologie erzählt werden, jedenfalls von den Teilen der Geschichte die heutzutage öffentlich bekannt sind. Denn Kryptologie ist bis heute und wohl mehr denn je ein aktives Forschungsgebiet, auf dem ein guter Teil der Forschung selbst im Verborgenen stattfindet. Eheleute, Firmen, Staaten,... - viele Mitspieler forschen auf dem Gebiet der Kryptologie und ein Erkenntnisvorteil gegenüber den anderen Mitspielern kann sich zu einem großen strategischen Vorteil entwickeln. Dafür hat die Geschichte rückblickend einige eindrucksvolle Beispiele hervor gebracht, auf die später noch eingegangen werden soll. Es ist wichtig von vorne herein zu betonen, dass die Kryptologie kein Teilbereich der linearen Algebra ist, nicht einmal per se eine mathematische Disziplin. Allerdings hat die Mathematik mit den Jahrhunderten einen immer größeren Einzug in die Kunst der Geheimschriften erhalten. Das erklärte Ziel dieses Abschnittes ist es, die Restklassenringe $\mathbb{Z}/n$ durch Anwendungen zu motivieren und dazu eignen sich die ausgewählten kryptographischen Themen sehr gut. Wir wollen diesen Abschnitt dann auch gleich damit beginnen, die Hauptdarsteller vorzustellen.

Definition 2.1.1. Sei $\mathbb{Z}/n$ die Menge der Restklassen bezüglich der auf den ganzen Zahlen definierten Äquivalenzrelation $\sim$, wobei $a, b \in \mathbb{Z}$ genau dann äquivalent im Sinne von $\sim$ sind, wenn die Differenz $a - b$ ein ganzzahliges Vielfaches von n ist.

Die Differenz zweier ganzer Zahlen ist genau dann ein Vielfaches von n, wenn die beide Zahlen beim ganzzahligen Teilen durch n den gleichen Rest r, mit $0 \leq r < n$, liefern. Auf diese Weise erkennt man, dass $\sim$ eine Partition von $\mathbb{Z}$ in genau n Teilmengen liefert. Mit anderen Worten, die Menge $\mathbb{Z}/n$ besteht aus n Elementen. Die folgenden beiden Lemmata besagen, dass die Äquivalenzrelation $\sim$ mit der üblichen Ringstruktur $(\mathbb{Z}, +, \cdot)$ auf eine gewisse Weise kompatibel ist, so dass sich Addition und Multiplikation auf die Restklassen übertragen lassen.

Lemma 2.1.2. *Seien $[a], [b] \in \mathbb{Z}/n$. Die Zuweisung*

$$[a] + [b] := [a + b]$$

ist wohldefiniert und definiert eine Additionsabbildung $+ : \mathbb{Z}/n \times \mathbb{Z}/n \to \mathbb{Z}/n$. Mit dieser Definition ist $(\mathbb{Z}/n, +, [0])$ eine abelsche Gruppe.

Beweis. Sofern ein Beweis dieser Aussage nicht in einer begleitenden Vorlesung niedergeschrieben wurde, ist dieser kurze Beweis eine gute Übungsaufgabe. □

Lemma 2.1.3. *Seien $[a], [b] \in \mathbb{Z}/n$. Die Zuweisung*

$$[a] \cdot [b] := [a \cdot b]$$

ist wohldefiniert und definiert eine Multiplikation $\cdot : \mathbb{Z}/n \times \mathbb{Z}/n \to \mathbb{Z}/n$. Mit dieser Definition ist $(\mathbb{Z}/n, \cdot, +, [1], [0])$ ein kommutativer Ring mit Eins.

Lemma 2.1.4. *Die Projektionsabbildung*

$$\mathbb{Z} \to \mathbb{Z}/n, n \mapsto [n],$$

ist ein Ringhomomorphismus.

Abbildung 2.1: Zahlenkreis zur Visualisierung von $\mathbb{Z}/n$

Beispiel 2.1.5. Rechnen in $\mathbb{Z}/n$
Ausnutzen der Eigenschaft Ringhomomophismus

An dieser Stelle soll zunächst nicht tiefer auf die allgemeine Theorie der modularen Arithmetik, des Rechnens in den Ringen $\mathbb{Z}/n$, eingegangen werden; das werden wir später erledigen, wenn die Anwendungen dies von uns verlangen. Stattdessen steigen wir direkt in die Kryptologie ein.

Wie in der Einleitung zu diesem Abschnitt bereits erklärt wurde, behandelt die Kryptologie die Untersuchung von Möglichkeiten der nicht-autorisierten Teilhabe an der Kommunikation zweier Entitäten. Um mathematisch tätig werden zu können, muss die Situation modelliert werden. Dazu verwenden wir das folgende naive Kommunikationsmodell aus Abbildung REF.[1]

Abbildung 2.2: Kommunikationsmodell - Kryptographie

In unserem Kommunikationsmodell gibt es einen Sender ('Alice'), der eine Nachricht über einen offenen Kanal an einen Empfänger ('Bob') schicken möchte. Die Namen von Sender und Empfänger entsprechen den geläufigen Bezeichnungen in der Fachliteratur.

Abbildung 2.3: Kommunikationsmodell - Kryptographie

Beispiele für reale Instanzen des in Abbildung 2.3 beschrieben abstrakten Kommunikationsmodells sind ...

Die Aufgabe der Kryptographie besteht nun darin, die Nachricht von Alice vor dem Versenden so zu verändern, dass Bob in der Lage ist die Nachricht zu lesen, andere Beobachter des Kommunikationskanals aber möglichst große Schwierigkeiten[2] haben, die Originalnachricht von Alice aus der beobachteten veränderten Nachricht zu rekonstruieren. Dieses Vorhaben erfordert es natürlich, dass Bob eine Möglichkeit gegeben ist, sich von den anderen Beobachtern zu unterscheiden. Der klassische Ansatz erlaubt es Alice und Bob deshalb ein Geheimnis abzusprechen, sozusagen im Voraus und unbeobachtet. Aus verschiedenen Gründen ist dieses Eingeständnis an Alice und Bob allerdings unglücklich und die moderne Kryptologie behandelt Verfahren, die auf solche Absprachen verzichten. So ist es zum Beispiel schlicht nicht realisierbar mit den Millionen von potentiellen Kommunikationspartnern im Internet jeweils im Voraus schon ein Geheimnis auszutauschen. Schon die 'alten Römer' kannten dieses Problem, wenngleich in einem kleineren Rahmen. Caesar hat deswegen zum Verbreiten seiner geheimen strategischen Anweisungen mit allen potentiellen Gesprächspartnern das gleiche Geheimnis abgesprochen und seine Kommunikationswelt damit in nur zwei größere Lager geteilt - Wissende und Unwissende. Zum Einstieg in die Kryptologie bietet die sogenannte Caesar-Verschlüsselung eine gute

[1] Verweis auf Shanons bahnbrechende Arbeit und sein Kommunikationsmodell.

[2] Das hier nicht von einer Unmöglichkeit gesprochen wird, hat gute Gründe, wie wir in (den Überlegungen zu perfekten Chiffren) sehen werden.

Gelegenheit, da sich verschiedene Prinzipien und Mechanismen gut anhand ihrer sehr schlichten Funktionsweise erläutern lassen. Wir müssen den Römern allerdings zugute halten, dass Schreiben und Lesen zu jener Zeit weit weniger verbreitet (und dazu noch erschwerlich) waren und für viele Feinde die römische Schrift wohl schon eine ausreichende Kryptographie darstellt hätte.

Ganz ohne mathematische Formalismen erklärt, besteht die Caesar-Verschlüsselung darin, jeden Buchstaben der Klartextnachricht durch den Buchstaben zu ersetzen, der im Alphabet drei Stellen später zu finden ist. Wenn man bei diesem Ersetzungsprozess einen der letzten drei Buchstaben des Alphabets zu ersetzen hat, dann zähle man hinter dem letzten Buchstaben des Alphabets wieder mit dem ersten Buchstaben weiter.

Veranschaulichen könnte man den Ersetzungsprozess also durch einen Buchstabenkreis:

Abbildung 2.4: Kommunikationsmodell - Kryptographie

Grafik zeigt Buchstabenkreis mit Addition 3 und verweist auf Ähnlichkeit zu ref(Zahlenrad mod n). Bemerkung, dass die Ähnlichkeit kein Zufall ist und in der folgenden Formalisierung aufgegriffen wird.

Wir wollen diesen Prozess nun etwas formaler fassen. Man mag einwenden, dass eine Formalisierung bei diesem einfachen Prozess gar nicht hilfreich ist, da sich alle Einzelheiten auch präzise mit natürlicher Sprache beschreiben und verstehen lassen und die alten Römer auch ohne die im folgenden verwendete Mathematik auskamen. Aber! Unsere Gründe für eine mathematisch-formalere Beschreibung sind vielfältig und überwiegen:

- Als Kryptograph erlaubt die Formalisierung uns direkt eine große Menge von Verallgemeinerungen der ursprünglichen Caesar-Chiffre,
- als Kryptoanalytiker erlaubt die Formalisierung uns Einsichten in die Struktur der Caesar-Chiffre und ihrer Verallgemeinerung, welche Voraussetzung für die Analyse ist und es uns ermöglicht den Aufwand einer Analyse abzuschätzen,
- als Studierende der Kryptologie dient uns diese Formalisierung als Aufwärmprogramm für den Umgang mit moderen Chiffren, die durchgängig mathematischer Natur sind und sich nur auf diesem Wege erschließen lassen und letztlich,
- als Studierende der Linearen Algebra, liefert diese Formalisierung eine erste Anwendung für das Arbeiten mit $\mathbb{Z}/n$ und ebnet gleichzeitig den Weg für eine spätere, moderne kryptographische Anwendung von modularer Arithmetik.

Es sei nebenbei bemerkt, dass die Gegensätzlichkeit der ersten beiden Stichpunkte der obigen Aufzählung unsere erste Begegnung mit dem ständigen Wettlauf zwischen Kryptographie und Kryptoanalyse ist. Dieser Aspekt hat die Disziplin seit jeher und bis heute geprägt und ist viel zu spannend um ihn neben der mathematischen Entdeckungsreise unbeobachtet zu lassen. Wir werden in

?? ein wenig ausführlicher darauf eingehen und dann auch auf einige der vielen vorzüglichen Bücher zu diesem Thema verweisen.

Nun aber wirklich zurück zur Caesar-Chiffre und ihrer mathematischen Beschreibung! Der Buchstabenkreis aus Abbildung 2.4 legt uns die Analogie mit dem Zahlenkreis aus Abbildung 2.1, den wir zur Visualisierung der Addition in $\mathbb{Z}/n$ eingeführt haben, direkt nahe. Um die beiden Kreise gedanklich übereinander zu legen betrachten wir $\mathbb{Z}/26$ und ordnen die 26 Buchstaben des Alphabets

$$\Omega = \{A, B, C, \ldots, X, Y, Z\}$$

(also ohne Beachtung von Umlauten oder Groß- und Kleinschreibung) in ihrer alphabetischen Reihenfolge "der Reihe nach" den Elementen in $\mathbb{Z}/26$ zu. Wir können hierbei nicht im mathematischen Sinne von einer Anordnung von $\mathbb{Z}/26$ sprechen, was in Übungsaufgabe **??** weiter thematisiert wird. Das soll uns aber nicht aufhalten und wir definieren einfach explizit eine bijektive Zuordnung

$$\begin{aligned} \varphi : \Omega \to & \mathbb{Z}/26, \text{ durch} \\ A \mapsto & [0] \\ B \mapsto & [1] \\ & \vdots \\ Z \mapsto & [25]. \end{aligned}$$

Dann können wir die Caesar-Verschlüsselung ebenfalls durch eine mathematische Abbildung $e : \Omega \to \Omega$ beschreiben, gegeben durch

$$e(\omega) = \varphi^{-1}(\varphi(\omega) + [3]).$$

Die Abbildung e wandelt einen Buchstaben ω also in eine Zahl, genauer in ein Element in $\mathbb{Z}/26$, um, addiert dann $[3]$ und gibt den Buchstaben zurück, den φ der Summe zuordnet. Wenn wir uns die Identifizierung von Ω und $\mathbb{Z}/26$ durch φ fest einprägen und auf eine explizite Unterscheidung im Folgenden zu verzichten bereit sind, dann lässt sich e einfach beschreiben als

$$e : \mathbb{Z}/26 \to \mathbb{Z}/26, \omega \mapsto \omega + [3].$$

Außerdem wollen wir beschließen, dass e durch buchstabenweise Anwendung auf eine Abbildung von Wörtern ausgedehnt wird. Caesars, natürlich verschlüsselt abgelegte, Sicherheitskopie von De Bello Gallico hätte also anstatt von

Gallia est omnes divisa in partes tres

mit den Worten

Jdoold hvw rpqhv glylvd lq sduwhv wuhv

begonnen. Zum Entschlüsseln des Geheimtextes benötigt es eine Abbildung $d : \mathbb{Z}/26 \to \mathbb{Z}/26$, mit der Eigenschaft

$$d \circ e = \mathrm{id}_{\mathbb{Z}/26} \tag{2.1.1}$$

Da es hier um Selbstabbildung endlicher Mengen handelt, ist d zwangsläufig das Inverse[3] zu e. Jedenfalls erfüllt die Abbildung $d : \omega \mapsto \omega - [3]$ genau diese Forderung und ist offensichtlich Caesars Entschüsselungsfunktion.

[3]Wie man sich leicht auch für allgemeine Bijektionen e überlegt, wenn man bereits etwas Kontakt mit den Begriffen Links- und Rechtsinverses hatte.

Natürlich ist die Verschiebung um exakt 3 in der Caesar-Verschlüsselung nicht entscheidend. Aus der Bedingung (2.1.1) folgt nur, dass e eine Bijektion sein muss. Die 24 anderen sinnvollen Additionen liefern in der gleichen Weise Verschlüsselungen und tatsächlich hat die Addition von [13] in bestimmten Kreisen des Internet eine nicht ganz ernst gemeinte Wiedergeburt unter dem Namen *rot*-13 gefeiert. Auch komplizierte Polynome liefern Veschlüsselungsfunktionen. Zum Beispiel ist für ein multiplikativ invertierbares Element $a \in \mathbb{Z}/26^*$ und ein beliebiges $b \in \mathbb{Z}/26$ die Abbildung

$$e_{a,b} : \mathbb{Z}/26 \to \mathbb{Z}/26, \omega \mapsto a \cdot \omega + b$$

ebenfalls eine Bijektion, eine sogenannte affine Chiffre. Insgesamt wäre die Menge $\mathrm{Bij}(\mathbb{Z}/26)$ der Bijektionen von $\mathbb{Z}/26$ groß genug für viele weitere Herrscher oder Feldherren, um sich ein für alle mal ihre eigene feste Verschlüsselungsbijektion auszusuchen. Dieses Vorgehen wäre natürlich unsinnig. Würde zum Beispiel eine gewählte Bijektion einmal in feindliche Hände fallen, dann wäre automatisch die gesamte Kommunikation, rückwirkend und fortwährend, offengelegt. Um mit Anderen verschlüsselt zu kommunizieren müsste die Bijektion den beabsichtigten Empfängern aber bekannt gemacht werden. Das Risiko wäre hoch, dass so mit der Zeit jemand Falsches die richtige Bijektion erfährt. Moderne kryptographische Verfahren ermöglichen es, diese Gefahren zu eliminieren. Um einen berühmten Angriff auf derartig aufgebaute Chiffren zu studieren, holen wir kurz etwas aus und diskutieren ein wichtiges Paradigma der jüngeren Kryptologie - das Kerckhoffs'sche Prinzip.

Definition 2.1.6 (Kerckhoffs Prinzip)**.** Eigentlich drittes Kerckhoffs'sches Prinzip? Formulieren

Ein bis zwei Sätze zur Motivation und Geschichte des Prinzips. Der technische Hintergrund zur Rechtfertigung von Kerckhoffs' Prinzip ist, dass sich aus Geräten oder Software, die heutzutage oft massenhaft zur Verschlüsselung eingesetzt werden, die Verschlüsselungsmechanismen durch *reverse engeneering* sehr genau bestimmen lassen könnten.

Nach dem Kerckhoffs'schen Prinzip können wir bei einer Kryptoanalyse von Caesars Chiffre zumindest annehmen, dass es sich sicherlich um eine *monoalphabetische Substitutions-Chiffre* auf einem kleinen Alphabet handelt. Für den Einstieg können wir sogar mal annehmen, wir wüssten, dass Caesar seine Nachrichten mit einer Translation verschlüsselt, also durch eine Abbildung $e_b : \mathbb{Z}/26 \to \mathbb{Z}/26$, die durch die Addition einer uns unbekannten Zahl b gegeben ist:

$$e_b(\omega) = \omega + b.$$

Dieses b möchten wir gerne bestimmen und nehmen dazu weiter an, dass wir eine verschlüsselte Nachricht Caesars abgefangen haben. Kryptoanalysen die unter dieser Prämisse stattfinden, nennt man *known chiphertext* Angriffe[4]. Monoalphabetische Substitutions-Chiffren wie e_b haben die Eigenschaft, dass jeder Buchstabe ω im Klartext die gleiche Häufigkeit hat, wie sein Chiffrat $e_b(\omega)$ im Geheimtext. Diese Beobachtung liefert den Ansatzpunkt für die Häufigkeitsanalyse (Entropieanalyse). In einem durchschnittlichen deutschen Text haben die

[4]In Abgrenzung zu *chosen plaintext* oder *known plaintext* Angriffen. In der Praxis kann es natürlich nicht-trivial sein, eine ausreichend große Menge an Geheimtext zu bekommen.

Buchstaben unterschiedliche relative Häufigkeiten und diese charakteristischen Häufigkeiten zeichnen sich schon bei überraschend kurzen Texten ab. So ist trotz der Fremdwörter auf dieser Seite der Buchstabe e der Häufigste - n von m Buchstaben auf dieser Seite sind ein e. Das macht eine relative Häufigkeit von n/m. Berechnen wir die relativen Häufigkeiten für jeden Buchstabe auf dieser Seite, so erhalten wir das folgende Diagramm.

Abbildung 2.5: Häufigkeitsverteilung auf Seite 13

Grafik zeigt Häufigkeitsverteilung der Buchstaben auf dieser Seite

Zum Vergleich betrachten wir die relativen Häufigkeiten mit denen aus einem relativ beliebigen längeren deutschen Text, zum Beispiel der **??**

Abbildung 2.6: Häufigkeitsverteilung der Referenz

Grafik zeigt Häufigkeitsverteilung der Buchstaben in einem Referenztext

dann sticht schnell ins Auge, dass die beiden Häufigkeitsdiagramme ziemlich gut übereinander passen. Wenn wir den folgenden Geheimtext abgefangen haben und für ihn eine Häufigkeitsanalyse durchführen, dann ergibt sich das Diagramm aus Abbildung 2.8.

Abbildung 2.7: Pergament mit einem Translationsgeheimtext

Allein an der Sonderstellung des Buchstaben ?X? in der Häufigkeitsverteilung des Geheimtextes erahnt man schon, dass ?X? gut das Chiffrat von e sein könnte. Verschiebt man also entsprechend die Abbildung 2.6 um 5 Stellen und legt sie dann über Abbildung 2.8, so erhält man eine überzeugende Übereinstimmung:

Bei dem obigen Angriff brauchte nur ein Buchstabe korrekt zugeordnet werden, um die ganze Verschlüsselung zu knacken, da eine Translation schon durch einen einzigen Funktionswert vollständig bestimmt ist. Um die etwas allgemeineren, ebenfalls oben erwähnten affinen Chiffren zu knacken, müssten wir also schon 2 Buchstaben des Geheimtextes korrekt mit ihren Klartextbuchstaben identifizieren. Entsprechend aufwendiger wird das Knacken von Permutationen höheren Grades. Um Angriffe über Häufigkeitsanalyse zu vermeiden, benutzt man verschiedene Techniken. Zum einen definiert man Substitutionen nicht auf Buchstabenebene, sondern auf Blöcken von Buchstaben, zum Beispiel auf $(\mathbb{Z}/26)^{32}$. Dadurch haben alle Blöcke von 'normalen'[5] Texten eine

[5] Bei computergenerierten oder militärischen Texten können jedoch sehr lange gleichartige Blöcke auftauchen, zum Beispiel in den Metadaten von Netzwerkkommunikation.

Abbildung 2.8: Häufigkeitsverteilung im Geheimtext

Abbildung 2.9: Vergleich der Häufigkeiten

relative Häufigkeit die sehr nahe an Null liegt. Zum anderen verwendet man verschiedene Techniken, die verhindern sollen, dass gleiche Klartextblöcke an verschiedenen Textstellen auf gleiche Geheimtextblöcke abgebildet werden. So gibt es zum Beispiel Verfahren, die einen Teil des Kryptogramms des Vorblocks wieder in den neuen Klartextblock einfließen lassen. Detaillierte Informationen zu letztgenannten Techniken bekommt man in den entsprechenden Vorlesungen zur Kryptologie oder durch eine Internetrecherche mit den Schlagworten chipher block chaining (CBC), electronic codebook mode (ECM), cipher feedback
REF (CFB), counter mode (CTR) und anderen.

Wir wollen den Sprung zu modernen Chiffren nicht vollziehen, ohne dabei auch ein bisschen auf die spannende Geschichte der Kryptologie einzugehen. Die Caesar-Chiffre war natürlich nicht der Anfang der Geheimschriften, wir haben sie nur deswegen als Einstieg gewählt - wie so viele andere einführende Werke auch - weil sie mathematisch elementar ist und zu unserem Fokus auf modulare Arithmetik passt. Die große Epoche der manuellen Kryptographie hat viele weitere Verschlüsselungsverfahren hervorgebracht. Jedes davon kommt mit einer spannenden Geschichte, von denen die das Verfahren erfanden, denen die es brachen und ihren jeweiligen Beweggründen. Einige gut erzählte Darstellungen dieser Geschichten findet man zum Beispiel in [**?**]. Angetrieben durch den ersten Weltkrieg und den industriellen Fortschritt fand im frühen 20. Jahrhundert eine Mechanisierung der Kryptographie statt. Die Verfahren zur Verschlüsselung wurden in dieser vergleichsweise kurzen Epoche nicht mehr per Hand durchgeführt, sondern in zunehmendem Maße an Maschinen übertragen. Berühmtestes Beispiel für eine solche Maschine ist sicherlich die schreibmaschinenähnliche Enigma. Ihr und der Geschichte ihrer Kryptoanalyse haben sich zahlreiche Bücher [**?**] und Filme [**?**] gewidmet. Etwas in den 1970er Jahren beginnt die dritte Epoche der Kryptologie, die Computerbasierte. Mit der Ausbreitung von computergestützter Kommunikation begann der Bedarf danach, eben diese auch geheim halten zu können. Eine unglaubliche Menge von Algorithmen wurde dazu erdacht, viele davon haben sich aber schon nach sehr kurzer Zeit als unhaltbar erwiesen und sind mehr oder weniger cleveren Angriffen zum Opfer gefallen. Andere Algorithmen galten als sicher genug, solange die Rechenkraft von Computern sich auf einem geringeren Niveau befand und sollten heute nur deswegen nicht mehr eingesetzt werden, weil alle möglichen Schlüssel in vertretbarer Zeit durchprobiert werden können. Dann gibt es viele Verschlüsselungsalgorithmen, die zwar gegen heutige Rechnermodelle als ausreichend sicher gelten, für deren Sicherheit aber kein Beweis besteht. Man muss bei solchen Aussagen stets im Hinterkopf haben, dass kryptologische Forschung zu einem großen Teil selbst im Verborgenen abläuft. So kommt es auch, dass der RSA-Algorithmus, den wir im Folgenden genauer besprechen wollen, den Namen seiner vermeintlichen
? Erfinder, Rivest, Shamir und Adleman trägt. Denn in den 1990er hat der britische Geheimdienst GCHQ die Geheimhaltung für ein Dokument aufgehoben, aus dem hervor geht, dass ein Mitarbeiter des GCHQ diesen Algorithmus im Wesentlichen schon einige Jahre früher erdacht habe. Und so sollte es wenig

verwundern, wenn die Literatur der kommenden Jahrzehnte über unsere heutige Kryptologie spannendere Geschichte zu erzählen weiss, als die Anekdoten aktueller Fachliteratur hergeben können.

Symmetrische und Asymmetrische Kryptosysteme

Bevor wir uns nun aber dem gerade erwähnten RSA-Kryptosystem zuwenden, müssen wir Rolle von Schlüsseln noch näher beleuchten, die wir bisher weitestgehend ausgeblendet haben. In (2.1.1) und den umliegenden Betrachtungen haben wir Kryptosysteme als Paare (e, d) von Abbildungen betrachtet, wobei die Verschlüsselungsabbildung $e : M \to C$ (encryption) von einem Klartext-Alphabet M in ein Geheimtextalphabet C abbildet. Die Entschlüsselungsabbildung $d : C \to M$ (decryption) invertiert die Abbildung e einseitig, was ja gerade der Inhalt der Forderung (2.1.1) war: Wendet man zuerst die Verschlüsselung e an und dann die Entschlüsselung d, dann soll der ursprüngliche Klartextbuchstabe wieder zum Vorschein kommen. Bei unserer Untersuchung der Caesar-Chiffre haben wir dann schon die Verschiebung im Tiefindex festgehalten und e_3 für Caesars Originalverschiebung geschrieben, $e_1 3$ als rot-13 kurz erwähnt und dann allgemeine Translationschiffren e_b und sogar die affinen Chiffren $e_{a,b}$ betrachtet. Diesem Vorgehen liegt die Einsicht zugrunde, dass die Verschlüsselungsvorschrift in den Fällen jeweils die Gleiche ist, nur ein Parameter als veränderliche Größe die Resultate beeinflusst. Dieser Parameter wird Schlüssel genannt und wir werden seine Veränderlichkeit von nun an expliziter berücksichtigen, indem wir unter einem Kryptosystem von nun an ein Paar (e, d) von Abbildungen verstehen, mit

$$\begin{aligned} e &: M \times K \to C \\ d &: C \times K \to M, \end{aligned}$$

wobei K die Menge der möglichen Schlüssel (keys) bezeichne, so dass für alle Schlüssel $k \in K$, ein Schlüssel $k' \in K$ existiert, für den

$$d(e(m, k), k') = m \tag{2.1.2}$$

gilt. Wir nennen ein Kryptosystem *symmetrisch*, wenn stets der gleiche Schlüssel zum Verschlüsseln und zum Entschlüsseln genutzt werden kann, wenn also für jedes $k \in K$ die Eigenschaft $k' = k$ gilt. Ein Kryptosystem heißt demnach *asymmetrisch*, wenn es Schlüssel $k \in K$ gibt, so dass $d(e(m, k), k) \neq m$ für mindestens ein $m \in M$ gilt.

Aufgabe?

Bemerkung 2.1.7. Jedes symmetrische Kryptosystem (e, d) lässt sich mittels einer Abbildung $f : K \to K$ zu einem asymmetrischen Kryptosystem umgestalten, indem man die Verschlüsselungsabbildung durch die Abbildung

$$e(-, f(-)) : M \times K \to C$$

ersetzt. Dadurch erzeugt man einen funktionalen Zusammenhang zwischen k und k', denn es gilt dann $k' = f(k)$. Diskutieren Sie, warum asymmetrische Kryptosysteme von diesem Typ nicht für Public Key Kryptographie (siehe **??**) eignen.

In der Praxis möchte man für ein Kryptosystem möglichst die zusätzliche Einschaft haben, dass für alle oder zumindest fast alle anderen Schlüssel, also $l \in K$ mit $k' \neq l$, nicht wieder der Klartext dargestellt wird, also

$$d(e(m,k),l) \neq m \tag{2.1.3}$$

gilt. Wir nehmen diese Forderung jedoch nicht mit in die Definition eines Kryptosystems auf. Auch ein schlechtes Kryptosystem soll zu dessen Studium ein Kryptosystem sein und ohnehin wäre (2.1.3) nicht die entscheidende Forderung, wie die Translationschiffren verdeutlichen. Stattdessen untersucht man Kryptosysteme auf eine Fülle von verschiedenen Sicherheitsanforderungen. Manche dieser Anforderungen sind so stark, dass sie zwar kaum erfüllt werden können, aber durch ihr Studium Hinweise entstehen, worauf beim Design von Chiffren oder beim Umgang mit ihnen geachtet werden muss. Um ein Gefühl dafür zu vermitteln, wovon hier die Rede ist, seien im Folgenden einige dieser Forderungen kurz skizziert:

- perfect secrecy/security
- ein 'genügend großer' Schlüsselraum K:
- (stongly) ideal secrecy/security
- IND CPA / IND CCA Ununterscheidbarkeit bei chosen plaintext attacks oder chosen ciphertext attacks.

Die obige Liste ist weit davon entfernt vollständig zu sein. In der kryptologischen Literatur existieren viele weitere Angriffsszenarien, die gleichzeitig auch eben die Sicherheitsanforderung definieren in jenem Szenario nicht anfällig für Angriffe zu sein. Und jede neue denkbare (und ausreichend bekannt gewordene[6]) Angriffstechnik definiert auf diese Weise wieder eine Sicherheitsanforderung.

...Wie zum Beispiel eine Kryptoanalyse-Technik - die lineare Kryptoanalyse(!) - gegen DES schon zu dessen Geburt bekannt gewesen zu sein scheint und daher nicht alle Angriffstaktiken die heute üblich sind, öffentlich bekannt sein dürften.

Einige Anmerkungen zur Public Key Kryptographie wären angebracht. Tatsächlich sollte ich PKK hier einführend besprechen und eine Referenz zur obigen Bemerkung einbauen. Es sollte abgegrenzt werden, wie asymmetrisch Kryptosysteme mindestens sein sollten, damit sie sich zur PKK eignen.

Ein perfektes Kryptosystem

Dieser Abschnitt definiert Perfektheit oder greift die Definition aus der Auflistung von Sicherheitsanforderungen oben auf, falls diese dort gegeben wurde. Dann kommt eine kurze Darstellung des One-Time pads und 3-4 Worte dazu, warum sich auf der Kenntnis dieses Kryptosystems nicht ausruhen lässt.

[6] Andeutungen in der Fachliteratur lassen durchblicken, dass viele der auf diesem Gebiet forschenden Akteure sogenannte non-disclosure agreements unterzeichnet haben, also Verträge, die ihn zumindest Einschränkungen bei der Veröffentlichung von Informationen machen. Eine solche Anmerkung findet man zum Beispiel in dem auch sehr lesenswerten Aufsatz 'The uneasy relationship between Mathematics and Cryptography' von Neal Koblitz [Kob07].

2.1.1 Das RSA-Kryptosystem

Erzeugung eines Schlüsselpaares:

1. Nehme zwei (große) Primzahlen p, q und setze $n = p \cdot q$.
2. Wähle ein $e \in \{2, \ldots, n-1\}$ mit $ggT(e, n) = 1$.
3. Bestimme das multiplikative Inverse d von e in $\mathbb{Z}/\varphi(n)$.

Dann ist der öffentliche Schlüssel (e, n) und der private Schlüssel (d, n). Die RSA-Verschlüsselungsfunktion ist dann einfach gegeben durch

$$E_{(e,n)} : \mathbb{Z}/n \to \mathbb{Z}/n, g \mapsto g^e.$$

Zur Entschlüsselung müssen wir das Rechnen in $\mathbb{Z}/n$ noch etwas genauer unter die Lupe nehmen. Dazu wollen wir die folgende allgemeine gruppentheoretische Aussage auf die obige Enschlüsselungssituation spezialisieren.

Lemma 2.1.8. *Sei G eine endliche Gruppe der Ordnung n. Dann gilt für jedes Element $g \in G$, dass $g^n = e$ ist.*

Beweis. □

Eulersche φ-Funktion einführen (Benutze ich ja oben schon!). Insb $\varphi(p) = p - 1$.

Korollar 2.1.9.

1. *('Kleiner Fermat') Sei $p \in \mathbb{N}$ eine Primzahl. Für jedes Element $a \in \mathbb{Z}/p$ gilt dann*
$$a^{p-1} = 1.$$
2. *(Satz von Euler-Fermat) Sei $n \in \mathbb{N}$ beliebig und $a \in \mathbb{Z}$ teilerfremd zu n. Dann gilt*
$$a^{\varphi(n)} = 1 \text{ in } \mathbb{Z}/n.$$

Beweis. Wegen der Vorbemerkung $\varphi(p) = p-1$ ist die erste Aussage offensichtlich ein Spezialfall der zweiten Aussage, welche direkt aus dem Lemma 2.1.8 folgt, wenn wir nachweisen, dass die Einheitengruppe $\mathbb{Z}/n^\times$ gerade die Ordnung $\varphi(n)$ hat. Tatsächlich repräsentiert $a \in \mathbb{Z}$, mit $0 \leq a \leq n-1$ genau dann ein invertierbares Element in $\mathbb{Z}/n$, wenn es teilerfremd zu n ist. HIER MÖCHTE ICH EIGENTLICH NICHT GERNE DEN EUKLIDISCHEN ALGORTIHMUS BRAUCHEN UND DARAUS FOLGERN, DASS TEILERFREMD GENAU DANN GILT, WENN 1 $\mathbb{Z}$-LINEAR KOMBINIERBAR IST. □

Zur Entschlüsselung der zum öffentlichen Schlüssel (e, n) gehörigen Verschlüsselungsabbildung $E_{(e,n)}$ definieren wir mit Hilfe des privaten Schlüssels (d, n) die Abbildung

$$D_{(d,n)} : \mathbb{Z}/n \to \mathbb{Z}/n, g \mapsto g^d.$$

Mit dem Satz von Euler-Fermat haben wir das nötige Werkzeug um die Entschlüsselung und damit zu sogenannte Korrektheit des RSA-Systems nachzuweisen:

Korrektheit von RSA

Lemma 2.1.10. *Seien $(e,n),(d,n)$ ein Schlüsselpaar wie in* **??**. *Dann gilt*

$$D_{(d,n)}(E_{(e,n)}(g)) = g, \text{ für alle } g \in \mathbb{Z}/n.$$

Beweis. In $\mathbb{Z}/n$ ist die Entschlüsselung des Schlüsseltextes gegeben durch

$$\begin{aligned} D_{(d,n)}(E_{(e,n)}(g)) &= D_{(d,n)}(g^e) \\ &= g^{ed} = g^{k\varphi(n)+1} \\ &= g^{k\varphi(n)} \cdot g = g \end{aligned}$$

wobei für die letzte Gleichheit der Satz von Euler-Fermat aus Korollar 2.1.9 benutzt wurde. □

Sicherheit von RSA

In diesem Abschnitt wird nachgewiesen, dass RSA höchstens so sicher ist, wie Faktorisieren. Ich möchte diskutieren, dass RSA in dieser Form nicht semantisch sicher ist, z.B. nicht IND-CPA und wie man das Problem mit padding behebt (Hinweis auf Abschnitt: Linearer Zufall). Außerdem möchte einen kleinen Angriff auf falsche Anwendungen von RSA vorzeigen, z.B. Chinesischer Restsatz, falls ich wirklich die Coppersmith Attacke durchspielen möchte.

2.1.2 RSA im Kartenzahlungsverkehr

Beschreibung von eines Bezahlvorgangs mit der ICC-Bezahlkarte und wie dabei RSA zum Einsatz kommt. Informationen dazu finden sich in [emv11, Chapter 7]. Außerdem möchte ich mit der Abbildung aus 'Chip and Pin is broken' beschreiben, wie eine fehlende Signierung ausgenutzt werden konnte, wie also die Verschlüsselung des Kommunikationskanals nicht paranoid ist.

2.1.3 RSA für sichere Internetverbindungen

Das RSA-Kryptosystem dient uns auch in einer anderen alltäglichen Situation als bei den oben beschriebenen Chipkarten-Zahlungen. Immer dann, wenn wir im Internet eine Webseite mittels *https* aufrufen, wenn wir also den jeweiligen Server zum Aufbau einer verschlüsselten Verbindung auffordern, dann benutzen wir RSA[7]. Zum Aufbau solcher verschlüsselten Verbindungen wird das TLS (transport layer security) Protokoll benutzt, welches die Art und Weise definiert, in der sich der Computer des Benutzers mit dem Webserver bekannt macht, die Details über die zu benutzende Verschlüsselung austauscht und schließlich, zum Beispiel für den Web-Browser, einen verschlüsselten Kommunikationskanal bereitstellt. Im Folgenden soll der gerade erwähnte Prozess des Aushandelns, der sogenannte TLS Handshake, etwas genauer beschrieben werden, natürlich mit einem Fokus auf die kryptologischen Vorgänge. Dieses Wissen ist heutzutage von gesellschaftlicher Relevanz, geht es dabei doch nicht nur darum die eigenen Bankgeschäfte beim Onlinebanking abzusichern, sondern sogar, wie die medialen Geschehnisse um den ?goldenen Frühling? und die jüngsten Geheimdiens-

REF

[7] Theoretisch lässt das Protokoll auch andere Public-Key Chiffren und Signaturverfahren zu, aber Stand 2013 spielen die zulässigen Alternativen praktisch keine Rolle.

tenthüllungen gezeigt haben, um die (Un-) Versehrtheit zentraler gesellschaftlicher Werte wie Freitheit und REF. Die technisch detaillierte Protokollbeschreibung findet sich in RFC2246 [DA99], die folgende Übersicht ist ihr entnommen.

Die Ziele des TLS-Protokolls für den Aufbau verschlüsselter TCP/IP Verbindungen sind

1. Authentifikation der Kommunikationspartner,
2. Aufbau eines verschlüsselten Kommunikationskanals und
3. Integrität der kommunizierten Daten.

Zunächst soll also gewährleistet werden, dass die Kommunikation tatsächlich zwischen den zwei beabsichtigten Computern stattfindet und beispielsweise die Zugangsdaten für das Onlinebanking nicht einem böswilligen Vermittler mitgeteilt werden, der gegenüber der Bank unsere Identität annimmt und uns gegenüber die Rolle der Bank vortäuscht - einem sogenannten Man-in-the-Middle. Würden wir die Banking-Session irrtümlicher Weise mit dem Vermittler initialisieren, dann würden wir ihm unsere Daten mit seinem öffentlichen Schlüssel verschlüsselt zukommen lassen. Dann wäre es nur ein schwacher Trost, dass andere, nur passiv lauschende Bösewichte, durch die Verschlüsselung mit leeren Händen da stehen würden. Wir werden in Bemerkung 2.1.11 darauf eingehen, dass sich in den Authentifikation die wohl größte Schwachstelle der TLS-Protokolls und der verschlüsselten Internetkommunikation befindet. Als zweites Ziel soll das Protokoll die stärkstmögliche Verschlüsselung aushandeln, mit der beide Seiten einverstanden sind und die dafür erforderlichen Schlüssel austauschen. Letztlich soll während der Übertragung der verschlüsselten Daten sichergestellt werden, dass dieser noch immer von der Kommunikationspartner kommen, mit dem die verschlüsselte Verbindung ursprünglich aufgebaut worden ist. Damit zum Beispiel ein böswilliger Lauscher nicht einfach den Auftrag zu einer zusätzlichen Überweisung in die Kommunikation einschleusen kann, dürfen nicht einfach nur Anweisungen an die Bank mit deren öffentlichem RSA-Schlüssel versendet werden.

Wir gucken uns nun genauer an, wie das TLS-Protokoll diese Ziele umzusetzen versucht:

Abbildung 2.10: Eine Grafik, ähnlich zu TLS-Handshake von Wikipedia erstellen.

Ablauf des TLS-Handshakes.

1. Das Auftaktsignal zum TLS gibt natürlich der Benutzer, der sich gerade zum Aufbau einer verschlüsselten Verbindung entschieden hat. Sein Browser, Emailprogramm oder ähnliches, im Folgenden einfach als Client bezeichnet, sendet unverschlüsselt eine Nachricht namens ClientHello an den Server. Diese Nachricht ist unverschlüsselt und enthält im Wesentlichen Informationen darüber, welche Krytosysteme der Client kann und bevorzugt, welche Methoden zur Datenkompression er kann und bevorzugt und außerdem 32 Byte lange Zufallszahl. Dieses Zufallsdatum wird

später dazu benutzt werden, die Integrität der Daten dieser Eingangsphase abzusichern, da die Verbindung ja noch unverschlüsselt ist und die beiden Kommunikationsteilnehmer nicht wissen, ob sie wirklich mit dem gewünschten Gegenüber sprechen. Die Authentifikation des Gegenübers steht noch aus.

2. Der Server beantwortet den Wunsch zum Verbindungsaufbau mit einer Nachricht namens ServerHello, welche aus den Listen der vom Client unterstützen Kryptosysteme und Kompressionsmethoden die jeweils möglichst stärkste auswählt, die der Server ebenfalls beherrscht. Auch das ServerHello enthält ein neues 32 Byte Zufallsdatum.

3. Als Nächstes sendet der Server dem Client sein sogenanntes Zertifikat. Dieser Zertifikat enthält den öffentlichen Schlüssel des Servers, falls sich die beiden Partien in ihren 'Hello'-Nachrichten zum Beispiel auf RSA geeinigt haben. Außerdem enthält dieses Zertifikat üblicher Weise eine Bestätigung von dritter Stelle, einer sogenannten Zertifizierungsauthorität (certificate authority), die durch eine Signatur bezeugen soll, dass der öffentliche Schlüssel tatsächlich dem gewünschten Gesprächspartner gehört.

4. Der Client überprüft jetzt das Zertifikat des Servers und fährt mit der Kommunikation nur fort, wenn er von der Authentizität der Daten und des Gegenübers überzeugt ist. In dem Fall sendet der Client üblicher Weise das sogenannte PreMasterSecret (46 Bytes), welches er erst an dieser Stelle frisch generiert und dann mit dem öffentlichen Schlüssel des Servers verschlüsselt.

5. Jetzt berechnen beide Seiten aus dem PreMasterSecret und den ausgetauschen Zufallszahlen das MasterSecret. Die Berechnungsvorschrift ist öffentlich bekannt und verwendet als Eingaben das PreMasterSecret, den Text „master secret“ und die beiden zu Beginn ausgetauschten Zufallszahlen.

6. Ab hier wechseln Server und Client auf eine mit dem MasterSecret symmetrisch verschlüsselte Kommunikation zur Übermittlung der eigentlichen Nutzdaten.

Nach der Beschreibung des Handshakes noch einen Satz über die Sicherung der Datenintegrität während der fortlaufenden Kommunikation verlieren.

Dazu muss ich den Diffie-Hellmann Schlüsselaustausch einführen und für die RSA-Signatur zum einen Signaturen überhaupt diskutieren, zum anderen die Man in the Middle Anfälligkeit besprechen, um DH(E)-RSA zu rechtfertigen.

Es ist eine aus verschiedenen Gründen plausible Annahme, dass die verschlüsselte Kommunikation zwischen einem Client und dem Server zwar vielleicht nicht live von einem Angreifer entschlüsselt werden kann, dafür aber vielleicht aufgezeichnet wird und es dem Angreifer gelingt, mit etwas zeitlichem Abstand an den geheimen Schlüssel des Servers zu kommen. Zum Beispiel könnte der Angreifer sich Zugang zu dem Server verschaffen und dort den geheimen Schlüssel auslesen. Selbst wenn das verschlüsselt kommunizierte Geheimnis dann längst nicht mehr auf dem Server wäre, könnte der Angreifer es nach dem oben

beschriebenen TLS-Handshake rekonstruieren, denn die Verschlüsselung basierte auf dem MasterSecret, welches wiederum aus dem PreMasterSecret errechnet wurde, welches der Client öffentlich, aber eben mit dem öffentlichen Schlüssel des Servers verschlüsselt, zum Server gesendet hat. Dieses nachgelagerte Risiko ist unnötig! Diffie und Hellmann haben 19?77? ein auf modularer Arithmetik basierendes Verfahren entwickelt, mit dem sich ein Schlüssel (z.B. das PreMasterSecret) so zwischen Client und Server austauschen ließe, dass ein passiver Angreifer ihn nicht erfahren kann. Und da wir in diesem Szenario einen zeitlichen Abstand angenommen haben, hat der Angreifer keine andere Wahl mehr, als passiv zu sein. Die gute Nachricht ist, dass der Diffie-Hellmann Schlüsselaustausch sogar als Möglichkeit im TLS-Handshake vorgesehen ist! Die etwas schlechtere Nachricht ist wiederrum, dass Stand 2013 kaum Server von dieser Möglichkeit Gebrauch machen[8]. DHE-RSA und perfect forward sercurity. Jahr?

Bemerkung 2.1.11. Die Rolle der Vertrauensnotwendigkeit von CAs sollte abschließend thematisiert werden.

[8] Link zum c't Artikel?

2.2 Codierungstheorie

In diesem Abschnitt wollen wir einen Einblick in die mathematische Disziplin names Coderiungstheorie erhalten. In der Codierungstheorie geht es, ähnlich wie in der Kryptologie, um die Untersuchung von Kommunikation, und ebenso ist es das Ziel, die Kommunikationsinhalte abzusichern. Jedoch ist damit in der Codierungstheorie nicht die Absicherung der Vertraulichkeit gemeint, sondern die Absicherung der Nachrichten gegen Rauschen, Fehler oder Unsicherheiten im Kommunikationskanal. Diese zunächst noch recht umschweifende Beschreibung gilt es mathematisch zu präzisieren und elementare Konzepte der linearen Algebra, wie (Unter-) Vektorräume oder Dimension, werden uns dabei sehr hilfreich unterstützen. Bevor wir unser Kommunikationsmodell aus Abschnitt 2.1 in modifizierter Form wieder aufgreifen, um die Ideen der Codierungstheorie zu formalisieren, wollen wir einen intuitiven Einstieg in das Thema begehen. Lesen Sie dazu langsam, aber unablässig den folgenden Text:

> Diseer Txet etnälht kien eizngies korrktees deuchtes Wrot udn trodtzem knan inh warhsceihnilch jedre lseen: ien ewtas komschies Biepseil frü Cordieungshtoerie.

Man könnte annehmen, dass der obige Text in dem Kommunikationskanal zwischen meinem Kopf und dem Kopf des Lesers irgendwo durch ein Rauschen beeinflusst wurde. Der wahre Sachverhalt ist natürlich, dass ein solches Rauschen vermutlich von verschiedenen codierungstheoretischen Algorithmen abgefangen worden ist und genau der Text oben in der Box zu lesen ist, der dort zu lesen sein sollte. Dieses Rauschen ist mutwillig, aber es demonstriert auf natürliche Weise, wie fehlererkennede, bzw -korrigierende, Codes funktionieren. Ihr Gehirn teilt Ihnen beim Lesen mit, dass mit den Worten in der Box irgendwas nicht stimmt, aber es ist sogar auch in der Lage, Ihnen sofort korrekte deutsche Wörter zu servieren. Das liegt daran, dass die in der Box begangenen Fehler nicht sehr schlimm sind, intuitiv gesehen ist der Abstand zu echten deutschen Wörtern nicht sehr groß. Es gibt viele Untersuchungen dazu, welche Fähigkeiten und welche Wirksamkeit das Gehirn beim Erkennen und Korrigieren von Fehlern in natürlicher Sprache hat **??**. Der Beispieltext in der Box wurde größtenteils nach dem wahrscheinlich bekanntesten Muster erzeugt, dass die Fehler nur zwischen korrekt Wortanfängen und Wortendungen erlaubt. Bei Fehlern nach diesem Muster liefert das Gehirn im Durchschnitt sehr gute Ergebnisse bei der Fehlerkorrektur **??**. Ähnliche Mechanismen haben wir in den vergangenen Jahren unseren Computern beigebracht. Falsche Wörter werden heute in vielen Fällen rot unterstrichen - hier ist also ein fehlererkenneder Code am Werk. Auf Smartphones und Tablets gehen wir mit den Codes oft noch einen Schritt weiter und lassen den erkannten Fehler oft direkt durch ein korrektes Wort ersetzen. Das ist natürlich deswegen naheliegend, da bei der Eingabe über berührungsempfindliche Bildschirme mit viel größerer Wahrscheinlich Fehler im Kommunikationskanal entstehen, denn die Finger treffen die richten Bereiche für die jeweiligen Buchstaben schlechter als auf klassische Tastaturen. Außerdem wird hauptsächliche natürliche Sprache in Smartphones oder Tablets eingegeben. Zwei weitere Schwierigkeiten, denen wir in der Behandlung von Codes wieder begegnen werden, lassen sich an diesem Beispiel auch schon aufzeigen. Es ist sehr schwierig Fehler zu erkennen, bei denen das Rauschen korrektes Code-

wort erzeugt hat. In diesem Satz zum Beispiel, steht ein falsches Elefant. Dieser Fehler ist erst auf dem Kontext erkennbar, denn das Wort 'Elefant' selbst ist korrekt. Solche Fehler finden auf einer höheren Ebene statt und sind durch die Algorithmen, die wir in diesem Kapitel behandeln nicht erkennbar. Die andere Schwierigkeit, die wir im Folgenden allerdings nicht unbeachtet lassen wollen, ergibt sich durch Fehler, die 'zu groß' sind.

Abbildung 2.11: Screenshot: falsche Korrektur

Ein weiblicher Crosant*Croissant heißt Elefantenkuh?
Google: how difficult write spellchecker filetype:pdf
Übungsaufgaben:

Aufgabe 2.2.1. Hier möchte ich gerne die sichere Totto-Wette einbauen, die in Huppert-Willems im Kapitel über Codes besprochen wird. Möchte ich als Hinweis gerne die Anschauung erklären, dass der Tippraum durch Bälle vom Radium der größten zugelassenen Fehltippzahl überdeckt werden soll? Codierungstheoretische Überlegungen liefern dann die Zahl der Tipps die zur Überdeckung nötig ist, aber keinen Algorithmus, der diese Zahl auch realisiert, richtig? Wie löst man das? Hat das Problem allgemeinere Instanzen?

Abschnittsinhalte

1. Idee & Modell
2. Vokabular (vgl LA vs Codierungstheorie)
3. ISBN und Hamming-Codes als einführende Beispiele
4. Reed-Salomon-Codes als 'modernes' Beispiel.
5. Historische Bemerkungen und wieso nicht wie in der Kryptologie viele Jahrhunderte zwischen dem einführenden und dem 'modernen' Beispiel liegen.
6. Abschnitt mit Grafik zu den konkurrierenden Anforderungen an Codes und Einordnung der Beispiele.
7. Praxisabschnitt. High end CRC-Codes auf dem ISBN-Beispiel aufbauen (USB-Bluetooth, ISDN,Ethernet,...Fehlererkennende Codes, Zyklische Codes). Data Matrix (Post), Aztec (DB) und Quick Response (Toyota/Überall) Codes als Anwendung von Reed-Salomon-Codes
8. Kleiner Ausblick auf Datenkompression (verlustfrei und verlustbehaftet) in Richtung mp3-Kapitel.
9. Einige interessante LA-nahe Anregungen finden sich bestimmt in [Lüt03].

Das McEliece-Verfahren

Dieser Abschnitt behandelt das sogenannte McEliece Verschlüsselungsverfahren. Für Leser die dieses Buch tatsächlich begleitend zum Studium der linearen Algebra lesen, stellt dieser Abschnitt vielleicht eine unerwünscht Abschweifung dar. Es werden keine Vorlesungsinhalte der linearen Algebra thematisiert, die nicht schon in den ersten beiden Abschnitten dieses Kapitels behandelt wurden. Für jene aber, die sich daran erfreuen konnten, nur auf der Grundlage ihrer Kenntnisse über lineare Algebra fußend in die Kryptologie und die Codierungstheorie Einblick erhalten zu haben, liefert die Verschmelzung dieser beiden Gebiete im McEliece Verfahren hoffentlich einen krönenden Abschluss dieses Kapitels, der des weiteren noch so nah an der aktuellen Forschung ist, wie wir sonst in diesem Buch nur selten wieder sein werden.

Angriff auf McEliece Public-Key Kryptosystem später als Anwendung von Markov-Ketten, vgl. [CS98]?

Kapitel 3

Lineare Abbildungen und Matrizen

3.1 Lineare Filter und Börsenkurse

An den großen Börsen vergeht kaum ein Tag, ohne dass Analysten und Berichterstatter sich über das Bild einer bestimmten linearen Abbildung unterhalten. Tatsächlich diskutieren Börsianer den Zusammenhang zwischen Elementen im Ziel- und Wertebereich dieser linearen Abbildung leidenschaftlicher als es in den meisten Vorlesungen zur Linearen Algebra der Fall sein dürfte. Sie sehen in den Werten dieser Abbildung eine 'moralische Unterstützung', sehen sie als Zeichen dafür, dass die Zeiten sich ändern oder sogar als das 'Maß aller Dinge'. Die Rede ist von der sogenannten *200-Tage-Linie*.

Abbildung 3.1: Beispielhafte 200-Tage-Linie

Natürlich ist der obige einführende Absatz zu diesem Abschnitt etwas zu boulevardesk für unser wissenschaftliches Anliegen, Anwendungen der Linearen Algebra zu studieren, und bedarf somit gleich einer gewissen Relativierung. Die 200-Tage-Linie ist kein Werkzeug mit eingebauter Gewinngarantie und die leidenschaftlichen Diskussionen der Börsianer haben in der Regel keine linearen oder überhaupt mathematischen Themen zum Inhalt, sondern sind eher psy-

chologisch spekulativer Natur. Die 200-Tage-Linie ist ein Versuch die Börsenweisheit "*The trend is your friend*" faktisch zugänglich zu machen. Trotz aller Relativierungen ist sie aber ein grundlegendes und viel beachtetes Instrument der Chart-Technik und ist auch aus mathematischer Sicht interessant. Zunächst fällt vielleicht auf, dass die beispielhafte 200-Tage-Linie aus Abbildung 3.1 kurvig ist und nicht dem erwarteten Aussehen einer linearen Abbildung entspricht. Den linearen Charakter dieser Abbildung herauszuarbeiten ist daher unser erster Ziel.

Definition 3.1.1. Eine reelle diskrete *Zeitreihe* ist eine Familie $(x_t)_{t\in T}$ reeller Zahlen, deren Indexmenge T durch diskrete Teilmenge $T \subset \mathbb{R}$ gegeben ist.

Das T steht für die zeitliche Komponente der Zeitreihe, oft fasst man T als Menge der Beobachtungszeitpunkte auf. Häufig ist $T = \mathbb{N}, \mathbb{Z}$ oder eine endliche Menge. Durch punktweise Addition und $\mathbb{R}$-Multiplikation erhält die Menge aller gleichartigen (gleiches T) Zeitreihen eine reelle Vektorraumstruktur:

$$\begin{aligned}(a_t)_{t\in T} + (b_t)_{t\in T} &:= (a_t + b_t)_{t\in T},\\ c \cdot (a_t)_{t\in T} &:= (c \cdot a_t)_{t\in T} \text{ für } c \in \mathbb{R}.\end{aligned}$$

Diese Vektorräume wollen wir mit $\mathbb{R}^T$ bezeichnen. Für eine endliche Menge T ist der Vektorraum endlichdimensional und isomorph zum $\mathbb{R}^{|T|}$. Für nichtendliche Mengen, wie $T = \mathbb{N}, \mathbb{Z}$, ist der Vektorraum $\mathbb{R}^T$ unendlichdimensional. Durch praktische Beobachtungen gewonnene Zeitreihen können immer nur eine endliche Indexmenge T haben. Der in Abbildung 3.1 abgebildete Deutsche Aktienindex DAX wurde seit seiner Einführung 1988, zumindest während der Handelszeiten, alle 15 Sekunden und später sogar sekündlich berechnet. Von seiner ersten Berechnung bis zur aktuellsten n-ten Berechnung gibt es endlich viele Werte

$$DAX_1, \ldots, DAX_n \in \mathbb{R}.$$

Um elegant generelle Aussagen machen zu können oder um auch zukünftige Entwicklungen der Zeitreihe im gleichen Modell mitberücksichtigen zu können, fasst man solche endlichen, über $[n] := \{1, \ldots, n\}$ indizierten, Zeitreihen gerne als Elemente in einem größeren Vektorraum auf. Für eine Inklusionsabbildung $\iota : T \hookrightarrow T'$ fasst man $a = (a_t)_t \in \mathbb{R}^T$ dann als Element in $\mathbb{R}^{T'}$ auf, indem man das Bild $\iota_*(a)$ unter der Abbildung

$$\iota_* : \mathbb{R}^T \hookrightarrow \mathbb{R}^{T'}, \quad (\iota_*(a)_t) = \begin{cases} a_s & \text{, falls } t = \iota(s) \\ 0 & \text{, sonst} \end{cases}$$

betrachtet. ι_* ist eine lineare Abbildung und daher ist das Bild von ι_* ein Untervektorraum von $\mathbb{R}^{T'}$.

Definition 3.1.2. Der gleitende linksseitige Mittelwert der Länge $r \geq 1$ ist definiert als

$$\phi_r : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}, \quad (a_t)_{t\in\mathbb{Z}} \mapsto \left(\frac{1}{r}\sum_{i=0}^{r-1} a_{t-i}\right)_{t\in\mathbb{Z}}.$$

Man verifiziert leicht das folgende Lemma:

Lemma 3.1.3. *ϕ_r ist eine lineare Abbildung.*

Die in der Einleitung dieses Abschnitts gefeierte 200-Tage-Linie ist definiert als das Bild $\phi_{200}(DAX)$ der DAX-Zeitreihe, die aus den Tagesschlussständen des DAX besteht. Das kurvige Erscheinungsbild der in Abbildung 3.1 rötlich dargestellten 200-Tage-Linie steht daher nicht im Widerspruch zur Linearität von ϕ_{200}.

Während in der Analysis die Untersuchung von Folgen maßgeblich auf Konvergenz beschränkt ist, interessiert man sich in der Zeitreihenanalyse in gleichem Maße für nicht in irgendeinem Sinne konvergente Folgen, bzw. Familien. Ein Ziel besteht darin eine gegebene Zeitreihe $(a_t)_{t\in T}$ in ihre konstituierenden Komponenten zu zerlegen. Damit kann im Fall von Messreihen eine Zerlegung in tatsächliche Messwerte und Messfehler angestrebt sein, um letztere dann heraus zu rechnen. Als allgemein üblich gilt eine Zerlegung in Komponenten $(b_t)_t, (s_t)_t, (r_t)_t$, so dass in einem sogenannten additiven Modell gilt

$$a_t = b_t + s_t + r_t,$$

wobei die Zeitreihe (b_t) in einem zu präzisierenden Sinne den Trend der Zeitreihe (a_t) beschreiben möge, die Zeitreihe (s_t) einen saisonalen oder periodischen Anteil von (a_t) beschreibt und gegebenenfalls nicht zuzuordnende Restterme in der Zeitreihe (r_t) abgelegt werden. Diese grobe Umschreibung der Anforderung an die Zerlegung kann nicht allgemein spezifiziert werden, sondern muss, je nach Art der Daten und der gewünschten Erkenntnisse, im Einzelfall präzisiert werden.

Beispiel 3.1.4. Betrachten wir als einfaches Beispiel die Zeitreihe $(a_t) \in \mathbb{N}^{34}$ der bis einschließlich des t-ten Spieltags geschossenen Tore des Hamburger SV in der Fußball-Bundesliga Saison 2013/2014:

$$(3, 4, 4, 8, 10, 10, 12, 17, 20, 23, 23, 26, 29, 30, 30, 31, 33, 33, 33, 33, 35, 38, 38, 39, 41, 41, 42, 43, 45, 46, 47, 48, 49, 51)$$

Der HSV hat in den 34 Spielen der regulären Saison 51 Tore erzielt und damit im arithmetischen Mittel genau $1,5$ Tore pro Spiel. Eine erste Visualisierung der Zeitreihe in Abbildung 3.2 suggeriert eine im Wesentlichen lineare Entwicklung. Wir vernachlässigen mögliche periodische Effekte, wie sie zum Beispiel durch Heim- und Auswärtsspiele erzeugt sein könnten, und untersuchen ein lineares Trend-Modell

$$a_t = f(t) + \epsilon_t,$$

mit einer (affin) linearen Funktion $f : \mathbb{R} \to \mathbb{R}, x \mapsto a \cdot x + b$, wobei $a, b \in \mathbb{R}$ noch so zu bestimmen sind, dass die Zeitreihe $(a_t)_t$ *möglichst gut* beschrieben wird. Die Residuen $(\epsilon_t)_t$ sollen bei einer solchen Beschreibung im arithmetischen Mittel 0 sein und möglichst wenig streuen. Der naive Schätzer

$$f(t) = 1,5 \cdot t$$

würde die Zeitreihe systematisch unterschätzen, die Residuen wären weder besonders klein, noch irgendwie gleichmäßig verteilt. Es lässt sich zeigen (vgl. Satz von Gauß-Markov [Geo04, Satz (12.15b)]), dass die Gaußsche Methode der kleinsten Quadrate eine in diesem Sinne beste Approximation liefert. Elementar beschrieben ermittelt man die Koeffizienten a und b von f dabei so, dass der

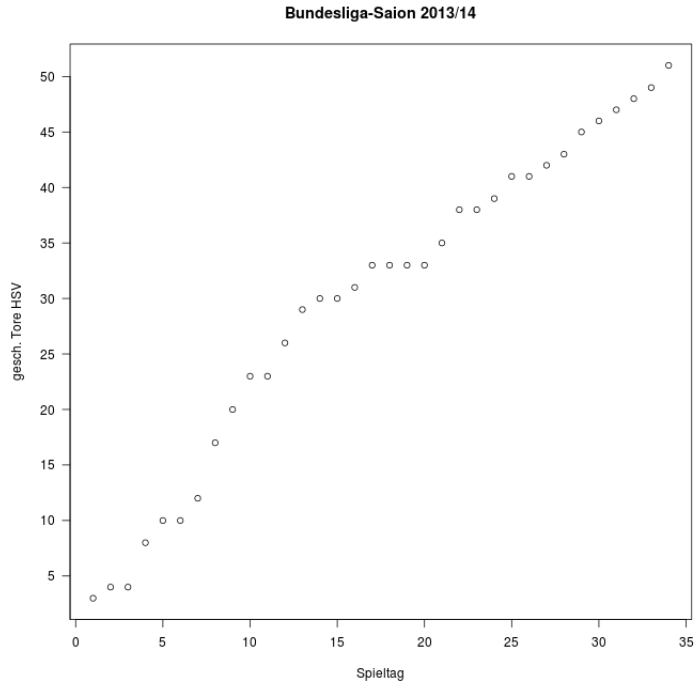

Abbildung 3.2: Zeitreihe der geschossenen Tore des Hamburger SV in der Fußball-Bundesliga Saison 2013/2014

Ausdruck

$$\sum_{t=1}^{34} (a_t - f(t))^2$$

minimiert wird. Der Fehler $a_t - f(t)$, den die Regressionsgerade f an der Stelle t macht, wir dabei quadratisch aufsummiert und die Summe der Quadrate ist durch eine geeignete Wahl von a und b zu minimieren - daher der Name der Methode. Diese Methode lässt sich mit den Mitteln Orthogonalität und Approximation, die in dieser Kombination erst in Kapitel 5 im Fokus stehen werden, sehr elegant beschreiben. Diese Chance soll hier, auch für den späteren Rückgriff, nicht ungenutzt verstreichen, wobei das Studium der Methode der kleinsten Quadrate in der folgenden Box beim ersten Lesen getrost übersprungen werden kann, wenn das Konzept der Orthogonalität oder der Zusammenhang zur Approximation noch nicht vertraut sind.

Die Methode der kleinsten Quadrate.
Das Ziel dieses kurzen Abschnittes ist es, eine endliche Zeitreihe $(a_t)_{t\in T}$ mittels der Methode der kleinsten Quadrate durch ein Polynom

$$f(X) = a_0 + a_1X + \ldots + a_dX^d$$

vom Grad $\leq d$ optimal zu approximieren. Die obige HSV-Beispiel angestrebte affin lineare Funktion entspricht dem Fall $d = 1$, aber der etwas allgemeinere Fall lässt sich dank der eingesetzten linearen Algebra ohne zusätzlichen Aufwand abhandeln. Es bezeichne $\mathbb{R}_{\leq d}[X]$ den $\mathbb{R}$-Vektorraum der Polynome vom Grad $\leq d$, der Vektorraum der Zeitreihen in T sei hier auf natürliche Weise mit $\mathbb{R}^{|T|}$ identifiziert. Die Einschränkung $f|_T$ einer Polynomfunktion $f : \mathbb{R} \to \mathbb{R}$ aus $\mathbb{R}_{\leq d}[X]$ definiert einen lineare Abbildung

$$\mathbb{R}_{\leq d}[X] \xrightarrow{\cdot|_T} \mathbb{R}^{|T|},$$

die für $d < T$ sogar injektiv ist und so $\mathbb{R}_{\leq d}[X]$ mit einem Untervektorraum von $\mathbb{R}^{|T|}$ identifiziert. Zu der vorgegebenen Zeitreihe $(a_t)_t$ gilt es nun ein Element $f|_T$ im Bild von $\cdot|_T$ zu finden, dass bezüglich der vom euklidischen Skalarprodukt induzierten Norm $\|\cdot\|_2$ den kleinsten Abstand zu $(a_t)_t$ hat, da dies aufgrund der Monotonie der Wurzelfunktion auch die Summe der Fehlerquadrate minimiert:

$$\operatorname*{argmin}_f \sum_{t\in T} (a_t - f(t))^2 = \operatorname*{argmin}_f \|(a_t)_t - f|_T\|_2$$

Dieses eindeutige minimierende Element lässt (vgl. Kapitel 5) sich durch die Orthogonalprojektion $\pi((a_t)_t)$ der Zeitreihe in den Untervektorraum $(\mathbb{R}_{\leq d}[X])|_T$ ermitteln:

$$\Rightarrow \operatorname*{argmin}_f \sum_{t\in T} (a_t - f(t))^2 = \pi((a_t)_t).$$

Dementsprechend betrachten wir für die lineare Regression der HSV-Zeitreihe das Bild der Monom-Basis

$$v_1 := 1|_T = (1, \ldots, 1)^t, \quad v_2 := X|_T = (1, 2, \ldots, 34)$$

und orthogonalisieren die beiden Vektoren v_1, v_2 zu einer orthogonalen Basis

$$u_1 := v_1, \quad u_2 := v_2 - \frac{\langle v_2, u_1\rangle}{\langle u_1, u_1\rangle} u_1 = v_2 - \overline{v_2}u_1$$

von $(\mathbb{R}_{\leq 1}[X])|_T$, wobei hier $\overline{v_2} = 17,5$ das arithmetische Mittel von v_2 bezeichne. Die orthogonale Projektion der Zeitreihe $(a_t)_t$ der geschossenen Tore auf den Untervektorraum berechnet sich als

$$\begin{aligned}
\pi((a_t)_t) &= \frac{\langle (a_t)_t, u_1\rangle}{\langle u_1, u_1\rangle} \cdot u_1 + \frac{\langle (a_t)_t, u_2\rangle}{\langle u_2, u_2\rangle} \cdot u_2 \\
&= \overline{(a_t)_t} \cdot v_1 + c \cdot (v_2 - \overline{v_2} \cdot v_1) \\
&= \left(\overline{(a_t)_t} - c \cdot \overline{v_2}\right) \cdot v_1 + c \cdot v_2 \\
&\simeq 5,13 \cdot 1|_T + 1,41 \cdot X|_T = (5,13 + 1,41X)|_T
\end{aligned}$$

Die errechnete Regressionsgerade ist in Abbildung 3.3 eingezeichnet. Zusätzlich zeigt sich in der Abbildung, dass die Residuen, ausgerichtet zur Skala an der rechten Seite, nicht all zu groß und vor allem zentriert sind. Ganz frei von Struktur scheint die Verteilung der Residuen allerdings nicht und obwohl diese Aussagekraft von der Statistik der geschossenen Tore nicht zwangsläufig zu erwarten ist, lässt sich sogar gerade die Verteilung der Residuen mit den guten und den schlechten Phasen des Hamburger SV in jener Saison in Einklang bringen.

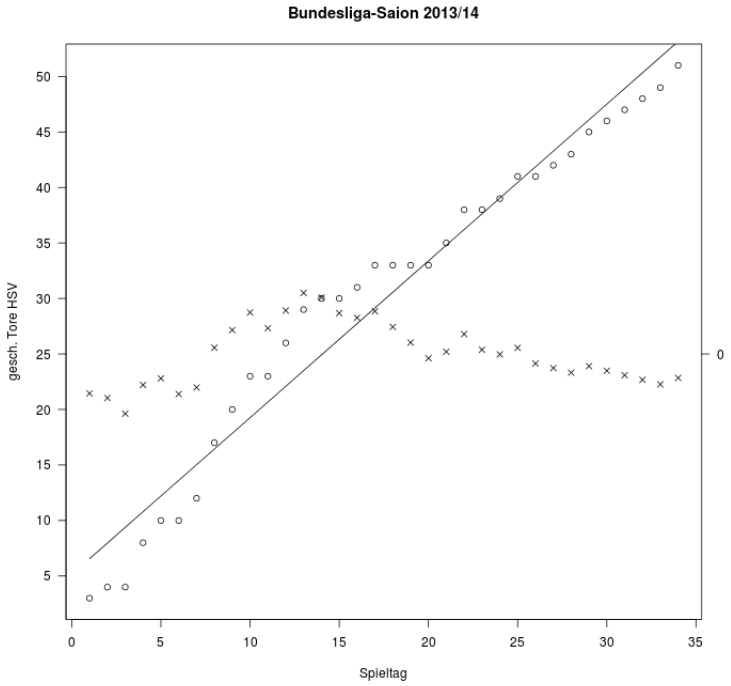

Abbildung 3.3: Lineare Regression und Residuen (×) zu der HSV-Zeitreihe.

In vielen Situationen ist es nicht sinnvoll die Trendkomponente (b_t), wie im vorangegangenen Beispiel, durch eine lineare Funktion $b_t = a \cdot t + b$ zu modellieren - seien die Koeffizienten a, b noch so geschickt gewählt, denn ein lineares Trend-Modell impliziert, dass der Trend sich unaufhaltsam und mit gleichbleibender Geschwindigkeit in alle Ewigkeit fortsetzt. Das ist, um auf unser Eingangsbeispiel zurück zu kommen, für Börsenkurse keine sehr sinnvolle Annahme, da sich dort immer wieder Phasen im wesentlichen steigender Kurswerte (sogenannte *Haussen*) und Phasen tendenziell fallender Kurse (*Baissen*) erkennen (oder zumindest empfinden) lassen. Diese Phasen sind allerdings nicht unbedingt durch eine monotone Kursentwicklung gekennzeichnet, sondern dadurch, dass abzüglich kleiner nervöser Schwankungen ein 'schwingen' in eine der beiden möglichen Richtungen vorliegt. Die 200-Tage-Linie ist dabei ein weit verbreiteter Versuch, diese kleinen Schwingungen zu eliminieren und die Kurse so zu glätten, dass längere Phasen monotoner Kursentwicklung sichtbar gemacht werden. Aus mathematischer Sicht ist es dann natürlich dringend erforderlich

zu untersuchen, inwiefern sich gleitende Mittelwerte, wie die 200-Tage-Linie und ihre Variationen, zu dieser Anwendung überhaupt eignen. In der Signalverarbeitung spricht man von einem Tiefpassfilter, wenn eine Abbildung hochfrequente Schwingungen aus einem Signal entfernt oder zumindest signifikant schwächt, während sie Tiefen, also Schwingungen mit niedriger Frequenz nahezu ungehindert erhält.

ϕ_N als Tiefpassfilter.

Um das Tiefpassverhalten von ϕ_N zu studieren importieren wir daher einige Begriffe aus der Signalverarbeitungstheorie. Eine Zeitreihe wird dort als Signal betrachtet und eine Abbildung von Signalen als System. Die gleitenden Mittelwerte ϕ_r sind Beispiele für Systeme. Man nennt $(\phi_r(a_t))_t$ das Ausgangssignal zum Eingangssignal $(a_t)_t$. Da ϕ_r nach Lemma 3.1.3 eine lineare Abbildung ist, spricht man in diesem Fall von eine linearen System. Des Weiteren erfüllt ϕ_r die Eigenschaft der *Zeitinvarianz*: Für jedes $t_0 \in \mathbb{Z}$ gilt

$$\phi_r \circ s_{t_0} = s_{t_0} \circ \phi_r,$$

wobei s_{t_0} ein Verschiebe-Operator ist, der dadurch definiert, dass in der t-ten Komponente von $s_{t_0}((a_t)_{t\in\mathbb{Z}})$ das Element a_{t-t_0} steht. Lineare zeitinvariante Systeme (LZI-Systeme) haben eine besonders zugängliche und gut verstandene Theorie. LZI-Systeme lassen sich vollständig durch ihre Antwort auf das Impulssignal

$$\delta = (\delta_t)_{t\in\mathbb{Z}}, \quad \delta_t := \begin{cases} 1 & \text{, falls } t = 0 \\ 0 & \text{, sonst,} \end{cases}$$

beschreiben. Das ist lediglich die signaltheoretische Umformulierung der Tatsache [Fis05, XXX], dass eine linear Abbildung durch ihre Werte auf einer Basis bestimmt ist. Die Standardbasis des Vektorraums $\mathbb{R}^{\mathbb{Z}}$ ist gegeben durch die Familie

$$\ldots, e_{-1}, e_0, e_1, \ldots, e_t, \ldots, \text{ jedoch ist } e_t = s_t(e_0) = s_t(\delta).$$

Die Zeitinvarianz sorgt daher dafür, dass die Auswertung der Abbildung an δ, die sogenannte Impulsantwort, schon die ganze Abbildung vollständig beschreibt.

Beispiel 3.1.5. Es sei ein $r \in \mathbb{N}$ fixiert und wir berechnen die Impulsantwort des gleitenden linksseitigen Mittelwerts ϕ_r der Länge r. Offensichtlich gilt

$$\phi_r(\delta)_t = \begin{cases} \frac{1}{r} & \text{, falls } 0 \leq t \leq r-1 \\ 0 & \text{, sonst.} \end{cases}$$

Für eine elegante Untersuchung des Antwortverhaltens von LZI-Systemen auf Schwingungen empfiehlt sich ein kurzzeitiger Übergang ins Komplexe. Wir definieren die *komplexe Exponentialfolge* $(e_t^{i\omega})_{t\in\mathbb{Z}}$, mit Frequenz $\omega \in \mathbb{R}$, durch

$$e_t^{i\omega} := e^{i\omega t} = \cos(\omega \cdot t) + i \cdot \sin(\omega \cdot t) \in \mathbb{C}.$$

Wendet man ein LZI-System auf Schwingungen wie die Exponentialfolgen an, dann erhält man als Bild im wesentlichen die gleiche Schwingung, lediglich komplex skaliert.

Lemma 3.1.6. *Sei $F : \mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}}$ ein LZI-System und $\omega \in \mathbb{R}$, dann ist*

$$F((e_t^{i\omega})_t) = H(\omega) \cdot (e_t^{i\omega})_t,$$

wobei $H(\omega) \in \mathbb{C}$ ein von ω abhängiger Skalar ist, der durch

$$H(\omega) = \textit{Z-transformierte der Impulsantwort an der Stelle } e^{i\omega}$$

gegeben ist.

Beweis. H korrekt hinschreiben. □

Bemerkung 3.1.7. Es kann beim ersten Lesen zunächst einfach zur Kenntnis genommen werden, dass sich die Aussage von Lemma Kontext von Eigenwerten und Eigenvektoren so formulieren lässt, dass die Exponentialfolgen stets Eigenvektoren, bzw Eigenfunktionen, von LZI-Systemen zu Eigenwerten $H(\omega)$ sind. Da wir Eigenwerte und Eigenvektoren bis zum Kapitel 4 nicht voraussetzen wollen, soll, außerhalb dieser Bemerkung, an dieser Stelle noch nicht sonderlich von diesen Formulierungen gebraucht gemacht werden. Für jene Leser, die mit Eigenwerten bereits vertraut sind, sei jedoch noch angemerkt, dass LZI-Systeme ein Beispiel dafür liefern, dass lineare Abbildungen auf unendlichdimensionalen Vektorräumen sehr große Mengen von Eigenwerten, sogenannte Spektren haben können. In einführenden Vorlesungen zur Linearen Algebra beschränkt man sich häufig auf das Studium endlichdimensionaler Vektorräume und lernt daher zunächst nur endliche Mengen von Eigenwerten eines Endomorphismus kennen. Betrachtet man, wie hier mit $\mathbb{R}^{\mathbb{Z}}$ und $\mathbb{C}^{\mathbb{Z}}$, unendlichdimensionale Vektorräume, dann brauchen die Spektren von Endomorphismen noch nicht mal diskret zu sein. Für eine weiterführende Behandlung dieses Phänomens sei auf Lehrbücher zur Funktionalanalysis verwiesen, z.B. [Wer07].

Die durch eine Laurent-Reihe gegebene Funktion H aus Lemma 3.1.6 beschreibt also das Dämpfungs- oder Resonanzverhalten, allgemeiner den *Frequenzgang*, eines LZI-Systems. Es soll daher H noch etwas genauer untersucht werden, um zumindest das Filterverhalten gleitender Mittelwerte wie der 200-Tage-Linie besser zu verstehen.

Definition 3.1.8. Sei $(x_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$. Die z-Transformation $X(z)$ von $(x_t)_t$ ist definiert als formale Laurent-Reihe

$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n} \in \mathbb{C}((z)).$$

Der Frequenzgang H eines LZI-Systems F wird durch Lemma 3.1.6 also identifiziert als z-Transformation der Impulsantwort von F, ausgewertet für die Frequenz ω an der Stelle $e^{i\omega}$.

Aufgabe 3.1.9. Falls das schön gelingt, würde ich hier gerne eine Aufgabe einstreuen, welche in Teil a) die z-Transformation mit der erzeugenden Funktionen zu einer Folge in Verbindung bringt und in Teil b) den Frequenzgang mit einer Fourierreihe verknüpft. Am Ende von Teil b) könnten man dann vielleicht einen Verweis auf Fourieranalyse einbauen und den in späteren Abschnitten (JPEG?) wieder aufnehmen?!

Um den Frequenzgang der gleitenden Mittelwerte ϕ_r zu studieren, berechnen wir die z-Transformation der Impulsantwort aus Beispiel 3.1.5.

Beispiel 3.1.10. Für den gleitenden linksseitigen Mittelwert ϕ_r der Länge r ist die z-Transformierte der Impulsantwort gegeben durch

$$\begin{aligned} H(z) =& \sum_{n=0}^{r-1} \frac{1}{r} z^{-n} = \frac{1}{r} \cdot \sum_{n=0}^{r-1} z^{-n} \\ =& \frac{1}{r} \cdot \frac{1 - z^{-r}}{1 - z^{-1}} \end{aligned} \tag{3.1.1}$$

Der Frequenzgang zeigt sich gemäß Lemma 3.1.6 in der Betrachtung des Betrags der Funktion

$$\mathcal{H} : [0, 2\pi] \to \mathbb{R}, \qquad t \mapsto H(e^{it}).$$

Für den gleitenden Mittelwert ϕ_3 ergibt sich folgendes Bild: Man erkennt in Ab-

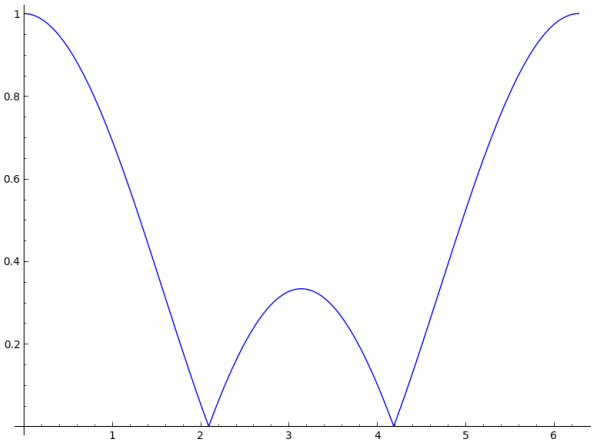

Abbildung 3.4: Dämpfung in Abhängigkeit der Frequenz

bildung 3.4, aber auch in Gleichung (3.1.1), dass ϕ_3 die Frequenz $\frac{2\pi}{3}$ vollständig aus dem Signal herausfiltert. Alle anderen von Null verschiedenen Frequenzen werden zumindest etwas gedämpft. Darüber hinaus fällt auf, dass oberhalb der vollständig herausgefilterten Frequenz ein Intervall existiert, in dem die Frequenzen ϕ_3 sogar mit zunehmender Intensität passieren können. Dieser Effekt ist für einen Tiefpassfilter natürlich wenig wünschenswert. Untersucht man den Frequenzgang für gleitende Mittelwerte unterschiedlicher Länge, in Abbildung 3.5 von $r = 1$ (blau) bis $r = 20$ (rot), so stellt man fest, dass sich dieser Effekt abschwächt, aber nie ganz verschwindet. Dies führt dazu, dass gleitende Mittelwerte nicht auf alle Zeitreihen pauschal eine glättende Wirkung haben und Entwicklungen lokal sogar umgekehrt darstellen können. Die Abbildung 3.5 lässt aber auch erkennen, dass sich gleitende Mittelwerte ϕ_N von größerer Länge zunehmend als Tiefpassfilter eignen, allerdings zu der Frequenz $\frac{2\pi}{N}$ auch immer früher abfallen und somit nur ein immer kleineres Frequenzband unbeschädigt passieren lassen.

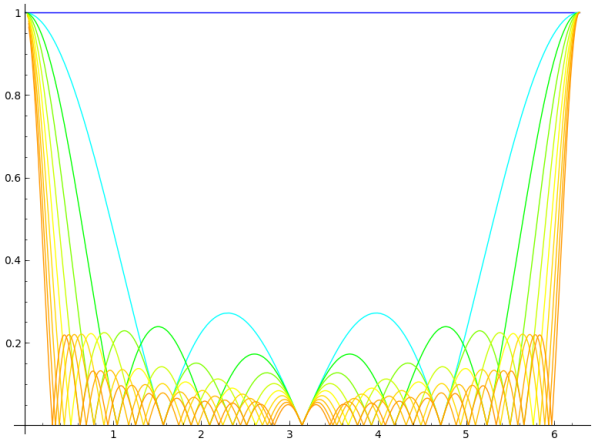

Abbildung 3.5: Dämpfung von ϕ_r für verschiedene Längen r.

Aufgabe 3.1.11. Die Zeitreihe $(a_t)_t$ beschreibe in monatlicher Abfolge erhobene Daten. Für welches $r \in \{1,2,3,4,5,6\}$ extrahiert die Zeitreihe $(a_t)_t - \phi_r((a_t)_t)$ die quartalsperiodische Entwicklung?

Um die verschiedenen Vor- und Nachteile der gleitenden Mittelwerte als Filter auszunutzen oder eben auszubessern, sind diverse abgewandelte Filter intensiv studiert worden. Variiert man Definition 3.1.2 geringfügig zu einer Abbildung

$$F_B : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}, \quad (a_t)_{t\in\mathbb{Z}} \mapsto \left(\sum_{i=-r}^{r} b_i \cdot a_{t-i} \right)_{t\in\mathbb{Z}},$$

wobei $B = (b_{-r}, \ldots, b_r) \in \mathbb{R}^{2r+1}$ ein "Gewichtsvektor" mit $\sum_{i=-r}^{r} b_i = 1$ sei, dann verallgemeinert man das Konzept des dort definierten linksseitigen gleitenden Mittelwerts zu einem allgemeineren Konzept des *gewichteten* gleitenden Mittelwerts, der je nach Wahl von B den Anschluss zu bedeutenden Filtern herstellt, z.B.

- den Binomial-Filtern oder diskreten Gauß-Filtern mit

$$b_i = \frac{1}{2^{2r}} \cdot \binom{i+r}{2r},$$

- oder den exponentiellen Filtern mit $\alpha \in [0,1]$, "$r = \infty$" und

$$b_i = \begin{cases} 0 & \text{, falls } i < 0 \\ \alpha(1-\alpha)^i & \text{, sonst,} \end{cases}$$

und öffnet so die Tür in das Gebiet des Filter-Designs, mit vielen beeindruckenden Anwendungen in der Zeitreihen- oder allgemeiner Datenanalyse und Bildverarbeitung, ein weiteres Stück.

3.2 Linearer Zufall

Lineare Schieberegister, Pseudozufallszahlen, Funktürschlüssel, Internethacks, Stromchiffren

Einführender Abschnitt

Folge von Zufallszahlen $(a_0, a_1, \ldots)$, Computer, deterministisch, endlich daher periodisch und in $\mathbb{F}_{2^n}$

Definition 3.2.1 (Linearer Kongruenzgenerator)**.** Sei a..., dann heißt die Folge $(a_n)_n$ von einem linearen Kongruenzgenerator erzeugt, wenn

$$a_n = a \cdot a_{n-1}$$

Schwäche von lineare Kongruenzgeneratoren diskutieren und zitieren, in welchen Programmiersprachen diese für Standardzufall verantwortlich sind.

zu besseren Pseudozufallsgeneratoren überleiten

Definition 3.2.2 (Linear rückgekoppeltes Schieberegister)**.**

Idee

Lineare Schieberegister, beziehungsweise linear rückgekoppelte Schieberegister (LFSR), haben vielfältige Anwendungen im Bereich von 'Pseudozufall'. Zunächst sollte in diesem Vortrag die Notwendigkeit und die Schwierigkeit der Erzeugung 'digitalen Zufalls' vermittelt werden. Es könnte etwas Warteschlangentheorie vorgeführt werden. Hier gehen Zufallszahlen in Simulationen ein, indem z.B. durch die Inversionsmethode aus einer Gleichverteilung andere Verteilungen (hier z.B. Poisson-Verteilung) erzeugt werden. Konkretes Beispiel könnte ein Wartezeitmodell für Schalter/Kassen/Toiletten sein. Natürlich sollte auch die Bedeutung von Zufallszahlen in der Kryptographie thematisiert werden. Je nach endgültiger Reihenfolge der Vorträge könnte auch intensiver die Verwendung von Pseudozufall in der Codierungstheorie (Scrambler, CDMA) beleuchtet werden. Als Anwendungen stehen hier die Übertragung von schwarzen Bildern via Digitalfernsehen und die Datengeschwindigkeit von mobilem Internet à la UMTS bereit. Hier entsteht auch eine Verknüpfungsmöglichkeit zum Vortrag 'aus Kupfer Gold' und DSL-Techniken (Vectoring,...). Unbedingt möchte ich in diesem Vortrag auch die Anwendung 'Funkautotürschlüssel' behandeln. Hier besteht die Problematik darin, dass nicht das statische Signal 'Tür öffnen' vom Schlüssel zum Auto übertragen werden darf, da sich dieses Signal von Eve (böse) einfach aufzeichnen und wieder abspielen lassen würde. Stattdessen soll möglichst konkret belegt werden, dass ein LFSR-'pseudozufälliges' Signal übertragen wird. Als interessanter Gag kann hier das Problem des synchronisationsverlusts zwischen Auto und Schlüssel eingestreut werden: Bedient man einen Funkschlüssel zu oft außerhalb der Reichweite des Autos, dann erzeugt des Schüssel Zufallszahlen außerhalb des Konfidenzintervalls des Autos und öffnet/schließt dieses nicht mehr. Weitere Anwendungsbeispiele von (nicht ausreichend zufälligem) Pseudozufall würde ich in der Informatik heraussuchen (prominente Hacks durch 'Erraten' der pseudozufälligen IP-Sequenznummer oder Address Space Layout Randomization als Gegenmaßnahme zu Hacks durch Pufferüberläufe).

3.2.1 Autoschlüssel

3.2.2 Scrambler

Einen Absatz über allgemeine Anwendungsgebiete von Scramblern. Dann darauf eingehen, wie Codemultiplexverfahren (Code Division Multiple Access - CDMA) eingesetzt werden können, um sich Übertragungskanäle zu teilen. Abgrenzung zu Frequenzmultiplexverfahren. Beispielsituationen in denen deutlich wird, dass solche Verfahren einen wertvollen praktischen Nutzen haben. Zentral herausarbeiten wie Pseudozufallsgeneratoren hier mitarbeiten können. Für Rückreferenzierung aus späterem Kapitel die Bedeutung der Orthogonalität in Zusammenhang mit Kovarianz und Korrelation darstellen.

3.3 Lineare Optimierung

Lineare Optimierung ist ein mathematisches Thema von großer wirtschaftlicher Bedeutung. Um diese These zu untermauern sollen 1-2 möglichst moderne Optimierungsprobleme vorgestellt und Lösungsansätze diskutiert werden.

Definition 3.3.1. Sei $z : \mathbb{R}^n \to \mathbb{R}$ eine Linearform, $A \in Mat(n \times m, \mathbb{R})$ und $b \in \mathbb{R}^m$. Die Aufgabe eine *optimale Lösung* $x \in \mathbb{R}^n$ zu bestimmen, so dass die *Nebenbedingungen*

$$A \cdot x \leq b \text{ und}$$
$$x \geq 0$$

erfüllt und die *Zielfunktion* z maximiert wird, d.h.

$$z(x) \geq z(x') \quad \forall x' \in \mathbb{R}^n,$$

heißt *lineares Programm.*

Unter der Fragestellung "Wie entscheidet man, ob die Nebenbedingungen erfüllbar sind?" Fourier-Motzkin-Elimination behandeln.

Lemma 3.3.2 (Farkas)**.** *Sei $A \in Mat(n \times m, \mathbb{R})$ und $b \in \mathbb{R}^m$, dann existiert entweder ein Element $x \in \mathbb{R}^n$, mit*

$$Ax \leq b \text{ und } x \geq 0,$$

oder es existiert ein Element $y \in \mathbb{R}^m$, mit

$$A^t y \geq 0 \text{ und } b^t \cdot y < 0.$$

Beweis. □

Vielleicht etwas über duale Lineare Programme sagen? Dann auch etwas über Lösung von LP mittel Fourier-Motzkin.

Polytop $P(A, b)$ und Ecken von Polytopen definieren.

Lemma 3.3.3. *Hat das durch (z, A, b) gegebene Lineare Programm überhaupt eine optimale Lösung, dann ist ebenfalls eine Ecke des Polytops $P(A, b)$ eine optimale Lösung von (z, A, b).*

Beweis. □

Das Simplexverfahren. Das Simplexverfahren soll ausführlich thematisiert werden.

Kurz die Existenz und Bedeutung anderer Verfahren thematisieren und Geschichte der Zugehörigkeit zu P oder NP linearer Programme benutzen.

3.4 Lineares Diskriminieren

Diskriminierung ist heikles Thema. In Artikel 14 der Europäischen Menschenrechtskonvention[1] ist gar ein Diskriminierungsverbot verankert. Das liegt zu einem gewissen Teil daran, dass dieses dem lateinischen *discriminare* entlehnte Wort in seiner deutschen Verwendung ein gleichzeitiges Benachteiligen oder Herabsetzen des vom Einen unterscheidbaren Anderen impliziert. Wenn wir uns nun im Folgenden mit der Diskriminanzanalyse beschäftigen wollen, dann ziehen wir uns zunächst auf ein Studium der Theorie eines mathematisch wertneutralen Unterscheidens zurück. Dennoch entsteht hier aus der Theorie ein mächtiges Instrument, dessen Einfluss auf unser Leben in dem Maße zunimmt, in dem es auf der Welle von *Big Data*, *Data Mining* und *Machine Learning* weiter in unzählige Anwendungen getragen wird. Der Mathematiker möge die Verantwortung dafür tragen, die Gesellschaft für diejenigen Stellen zu sensibilisieren, in denen sein Instrument, für andere möglicherweise undurchsichtig, auf eine unmoralische Weise verwendet wird oder werden soll. Gleichzeitig beinhaltet diese Theorie das Potential Teil von Anwendungen mit großem allgemeinen Nutzen zu sein, beispielsweise bei der Früherkennung von Krankheiten in der medizinischen Forschung.

In der Diskriminanzanalyse geht es darum, eine Theorie des formalen induktiven Lernens zu etablieren, also abgeleitet nach [BBL04] darum, ein Phänomen in einigen Instanzen zu beobachten, ein allgemeines formales Modell zu jenem Phänomen zu bilden und aus dem Modell möglichst zutreffende Vorhersagen über weitere Instanzen des Phänomen abzuleiten. Mit anderen Worten soll aus Wissen über Spezialfälle ein allgemeineres Wissen generalisiert werden. Dafür ist es offenbar grundlegend Unterschiede und Gemeinsamkeiten zwischen verschiedenen Instanzen erkennen zu können. Es bezeichne I eine Menge von "Instanzen", denen mittels einer Abbildung

$$x : I \to \mathbb{R}^n$$

Merkmalsvektoren zugeordnet seien. Außerdem haben alle Instanzen einen "Typ"

$$t : I \to T = \{t_0, \ldots, t_n\}.$$

Die Situationen soll nun so interpretiert werden, dass die Merkmalsvektoren $x(i)$ für alle Instanzen beobachtbar sind, während der Typ $t(i)$ nicht allgemein zugänglich ist und uns nur für eine Teilmenge $I' \subset I$ bekannt ist. Das Ziel

[1] Einsehbar unter `http://www.echr.coe.int/Documents/Convention_DEU.pdf`

besteht nun darin, aus den Merkmalsvektoren $x(i)$ der bekannten Spezialfälle $i \in I'$ eine *Diskriminanzfunktionen*

$$d : x(I) \to T$$

abzuleiten, so dass im Optimalfall $d \circ x = t$ gilt oder der Fehler in einem dann zu präzisierenden Sinn möglichst klein ist. Der einfacheren Darstellung halber soll im Folgenden angenommen werden, dass nur es nur zwei verschiedene Typen $(n = 1)$ gibt.

Beispiel 3.4.1. Eine typische Problemstellung in der Lebensmittelchemie ist die Herkunftskontrolle von Produkten. Das Lebensmittel- und Futtermittelgesetzbuch[2] verbietet in §11, Abs. 1, Nr. 1 irreführende Angaben zur Herkunft von Lebensmitteln. Bei Pistazien beispielsweise liegen die Hauptanbaugebiete im Mittleren Osten und in den USA, so dass sich der Pistazienweltmarkt im Wesentlichen in Pistazien dieser beiden Typen unterteilen lässt. Pistazien aus dem Mittleren Osten enthielten in der Vergangenheit immer wieder hohe Konzentrationen der als krebserregend eingestuften Aflatoxine, weshalb sie bei einer Einfuhr in die EU getestet werden müssen und auf dem Weltmarkt schlechtere Preise erzielen. Infolge des Preisunterschieds kommt es gelegentlich zu Umetikettierungen und man möchte zu einer gegebenen Pistazie gerne das Ursprungsland verlässlich ermitteln können. In einer tatsächlichen Anwendung ist I also ein Modell für die Menge der Pistazien und es kommt darauf an, einzelnen Pistazien einen Merkmalsvektor zuzuordnen, der eine verlässliche und relativ täuschungssichere Herkunftsbestimmung ermöglicht. Naheliegende Merkmalskomponenten wie Größe, Gewicht oder Volumen einer Pistazie sind offensichtlich ungeeignet. Mineral- oder Nähstoffgehalt sind vermutlich zu stark abhängig von der Bodenbeschaffenheit, dem Klima während der Wachstumsphase und der Variation unter den über 60 Sorten, um für eine stabile Unterscheidbarkeit zu sorgen.

In der Lebensmittelchemie hat es sich in den letzten Jahren bewährt die Merkmalsvektoren aus der relativen Häufigkeit bestimmter stabiler Isotope von Elementen zusammenzusetzen, also jenen Atomen mit gleich vielen Elektronen und Protonen, aber einer abweichenden Zahl von Neutronen. Zum Beispiel existieren für das Element Sauerstoff (O) mit seinen je 8 Elektronen und Protonen 3 stabile Isotope. Das häufigste natürliche Vorkommen besitzt Sauerstoff-16 (^{16}O) mit 8 Neutronen im Atomkern, wobei die sogenannte Massenzahl 16 die Summe der Neutronen und Protonen im Atomkern bezeichnet. Die anderen beiden stabilen Isotope von Sauerstoff sind ^{17}O und ^{18}O, mit einem beziehungsweise zwei zusätzlichen Neutronen im Atomkern. Mit einem Massenspektrometer lassen die Häufigkeiten der verschiedenen Isotope in einer Pistazie bestimmen. Man normiert und zentriert diese relativen Häufigkeiten und setzt zum Beispiel

$$\delta^{18}O = \frac{\frac{^{18}O}{^{16}O}}{n_O} - 1,$$

wobei das im Zähler stehende gemessene Verhältnis der Probe durch das standardisierte "natürliche" Verhältnis n_0 geteilt wird. In Abbildung 3.6 sind auf der vertikalen Achse die Summe der Isotopenverhältnisse von Kohlenstoff-13 und Stickstoff-15 und auf der horizontalen Achse das Isotopenverhältniss von Sauerstoff-18 aufgetragen.

[2] Eigentlich: Lebensmittel-, Bedarfsgegenstände- und Futtermittelgesetzbuch, `www.gesetze-im-internet.de/lfgb/`

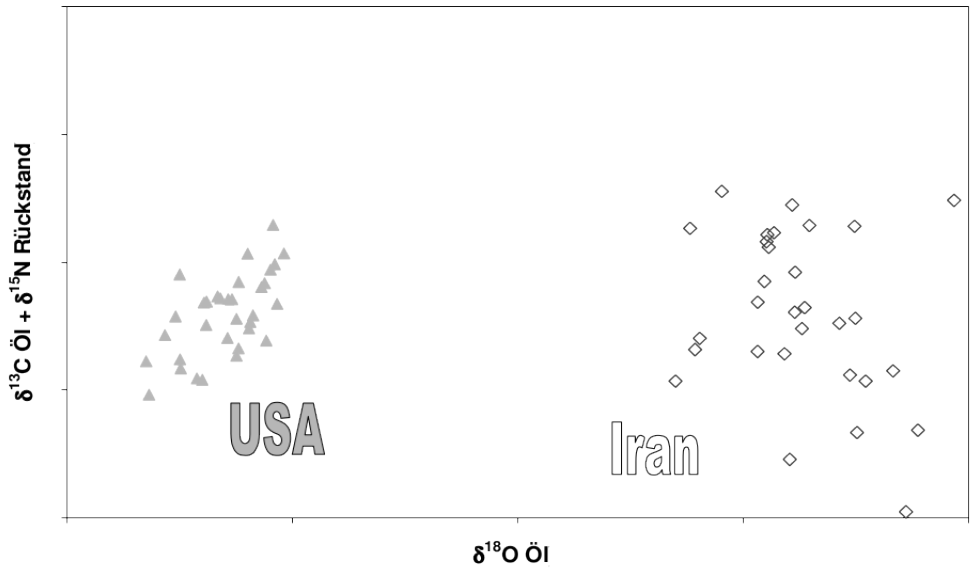

Abbildung 3.6: Isotopenverhältnisse in Stichproben von Pistazien, abgewandelt aus [Hei06] entnommen.

Es zeigt sich, dass Pistazien aus den USA und aus dem Iran sich bezüglich dieser Werte gut unterscheiden lassen. Eine detaillierte Untersuchtung, inklusive einer entsprechenden Diskriminanzanalyse von Pistazien, die ganz wesentlichen die hier vorgestellten Methoden verwendet, findet sich in [Hei06]. Dieses Beispiel verdeutlicht schon bevor wir uns mit der Mathematik des Unterscheidens befassen, dass eine geeignete Konstruktion von Merkmalsvektoren eine entscheidende Bedeutung in diesem Prozess hat.

Definition 3.4.2. Ein Paar von Punktmengen (M_+, M_-), mit $M_+, M_- \subset \mathbb{R}^n$, heißt *linear trennbar*, wenn es ein $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ und ein $r \in \mathbb{R}$ gibt, so dass

$$m_1 \cdot v_1 + \ldots + m_n \cdot v_n \geq r \tag{3.4.1}$$

für alle $m = (m_1, \ldots, m_n)$ aus M_+ gilt und

$$m_1 \cdot v_1 + \ldots + m_n \cdot v_n \leq r$$

für alle $m = (m_1, \ldots, m_n)$ aus M_- gilt, aber nicht für alle $m \in M_+ \cup M_-$

$$m_1 \cdot v_1 + \ldots + m_n \cdot v_n = r$$

gilt. Sind beide Ungleichungen strikt, so heißt das Paar *strikt linear trennbar*.

Sei I eine Menge von Instanzen mit Typen $T = \{+, -\}$ und einer Merkmalszuordnung $x : I \to \mathbb{R}^n$, so dass die Typen linear trennbar sind, womit gemeint ist, dass die Mengen

$$x(t^{-1}(+)) \text{ und } x(t^{-1}(-))$$

linear trennbar sind. Die durch ein v gemäß (3.4.1) definierte Linearform

$$\langle \cdot, v \rangle : \mathbb{R}^n \to \mathbb{R}$$

liefert durch entsprechende Verschiebung und Komposition mit der Vorzeichenabbildung $sgn : \mathbb{R} \to \{+, -\}$ eine Diskriminanzfunktion

$$d := sgn \circ (\langle \cdot, v \rangle - r) : \mathbb{R}^n \to T.$$

Der affine Unterraum $Ker(\langle \cdot, v \rangle - r)$ heißt *Diskriminanzhyperebene* und bildet die Grenze zwischen Elementen vom Typ M_+ und M_-. Mit dem Ziel die Merkmalsvektoren nach ihren Typen linear zu trennen, stellt sich zunächst die Frage, ob eine trennende Hyperebene überhaupt existiert.

Definition 3.4.3. Für eine Menge $M \subset \mathbb{R}^n$ bezeichne $K(M) \subset \mathbb{R}^n$ die konvexe Hülle von M, definiert als

$$K(M) = \left\{ \sum_{i=1}^{m} \lambda_i \cdot m_i \mid \sum_{i=1}^{m} \lambda_i = 1, \lambda_i \geq 0, \quad m_i \in M \right\}.$$

Eine Teilmenge $M \subset \mathbb{R}^n$ heißt konvex, wenn für alle $x, y \in M$ und jedes $\lambda \in [0, 1]$ das Element $\lambda \cdot x + (1 - \lambda) \cdot y$ ein Element von M ist. Für eine konvexe Teilmenge M des $\mathbb{R}^n$ gilt offenbar $M = K(M)$. Eine Menge $P \subset \mathbb{R}^n$ heißt *konvexes Polytop*, falls $P = K(M)$ für eine endliche Menge M gilt.

Satz 3.4.4. *Seien M_0 und M_1 nicht-leere Teilmengen des $\mathbb{R}^n$.*

1. *Zwei konvexe Polytope $P_0 = K(M_0)$ und $P_1 = K(M_1)$ sind genau dann strikt linear trennbar, wenn sie disjunkt sind, also*

   $$P_0 \cap P_1 = \emptyset$$

   gilt.

2. *Sind konvexe Mengen M_0 und M_1 disjunkt, dann sind sie linear trennbar.*

Beweis. Wenn die Polytope P_0 und P_1 von einer durch

$$\{x \in \mathbb{R}^n \mid \langle v, x \rangle = r\}$$

gegeben Hyperebene strikt getrennt werden, dann sind sie disjunkt, da sonst für ein x im Schnitt der Polytope $r < \langle v, x \rangle < r$ gelten würde.

Umgekehrt seien P_0 und P_1 disjunkt. Die Minkowski-Differenz

$$P = P_0 - P_1 = K(M_0 + M_1) = K(\{m_0 - m_1 \mid m_0 \in M_0, m_1 \in M_1\})$$

ist dann ein konvexes Polytop P, welches die Null nicht enthält. Betrachtet man das Bild von P unter der Einbettung

$$i : \mathbb{R}^n \to \mathbb{R}^{n+1}, \quad x \mapsto \begin{pmatrix} x \\ 1 \end{pmatrix},$$

dann liegt folglich $(0,\ldots,0,1)^t$ nicht in $K(\{\binom{m}{1} \mid m \in P_0 - P_1\})$ und daher auch nicht im Kegel

$$Cone\left(\left\{\binom{m}{1} \mid m \in P_0 - P_1\right\}\right) = \left\{\begin{pmatrix} \vdots & & \vdots \\ m_1 & \ldots & m_l \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \cdot x \mid x \geq 0\right\}.$$

Nach Farkas' Lemma 3.3.2 existiert daher ein $y \in \mathbb{R}^{n+1}$, mit

$$y^t \cdot \begin{pmatrix} \vdots & & \vdots \\ m_1 & \ldots & m_l \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \geq 0 \wedge y^t \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} < 0.$$

Dies zeigt, dass $i(P)$ von $i(0)$ strikt linear trennbar ist. Das Urbild einer strikt trennenden Hyperebene unter i gibt dann eine P und 0 strikt trennende Hyperebene, welche dann auch P_0 und P_1 strikt trennt.

Die zweite Aussage des Satzes ist eine Folgerung des Satzes von Hahn-Banach [Wer07, Abschnitt III.2] aus der Funktionalanalysis, die wir hier nur zur Ausweitung der Intuition mit aufgeführt haben. □

Die lineare Trennbarkeit zweier Punktmengen ist offensichtlich äquivalent zur linearen Trennbarkeit der beiden jeweils aufgespannten Polytope, welche nach obigem Satz wiederum zur Disjunktheit der Polytope äquivalent ist. Zu entscheiden, ob zwei Polytope disjunkt sind oder kollidieren, ist eine wichtige Standardaufgabe in der Computergrafik, bei Physik-Simulationen und in der Bewegungsplanung von Robotern. Glücklicherweise hat die mathematische Forschung sehr effiziente Algorithmen, wie zum Beispiel den GJK-Algorithmus [GJK88], für dieses Problem hervorgebracht. Die frage der linearen Trennbarkeit ist also selbst für große Datenmengen effizient zu beantworten und wir können uns dem nächsten Schritt zuwenden. Was tun, wenn, wie in Abbildung 3.6 aus dem Pistazienbeispiel, die Stichprobe linear trennbar ist?

Betrachten wir eine Teilmenge von Instanzen $I' \subset I$, mit linear trennbaren Typen M_+ und M_-, dann stellt sich die Frage, wie aus den potentiell vielen trennenden Diskriminanzhyperebenen $v^\perp$ eine gefunden werden kann, die beste Chancen hat auf die Grundgesamtheit I zu generalisieren. Das Ziel einer trennenden Hyperebene in Beispiel 3.4.1 soll es sein, die Herkunft zukünftiger Pistazienproben, allein anhand der Laborwerte, dadurch zu bestimmen, ob sie, bildlich gesprochen, links oder rechts von jener trennenden Hyperebene liegen. Abbildung 3.7 verdeutlicht, dass es für die Lage der Hyperebene einigen Spielraum gibt.

Natürlich ist auch die Frage berechtigt, warum überhaupt eine Hyperebene, bei einem zweidimensionalen Merkmalsraum also eine Gerade, zum Trennen der Instanzen gewählt werden soll. Dafür sind zwei Gründe wesentlich. Zum einen lässt sich linear besonders gut rechnen. Das klingt banal, ist aber ein gewichtiger Grund. Lineare Konstruktionen haben ein sehr gut etablierten theoretischen Rahmen und sind somit nicht nur numerisch zu untersuchen, sondern auch für qualitative Erkenntnisse gut geeignete. Aber auch auf numerischer Ebene ist das

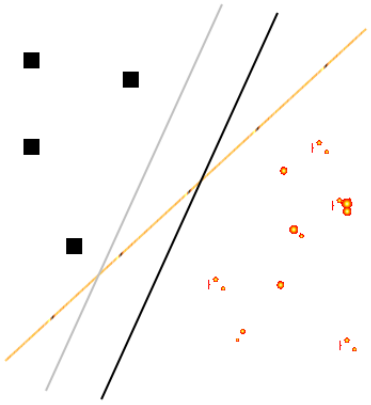

Abbildung 3.7: Viele Diskriminanzhyperebenen.

'einfache' lineare Rechnen ein Vorteil, da die Diskriminanzanalyse häufig mit großen Datenmengen konfrontiert ist und eine auf den ersten Blick nicht sehr aufwendige Rechnung dann den Unterschied zwischen einer Analyse beinahe in Echtzeit, einer machbaren Analyse und einer mit heutiger Rechenkraft und Speichergröße nicht zugänglichen Analyse bedeuten kann.

Zum anderen gibt es eine stochastische Rechtfertigung. Die Stochastik kombiniert als mathematische Disziplin die Methoden der Wahrscheinlichkeitstheorie und der Statistik zu einer Wissenschaft des Schätzens. Da es in der Diskriminanzanalyse darum geht, den Typ einer Instanz qualifiziert zu schätzen, gibt es natürlich auch stochastische Aussagen darüber, was wann wahrscheinlich ein guter Schätzer ist. Wer dem folgenden Exkurs über eine stochastische Argumentation nachgehen möchte, wird Grundlagen der Wahrscheinlichkeitstheorie benötigen. Für Leser die stochastische Vorkenntnisse zunächst nicht voraussetzen möchten, lautet die Argumentation zusammengefasst, dass unter der stark einschränkenden aber üblichen Annahme, die Typen seien mit einer ähnlichen Streuung unter einer Gaußschen Normalverteilung im Merkmalsraum verteilt, die wahrscheinlichste Zuordnung durch eine Hyperebene, also eine lineare Trennung, realisiert wird.

Spezialfall eines linearen naiven Bayes-Klassifikators

- Angenommen die Instanzen seien vom Typ + oder − und im Merkmalsraum normalverteilt mit gleich gleicher Kovarianzmatrix Σ, d.h.

$$P(X \mid t(X) = +) \text{ ist } \mathcal{N}(\mu_0, \Sigma)\text{-verteilt und}$$
$$P(X \mid t(X) = -) \text{ ist } \mathcal{N}(\mu_1, \Sigma)\text{-verteilt,}$$

  wobei die Zufallsvariable X der Einfachheit halber mit gleicher Wahrscheinlichkeit Instanzen vom Typ + wie vom Typ − sei.

- Man bedient sich nun der Methode der Hypothesentests und testet die *Nullhypothese* 'Instanz mit Merkmalsvektor x ist vom Typ +' gegen die *Alternativhypothese* 'Instanz mit Merkmalsvektor x ist vom Typ −'.

- Der *Satz von Bayes*,

$$P(t(X) = + \mid X = x) = \frac{P(X = x \mid t(X) = +) \cdot P(t(X) = +)}{P(X = x)}$$

  verwandelt eine Abschätzung

$$P(t(X) = + \mid X = x) > P(t(X) = - \mid X = x)$$

  der Typ-Wahrscheinlichkeiten einer Realisierung x um, in einen Likelihood-Quotienten-Test.

- Der *Likelihood-Quotient* liefert nach dem *Lemma von Neyman-Pearson* [Geo04, S:???] einen Hypothesentest mit optimaler Trennschärfe. Durch die Annahme der gleichen Kovarianz Σ entpuppt sich das durch den Likelihood-Quotienten gegebene Trenn-Kriterium als eine Hyperebene:

$$\begin{aligned}
&\frac{\exp(-\frac{1}{2}(x-\mu_1)^t \cdot \Sigma \cdot (x-\mu_1))}{\exp(-\frac{1}{2}(x-\mu_0)^t \cdot \Sigma \cdot (x-\mu_0))} && < t \\
\Leftrightarrow &\exp((x-\mu_0)^t \cdot \Sigma \cdot (x-\mu_0) - (x-\mu_1)^t \cdot \Sigma \cdot (x-\mu_1)) && < t' \\
\Leftrightarrow &(x-\mu_0)^t \cdot \Sigma \cdot (x-\mu_0) - (x-\mu_1)^t \cdot \Sigma \cdot (x-\mu_1) && < t'' \\
\Leftrightarrow &(2\mu_0^t \cdot \Sigma - 2\mu_1^t \cdot \Sigma) \cdot x + (\mu_0^t \Sigma \mu_0 - \mu_1^t \Sigma \mu_1) && < t'' \\
\Leftrightarrow &((\mu_0 - \mu_1)^t \cdot \Sigma) \cdot x && < t''',
\end{aligned}$$

  wobei wir eine genaue Betrachtung des Schwellenwertes t vernachlässigen, um nur den Charakter einer geeigneten Konstanten zu betonen. Eine präzise Wahl von t wäre davon abhängig, mit welcher Wahrscheinlichkeit man einen *Fehler erster, bzw. zweiter Art* ausschließen möchte.

In der Praxis zeigt sich häufig, dass Verfahren, die unter bestimmten wahrscheinlichkeitstheoretischen Voraussetzungen optimal funktionieren, auch dann noch sehr gut brauchbar sind, wenn diese Voraussetzungen nicht erfüllt sind [Ris01]. Wir nehmen dies als Ermutigung, uns losgelöst von Annahmen über hin-

tergründige Wahrscheinlichkeitsverteilungen nach einer sinnvoll im Merkmalsraum liegenden Hyperebene zu suchen. Liegt eine trennende Hyperebene nahe an den Instanzen $i \in I'$ von Typ M_+, dann steigt das Risiko eine neue Instanz fälschlich dem Typen M_- zuzuordnen, obwohl sie eigentlich vom Typ M_+ und das umgekehrte Risiko steigt, je näher die Hyperebene an den Instanzen vom Typ M_- liegt. Daher ist es im Allgemeinen, wenn beide Fehlzuordnungen gleich schädlich sind, sinnvoll, die trennende Hyperebene möglichst gleich weit entfernt von den Lerninstanzen aus I' zu konstruieren.

Wie berechnet man konstruktiv ein passendes v: Stützvektoren definieren und QP herleiten ...

Stützvektormaschinen (SVM)

Was tun, wenn die Instanzen nicht linear trennbar sind...

Abbildung 3.8: Nicht linear trennbare Instanzen und eine linear trennbare Einbettung in einen höher-dimensionalen Raum.

Etwas aus der Originalarbeit [CV95] lernen und Beispiel mit trennendem Kreis oder Ellipse durchführen.

Den Kernel-Trick für das Ausrechnen von Skalarprodukten im niedrig-dimensionalen Raum erklären.

Als Motivation möchte ich u.a. gefälschte Likes und Follower in sozialen Netzwerken nehmen und mit [WWZZ14] verknüpft an maschinelles Lernen anschließen und Sützvektormaschninen und lineare Diskriminanzfunktionen darstellen.

3.5 Input-Output Analyse

Ersetzt den nach hinten gerückten Abschnitt 4.2 über Markov-Ketten. Ich finde dieses Thema zu einem frühen Zeitpunkt angemessener, da Einstiegsbetrachtungen gut an eine Behandlung des Themas anknüpfen können, wie ich sie in einem Buch aus der Schroedel-Reihe Neue Wege Mathematik gefunden habe (NWM 11/12, Nds, S.369). Interessante Anhaltspunkte könnte auch eine kritische Analyse des Leontief-Kapitel aus [Roe88] geben. Klassiker: [Hup90, Kapitel IV.2].

2Bibtex

3.6 Spieltheorie

Der Begriff Spieltheorie erlebt in den letzten Jahren in Deutschland eine steigende Prominenz, wie eine kurze Suche bei FAZ, Zeit oder Google News [3] belegen dürfte. In der Spieltheorie geht es um die Analyse von Entscheidungssituationen, an denen mehrere Teilnehmer, mit unter Umständen konkurrierenden Absichten, beteiligt sein können. Wie man sieht ist der Begriff des Spiels hier also weit gefasst und umschreibt viel mehr als nur Gesellschaftsspiele. Viele Nobelpreise[4]

[3] Copyrights?

[4] Genauer: Von der schwedischen Reichsbank in Erinnerung an Alfred Nobel gestifteter Preis für Wirtschaftswissenschaften.

sind für Arbeiten mit spieltheoretischem Bezug verliehen worden und zahlreiche Wissenschaften greifen auf Ideen der Spieltheorie zurück. Dennoch hat nimmt die Spieltheorie in diesem Buch eine Sonderrolle dahingehend ein, dass Anwendungen der Spieltheorie stets einen theoretischen und modellierenden Charakter haben. So schreibt zum Beispiel Reinhard Zintl in seinem Festvortrag [Zin95] am Max-Planck-Institut für Gesellschaftsforschung „Sozialwissenschaftliche Anwendungen der Spieltheorie bestehen darin, die Spieltheorie als ein Instrument der Theoriebildung (...) einzusetzen(...). Die Spieltheorie ist 'immer' reine Spieltheorie.“. Die Auseinandersetzung mit der Frage, ob die Spieltheorie denn in irgendeinem strengen Sinne Anwendungen produziert, soll und kann aber nicht in diesem Buch behandelt werden. Wir wollen uns vielmehr auf die eleganten Ideen stürzen, die diese Disziplin hervorgebracht hat und auf den Beitrag, den die Lineare Algebra dazu leisten kann. Im Anschluss an die Diskussion von Nash-Gleichgewichten wollen wir dennoch eine/ein Paar interessante Anwendungen erleutern, ohne uns dabei auf die Strenge des Anwendungsbegriffes zu konzentrieren.

Definition 3.6.1. Spielbegriff 2-Personen Spiele Null-/Kostantsummenspiele

Diskussion kooperativ/Nichtkooperativ

3.6.1 Nash-Gleichgewichte

Als lineares Komplementaritätsproblem [Sch08].

Wir wollen Nash-Gleichgewichte auf das allgemeinere lineare Komplementaritätsproblem zurückführen und dann zeigen, wie diese gelöst werden kann.

Definition 3.6.2. Seien $M \in \mathrm{M}(n \times n, \mathbb{R})$ und $q \in \mathbb{R}^n$. Das lineare Komplementaritätsproblem zu (M, q) wird gelöst durch die Bestimmung der Menge $LCP(M, q)$ aller Vektoren $x, y \in \mathbb{R}^n$, mit den Eigenschaften

$$y = Mx + q, \tag{3.6.1}$$

$$0 \leq x, y \text{ und} \tag{3.6.2}$$

$$0 =< x, y > \tag{3.6.3}$$

Satz 3.6.3. *Einen Satz formulieren, der Lösungen eines bestimmten LCP mit Nashgleichgewichten identifiziert.*

Beweis. □

Mit Referenz auf das relativ gut lesbare Paper von Nash hier einen Hinweis dazu geben, dass Nash eigentlicher Beweis für die Existenz von Nash-Gleichgewichten nicht konstruktiv ist, sondern topologische Methoden benutzt. Kann man zu dem Fixpunktsatz topologisch sagen?

Dann hier den Lemke-Howson Algorithmus einführen. Inwiefern ist dieser an eine homotopietheoretische Idee angelehnt? (Siehe 'homotopy' Bemerkung auf der englischen Wikipedia.

Korollar 3.6.4. *Existenz von Nash-Gleichgewichten*

3.6.2 Eine 'Anwendung' der Spieltheorie

Welche Themen könnte man hier besprechen? Sichtweise als Spieler? Sichtweise als Mechanismus-Designer. Auktionstheorie?

Kapitel 4

Eigenwerte

Wie allgemein gibt es Resultate über Singulärwertzerlegungen? Kann ich in diesem Kaptiel schon darauf eingehen oder brauche ich dafür noch Normen und lande daher in Kapitel 5? Wahrscheinlich wird es schließlich betonenswert, dass Eigenwerte in viel mehr Anwendungen einen wichtige Rolle spielen, als es beim Blick auf die Länge dieses Kapitels zunächst den Anschein haben könnte, aber dass diese Anwendungen oft im Zusammenspiel mit Normen daher kommen und daher, sofern überhaupt, erst in Kapitel 5 präsentiert werden.

Für die Abschnitte über Stabilitätslagen und Eigenschwingungen könnten interessante Inspiration aus Dankert & Dankerts Buch [DD04] über Technische Mechanik erhalten. (Einblick besorgen!)

4.1 Entscheidungstheorie

Wir wollen im Folgenden Methoden der Linearen Algebra einsetzten um mit dem sogenannten *Analytic Hirarchy Process* [Saa90] ein Verfahren aus der Entscheidungstheorie[1] zu ergründen. Ohne hier den philosophischen und psychologischen Aspekten der Entscheidungstheorie gerecht werden zu können, wollen wir, eben in einem mathematischen Sinne auf Rationalität beschränkt, einen Einstieg in das Gebiet wagen.

Um eine möglichst objektive Betrachtung in der folgenden Diskussion zu erleichtern, fokussieren wir nicht auf eine der vielen ernsten wirtschaftlichen oder emotional aufgeladenen Entscheidungen, die zweifellos ebenfalls mit den hier dargestellten Techniken zu analysieren sind, sondern auf Beispielprobleme wie

- das leidliche Problem im Restaurant ein Gericht von der Karte zu wählen, dass lecker ist, den Hunger stillt, eine reichhaltige Versorgung mit den wichtigsten Nährstoffen sichert und das Portemonnaie nicht überstrapaziert ohne dabei Einschränkungen in Bezug auf Allergene, Art oder Menge tierischer Inhaltsstoffe, Kohlehydrate oder Kalorien zu verletzen.

- Das Problem als Datingportal seinen Nutzern eine Bandbreite an mögli-

[1] Oder vielleicht besser, aus der Theorie der rationalen Entscheidungen, *engl. rational choice theory.*

chen Dates vorzuschlagen[2], die in gewissen Merkmalen möglichst ähnliche Präferenzordnungen erzeugt haben ('Gleich und Gleich gesellt sich gern'-Merkmale) und in anderen ('Gegensätze ziehen sich an'-) Merkmalen gerade differente Muster aufweisen und dabei die oft leidenschaftslosen Kriterien, wie die Lieblingsfarbe, so zu gewichten, dass ein glücklicher gemeinsamer Lebensentwurf entspringt oder zumindest eine Zufriedenheit mit der Funktionsweise des Portals erzeugt wird. Oder

- das Problem als Onlineshop über die Produktempfehlungen für den jeweiligen Benutzer zu entscheiden, so dass in Konsistenz mit den bisherigen Käufen des Kunden eine Affinität zum jeweiligen Produkt zu erwarten ist, für den Shop-Betreiber die Umsatzrendite gut ist, die Bewertungen der anderen User eine gute Zufriedenheit mit dem Produkt erwarten lassen und dem Kunden nach dem Kühlschrankkauf in der Vorwoche nicht leichtfertig ein wiederholtes Interesse an Kühlschränken unterstellt wird.

Diese Entscheidungssituationen haben gemeinsam, dass die auszuführende Entscheidungshandlung klar definiert ist, die Menge der Alternativen endlich ist und die für die Entscheidung zu berücksichtigenden Kriterien endlich aber vielzählig sind. Außerdem wird deutlich, dass häufig unklar ist, wie die verschiedenen Kriterien für die Entscheidungsfindung zu gewichten sind und wie die Alternativen in Bezug auf die Erfüllung der einzelnen Kriterien gemessen werden können. Letzteres liegt zum einen daran, dass für viele Kriterien keine absolute Bewertungsskala sinnvoll existieren kann. Lassen sich Zucker- oder Fettgehalt gut in Gramm messen, so ist das Kriterium lecker zu sein viel zu subjektiv und zu unpräzise definiert. Zum anderen lassen bestimmte Situationen oft keine absoluten Bewertungen zu, selbst wenn diese theoretisch möglich wären. Wer einen Stein in der Hand hält wird ohne weiteres kaum das exakte Gewicht benennen können. Wer aber einen zweiten Stein erhält, kann häufig beurteilen, welcher der beiden Steine schwerer ist als der Andere. Wir kommen nun endlich zum Analytic Hirarchy Process (AHP), der sehr gut in solchen Situationen anwendbar ist.

Den Kern des AHP bilden paarweise Vergleiche und eine Methode, um aus diesen paarweisen Vergleichen eine Gesamtbewertung zu konstruieren. Stellen wir uns zum Einstieg einen Experten für intuitive Gewichtsvergleiche vor, der die Aufgabe hat das Gewicht von drei Steinen zu bewerten. Dazu führt er zwei bis drei Vergleiche durch. Anfangs nimmt er Stein 1 und Stein 2 in die linke bzw. rechte Hand und schätzt das Gewicht von Stein 2 als doppelt so schwer wie das Gewicht von Stein 1 ein. Dann legt er Stein 1 ab und nimmt stattdessen Stein 3 auf. Dessen Gewicht schätzt er als dreifaches Gewicht von Stein 2 ein. Diese beiden Bewertungen können in einer Matrixschreibweise als

$$\begin{pmatrix} 1 & & \\ 2 & 1 & \\ & 3 & 1 \end{pmatrix}$$

festgehalten werden und zu einer

1. *reziproken*, d.h. einer Matrix $A = (a_{ij}) \in M(n \times n, \mathbb{R})$ mit $a_{ij} = a_{ji}^{-1}$ für alle $i, j \in \{1, \ldots, n\}$, und

[2] Vergleiche dazu den humorvollen inversen Ansatz von Amy Webb in *How I hacked online dating.* `www.ted.com/talks/amy_webb_how_i_hacked_online_dating.html`

2. *konsistenten*, d.h. für alle $i, j, k \in \{1, \ldots, n\}$ gilt $a_{ik} = a_{ij} \cdot a_{jk}$, Matrix ergänzt werden:

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{6} \\ 2 & 1 & \frac{1}{3} \\ 6 & 3 & 1 \end{pmatrix}$$

Der Eintrag a_{ij} in dieser Bewertungsmatrix besagt also, dass die Merkmalsausprägung der i-ten Alternative a_{ij}-mal so stark bewertet wird, wie die der j-ten Alternative. Könnten wir von einer absoluten numerischen Bewertung $b_1, \ldots, b_n \in \mathbb{R} \setminus 0$ von n Alternativen ausgehen, dann würden die Verhältnisse $\frac{b_i}{b_j}$ zu einer Bewertungsmatrix

$$B = \begin{pmatrix} \frac{b_1}{b_1} & \frac{b_1}{b_2} & \cdots & \frac{b_1}{b_n} \\ \frac{b_2}{b_1} & \frac{b_2}{b_2} & \cdots & \frac{b_2}{b_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{b_n}{b_1} & \frac{b_n}{b_2} & \cdots & \frac{b_n}{b_n} \end{pmatrix}$$

führen. Die Spalten von B sind offenkundig alle skalare Vielfache voneinander und daher gilt rang $B = 1$. Folglich hat B höchstens einen von Null verschiedenen Eigenwert, der sich in

$$B \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = n \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

gleich mit einem interessanten Eigenvektor präsentiert. Alle Eigenvektoren zum Eigenwert n der Matrix B sind also Skalierungen des aus den absoluten Bewertungen zusammengesetzten Vektors. Nun wäre noch keine Lineare Algebra nötig gewesen, um zu erkennen, dass unser erdachter Experte für intuitive Gewichtsvergleiche das exakte Gewicht aller Steine bestimmen könnte, wenn er das exakte Gewicht eines einzigen Steines kennt. Interessanter ist, dass diese Methode es auch ohne ein exaktes Gewicht, aber dafür unter Kenntnis zusätzlicher statistischer Steuungsmerkmale erlaubt, eine fundierte Schätzung der Einzelgewichte abzugeben.

In vielen Situationen werden paarweise Vergleiche zu Bewertungsmatrizen führen, die nicht länger konsistent sind. Wie oben bereits erwähnt wurde, eignet sich AHP gut, um auch qualitative Merkmalsausprägungen zu berücksichtigen. Dazu werden von Saaty Bewertungsskalen von 1 für eine gleich starke Ausprägung, bis 9 für eine viel stärkere Ausprägung empfohlen [Saa90, p. 15]. Es ist klar, dass schon sehr kleine Bewertungsmatrizen bei unterscheidbaren Alternativen dann nicht mehr konsistent sein können.

Beispiel 4.1.1. Die lokale Uni-Mensa bietet täglich drei Gerichte an. Aus der tagesaktuellen Auswahl sollen Milchreis, Gemüseeintopf und Veggie-Burger bezüglich des Kriteriums 'lecker' bewertet werden. Da der Autor eine Abneigung gegen Milchreis besitzt, bewertet er den Gemüseeintopf als deutlich leckerer (5) als den Milchreis und findet den Veggie-Burger noch mal ein gutes Stück leckerer (3) als den Gemüseeintopf:

$$\begin{pmatrix} 1 & \frac{1}{5} & \\ 5 & 1 & \frac{1}{3} \\ & 3 & 1 \end{pmatrix}$$

Da sich diese Matrix in der vorgeschlagenen Skala nicht mehr konsistent bewerten lässt, wählt der Autor einfach die maximale Merkmalsausprägung und landet, obwohl er ein großer Mensa-Experte ist, bei der inkonsistenten Bewertungsmatrix

$$\begin{pmatrix} 1 & \frac{1}{5} & \frac{1}{9} \\ 5 & 1 & \frac{1}{3} \\ 9 & 3 & 1 \end{pmatrix} \tag{4.1.1}$$

In komplexeren Situationen tritt Inkonsistenz der Bewertungsmatrix aus viel inhärenteren Gründen und unabhängig vom Skalenintervall auf. Es wird sich zeigen, dass Inkonsistenz ein messbares und in gewissem Rahmen handhabbares Problem darstellt.

Zur weiteren Untersuchung führen wir nun den bedeutenden Satz von Perron-Frobenius[3] ein. Eine Matrix oder einen Vektor heißt *positiv*, bzw. *nicht-negativ*, wenn alle Einträge positiv, bzw. nicht-negativ, sind.

Satz 4.1.2 (Perron-Frobenius)**.** *Sei A eine nicht-negative reelle Matrix, dann ist der Spektralradius von A ein Eigenwert von A und es gibt einen nicht-negativen Eigenvektor. Gibt es darüber hinaus ein $k \in \mathbb{N}$, so dass A^k positiv ist, dann wird der Eigenraum zum Spektralradius von einem positiven Eigenvektor erzeugt.*

Beweis. In modernen Lehrbüchern wie in [HW06, Hauptsatz 6.3.3] findet sich häufig ein Beweis des Satzes nach Wielandt. Perrons Originalarbeit [Per07] ist schon früh im Studium gut lesbar, doch sein Beweis benutzt auch analytische Argumente. Wer früh den Reiz verspürt eine Forschungsarbeit zu lesen, dem sei daher empfohlen Wielandts digital erhältliche Arbeit [Wie50] im Original zu Rate zu ziehen. □

Der *Spektralradius* $\rho(A)$ einer Matrix $A \in \mathrm{Mat}(n \times n, \mathbb{C})$ ist definiert als

$$\rho(A) := \max\{|\lambda| \mid \lambda \in \mathbb{C} : \det(A - \lambda E_n) = 0\}.$$

Die Bewertungsmatrix (4.1.1) aus dem Mensa-Beispiel hat einen Spektralradius von etwa $3,029$ und tatsächlich ist gilt für den Eigenwert λ_{max} einer reziproken Matrix aus dem Satz 4.1.2:

Lemma 4.1.3. *Sei $A \in \mathrm{Mat}(n \times n, \mathbb{R})$ eine positive reziproke Matrix. Dann gilt $\lambda_{max} \geq n$, wobei Gleichheit genau dann eintritt, wenn A konsistent ist.*

Beweis. Nach Satz 4.1.2 existiert ein positiver Eigenvektor v von A zum Eigenwert λ_{max}. Definiere

$$b_{ij} := a_{ij} \cdot \frac{v_j}{v_i}.$$

Dann ist die Matrix $(b_{ij})_{i,j}$ reziprok, es gilt $b_{ij} = b_{ji}^{-1}$ und

$$\sum_{j=1}^{n} b_{ij} = \frac{\sum_{j=1}^{n} a_{ij} v_j}{v_i} = \frac{(Av)_i}{v_i} = \frac{\lambda_{max} \cdot v_i}{v_i} = \lambda_{max}.$$

[3] Frobenius verallgemeinerte in [Fro12] den Satz von Perron [Per07, S. 261] mit eben jener Aussage von positiven auf nicht-negative Matrizen.

Daher ist

$$\begin{aligned} n \cdot \lambda_{max} &= \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \\ &= \sum_{i=1}^{n} a_{ii} + \sum_{i>j} b_{ij} + b_{ji} \\ &= n + \sum_{i>j} b_{ij} + b_{ij}^{-1} \end{aligned}$$

Und da für $x > 0$ die Ungleichung $x + \frac{1}{x} \geq 2$ äquivalent ist zu der offensichtlich gültigen Ungleichung $(x-1)^2 \geq 0$, ist die Ungleichung wahr und es gilt genau dann Gleichheit, wenn $x = 1$ ist. Daher können wir die Summe weiter abschätzen durch

$$\begin{aligned} &\geq n + \frac{n^2 - n}{2} \cdot 2 \\ &= n^2. \end{aligned}$$

Also gilt $\lambda_{max} \geq n$ mit Gleichheit genau dann, wenn $b_{ij} = 1$ für alle $i, j = 1, \ldots, n$, bzw. wenn $a_{ij} = \frac{v_i}{v_j}$ reziprok ist. □

Demnach kontrolliert also der Wert $\lambda_{max} - n \geq 0$ einer positiven reziproken Matrix die Abweichung von einer konsistenten Matrix. Man definiert den *Konsistenzindex* von A als

$$\mu(A) := \frac{\lambda_{max} - n}{n - 1}.$$

Aufgabe 4.1.4. Sei A, wie im Text, eine positive reziproke Matrix und seien $\lambda_{max}, \lambda_2, \ldots, \lambda_n$ die Eigenwerte von A. Man zeige, dass $\mu(A)$ das arithmetische Mittel der $n-1$ kleinsten Eigenwerte $\lambda_2, \ldots, \lambda_n$ von A ist.

Man betrachtet den Konsistenzindex, um die Qualität einer Bewertung beurteilen zu können. Da man sich, wie in Beispiel 4.1.1 erwähnt, nicht auf konsistente Bewertungsmatrizen beschränken kann, muss man ein gewisses Maß an Inkonsistenz zulassen. Hierzu kann man sich an der mittleren Inkonsistenz einer zufälligen reziproken Matrix gleicher Größe orientieren (vgl. [Saa08, p. 265]), worauf an dieser Stelle aber nicht weiter eingegangen werden soll.

IN EINEM ABSATZ HIER die Störungstheorie von einfachen Eigenwerten thematisieren und deren Stabilität in Bezug zu inkonsistenten reziproken Matrizen setzen. todo

Auf die bis hierhin beschriebene Weise lässt sich also eine Beurteilung von verschiedenen Alternativen bezüglich *eines* Kriteriums ermitteln. Der AHP besteht, als Hilfsmittel zur mulikriteriellen Entscheidungsfindung, nun aus einer baumartig geschachtelten Wiederholung der oben beschriebenen Eigenwert-Technik. Dabei steht das Ziel in höchsten Hierarchieebene und die Handlungsalternativen in der untersten Ebene. In den mittleren Ebene erstellt man eine Baumartige Abhängigkeitsstruktur aus Kriterien und Unterkriterien. Bei der Gestaltung der unterschiedlichen Ebenen ist darauf zu achten, dass sich alle Kriterien der einer Ebene, mit gleichem Anschlussknoten in der übergeordneten

Ebene, miteinander in Bezug auf ihre Relevant für den gemeinsamen übergeordneten Knoten vergleichend bewerten lassen. Zu jedem Knoten aus den oberen Ebenen wird dann durch paarweise vergleiche eine Bewertungsmatrix gebildet und ein Eigenvektor zum Spektralradius (vgl. Satz 4.1.2) bestimmt. Diesen Eigenvektor skaliert, bzw. wählt, man in der Regel so, dass er bezüglich seiner 1-Norm normiert ist, also die Summe aller seiner Einträge eins ist, um eine gleichmäßige Aufteilung der Gesamtrelevanz eines Knotens in seine Nachfolger zu ermitteln.

Beispiel 4.1.5. Für das erste Qualifikationsspiel der Saison möchte der Trainer einer Fußballnationalmannschaft einen 'optimalen' Torwart bestimmen. Der kommende Gegner wird als etwas schwächer eingeschätzt, weshalb dem Trainer im Torwartspiel ein modernes Stellungsspiel und ein starkes Verhalten in 1-gegen-1 Konter-Situationen wichtig ist. Neben dem Torwartspiel ist dem Trainer wichtig, dass Torhüter über ein aktuell gutes Selbstvertrauen verfügt und dies für den Gegner möglichst auch sichtbar ist. Sein Bauchgefühl möchte der Trainer durch den AHP absichern und identifiziert 'Ausstrahlung/Selbstvertrauen' und 'Torwartspiel' als Hauptkriterien seiner Entscheidung. 'Modernes' Stellungsspiel und '1-gegen-1'-Stärke betrachtet er als Unterkriterien des Torwartspiels. Als Torhüter hat er die Spieler 'Alter', 'Falke' und 'Holzfäller' berufen.

Abbildung 4.1: Torhüterwahl mittels AHP

Das Kriterium 'Torwartspiel' ist dem Trainer für die optimale Torhüterwahl gegen den kommenden Gegener ein ganzes Stück wichtiger als 'Ausstrahlung/Selbstvertrauen'. Für das Torwartspiel findet er das moderne Stellungsspiel etwas wichtiger als die Stärke in den ohnehin schwierigen 1-gegen-1 Situationen:

$$\begin{array}{cc} \text{Optimaler TW} & \text{Torwartspiel} \\ \begin{pmatrix} 1 & 5 \\ \frac{1}{5} & 1 \end{pmatrix} & \begin{pmatrix} 1 & 3 \\ \frac{1}{3} & 1 \end{pmatrix} \end{array}$$

Die Kriterien der jeweils untersten Ebene werden nun herangezogen, um die Torhüter an Hand von ihnen zu vergleichen.

$$\begin{array}{ccc} \text{Modernes Stell.} & \text{1-gegen-1 Stärke} & \text{Ausstrahlung} \\ \begin{pmatrix} 1 & 5 & 3 \\ \frac{1}{5} & 1 & 3 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix} & \begin{pmatrix} 1 & 3 & 1 \\ \frac{1}{3} & 1 & \frac{1}{3} \\ 1 & 3 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 9 & 3 \\ \frac{1}{9} & 1 & \frac{1}{3} \\ \frac{1}{3} & 3 & 1 \end{pmatrix} \end{array}$$

Daraus ergibt sich durch ein gewichtetes Aufsummieren der Kriterienbewertungen die Gesamtbewertung

$$\begin{aligned} \text{Bewertung(Alter)} &= \frac{5}{6} \cdot \left(\frac{3}{4} \cdot 0,65 + \frac{1}{4} \cdot 0,43 \right) + \frac{1}{6} \cdot (0,69) && = 0,61 \\ \text{Bewertung(Falke)} &= \frac{5}{6} \cdot \left(\frac{3}{4} \cdot 0,22 + \frac{1}{4} \cdot 0,14 \right) + \frac{1}{6} \cdot (0,08) && = 0,18 \\ \text{Bewertung(Holzfäller)} &= \frac{5}{6} \cdot \left(\frac{3}{4} \cdot 0,13 + \frac{1}{4} \cdot 0,43 \right) + \frac{1}{6} \cdot 0,23 && = 0,21 \end{aligned}$$

woraufhin der Trainer Alter in die Startaufstellung bringt, Holzfäller auf die Bank nimmt und Falke den Platz auf der Tribüne zugewiesen bekommt.

Zusammengefasst beschreibt sich das Vorgehen im AHP wie folgt:

1. Die Problemstellung definieren und die möglichen Alternativen herausfinden.
2. Die relevanten Kriterien zur Entscheidungsfindung bestimmen und hierarchisch ordnen.
3. Bewertungsmatrizen aus paarweisen Vergleichen für jeden Knoten im Baum erzeugen und die Bewertung nach der obigen Eigenwertmethode errechnen.
4. Gesamtbewertung durch entsprechend gewichtetes Aufsummieren aller Teilbewertung errechnen.

In natürlicheren Situationen sind in der Regel die Anzahl der Kriterien oder Alternativen größer als in den Beispielen 4.1.1 und 4.1.5. Um in diesen Fällen einen Zugang zum Eigenraum des dominanten Eigenwertes der Bewertungsmatrix zu haben, ist es hilfreich die *Potenzmethode* zu kennen. Diese ist ein iteratives Verfahren zur Bestimmung eines Eigenvektors zum betragsgrößten Eigenwert λ_1 einer Matrix $A \in \mathrm{Mat}(n \times n, \mathbb{C})$, falls dieser Eigenwert dominant ist, in dem Sinne, dass

$$|\lambda_1| > |\lambda_2|, \ldots, |\lambda_r|$$

für alle anderen Eigenwerte $\lambda_2, \ldots, \lambda_r$ von A gilt. Sei x_0 ein geeigneter[4] Startvektor und

$$x_{i+1} := \frac{1}{|A \cdot x_i|} \cdot A x_i.$$

Dann ist $(x_k)_{k\in\mathbb{N}}$ eine Folge von normierten Vektoren, die durch iteriertes Anwenden von A gegeben ist, genauer

$$x_k = \frac{1}{|A^k x_0|} \cdot A^k x_0.$$

Die Idee der Potenzmethode ist, dass die Eigenraumkomponente des dominanten Eigenwertes durch iteriertes Anwenden von A auf x_0 schneller wächst als der Rest von x_0 und daher durch iteriertes Normieren alle anderen Komponenten verschwinden. Präzise gilt:

Satz 4.1.6. *Sei A eine positive Matrix und $(x_k)_{k\in\mathbb{N}}$ wie oben. Dann konvergiert die Folge $(x_k)_{k\in\mathbb{N}}$ gegen einen Eigenvektor x zum dominanten Eigenwert λ_1 von A und die Folge der* Rayleigh-Quotienten

$$R(A, x_k) := \frac{< x_k, Ax_k >}{< x_k, x_k >} \tag{4.1.2}$$

konvergiert gegen λ_1.

[4] Es zeigt sich im folgenden Beweis, dass ein Startvektor geeignet ist, wenn er eine nicht verschwindende Eigenraumkomponente zum dominanten Eigenwert hat, wenn also sein Anteil in $\mathrm{Eig}(A, \lambda_1)$ nicht null ist. Das ist nach Satz 4.1.2 für positive Matrizen fast-sicher der Fall.

Beweis. Nach [Fis05, Satz XXX?] gibt es eine Basis von $\mathbb{C}^n$ aus Hauptvektoren von A, wenn A als komplexe Matrix aufgefasst wird. Daher lässt sich x_0 schreiben also

$$x_0 = h_1 + \ldots + h_k,$$

wobei h_i ein Hauptvektor der Stufe r_i zum Eigenwert λ_i ist, für $i = 1, \ldots, k$ und λ_i die paarweise verschiedenen Eigenwerte von A sind. Um das Konvergenzverhalten von $(x_s)_{s\in\mathbb{N}}$ zu verstehen, betrachten wir zunächst unskaliert die Vektoren

$$\begin{aligned} A^s x_0 = A^s(h_1 + \ldots + h_k) &= \sum_{i=1}^{k} (A - \lambda_i E_n + \lambda_i E_n)^s h_i \\ &= \sum_{i=1}^{k}\sum_{j=0}^{s} \binom{s}{j} \lambda_i^{s-j} (A-\lambda_i)^j h_i, \end{aligned}$$

wobei $(A-\lambda_i E_n)^j h_i = \begin{cases} v_i & j = r_i - 1 \\ 0 & j \geq r_i \end{cases}$, für einen Eigenvektor v_i zum Eigenwert λ_i. Also bricht für große s die innere Summe stets vorzeitig ab:

$$\begin{aligned} A^s x_0 &\simeq \sum_{i=1}^{k}\sum_{j=0}^{r_i-1} \binom{s}{j} \lambda_i^{s-j} (A-\lambda_i)^j h_i \\ &= \lambda_1^s \sum_{i=1}^{k}\sum_{j=0}^{r_i-1} \binom{s}{j} \frac{\lambda_i^{s-j}}{\lambda_1^s} (A-\lambda_i)^j h_i \end{aligned}$$

Für $i \geq 2$ ist

$$\lim_{s\to\infty} \binom{s}{j} \frac{\lambda_i^{s-j}}{\lambda_1^s} = 0,$$

wie man beispielsweise mittels Quotientenkriterium leicht einsieht: Da $\lambda_1 > |\lambda_i|$, gilt

$$\frac{\binom{s+1}{j} \frac{|\lambda_i|^{s+1-j}}{\lambda_1^{s+1}}}{\binom{s}{j} \frac{|\lambda_i|^{s-j}}{\lambda_1^s}} = \frac{(s+1)}{(s-k+1)} \cdot \frac{|\lambda_i|}{\lambda_1} \xrightarrow{s\to\infty} \frac{|\lambda_i|}{\lambda_1} < 1.$$

Für $i = 1$ ist gilt asymptotisch, also für große s, dass

$$\binom{s}{0} \frac{\lambda_1^s}{\lambda_1^s} < \ldots < \binom{s}{r_1} \frac{\lambda_1^{s-r_i}}{\lambda_1^s},$$

weshalb die Folge $(x_s)_{s\in\mathbb{N}}$ gegen ein skalares Vielfaches von

$$(A-\lambda_1)^{r_1} h_1 = v_1$$

konvergiert.

Zum Nachweis der Konvergenz der Rayleigh-Quotienten-Folge $R(A, x_k)$ gegen den Eigenwert λ_1 betrachten wir das Skalarprodukt

$$< R(A,x_k)x_k - Ax_k, x_k >= \frac{< x_k, Ax_k >}{< x_k, x_k >} < x_k, x_k > - < Ax_k, x_k >= 0 \tag{4.1.3}$$

und folgern

$$R(A, x_k) = \underset{\lambda \in \mathbb{R}}{\operatorname{argmin}} |\lambda x_k - A x_k|.$$

Wie bereits gezeigt wurde, ist $x := \lim_{k\to\infty} x_k$ ein Eigenvektor (zum Eigenwert λ_1) von A. Daher gilt

$$|R(A, x_k)x_k - Ax_k| \xrightarrow{k\to\infty} 0.$$

Daher, und wegen der Stetigkeit von A, existiert zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so dass

$$|R(A, x_k)x_k - Ax_k|, |Ax - Ax_k|, |\lambda_1 x_k - \lambda_1 x| < \frac{\varepsilon}{3},$$

für alle $k \geq N$. Somit ist

$$\begin{aligned}
|R(A, x_k) - \lambda| &= |R(A, x_k) - \lambda| \cdot |x_k| = |R(A, x_k)x_k - \lambda x_k| \\
&= |R(A, x_k)x_k - Ax + \lambda_1 x - \lambda x_k| \\
&= |R(A, x_k)x_k - Ax_k + Ax_k - Ax + \lambda_1 x - \lambda x_k| \\
&\leq |R(A, x_k)x_k - Ax_k| + |Ax_k - Ax| + |\lambda_1 x - \lambda x_k| \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \text{ für alle } k \geq N.
\end{aligned}$$

□

Bemerkung 4.1.7. Ist $v \neq 0$ ein Eigenvektor von A zu einem Eigenwert λ, dann ist

$$R(A, v) = \frac{< v, Av >}{< v, v >} = \frac{< v, \lambda v >}{< v, v >} = \lambda.$$

Ist $v \neq 0$ aber nur eine Approximation eines solchen Eigenvektors, dann nimmt die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \quad r \mapsto |r \cdot v - Av|, \tag{4.1.4}$$

ihr Minimum in $R(A, v)$ an, wie in (4.1.3) mittels Lotfußpunkt-Ansatz gezeigt wurde. Die in Gleichung (4.1.2) definierte Folge von Rayleigh-Quotienten konvergiert also nicht nur gegen den dominanten Eigenwert der Matrix, sondern liefert in jedem Folgenglied eine, im Sinne der Abstandsfunktion f, beste Approximation des Eigenwertes. Man sagt, der Rayleigh-Quotient löst das durch (4.1.4) gegebene Ausgleichsproblem. Die Thematik linearer Ausgleichsprobleme werden wir in Kapitel 5 ausführlicher aufgreifen.

Korollar 4.1.8. *Ist A reziprok und $v_0 = (1, \ldots, 1)$, liefert dann v_1 wegen der linearen Konvergenz schon die richtige Tendenz?*

Beweis. □

Soll der Analytic Hirarchy Process von einer Gruppe zur Entscheidungsfindung genutzt werden, so ist zu beachten, dass das arithmetische Mittel von Bewertungsmatrizen nicht reziprok ist. Um den AHP auf der Basis von Gruppenbewertungen nutzen zu können, kann ein (gewichtetes) geometrisches Mittel eingesetzt werden: Für $a_1, \ldots, a_n > 0$ mit $a_1 + \ldots + a_n = a$ nennt man

$$f(b_1, \ldots, b_n) := \sqrt[a]{b_1^{a_1} \cdot \ldots \cdot b_n^{a_n}}$$

das *gewichtete geometrische Mittel* der positiven reellen Zahlen $b_1, \ldots, b_n$. Ist $a_i = 1$ für $i = 1, \ldots, n$, so erhält man das übliche geometrische Mittel. Die Gewichte erlauben es an dieser Stelle, die Urteile Einzelner, z.B. von Experten oder Vorgesetzten, stärker in die gemeinsame Bewertung einfließen zu lassen.

Aufgabe 4.1.9. Es seien reziproke Bewertungsmatrizen $B^{(1)}, \ldots, B^{(n)}$ gegeben, mit Einträgen $B^{(k)} = (b_{ij}^{(k)})_{ij}$ für $k = 1, \ldots, n$. Eine gemeinsame Bewertungsmatrix $B = (b_{ij})_{ij}$, soll durch

$$b_{ij} = f(b_{ij}^{(1)}, \ldots, b_{ij}^{(n)})$$

definiert werden, wobei f ein beliebig gewichtetes geometrisches Mittel bildet. Man zeige, dass auf diese Weise B eine reziproke Matrix darstellt.

Unter der Voraussetzung, dass eine gemeinsame Bewertung gewisse Fairness-Bedingungen befolgen soll, lässt sich sogar zeigen, dass geometrisches Mitteln die einzig zulässige Funktion zum Kombinieren der Einzelbewertungen darstellt [Saa08, Theorem 2]. Diese Überlegungen führen direkt einen in die Nähe von Arrows berühmten Satz über die Unmöglichkeit einer Gruppenrangordnungsfunktion die gewissen, intuitiv plausiblen Ansprüchen genügt [Arr12]. Saaty und Vargas haben den AHP in Gruppenprozessen auch vor diesem Hintergrund untersucht und können zeigen, dass geometrisches Mittel und die Eigenwertmethode des AHP Gruppenrangfunktion führen, die Arrows Bedingungen genügt, also insbesondere nicht diktatorisch ist [SV12]. Dieses Resultat steht nicht im Widerspruch zum Satz von Arrow, da in diesem Prozess nicht nur die ordinalen individuellen Präferenzen, sondern auch die Ausprägungen dieser Präferenzen berücksichtigt werden können.

4.2 Markov-Ketten

Die Theorie der Markov-Ketten ermöglicht es, weitgehend auf den Methoden der Linearen Algebra fußend, einen gehörigen Einblick in ein aktives und modernes Kapitel namens *Stochastische Prozesse* aus dem Buch der Wahrscheinlichkeitstheorie zu erhalten.

Beschreibung von stochastischen Prozessen und Einschränkung auf Markov-Prozesse. Wahrscheinlichkeitstheoretische Vorbemerkungen und Einführungen, möglichst knapp.

Definition 4.2.1. Markov-Prozess

4.2.1 Pagerank

Googles Pagerank Algorithmus in den Funktionsablauf einer Suchmaschine einordnen und auf den entsprechenenden Abschnitt zu Information Retrieval in Kapitel 5 verweisen. Die Ordnungsidee als stationäre Verteilung in einem Markov-Prozess beschreiben. [PBMW99]

4.2.2 Markov-Chain Monte Carlo

Das Einstiegsbeispiel aus [Dia09] passt hier wahrscheinlich ganz nett, wenn ich es schaffe einen sinnvollen Anschluss an Krypto-Überlegungen aus Kapitel 2.1 zu erzeugen. Vielleicht sogar Idee aus [CS98] aufgreifen?

4.2.3 Hidden Markov-Model

4.3 Stabilitätslagen

4.4 Schwingungen, Eigenschwingung

Kapitel 5

Normierte Vektorräume

Die Welt ist in vielen Situationen zu detailliert, zu komplex, um sie überhaupt oder gar effizient zu verarbeiten. Wir sind daher häufig daran interessiert ein unter bestimmten Gesichtspunkten bestmögliches reduziertes Abbild von komplexen Situationen zu erstellen. Dieser Abschnitt soll zeigen, dass sich in diesem Vorhaben eine Anwendungsmöglichkeit für lineare Algebra finden lässt. Das aus der Schulmathematik bekannte Lotfußpunkt-Verfahren zur Abstandsbestimmung liefert, aus einem neuen Blickwinkel und dank der Abstraktion des Vektorraumkonzepts, hierzu ein fundamentales Werkzeug und das Konzept der Orthogonalität erscheint plötzlich mit zahlreichen Anwendungen. Beabsichtigt ist, einige dieser Approximations-Anwendungen (vielleicht etwas knapper gehalten) vorzustellen und dafür Bildverarbeitung am Beispiel von JPEG sogar theoretisch und experimentell zu behandeln. Aufgrund des vermutlich besonders ausgeprägten Interesses bei Studenten und Schülern und der interessanten Materialien ist ein Unterabschnitt über Fourieranalyse und das mp3-Format angedacht.

Interessant ist hierzu natürlich [Hei10].

5.1 Computertomographie

Eine zielgruppentaugliche Darstellung findet sich in [HS02].

5.2 Vom Bitmap zum JPEG

auch MPEG?

5.3 Fourieranalyse

5.4 Das mp3-Format

In diesem Abschnitt sollen die Ideen von Orthogonalität und Approximation erneut aufgegriffen werden und gezielt auf das mp3-Format gemünzt werden. Diese Anwendung ist in vielfacher Hinsicht sehr interessant. So handelt es sich

um einen deutschen Exportschlager, um eine Anwendung die einem während jeder Fahrt mit der U-Bahn mehrfach begegnet und gewaltige ökonomische Auswirkungen hat. Das Frauenhoferinstitut stellt hierzu interessante Materialien für den schulischen Mathematikunterricht im Internet zur Verfügung. Vorhergegangene Vorträge zu Approximation und Codierungstheorie erlauben weitere Aspekte des mp3-Formats auf einer ausgebauten theoretischen Grundlage zu besprechen. Außerdem sollen interdisziplinäre Aspekte ausdrücklich Berücksichtigung finden (Schallwellen, Wahrnehmung von Musik).

5.5 DSL-ISDN und Vectoring

Idee

Oben genannten Ideen sorgen in der Signalverarbeitung dafür, dass die alten Kupferkabel im Netz der deutschen Telekom wieder konkurrenzfähig zu den modernen, aber teuer neu zu verlegenden, Glasfaserkabeln werden. Die Unterschiede der Techniken, die mathematischen Grundlagen und der hier zu Tage tretende finanzielle Wert von Mathematik könnten diskutiert werden. Ganz aktuell ist eine Technik unter dem Stichwort 'Vectoring' von großer Bedeutung.

5.6 Informationsgewinnung

Auf den ersten Blick ist die Diplomarbeit [Nie03] von Jörg Niehoff ein netter Einstieg in das Thema. Die dort aufgeführte Literatur liefert weitere Anhaltspunkte.

Kapitel 6

Bilineare Algebra und Geometrie

6.1 Navigation und Kegelschnitte

Mögliche Themen: Multilateration, Trilateration vs. Triangulation, GPS, GLONASS, GALILEO et al.

6.1.1 Hyperbelnavigation

Einführend könnte man das aus heutiger Sicht beinah als Low-Tech einzustufende Systeme wie Decca oder Loran betrachten. Dazu wäre natürlich ein moderner Einsatzzweck unbedingt wünschenswert.

6.1.2 Satellitengestützte Navigation

[RSAS12, Kapitel 1] hat einiges über GPS zu sagen.

Notizen über Kreuzkorrelation bei GPS-Empfängern einbauen.

Kapitel 7

Anhang: Lineare Algebra als offenes Forschungsgebiet

Idee

Dieser Vortrag unterscheidet sich von den anderen Vorträgen dadurch, dass keine Anwendung im Mittelpunkt steht. Stattdessen soll aufgezeigt werden, dass auch die (endlichdimensionale) lineare Algebra kein vollständig verstandenes Gebiet der Mathematik ist. Exemplarisch könnten 2-3 ältere und doch bis heute aktuelle Themengebiete der linearen Algebra (Matrizen-Büschel, Paare sich wechselseitig auslöschender Transformationen, Vier-Unterraum-Problem) und deren typische Fragestellungen vorgestellt werden. Außerdem sollten 2-3 aktuelle 'Forschungsbaustellen' der linearen Algebra (z.B. Darstellungstheorie, Kategorifizierung, quantentheoretische Aspekte) mitsamt typischer Fragestellungen behandelt werden, so dass insgesamt klar wird, dass lineare Algebra über die in der Vorlesung vorgestellten Ausmaße weit hinaus geht und an vielen Stellen Fragen enthält, die wir mit heutigem Wissen oder auch prinzipiell nicht beantworten können.

In einem Briefwechsel zwischen Ziegler und Ringel schreibt Ringel in Bezug auf eine Interviewaussage Zieglers:

```
Leider vermitteln fast alle Buecher zur LA (und moeglicherweise auch die
entsprechenden Vorlesungen?) den voellig irrefuehrenden Eindruck, dass es sich
hier um eine abgeschlossene Theorie handelt - und das, obwohl nicht einmal das
Wissen Kroneckers (Klassifikation der Matrizenbueschel, sehr wichtig fuer das
Loesen von Differentialgleichungen, siehe Gantmacher) oder Dedekinds (Struktur
des freien modularen Verbands in 3 Erzeugenden) vermittelt wird, geschweige denn
das, was etwa Gelfand und Ponomarev in den 60er Jahren erreicht haben (Paare
sich annullierender Operatoren, Loesung des Vierunterraumproblems,...).

Gabriel hat schon vor langer Zeit betont, dass man die Kategorie der
endlich-dimensionalen Vektorraeume sich so vorstellen sollte: Objekte sind die
natuerlichen Zahlen, Morphismen sind die nxm-Matrizen (in der Sprache der
Kategorientheorie: das ``Geruest''); in dieser Weise ist die LA etwas, was man
heute gerne ``Kategorifizierung der natuerlichen Zahlen'' nennt, und damit eine
nicht-kommutative Version der natuerlichen Zahlen. Dieser Gesichtspunkt ist
```

bisher noch gar nicht ausgereizt, bildet aber einen Ausgangspunkt fuer viele Aspekte der ‘‘nicht-kommutativen Geometrie’’ (man denke nur an die Quantenzahlen, definiert durch Gauss-Polynome, und viele andere q-Phaenomene).

Selbst Ihre Neuformulierung: Nicht einmal in der Linearen Algebra von endlichdimensionalen Vektorraeumen ist alles erforscht. erscheint mir viel zu schwach! Im Gegensatz etwa zur Zahlentheorie, wo schon durch Euler und spaestens Gauss ziemlich alle elementaren Fragen abschliessend behandelt wurden, gibt es zum Beispiel in der Darstellungstheorie endlich-dimensionaler Algebren (und was ist dies anderes als LA?) Unmengen an ganz einfach formulierbaren, aber voellig unerforschten Fragen, an denen sich auch schon ein Student die Zaehne ausbeissen kann... Um wenigestens ein Beispiel explizit zu erwaehnen, moechte ich auf die (Einleitung zur) Arbeit ‘‘Invariant subspaces of nilpotent operators‘‘ verweisen, die im Crelle Journal erscheinen wird (ArXiv math.RT/0608666).

Möchte ich den original Wortlaut hier überhaupt zitieren? Dann muss ich den Link wieder heraussuchen.

Kapitel 8

Anhang: Was Mathematik eigentlich ist

Frage natürlich nicht beantworten (bessere Überschrift finden?) aber ein Bild von Mathematik mit seinen Disziplinen skizzieren, in das die vorherigen Kapitel eingearbeitet werden. Historische Dimension.

Literaturverzeichnis

[Arr12] Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

[BBL04] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

[CS98] A. Canteaut and N. Sendrier. Cryptanalysis of the original McEliece cryptosystem. In *Advances in Cryptology - ASIACRYPT'98*, pages 187–199. Springer, 1998.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-Vetor Networks. *Machine Learning*, 20:273–297, 1995.

[DA99] T. Dierks and C. Allen. The TLS Protocol. Technical Report 2246, Network Working Group, January 1999.

[DD04] Jürgen Dankert and Helga Dankert. *Technische Mechanik*. Springer, 2004.

[Dia09] Persi Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[emv11] *EMV Integrated Circuit Card Specifications for Payment Systems*, volume 2. EMV Co., November 2011.

[Fis05] Gerd Fischer. *Lineare Algebra (vieweg studium; Grundkurs Mathematik)*. Vieweg+Teubner Verlag, 15, verb. aufl. 2005 edition, 9 2005.

[Fro12] Ferdinand Georg Frobenius. *Über Matrizen aus nicht negativen Elementen*. Königliche Akademie der Wissenschaften, 1912.

[Geo04] Hans-Otto Georgii. De Gruyter, 2004.

[GJK88] Elmer G Gilbert, Daniel W Johnson, and S Sathiya Keerthi. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE Journal of Robotics and Automation*, 4(2):193–203, 1988.

[Hei06] Anke Heier. *Nachweis der geographischen Herkunft von Pistazien anhand der Stabilisotopenverhältnisse*. PhD thesis, TU Berlin, 2006.

[Hei10] J. Heitzer. *Orthogonalität und beste Approximation.* PhD thesis, Universitätsbibliothek, 2010.

[HS02] M. Hochbruck and J.M. Sautter. Mathematik fürs Leben am Beispiel der Computertomographie. *Mathematische Semesterberichte*, 49(1):95–113, 2002.

[Hup90] Bertram Huppert. *Angewandte lineare Algebra.* Walter de Gruyter, 1990.

[HW06] B. Huppert and W. Willems. *Lineare Algebra.* B.G.Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden (GWV), 2006.

[Kob07] Neal Koblitz. The uneasy relationship between mathematics and cryptography. *Notices of the AMS*, 54(8):972–979, 2007.

[Lüt03] Werner Lütkebohmert. *Coderierungstheorie.* Vieweg Studium - Aufbaukurs Mathematik. Vieweg, 2003.

[Mut06] Herbert J. Muthsam. *Lineare Algebra und ihre Anwendungen.* Elsevier, 2006.

[Nie03] Jörg Niehoff. Informationsgewinnung im Vektorraum-Modell. Master's thesis, Heinrich Heine Universität Düsseldorf, 2003.

[ODGG09] C.P. Ortlieb, C.V. Dresky, I. Gasser, and S. Günzel. *Mathematische Modellierung.* Vieweg+ Teubner Verlag— GWV Fachverlage GmbH, Wiesbaden, 2009. Eine Einführung in zwölf Fallstudien.

[PBMW99] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. 1999.

[Per07] Oskar Perron. Zur Theorie der Matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[Ris01] Irina Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[Roe88] John E Roemer. *Analytical foundations of Marxian economic theory.* Cambridge University Press, 1988.

[RSAS12] C. Rousseau, Y. Saint-Aubin, and M. Stern. *Mathematik und Technologie.* Springer, 2012.

[Saa90] Thomas L Saaty. How to make a decision: the Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1):9–26, 1990.

[Saa08] Thomas L Saaty. Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. *RACSAM-Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 102(2):251–318, 2008.

[Sch08] Uwe Schäfer. *Das lineare Komplementaritätsproblem.* Springer-Lehrbuch. Springer, 2008.

[SV12] Thomas L Saaty and Luis G Vargas. The possibility of group choice: pairwise comparisons and merging functions. *Social Choice and Welfare*, 38(3):481–496, 2012.

[Wer07] Dirk Werner. *Funktionalanalysis.* Springer-Lehrbuch. Springer, Berlin [u.a.], 6., korrigierte aufl. edition, 2007.

[Wie50] Helmut Wielandt. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift*, 52(1):642–648, 1950.

[WWZZ14] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers. *The 23rd USENIX Security Symposium (Usenix Security 2014)*, 2014.

[Zin95] Reinhard Zintl. Der Nutzen unvollständiger Erklärungen: Überlegungen zur sozialwissenschaftlichen Anwendung der Spieltheorie. Technical report, MPIfG discussion paper, 1995.

Index